Journal of Modern Education Review, ISSN 2155-7993, USA
February 2020, Volume 10, No. 2, pp. 72–83
Doi: 10.15341/jmer(2155-7993)/02.10.2020/002

http://www.academicstar.us

Kappa Index as Reliability Test of Assessment Instruments Based on 2013 Curriculum in Indonesia

Didik Setyawarno, Atik Kurniawati
(Yogyakarta State University, Indonesia)

Abstract: This article is intended for teachers, university students, and educational observers in Indonesia. This article examines the concept of assessment and testing of instrument reliability based on the 2013 curriculum in Indonesia. The concept of assessment covers aspects of assessment, assessment processes, and assessment techniques. Testing the reliability of the assessment instrument in accordance with the 2013 curriculum in Indonesia namely Kappa index. The study methods used in this article are literature studies from a variety of sources both in books and in journal articles. It also uses data simulation methods to determine the reliability of the assessment instrument. The results of this study indicate that (1) Indonesia's assessment based on the 2013 curriculum is an authentic assessment that includes cognitive, affective, and psychomotor aspects. The assessment process includes daily, midterm, end-of-year, and school exams, as well as national school exams. Assessment techniques are performed with tests and non-tests based on the aspects evaluated. (2) The Kappa Index is used to measure the level of reliability of the assessment instrument in accordance with the criterion assessment approach. Model one administration by calculating the z value combined with traditional reliability values (KR 1 or KR2 and Alpha Cronbach). (3) The Kappa index for the two-test model by calculating for category of the success and failure with the Kappa index equation.

Key words: reliability test, Kappa Index, Curriculum 2013 in Indonesia

1. Introduction

The 2013 curriculum in Indonesia is a curriculum that runs in schools both elementary, junior high, and high school and equivalent. The curriculum is competency based, meaning there are certain competencies that must be mastered by students when they finish learning. The 2013 curriculum refers to the Competency Based Curriculum (CBC) as a guide for the implementation of education to develop various domains of education which include knowledge, skills and attitudes (Mahmuda Kartika & Oktova, 2017, p. 9). The existence of competence as a learning outcome, then a teacher or educator is expected to use assessment instruments that have good reliability in accordance with the applicable curriculum. In addition the 2003 National Education System Law states that the assessment of learning outcomes by educators is the process of gathering information/data about student learning outcomes in aspects of attitudes, aspects of knowledge, and aspects of skills that are carried out in a planned and systematic manner undertaken to monitor the process, learning progress , and improvement of learning outcomes

Didik Setyawarno, M.Pd, Yogyakarta State University; research area: science education. E-mail: didiksetyawarno@uny.ac.id.
Atik Kurniawati, M.Pd, Yogyakarta State University; research area: biology education. E-mail: atik_kurniawati@uny.ac.id.

through the assignment and evaluation of learning outcomes (Kemdikbud, 2015). The results of monitoring and evaluating the implementation of the 2013 curriculum at the junior high school level in 2014 showed that one of the difficulties of educators in implementing the 2013 curriculum was assessment. Approximately 60% of educator respondents stated they have not been able to design, implement, process, report, and utilize the results of the assessment properly (Tim Direktorat Pembinaan SMP, 2017). Important things to consider when carrying out assessments in the 2013 Curriculum are the Mininal Completeness Criteria, predicate, remedial, and enrichment.

During this time, the minimum completeness criteria or (cut of score) is determined using policy (Mardapi, Hadi & Retnawati, 2015). For example, on the national exam in Indonesia. In 2005, the graduation limit was 4.01. For 2006 and 2007, the graduation limit was increased to 4.26, which in 2009 was further increased to 5.26. Value of 4.01, 4.26 and 5.26 is a relatively low limit compared to the graduation limit of other countries. However, the community responded with great anxiety and anxiety, and this limit was considered too high. Regulation of the Minister of Education and Culture of the Republic of Indonesia Number 53 of 2015 states that the minimum completeness criteria, hereinafter referred to as KKM, is a learning completeness criterion determined by the Education Unit that refers to graduation competency standards, taking into account the characteristics of students, the characteristics of subjects, and the condition of the Education Unit. Mininal Completeness Criteria (MCC) is very influential on student learning systems. Student learning outcomes are measured using an assessment instrument that is applied in the learning evaluation process. An instrument is said to be good if it is valid and reliable (Matondang, 2009). The assessment instruments that will be used in the evaluation of learning need to meet these requirements. The reliability of the instrument in the measurement language is known as reliability. A reliable instrument will produce good measurement results. Education in Indonesia based on the 2013 curriculum demanding the existence of material learning that is measured by the mastery test assessment instrument which is realized by the existence of value limits from Mininal Completeness Criteria.

The existence of MCC scores students inevitably have to be prosecuted in order to get a minimum grade according to the MCC that has become the school's own provisions. MCC is made to improve the quality, quality of education which later is useful to be used as a reference by each subject teacher. Given the competency-based teaching objectives, the appropriate assessment approach is the criterion reference assessment (Wakseno, 1985). Analysis of items is an important step because to determine the quality of the questions so that the questions can be used or not. For example, a good multiple choice question quantitatively needs to be considered validity, reliability, level of difficulty, distinguishing features of the problem, and the effectiveness of deception based on classical theory. The reliability of the instrument or reliability in other languages is called the internal consistency index, which for scoring politomus is the Cronbach alpha index and for dichotomous scoring is the KR-20 index. (Adams & Kho, 1996). That reliability applies to tests that have a selection function, not to measure achievement. To measure achievement or learning outcomes in schools, it needs to be converted into a Kappa index. Practices in developing learning outcomes assessment instruments currently many educators or researchers still have a tendency to use item quality analysis classically. Moreover, the reliability analysis used in the classical approach is very suitable for item analysis that will be used in the evaluation of learning outcomes based on the norm reference assessment approach. The criterion reference assessment approach has different principles, so this article discusses the relationship between criterion reference assessment with the 2013 curriculum, Kappa index for two administration model, and procedure conversion of Kappa index for one administration model from KR 20 or KR 21 and Cronbach’s alpha. Based on the description above, this article intends to examine the concept of instrument

reliability based on the 2013 curriculum in Indonesia. The concept includes aspects of evaluation and techniques. The reliability test of the assessment instrument is in accordance with the 2013 curriculum in Indonesia.

2. Research Methods

The research method used is study of literature based on the 2013 curriculum and data simulation for calculating Kappa index. Details of the research method are as follows.

- The study of literature from various sources, both books and journal articles, to examine the concept of the assessment system based on the 2013 curriculum in Indonesia.
- Data simulation method from the 2015 Physics National Examination results at Ciparay High School in West Java, Indonesia in determining the reliability of assessment instruments in accordance with the 2013 curriculum for the determination of Kappa index from one administration model which was analyzed with application of Iteman 4.
- Data simulation methods from pretest and posttest at SMP N 2 Mlati Yogykarta, Indonesia in determining the reliability of assessment instruments in accordance with the 2013 curriculum for determining the Kappa index for two administration model.

The methods are chosen because they are considered capable of discussing the main objects that will be discussed in this article.

3. Results and Discussion

3.1 Assessment Based on 2013 Curriculum in Indonesia

Every learning conducted by educators, teachers, or researchers requires an assessment and evaluation system that includes assessment methods and instruments and analysis procedures in accordance with the learning curriculum used. The curriculum applied in Indonesia is the 2013 curriculum. The 2013 curriculum is basically a competency-based curriculum with basic competencies as a minimum competency that must be achieved by students (Tim Direktorat Pembinaan SMP, 2017, p. 7). Learning evaluation activities become an integrated part of class activities. This is confirmed again that assessment as part of classroom activities is a fundamental process required to promote learning and ultimately achievement (Jones, 2005). Assessment activities are activities that are integrated in every learning process in the class conducted by the teacher. This is in line with define assessment, namely the process of gathering and processing information to measure the achievement of student learning outcomes (Muhammad, 2015).

The implementation of assessment in Indonesian schools refers to the Education Assessment Standards and other relevant assessment rules, namely criteria regarding the scope, goals, benefits, principles, mechanisms, procedures, and instruments of student learning outcomes assessment used as a basis for evaluating student learning outcomes in primary and secondary education. The 2013 curriculum which has been put in place has also undergone various revisions several times which are expected to become one of the comprehensive assessment breakthroughs. Minister of Education and Culture Regulations number 66 of 2013 states that assessment must guarantee:

- Planning assessment of students in accordance with the competencies to be achieved and based on the principles of assessment;
- Implementation of assessment of students in a professional, open, educative, effective, efficient, and in

accordance with the socio-cultural context; and

- Reporting the results of assessment of students in an objective, accountable, and informative (Kemdikbud, 2013).

This is because the assessment format that is emphasized includes: (1) the approach used is complete learning, (2) measuring what students can do, (3) carried out continuously, (4) using a variety of assessment techniques, and (5) the reference of evaluation is the criteria in the form of Basic Competence (Wuryani & Irham, 2014). The assessment component becomes a part that should not be left behind in the learning process. The assessment format is an indicator of authentic assessment. Authentic assessment is a form of assessment that requires students to display attitudes, use knowledge and skills gained from learning in carrying out tasks in real situations. Authentic assessment based on the 2013 curriculum includes the assessment of attitude competencies, knowledge competencies, and skills competencies (Wildan, 2017).

The 2013 curriculum in Indonesia basically implements activity-based learning, which is expected to produce productive, creative, innovative and affective Indonesian people through the strengthening of integrated attitudes, knowledge and skills. Implications for the implementation of the assessment include the assessment of attitudes, knowledge, and skills, which are carried out using a variety of ways, including observations, project assessments, and portfolios (Muhammad, 2015). Existing assessments at school as in the 2013 curriculum assessment guidelines include daily assessments, midterm assessments, end of semester assessments, year end assessments, and school exams, as well as national standard school exams. The types of assessment in practice use measuring instruments or instruments for assessing learning outcomes.

Reiterated again in the Minister of Education and Culture Regulations Number 32 of 2013, amendments to Government Regulation number 19 of 2005 concerning National Education Standards Article 22 paragraph 2 states that assessment techniques as referred to in paragraph (1) can be in the form of written tests, observations, practical tests, and assignments individuals or groups. With the assessment carried out through various means it is possible to be able to obtain comprehensive results, where the teacher can explore various information from students, which is then known as classroom assessment (Wildan, 2017). The assessment conducted is directed to measure the achievement of Basic Competencies in the Core Competencies that exist in the 2013 curriculum. These competencies include spiritual, social, knowledge, and skills competencies. The practice of assessment in class needs to be carried out through three approaches namely assessment of learning, assessment for learning, and assessment as learning (Muhammad, 2015). Assessment of learning is carried out to measure students' achievement of competencies that have been determined based on basic and core competencies that have been developed into assessment indicators. Assessment for learning allows teachers to use information about the condition of students to improve learning, while assessment as learning allows students to see the achievements and progress of learning to determine learning targets.

The experience from the previous curriculum in Indonesia is that the assessment is still conventional, the assessment of learning is more dominant than the assessment for learning and assessment as learning. Assessment practices in 2013 Curriculum ideally should prioritize assessment as learning and assessment for learning compared to assessment of learning. Assessment of learning is an assessment carried out after the learning process is completed with the aim of knowing the achievement of learning outcomes after students have completed the learning process that can be done on a daily, midterm, or end of semester basis. Assessment for learning in the 2013 curriculum is carried out throughout the learning process and is used as a basis for improving the learning process. The learning that is used based on the 2013 curriculum is learning that has a scientific approach approach,

so that assessment for learning is able to see the quality of learning carried out by teachers in the classroom. Various forms of formative assessment, such as class assignments, presentations, and quizzes, are examples of assessment for learning. Assessment as learning is not much different from assessment for learning that is carried out during the learning process. The main difference from the assessment is the assessment as learning actively involves students in the assessment activities (Muhammad, 2015). Self assessment and peer assessment are examples of assessment as learning. In assessment as learning students can also be involved in formulating assessment procedures, criteria, and rubrics/guidelines for assessment so that they know exactly what needs to be done in order to obtain maximum learning outcomes.

One of the assessment models that is in line with the 2013 curriculum is an authentic assessment. Authentic assessment is the process of gathering information about the development of teachers and the achievement of learning undertaken by learners through a variety of techniques that are able to reveal, or show exactly proving that the learning objectives have been completely mastered and achieved. This assessment includes four types of assessment portfolio assessment work, project assessment and written assessment. This kind of assessment is able to describe the increase of students in learning outcomes, both in order to observe, reason, try, build networks, and others. Aside from the approaches and models of assessment in learning that are relevant to the 2013 curriculum in Indonesia, student learning outcomes are processed using a criterion reference assessment approach that is by comparing students' achievements with determined competency criteria (Crocker & Algina, 2008). The results of assessment of students, both formative and summative, are not compared with the results of other students but are compared with the mastery of competencies required. To determine the achievement of Basic Competence, the teacher must formulate a number of indicators as a reference for assessment and the school must also determine minimum learning completeness or minimum completeness criteria (MCC) to decide whether a student is complete or not. MCC illustrates the quality of education units, therefore MCC needs to be evaluated annually and is expected to gradually increase in MCC.

Based on the description above it can be stated that the concept of assessment based on the 2013 curriculum in Indonesia uses an authentic assessment model, aspects of the assessment include the assessment of attitudes, knowledge, and skills. The assessment technique can be done with tests and non-tests that are adjusted to the assessment indicators described in the 2013 curriculum. Forms of assessment used include work assessments, portfolio assessments, project assessments, and written assessments. The results of the combination of all assessments will better reflect a more comprehensive assessment to look at children’s abilities objectively. In addition, the assessment approach that is relevant to the 2013 curriculum applied is more dominant in the assessment approach to learning and assessment as learning than the assessment approach to learning. Processing student learning outcomes used by teachers is criterion reference assessment that has been determined by each school.

3.2 Kappa Index for One Administration Model

An important form of measurement in behavioral and social sciences is nominal classification, that is, the assignment of subjects to qualitative categories, as in psychiatric diagnosis (Warrens, 2015). Learning evaluation activities in class refer to the 2013 curriculum that is applicable in Indonesia. The 2013 curriculum in Indonesia is a competency-based curriculum, so the assessment instruments are also adjusted to the demands of the curriculum. One of the requirements used in the student learning outcomes assessment instrument is that the instrument must have reliable power or reliability. The reliability test on the norm reference assessment is done by looking at the

internal consistency value of the measuring instrument can be calculated using the Alpha-Cronbach and Kuder-Richardson coefficient formula (KR-20 or KR-21). The norm reference assessment compares the results of one student with other students in a class that is generally divided into several groups namely upper, middle, and lower groups. Competency-based assessment instrument reliability tests differ in their analysis by looking at the Kappa index obtained in the instrument test results (Subali, 2010). Actually the kappa statistic was introduced by Cohen in 1960 (Cohen J, 1960). Determination of the Kappa index with one administering test can be done using a table of estimated values of the coefficients of the Kappa Coefficients as in Table 1 (Subkoviak, 1988).

Table 1 Estimated Value of the Kappa Coefficient

|z|	Traditional Reliability (r)								
	.10	.20	.30	.40	.50	.60	.70	.8O	.90
.00	.06	.13	.19	.26	.33	.41	.49	.59	.71
.10	.06	.13	.19	.26	.33	.41	.49	.59	*.71*
.20	.06	.13	.19	.26	.33	41	.49	.59	.71
.30	.06	.12	.19	.26	.33	.40	.49	.59	.71
.40	.06	.12	.19	.25	.32	.40	.48	.58	.71
.50	.06	.12	. 1 8	.25	.32	.40	.48	.58	.70
.60	.06	.12	.18	.24	.31	.39	.47	.57	.70
.70	.05	.11	.17	.24	.31	.38	.47	.57	.70
.60	.05	.11	.17	.23	.30	.37	.46	.56	.69
.90	.05	.10	.16	.22	.29	.36	.45	.55	.68
1.00	.05	.10	.15	.21	.28	.35	.44	.54	.68
1.10	.04	.09	.14	.20	.27	.34	.43	.53	.67
1.20	.04	.08	.14	.19	.26	.33	.42	.52	.66
1.10	.04	.08	.13	.18	.25	.32	.41	.51	.65
1.40	.03	.07	.12	.17	.23	.31	.39	.50	.64
1.50	.03	.07	.11	.16	.22	.29	.30	.49	.63
1.60	.03	.06	.10	.15	.21	.28	.37	.47	.62
1.70	.02	.05	.09	.14	.20	.27	.35	.46	.61
1.90	.02	.05	.08	.13	.16	.25	.34	.45	.60
1.90	.02	.04	.08	.12	.17	.24	.*32*	.43	.59
2.00	.02	.04	.07	.11	.16	.22	.31	.42	.58

The procedure must be carried out by the one-administration model namely:

- Calculates the standard score or z score (z) of the cutoff score from the test which is stated as a benchmark score/graduation criteria. The value (z) can be calculated using the following equation.

$$|z| = \frac{c - 0{,}5 - \bar{x}}{SD}$$

Explanation:

c = raw cut score from the test

$\bar{x}$ = average score obtained by students

SD = standard deviation

The value of 0.5 in the above equation is a continuous correction built from facts in the Approximate Value

table of the Kappa Coefficient obtained by estimating the test scores following a discrete distribution to be converted into a continuum normal distribution. The z distribution of calculation results is in the form of a normal distribution and uses an absolute sign, so that the magnitude of the coefficient of agreement or the kappa coefficient is always positive. So, by giving an absolute sign for z can use Table 1. The reliability of the r test score, which appears in Table 1 can be obtained using traditional reliability indices such as the Kuder-Richardson reliability coefficient (K-R) or alpha-Cronbach (Subkoviak, 1988).

- Calculating the value of the reliability of the traditional Alpha-Cronbach, Kuder-Richardson (KR-20 or KR-21) from the analysis of the learning outcomes of the norm approach.
- Look for the meeting point between the standard score or z score (z) with the traditional reliability value of Alpha-Cronbach, Kuder-Richardson (KR-20 or KR-21).

For example, the interpretation of the kappa coefficient is the result of the traditional reliability analysis using the Iteman 4 application as shown in Table 2.

Tabel 2 Analysis of Reliability

Score	Alpha	Split-Half (Random)	Split-Half (First-Last)	Split-Half (Odd-Even)
Scored items	0.811	0.823	0.838	0.854

Calculation of standard score or z score (z) obtained 1.20, so for the Kappa index by finding the meeting point of the standard score or z score (z) with a reliability value of 0.811 (rounded 0.80) obtained a Kappa index of 0.52. Interpretation of instrument quality from the Kappa index as shown in Table 3.

Table 3 Interpretation of Kappa's Index Coefficients

Kappa Index	Agreement
< 0.20	Poor
0.21 – 0.40	Fair
0.41 – 0.60	Moderate
0.61 – 0.80	Good
> 0.80	Very good

In addition, a test that is applied in a class over a full period (for example 1 semester) can use a kappa coefficient of 0.35 to 0.50 or more (Subali, 2010).

3.3 Kappa Index for Two Administration Model

The recommended reliability index for criterion reference assessment is the Kappa index. Kappa index is a widely used index for assessing agreement between raters (Tang, Hu, Zhang, Wu, & He, 2015). These two coefficients measure the consistency of the mastery-non-mastery classification between the two tests or the administration of the test (Viera & Garrett, 2005); Warrens, 2015). The index requires a different interpretation than the traditional reliability coefficient KR 20, KR 21, or Alpha-Cronbach which uses the item score correlation coefficient with the total score (Subkoviak, 1988). Successful fail-classifications in both test execution and test administration can be summarized as in Table 4.

Table 4 Classification of Testee in Two Administration

		Administration 2		Total
		Mastery	Non Mastery	
Administration 1	Mastery	*A*	*B*	*(A + B)*
	Non mastery	*C*	*D*	*(C + D)*
	Total	*(A + C)*	*(B + D)*	*N*

The coefficient of agreement is the proportion of test takers who consistently belong to the groups that succeed and fail from both administering test tests. The coefficient of approval designates the po value calculated by the following formula (Cohen J, 1960; Warrens, 2015; Viera & Garrett, 2005).

$$\boldsymbol{p_o} = \frac{(\boldsymbol{A} + \boldsymbol{D})}{\boldsymbol{N}}$$

Explanation:

A = the total of mastery group test takers in both test administrators

D = the total of non mastery group test takers in both test administrators

N = the total of testee

The upper limit of the approval coefficient is 1.00, which is reached if the classifications in the two test administrators are consistent for all test takers in each group. The lower limit of the coefficient of agreement can be calculated by the following formula (Cohen J, 1960; Warrens, 2015; Viera & Garrett, 2005).

$$\boldsymbol{p_{chance}} = \frac{[(\boldsymbol{A} + \boldsymbol{B})(\boldsymbol{A} + \boldsymbol{C}) + (\boldsymbol{C} + \boldsymbol{D})(\boldsymbol{B} + \boldsymbol{D})]}{\boldsymbol{N^2}}$$

Explanation:

B = the number of group test takers that mastery in administering the first test but non mastery in administering the second test

C = the number of group test takers that non mastery in administering the first test but mastery in administering the second test

The lower limit (p_{chance}) is the proportion of classifications that are coincidentally consistent if the results of the mastery-nonmastery on the second test administration are completely independent of the first test administrator where the value $p_{chance} \geq 0.50$. The lower limit the kappa coefficient (K) can be calculated using the following formula.

$$\boldsymbol{K} = \frac{\boldsymbol{p_o} - \boldsymbol{p_{chance}}}{\boldsymbol{1} - \boldsymbol{p_{chance}}}$$

The kappa coefficient is the proportion of consistent classifications that are in line with expectations and which are accidental. The upper and lower limits of the kappa coefficient are 1.00 and 0.00, which occur when the results on both administering the test are consecutively in complete or completely free agreement. Examples of application to the two-administration model are administration 1 and administration 2 with data as shown in Table 5.

Table 5 Item Answer Two Administration Model

Testee	Answers to Administration 1					Testee	Answers to Administration 2				
	1	2	3	4	5		1	2	3	4	5
1	0	1	1	0	1	1	1	1	1	1	1
2	1	1	1	0	1	2	1	1	1	1	1
3	0	1	0	1	0	3	1	1	1	1	1
4	1	0	1	1	1	4	1	1	1	1	1
5	1	1	1	1	1	5	0	1	1	0	1

Explanation: 0 = non mastery; 1 = mastery.

The analysis procedure is as follows.

- Grouping the categories of each item according to Table 4. The illustration can be seen in Table 6.

Table 6 Classification of Answers

Testee	Item	Administration 1	Administration 2	Category
1	1	0	1	C
2		1	1	A
3		0	1	C
4		1	1	A
5		1	0	B
1	2	1	1	A
2		1	1	A
3		1	1	A
4		0	1	C
5		1	1	A
1	3	1	1	A
2		1	1	A
3		0	1	C
4		1	1	A
5		1	1	A
1	4	0	1	C
2		0	1	C
3		1	1	A
4		1	1	A
5		1	0	B
1	5	1	1	A
2		1	1	A
3		0	1	C
4		1	1	A
5		0	0	D

Summing each category of all items. Based on Table 6 the number of each category is obtained for each item as in Table 7.

Table 7 Recapitulation Item

Item	Category	Total
1	A	2
	B	1
	C	2
	D	0
2	A	4
	B	0
	C	1
	D	0
3	A	4
	B	0
	C	1
	D	0
4	A	2
	B	1
	C	2
	D	0
5	A	3
	B	0
	C	1
	D	1

- Input the value of each category to Ms. Excel to calculate the value of P_o, P_{chance}, and K. The results of the analysis of point (b) as in Table 8.

Table 8 Number of Categories

Category	Item					Total
	1	2	3	4	5	
A	2	4	4	2	3	15
B	1	0	0	1	0	2
C	2	1	1	2	1	7
D	0	0	0	0	1	1

Explantion: each testee does 5 items, so 5 testees do 25 items (N = 25 and N2 = 625).

$$p_o = \frac{(A+D)}{N} = \frac{(12+1)}{25} = 0.64$$

$$p_{chance} = \frac{[(A+B)(A+C)+(C+D)(B+D)]}{N^2}$$

$$p_{chance} = \frac{[(15+2)(15+7)+(7+1)(2+1)]}{25^2}$$

$$p_{chance} = 0.6368$$

$$K = \frac{p_o - p_{chance}}{1 - p_{chance}} = \frac{0.64 - 0.6368}{1 - 0.6368} = 0.008811$$

The results of the analysis obtained K = 0.008811 with low instrument reliability interpretation.

4. Conclusion

Based on the results of the discussion, the conclusions are (1) assessment based on the 2013 curriculum in Indonesia uses an authentic assessment model, aspects of the assessment include attitudes, knowledge, and skills. The assessment technique can be done with tests and non-tests that are adjusted to the assessment indicators described in the 2013 curriculum. Forms of assessment used include work assessments, portfolio assessments, project assessments, and written assessments. The results of the combination of all assessments will better reflect a more comprehensive assessment to look at student's abilities objectively. In addition, the assessment approach that is relevant to the 2013 curriculum applied is more dominant in the assessment approach for learning and assessment as learning than in the assessment of learning approach. The processing of student learning outcomes used by teachers is a criterion reference assessment that have been determined by each school. (2) Kappa index is used to measure the level of reliability of the assessment instrument in accordance with the criterion reference assessment approach. One administration model can be done by calculating the z value combined with traditional reliability values (KR 1 or KR2 and Alpha Cronbach). (3) Kappa index for the two administration models by calculating the mastery and non mastery category with the Kappa index equation.

References

Adams R. J. and Kho S. T. (1996). Acer quest version 2.1. Camberwell, Victoria: he Australian Council for Educational Research.

Cohen J. (1960). "A coefficient of agreement for nominal scales", *Educational and Psychological Measurement*, Vol. 20, pp. 37–46.

Crocker L. M. and Algina J. (2008). *Introduction to Classical and Modern Test Theory*, Mason (OH): Cengage Learning.

Jones C. A. (2005). *Assessment for Learning*, London: Learning and Skills Development Agency.

Kemdikbud (2015). *Peraturan Menteri Pendidikan dan Kebudayaan Republik Indonesia Nomor 53 tahun 2015 Tentang Penilaian Hasil Belajar oleh Pendidik dan Satuan Pendidikan Pada Pendidikan Dasar dan Pendidikan Menengah*, Jakarta: Kemdikbud.

Kemdikbud (2013). *Peraturan Menteri Pendidikan dan Kebudayaan Republik Indonesia No. 66 Tahun 2013 tentang Standar Penilaian Pendidikan*, Jakarta: Kemdikbud.

Mahmuda A., Kartika I. and Oktova R. (2017). "Pengembangan dan Uji Coba Instrumen Penilaian Hasil Belajar IPA SMP/MTS Kelas VII Pada Materi Karakteristik Zat", *Berkala Fisika Indoneia*, Vol. 9, No. 1, pp. 8–15.

Mardapi D., Hadi S. and Retnawati H. (2015). "Menentukan Kriteria Ketuntasan Minimal Berbasis Peserta Didik", *Jurnal Penelitian dan Evaluasi Pendidikan*, Vol. 19, No. 1, pp. 38–45, doi: https://doi.org/10.21831/pep.v19i1.4553.

Matondang Z. (2009). "Validitas dan Reliabilitas Suatu Instrumen Penelitian", *Jurnal Tabularasa, PPS UNIMED*, Vol. 6, No. 1, p. 11.

Muhammad H. (2015). *Panduan Penilaian Untuk Sekolah Menengah Atas*, Jakarta: Kemdikbud.

Subali B. (2010). *Panduan Praktikum Penilaian, Evaluasi, dan Remediasi Hasil Belajar Biologi*, Yogyakarta: FMIPA UNY.

Subkoviak M. J. (1988). "A practitioner's guide to computation and interpretation of reliability indices for mastery tests", *Journal of Educational Measurement*, Vol. 25, pp. 47–55, doi: https://doi.org/10.1111/j.1745-3984.1988.tb00290.x.

Tang W., Hu J., Zhang H., Wu P. and He H. (2015). "Kappa coefficient: A popular measure of rater agreement", *Biostatistics in Psychiatry*, Vol. 27, No. 1, pp. 62–67.

Tim Direktorat Pembinaan SMP. (2017). *Panduan Penilaian oleh Pendidik dan Satuan Pendidikan*, Jakarta: Kementerian Pendidikan dan Kebudayaan.

Viera A. J. and Garrett J. M. (2005). "Understanding interobserver agreement: The kappa statistic", *Family Medicine*, Vol. 37, No. 5, pp. 360–363.

Wakseno I. (1985). "Penelaahan Kembali Startegi Penilaian Acuan Norma (PAN) dan Penilaian Acuan Patokan (PAP) sebagai Pendekatan dalam Penilaian Hasil Belajar", *Cakrawala Pendidikan*, Vol. IV, No. 1, pp. 22–37.

Warrens M. J. (2015). "Five ways to look at Cohena's Kappa", *Journal of Psychology & Psychotherapy*, Vol. 5, No. 4, doi: https://doi.org/10.4172/2161-0487.1000197.

Wildan W. (2017). "Pelaksanaan penilaian autentik aspek pengetahuan, sikap dan keterampilan di sekolah atau madrasah", *Jurnal Tatsqif*, Vol. 15, No. 2, pp. 131–153, doi: https://doi.org/10.20414/jtq.v15i2.3.

Wuryani W. and Irham (2013). Penilaian dalam Perspektif Kurikulum 2013, in: *Sistem Penilaian dalam Kerangka Pendidikan Karakter Sebagai Implementasi Kurikulum 2013 di Kota Rantau*, Kabupaten Tapin, Kalimantan Selatan, p. 19.