International Journal of Instruction *April 2020 ● Vol.13, No.2*
e-ISSN: 1308-1470 ● www.e-iji.net *p-ISSN: 1694-609X*
pp. 507-518

Received: 27/03/2019
Revision: 22/11/2019
Accepted: 28/11/2019
OnlineFirst:24/01/2020

The Quality of Test Instruments Constructed by Teachers in Bima Regency, Indonesia: Document Analysis

Syahrul Ramadhan
Doctoral Program of Educational Research and Evaluation Department, Yogyakarta State University, Indonesia, *syahrul.ramadhan2015@student.uny.ac.id*

Rudy Sumiharsono
Prof, IKIP PGRI Jember, JL. Jawa 10 Jember, Indonesia, *rudysumiharsono@gmail.com*

Djemari Mardapi
Prof., Educational Research and Evaluation Department, Yogyakarta State University, Indonesia, *djemari.@uny.ac.id*

Zuhdan Kun Prasetyo
Prof., Faculty of Science, Yogyakarta State University, Indonesia, *zuhdan@uny.ac.id*

The analysis of the Test Instruments' quality is a crucial thing needs to be conducted. The test instruments made by teachers must fulfil the requirements (validity, reliability, and standard error of measurement) until the measurement result obtained can describe the students' actual abilities. This research aims to analyse the content validity (quantitative and qualitative), empirical validity, reliability, and standard error of measurement of semester final exam test instruments on Physics Grade XII Senior High School academic year 2017/2018 designed by teachers. The data analysis was based on 5 question documents (135 items) made by teachers and 555 answer sheets of the students at five schools. The research results show that teachers' ability in making the test instruments of the semester final exam is still limited. It is proven as a problem found through the representation of the analysis result based on the test of content validity, empirical validation, reliability and standard errors of measurement. It is suggested that school principals or the Educational Authority should hold effective training regularly and invite teachers to participate in it to sharpen their skills in making effective test instruments.

Keywords: instruments' quality, document analysis, test instruments, teachers, test

INTRODUCTION

Success in the learning process can be seen through the assessment process. In the effort to assess student achievement in learning, teachers usually conduct measurements in the

Citation: Ramadhan, S., Sumiharsono, R., Mardapi, D., & Prasetyo, Z. K. (2020). The Quality of Test Instruments Constructed by Teachers in Bima Regency, Indonesia: Document Analysis. *International Journal of Instruction*, 13(2), 507-518. https://doi.org/10.29333/iji.2020.13235a

form of written, practical and oral tests. A good instrument is needed to conduct measurements. A good instrument is an instrument that has high validity and reliability and has the smallest possible errors in capturing information about the success of the learning process. A good instrument will produce accurate measurements in obtaining information about the success of the learning process (Azwar, 2012). In line with this opinion, Mardapi (2008) states that to produce accurate information, the instruments in measurement must be reliable, so that the instrument is able to produce the measurement errors as small as possible.

In order to get a good instrument, it is necessary to do a validity test. Validity is an important thing that must be considered by every teacher who designs a test for measuring student achievement. The test is said to be valid if the test has measured the actual abilities possessed by students through learning activities (Ramadhan, Mardapi, Prasetyo, & Utomo, 2019). When the test used does not represent the ability of students, then information about students' abilities is difficult to obtain fully. Therefore to get information about the abilities obtained by students through learning, the teacher must design the test as well as possible.

Azwar (2010) states that a test instrument is said to have high reliability if the test score is highly correlated with its own pure score. From this statement, reliability can be interpreted as how high the correlation between scores appears on two parallel tests. In line with this opinion, Allen & Yen (1979) also states that a test set is said to be reliable if the observed score has a high correlation with the actual score. This means that when a test instrument that has a good reliability value is used to measure student learning achievement, the test instrument will produce visible values that are close to the actual value, so that the standard measurement error owned by the measuring instrument is very small.

Therefore, to minimize errors in measurement, a good instrument is needed. From the results of examinations done by students, the teacher gives a score that is usually in the form of numbers. However it raises a question, whether the score obtained from the test results is the actual student score? Wright (2007) states that *True score = observed score ± measurement error.* (Actual score = measurement results score ± measurement error). From that equation, two possibilities will occur. First, the score may be lower than the actual score. Second, the score of measurement results may be higher than the actual score. If one of the two possibilities occurs, then an error has occurred in the measurement process.

Removing all sources of measurement errors is difficult, but measurement errors can be minimized so that the scores can reflect the actual ability of test participants. Among the sources of measurement error, it seems that the easiest to control is the factor used to measure (Ramadan & Mardapi, 2015). Therefore, to minimize measurement errors is by focusing on the making of good measurement tools or instruments (test questions). The fact is that there are still many teachers who have not been able to make instruments (questions) properly. That was strengthened by Miller's (2008) statement that *"Most teachers, administrators, and career guidance personnel acknowledge that tests (developed by teachers, commercially developed and standardized) are not fully valid or reliable"*.

In Indonesia, the Final Semester Examination (UAS) is an educational evaluation conducted every six months. The evaluation is based on a system that is implemented based on the semester system. For teachers of grade XII at Senior High School, the Final Semester I am expected to know the readiness of students in facing the National Examination (UN) because grade XII material is 50% of the total UN question material in Senior High School. To make the results of the National Examination able to describe the actual learning results obtained by students, the question instrument used must meet all the requirements for a good measurement instrument and tested in various aspects. The problem that arises now is whether the Final Semester Examination test instruments are really a good measurement tool that is able to reflect students' abilities.

This research was conducted in Bima Regency, West Nusa Tenggara Indonesia. Bima Regency was chosen as the research location because it was considered as representative of eastern Indonesia. Besides that, similar research has never been carried out. The field of study selected in this research is Physics. Making physics test instruments is considered quite complicated compared to other subjects because it is not only equations or mathematical calculations but also involves the ability to analyze in the form of questions related to concepts and laws of physics.

Based on the preliminary observations result, it was found out that the teachers in Bima Regency generally did not analyse the validity, reliability and the measurement errors in the measurement process. This is not a good thing considering the analysis is very fundamental because it is used to find out whether the test instrument has fulfilled good characteristics so that the measurement results can describe the actual abilities of students. So, the purpose of this research are;

- To find out the contents validity (quantitative or qualitative) of Senior High School Final Semester Examination test instruments for Physics subject at grade XII in the academic year2017/2018.
- To find out empirical validity (Item Difficulty, Item Discrimination) Senior High School Final Semester Examination test instruments for Physics subject at grade XII in the academic year2017/2018.
- To find out the reliability of Senior High School Final Semester Examination test instruments for Physics subject at grade XII in the academic year 2017/2018, and
- To find out the estimation of standard errors measurement of Senior High School Final Semester Examination test instruments for Physics at grade XII in the academic year2017/2018.

Through this research, it is expected that it can provide comprehensive information for the teachers, schools or educational institutions in Bima Regency specifically and generally in Indonesia about analysis of the test instruments quality that have been made. The hope is that through the information from the research result, teachers can conduct self-evaluation in order to become more serious and skilled in developing the test instruments so that they can measure the competence of students correctly and appropriately.

METHOD

Type of Research

This research uses quantitative methods and is categorised as research *Ex post facto* because this research describes the events that have already occurred.

Research Sample

The objects in this research are the test instruments made by teachers in the form of multiple choices and all student answer sheets of Senior High School Physic subject at grade XII Semester Final Examination in the academic year of 2017/2018. There are five sets of multiple choice made by the teachers. While the students answer sheets used in this research were 555 sheets taken from 5 high schools located in Bima Regency Indonesia. Each school that becomes the sample has different test instruments but has the same form, that is multiple choice. The question sheet was analysed by the expert to see content validity. Student answer sheets were examined to see empirical validity, reliability and the amount of the standard error of measurement that occurred in the test instruments made by teachers of each high school.

Data Collection Techniques

The data collection technique in this research is documentation. The researcher directly came to the high schools which became the object of study in Sape Sub-district and then met with the principal to ask permission to take the data in the school. Furthermore, the researcher met the Physics subject teachers to get information about the test instruments and student answer sheets that had been tested in the academic year 2017/2018. The number of items in each school and the number of student answer sheets can be seen in the following table:

Table 1
Number of Test Instruments and Student Answer Sheets

No	School	N of Items	N of answer sheets
1	State Senior High School 1 of Sape	25	215
2	State Senior High School 2 of Sape	25	101
3	State Senior High School 3 of Sape	35	65
4	Senior High School of PGRI	25	55
5	Senior High School of Muhammadiyah	25	119
Total		135	555

Data Analysis

Content validation

The validation is determined using expert agreement. The agreement on the subject of the study or often known as the domain measured determines the level of validity (content related) (Retnawati, 2016). This is due to measurement instruments, such as tests or questionnaires are proven valid if the expert believes that the instrument is able to measure the ability defined in the measured domain. In this research, it used analysis assistance from 3 experts, consisting of physicists, educational research and evaluation.

Validity analysis on the question sheet refers to the Republic of Indonesia's Minister of National Education Regulation Number 20 of 2007 concerning Educational Assessment Standards (Indonesia, 2015). Learning result assessment instruments used by educators fulfilled the requirements of (a) substance, is representing competencies assessed, (b) construction, is fulfilling technical requirements in accordance with the form of instruments used, and (c) language, is using good and correct language also communicative in accordance with the level of students development. Besides those

these aspects, the researcher added the fourth requirement, that is (d) the level of thinking based on the characteristics of the Higher Order Thinking Skill from Bloom's Taxonomy. The cumulative results of the four components (substance, construction, language and thinking level) will be the scores that will later be used to calculate the content validity using Aiken formula (Aiken, 1985).

Aiken formulated a formula Aiken's "V" tocalculate *content-validity coefficient* which is in accordance to the results of an expert panel assessment of n people towards an item in terms of how far the item represents the measured construct. The formula proposed by Aiken is as follows (Azwar, 2012; Ramadhan, Mardapi, Prasetyo, & Utomo, 2019; Ramadhan, Mardapi, Sahabuddin, & Sumiharsono, 2019):

$$V = \frac{\Sigma s}{n(c-1)} \quad \text{.....(1)}$$

Formula Description:

V : item validity index
s : the score set for each rater is reduced by the lowest score in the category used ($s=r-l_o$, → r = choice score and rater l_o = lowest score in the scoring category)
n : the amount of rater
c : number of ratings / criteria.

Reliability

The instrument's reliability is intended to see the consistency of the tests made by the teachers if the observation is repeated. The level of instrument's reliability empirically proven by the amount of the reliability coefficient which is in the range of 0 to 1 (Mardapi, 2008, 2012). The higher the coefficient value means the higher the reliability and vice versa. The coefficient formula of *Alpha Cronbach*'s used to estimate and calculate the test reliability is by using *Iteman 4.3* computer program. The reliability estimation is basedon the index of instrument reliability that is good if > 0,7 (Mardapi, 2008, 2012; Retnawati, Kartowagiran, Arlinwibowo, & Sulistyaningsih, 2017).

Question characteristic analysis

Quantitative analysis of the characteristics of test instruments in the Final Examination Semester of High School in physics was conducted based on the classical Test Theory approach. The researcher analyzes student response patterns (based on student answer sheets) to see information about test instruments that are feasible and inappropriate to be tested based on item parameters, namely Difficulty Items, Discrimination Items, and the distractors (Ramadhan, Mardapi, Prasetyo, & Utomo, 2019; Ramadhan, Mardapi, Sahabuddin, & Sumiharsono, 2019).

The analysis of the Item Difficulty of Senior High School Final Semester Examination in Physic test was done using *Iteman 4.3* computer program. The Item Difficulty of the question can be seen in the column *Prop. Correct.* Item questions that have Item Difficulty are at intervals of 0.3 to 0,8.

The Item Discrimination analysis of Senior High School Final Semester Examination in Physic Subject test can be seen in the *Point Biserial* column conducted using the *Iteman 4.3* computer program. The criteria for good question is if it has a value of $D \geq 0.3$, while question that hasa value of $D \leq 0.3$ need to be revised or replaced with new items.

The information about the distractors can also be taken from *Iteman 4.3* computer programs, which is in the column *Prop Endorsing*. The distractors are said to function if the value of *PropEndorsing* in each multiple choice has a value greater than 0.05. The proportion value of each question that has a value smaller than the value of *PropEndorsing*, means that the distractor needs to be revised.

The standard error of measurement

Standard errors of measurement analysis were performed using the Feldt model (Feldt, Steffen, & Gupta, 1985). The first step to finding standard measurement errors with the Feldt model is by creating a sectional distribution table of the test results given to the research subjects once then proceed by dividing the test randomly into two parts that are not equal in length and the contents are still homogeneous (conjugate). Thus, two distributions are obtained. Each part consists of various numbers of questions. To obtain the reliability test index using the division above, Feldt proposed equation as follows:

$$rxx' = \frac{4(Sx_1 x_2)}{s^2x - [\frac{S^2_{X_1} - S^2_{X_2})}{sx}]^2} \quad \text{..... (2)}$$

Formula Description:

Rxx' : Test Reliability

S^2_{X1} : Score Variant on part 1

S^2_{X2} : Score Variant on part 2

$S_{X1\,X2}$: Score Co Variant on part 1 and 2

Sx : Standard deviation of test scores

Thus to look for error variants, first it must be searched the pure score variants. To find pure score variants, pure classical score theory equations can be used, $S^2_t = S^2_x$. Rxx'. Therefore the estimation of error variants can be determined as follows:

$$S^2_e = S^2_x - S^2_t \quad \text{..... (3)}$$

Formula Description:

S^2_e : Variant Estimation of measurement error

S^2_x : ObservationScore Variant

S^2_t : Pure score variant

Standard error of measurement is

$$S_e = \sqrt{S^2_e} \quad \text{..... (4)}$$

FINDINGS

Content Validity

Expert judgment in this study was conducted by 3 (three) experts in the subjects of physics, education research and evaluation. An instrument is said to be valid if the expert believes that the instrument measures the things to be measured. The expert judgment gives an assessment that will be used to prove the content validity. The expert assessment results are then calculated using the Aiken equation. The Aiken calculation results for each question are then compared to the Aiken table. The results are shown in the Table 2:

Table 2
The Result of the Content Validation Analysis using Aiken Formula

No	School	N of Items	Standard of Aiken Table	Item Characteristics	
				Valid	Invalid
1	State Senior High School 1 Sape	25	0,66	20	5
2	State Senior High School 2 Sape	25	0,66	18	7
3	State Senior High School 3 Sape	35	0,60	30	5
4	Senior High School of PGRI	25	0,66	23	2
5	Senior High School of Muhammadiyah	25	0,66	24	1
Total		135		115	20

Based on Table 2, it can be seen that there are 20 invalid items out of a total of 135 items. It means that there are about 15% (20 items) of the total overall items that are assessed to show a quality that is not good (invalid). The total of valid items is quite a lot based on the expert's assessment of 85% (115 items). Question analysis by experts then quantified based on substance, construction, language and thinking level aspects. Based on the results of the substance analysis which includes the suitability of the questions with indicators and the suitability of the material with competence, the results of 75% are fulfilled. Based on the aspect of construction includes the characteristics of the question form, the main question idea is not confusing, and the function of the distractor shows the result of 80%. Language aspects which include the use of appropriate language, communicative language and not using bad language show 90% results. On the aspect of thinking level shows the lowest value that is 65%. This indicates that teachers have not yet been fully expert in making test instruments with a useful category of high thinking levels. Aspects of the level of thinking include the use of new questions (Different from previous examples or assignments), having a stimulus, based on contextual problems and using the top 3 levels of Bloom's revised taxonomy (Analyzing, evaluating and creating).

Reliability

Reliability Analysis was calculated using the *Iteman4.3* program. Reliability estimates are based on index of instrument reliability said to be good if > 0,7.

Table 3
Reliability of the Test Set

School	Reliability	Criteria	Decision
State Senior High School 1 of Sape	0,667		Not Reliable
State Senior High School 2 of Sape	0,695		Not Reliable
State Senior High School 3 of Sape	0,667	> 0,7	Not Reliable
Senior High School of PGRI	0,815		Reliable
Senior High School of Muhammadiyah	0,765		Reliable

From the results of reliability estimation, it is known that there are unreliable assessments in three high schools, that are State Senior High School 1 of Sape, State Senior High School 2 of Sape and State Senior High School 3 of Sape with a reliability coefficient of less than 0.7. Other results show reliable ratings with reliability coefficients of more than 0.7 found only in PGRI Senior High School and Muhammadiyah Senior High School.

Characteristic Analysis

Quantitative analysis of the test instruments characteristics based on the Classical Test Theory is done by using the help of computer program *Iteman4.3*. An item is said to have good characteristics if the item meets the criteria of Item Difficulty, Item Discrimination and Distractor. Items that have a good level of difficulty are in the interval 0.3 to 0.8. Criteria for good items have distinguishing power values D≥0.3, while Distractors or fraudsters are said to function if the value of Prop Endorsing in each multiple choice has a value greater than the value of 0.05. The analysis results of essential information about the characteristics of test instruments are shown in Table 4.

Table 4
Characteristic of the Test Set

School	Characteristics		N of Items	N of Good Item	N of Poor Item
	Item Difficulty	Item Discrimination			
State Senior High School 1 of Sape	0,414-0,800	0,175-0,503	25	12	13
State Senior High School 2 of Sape	0,079-0,911	-0.173-0,659	25	9	16
State Senior High School 3 of Sape	0,077-0,892	-0,277-0,718	35	21	14
Senior High School of PGRI	0,193-0,689	0,115-0,684	25	20	5
Senior High School of Muhammadiyah	0,055-0,80	-0,115-0,729	25	18	7

The test instruments used by State Senior High School 1 were 25 items. From these items, there were 14 items (56%) that had a good Item Difficulty criterion, because they had a *prop correct* value of 0.3-0.8. Based on the Item Discrimination analysis, from 25 items there were 19 questions (76%) which had good criteria because they had a pointer's value of more than 0.3. Based on the analysis of the function of distractors, there were 19 questions (76%) that had good distractors. The question that has a distractor is not good because one or more alternative answers do not have negative correlative values and the other alternative answers used are more appropriate than the key answers that have been determined. Some items that do not meet these requirements can be revised or discarded. The results of the final analysis showed 12 items (48%) that were good and 13 items (52%) whose characteristics were not good.

The test instruments of Physics in Final Semester Examination Test of State Senior High School 2 Sape used 25 questions in the form of multiplechoice. Based on the results of the analysis, there were 12 questions (48%) that had a good level of difficulty criteria, because they had a *prop correct* value of 0.3-0.8. While for theItem Discrimination, there were 16 questions (64%) which had good criteria because they had a point value of more than 0.3. Based on the analysis of the function of distractors, there were 19 questions (76%) that had good distractors. The results of the final analysis showed 9questions (36%) that were good and 16 items (64%) whose characteristics were not good.

The next test instruments that were analysed was the Final Semester Examination test instruments of State Senior High School 3 Sape, amounting to 35 multiple choice questions. The results of the analysis showed that there were 22 questions (62.86%) that had good Item Difficultycriteria because they had a *prop correct* value of 0.3-0.8. While for distinguishing power, there were 20 questions (57.14%) which had good criteria because they had a point value greater than 0.3. Based on the analysis of the distractor function, there were 19 questions (54.26%) that had a good distractor. The results of the

final analysis showed 21 questions (60%) that were good and 14 questions (40%) whose characteristics were not good.

Twenty-five items of PGRI Senior High School Sape Physics Final Semester Examination Test set. 20 items had a good Item Difficultycriteria because they had a *prop correct* value of 0.3-0.8. While for Item Discrimination, 16 questions had good criteria because they had a point value greater than 0.3. Based on the analysis of the distractors function, there were 19 questions (54.26%) that had a good distractor. The results of the final analysis showed 20 questions (60%) that were good and five questions (40%) whose characteristics were not good.

From the results of the analysis on 25 items about the Physics test instruments used inFinal Semester Examination in Muhammadiah Sape Senior High School Physics can be concluded that there were 22 questions (88%) that had good Item Difficultycriteria because they had a *prop correct* value of 0.3-0.8. The results of the analysis for Item Discriminationshowed that there were 18 questions (72%) which had good criteria because they had a point value greater than 0.3. Based on the analysis of the distractors function, there were 19 questions (76%) that had good distractors. The final analysis results showed 18 questions (72%) that were good and seven questions (28%) whose characteristics were not good.

Standard Error Measurement Analysis

The standard error estimation of the Senior High School Physics Test instruments measurement was done based on the estimation method from Feldt. The calculation results of the standard error of measurement can be seen in Table 5.

Table 5
SEM Analysis Result

School	Feldt
State Senior High School 1 of Sape	2,507
State Senior High School 2 of Sape	2,230
State Senior High School 3 of Sape	3,190
Senior High School of PGRI	2,058
Senior High School of Muhammadiyah	2,662

The estimated value of the measurement standard error is substituted in the Classical Test Theory Formula to find out the true value range taken by each student through the measurement process. The estimation of standard error measurement for the Physics Final Semester Examination test instruments of State Senior High School 1 isbased on the Feldt Method results in an estimate of the standard error of measurement of 2.507. The true value obtained by students is based on standard measurement errors of 2.507 with a confidence level of 95% having a range between $X - 4{,}914 \leq T \leq X + 4{,}914$.

The estimation of measurement standard error of State Senior High School 2 Physics Final Semester Examination test instruments is based on the Feldt Method of 2.230, the true value obtained by students based on the standard error of measurement is 2.230 with a confidence level of 95% having a range between $X - 4{,}371 \leq T \leq X + 4{,}371$.

The estimation of measurement standard error of State Senior High School 3 Physics Final Semester Examination test instruments is based on the Feldt Method is 3,190, the

true value obtained by students based on standard error of measurement is 3,190 with a confidence level of 95% having a range between $X - 6{,}252 \leq T \leq X + 6{,}252$.

The estimation of measurement standard error of the PGRI Senior High School Physics Final Semester Examination test instruments is based on the Feldt Method is 2.058, the true value obtained by students based on the standard error of measurement is 2.058 with a confidence level of 95% having a range between $X - 4{,}034 \leq T \leq X + 4{,}034$.

The estimation of measurement standard error of Muhammadiah Senior High School Physics Final Semester Examination test instruments isbased on the Feldt Method is 2.662, the true value obtained by students based on the standard error of measurement is 2.662 with a confidence level of 95% having a range between $X - 5{,}218 \leq T \leq X + 5{,}218$.

DISCUSSION

Making valid and reliable test instruments and small measurement standard errors are one of the abilities that must be possessed by a teacher. However, in reality, there are still many teachers who have not been able to make valid and reliable of test instruments, the results of this research also strengthened these findings.

The results of the expert's assessment analysed using the Aiken equation showed that there are 20 invalid items out of a total of 135 items. This means that there are around 15% (20 items) of the total overall items that are judged to show a quality that is not good (invalid). The number of valid items is quite a lot based on the expert's assessment of 85% (115 items). Based on the substance aspect, it is known that 75% of the item questions have been fulfilled, 80% of the item questions have fulfilled the construction aspect, 90% of the questions have fulfilled the language aspect and in the thinking level aspect shows only 65% of the questions have fulfiled the criteria.

These findings generally indicate that the expert agrees that teachers have made a good test instrument based on the aspects of substance, construction and language, but teachers still have difficulty in compiling the high-level thinking test instrument. The low level of thinking shows that teachers are not yet skilled in making test instruments with good high level thinking categories. The component level of thinking includes the use of a new question (Different from previous examples or tasks), has a stimulus, based on contextual issues and uses the top 3 levels of Bloom'srevised taxonomy (Analysis, evaluation, and creating). In addition to teachers who are not skilled at making HOTS level questions, the results of the analysis also indicate that there were indications that students were not too skilled in working on the high-level type questions (HOTS).

Quantitative analysis of item characteristics about test instruments based on Classical Test Theory was performed using the help of *Iteman's 4.3* computer program. This analysis yields important information about item characteristics, namely, reliability, Item Difficulty, Item Discrimination and distractor. The assessment results showed that there were 55 (40.74%) items that were not valid from a total of 135 items.

The Item Difficulty of the Final Semester Examination test instrumentsis not in accordance with the ability of the participants, the Item Discrimination and the distractor in the Final Semester Examination has not been able to function properly. The number of the Item Difficulty and the Item Discrimination of the test affect thetest reliability

index. The easy or difficulty of the tests tend to have small reliability values. This condition implies a fairly high measurement error.

Valid and reliable of test instruments are important for teachers to get the true results from the success of the learning process. Problems/instruments that are not good will make it difficult for teachers to get information about the success of learning that has been done and cause a standard error of measurement. The teacher must try, learn, and take part in training to improve their competence in making good instruments so that the questions they designed have high validity and reliability and have the as smallest as possible standard errors of measurement. Therefore teachers should be aware of these weaknesses so that they can correct their weaknesses.

Purnomo's research (2007) shows that the results of the analysis of test instruments tested at UAS or Final Examination in three elementary schools in Gajahmungkur Sub-District Semarang, apparently most of them cannot be used because they do not meet the requirements of validity, reliability, level of difficulty and distinguishing power of questions. This indicates that the teacher has not been able to completely compile the questions of final examination.

Feldt et al. (1985) state that matching the form of test construction is basically a process of choosing a multilevel item rather than a sample that is completely randomised from a population of items. Strengthened by the opinion of Azwar (2012) which states that the greater the variability means that the scores in the distribution are increasingly diverse, where if the variability is small means that the scores in the distribution tend to be the same or called homogeneously. Based on these statements, it can be concluded that a homogeneous score will produce small variability, while a different score will produce considerable variability.

CONCLUSION

The results of the research show that the teacher's ability to prepare final semester exam test instruments is still limited. It is proven as a problem found through the representation of the analysis result based on the test of content validity, empirical validation, reliability and standard errors of measurement. Therefore, the researcher gave some suggestions and recommendations including the Education Official in Bima Regency as well as from the relevant school parties. It is hoped that they will hold competency enhancement training for teachers, especially training in making good test instruments, that is test instruments that are valid, reliable, and has standard errors as small as possible. The school or education official should invite speakers from recognised academic institutions to hold the training for teachers on a regular basis. The teacher's knowledge of standard measurement error estimation is still lacking, therefore other researchers also have the opportunity to conduct the same analysis in different or further locations. Besides, other researchers need to develop effective training models to sharpen the teacher's skills in compiling effective questions.

ACKNOWLEDGMENTS

This work was financially supported by Lembaga Pengelola Dana Pendidikan (LPDP). We would like to express our institution, Yogyakarta State University that has provided

many experiences in term of research.The deepest gratitude is for students who participated in this experimental research and also those who have contributed to our empirical data collection. Thank you to high school teachers for their helps in this research

CONFLICT OF INTERESTS

The authors declare that there is no conflict of interest.

REFERENCES

Aiken, L. R. (1985). Three coefficients for analyzing the reliability and validity of ratings. *Educational and psychological measurement, 45*(1), 131-142.

Allen, M. J., & Yen, W. M. (1979). Introduction to measurement theory. Monterey, CA: Brooks: Cole Publishing Company.

Azwar, S. (2012). Reliabilitas dan validitas. *Yogyakarta: Pustaka Pelajar*.

Azwar, S. (2010). Reliabilitas dan validitas alat ukur. *Yogyakarta: Pustaka Pelajar*.

Feldt, L. S., Steffen, M., & Gupta, N. C. (1985). A comparison of five methods for estimating the standard error of measurement at specific score levels. *Applied psychological measurement, 9*(4), 351-361.

Indonesia, P. M. P. N. R. (2015). Nomor 20 Tahun 2007 tentang Standar Penilaian Pendidikan. *Jakarta: Badan Standar Nasional Pendidikan.*

Mardapi, D. (2008). Teknik penyusunan instrumen tes dan nontes: Yogyakarta: Mitra Cendikia Press.

Mardapi, D. (2012). Pengukuran penilaian dan evaluasi pendidikan. *Yogyakarta: Nuha Medika*.

Miller, P. W. (2008). *Measurement and teaching*: Patrick W. Miller & Associates.

Purnomo, A. (2007). Kemampuan guru dalam merancang tes berbentuk pilihan ganda pada mata pelajaran ips untuk ujian akhir sekolah (UAS). *Lembaran Ilmu Kependidikan, 36*(1).

Ramadhan, S., & Mardapi, D. (2015). Estimasi Kesalahan Baku Pengukuran Soal-Soal UAS Fisika Kelas XII SMA di Kabupaten Bima NTB. *J Evaluasi Pendi., 3*(1), 90-98.

Ramadhan, S., Mardapi, D., Prasetyo, Z. K., & Utomo, H. B. (2019). The development of an instrument to measure the higher order thinking skill in Physics. *European Journal of Educational Research, 8*(3), 743-751.

Ramadhan, S., Mardapi, D., Sahabuddin, C., & Sumiharsono, R. (2019). The estimation of standard error measurement of Physics final examination at senior high schools in Bima Regency Indonesia. *Universal Journal of Educational Research, 7*(7), 1590-1594.

Retnawati, H. (2016). *Validitas reliabilitas dan karakteristik butir*. Yogyakarta: Parama.

Retnawati, H., Kartowagiran, B., Arlinwibowo, J., & Sulistyaningsih, E. (2017). Why Are the Mathematics National Examination Items Difficult and What Is Teachers' Strategy to Overcome It? *International Journal of Instruction, 10*(3), 257-276.

Wright, R. J. (2007). *Educational assessment: Tests and measurements in the age of accountability*: Sage Publications.