JOURNAL of Mathematics Education
Website: http://usnsj.com/index.php/JME
Email: *ahmad.rustam1988@gmail.com*

INTER-RATER RELIABILITY OF THE MATHEMATICS TEST INSTRUMENT AT JUNIOR SCHOOL

AUTHORS INFO

Destiniar
Universitas PGRI Palembang
destiniarpgri@yahoo.co.id

ARTICLE INFO

ISSN: 2528-2026
Vol. 2 No. 1, June 2017
URL: http://usnsj.com/index.php/JME/article/view/2.1.17-23

Abstract

The purpose of this research was to know whether or not there was difference of interrater reliability coefficient about mathematics test which evaluated by 18 raters and 12 raters and analyzed using Fleiss Kappa method. Rater in this study as many as 60 people consisting of 20 mathematics teachers of the Junior School (SMP), 20 lecturers of Mathematics Education, and 20 lecturers of FMIPA mathematics. Instruments assessed in the form of mathematical test instrument for class IX with a simple multiple choices. The reserach method used in the form of experimental methods and data obtained were analyzed by using t-test. The result obtained there was difference of interrater reliability coefficient about mathematics test which assessed by 18 raters and 12 raters. The interateral reliability coefficient of the class IX mathematics test analyzed using the Fleiss Kappa method and rated by 18 raters was higher than that assessed by 12 raters.

Keywords: *inter-rater reliability, Fleiss Kappa method, instrument for mathematics, test for instruments*

A. Introduction

In the field of education, measurement is very important. Measurement of learning outcomes will be required for the test result instrument. Furthermore, a test instrument of learning outcomes is considered good when it meets the valid, reliable, and usable criteria. (Kusaeri, 2014: 50). Widoyoko (2015: 129) said that a test instrument is said to be valid if the test measures exactly what to measure. Djanuarsih (2012) said the reliability of measuring instruments is the accuracy or precise measuring instrument in measuring what should be measured. A test is said to be reliable if it is used repeatedly with relatively similar conditions, then the results obtained will also remain the same (consistent). Kusaeri (2014: 51) a usable instrument has the meaning that the valuation used is practically the procedure.

The most commonly used test form is a multiple-choice test because it is easy to measure various aspects. Widoyoko (2015: 129) says that a test instrument is said to be valid if the test measures exactly what it wants to measure. A test is said to be reliable if it is used repeatedly with relatively similar conditions, then the results obtained will also remain the same (consistent).

Fleming & Judith (2004: 39) states that inter-rater reliability involves two or more assessors to assess the same instrument. The reliability coefficient obtained from the results of the assessment of the rater is more meaningful to the consistency of the rater (inter-rater reliability). The consistency of the rater in assessing the suitability of an instrument called the interrater reliability coefficient is more referring to the validity of content or content validity.

In a previous study, many raters were used differently. In the study of van Daalen et al. (2009) and Tsuchiya et al. (2013) showed the using of two observers (raters), a study conducted by Hansson, E., et al. (2014), research conducted by Khaliq (2011) using three raters, research by Afrizal et al. (2015) using four raters, while research conducted by Joseph L. Fleiss using six raters. Research conducted by Ahmad (2015) using 20 raters. So, it appears a problem if many raters are used differently, whether the interrater coefficient of reliability is also different. By considering the previous researches, the current research decided to use 12 raters and 18 raters. Determination using 12 raters and 18 raters is only the researcher's judgment.

Nitko (1996:72-73) said that the method of estimating the coefficient of reliability should also be considered. The method that can be used to calculate the inter-rater reliability coefficient is some. Multon (2010: 2) recommends Fleiss Kappa method to get more than two consistent reviewers. The objective in this current research is to investigate whether or not there any difference of inter-rater reliability coefficient between 18 raters and 12 raters analyzed using Fleiss Kappa method.

B. Literature Review

Widoyoko (2015: 45) states that the test is one tool to make measurements, namely tools to collect information characteristics of an object. Meanwhile, according to Kerlinger (1973: 492) says that the test is a set of stimuli given to a person with the intention to get answers that can be used as a basis for determining the numbers or be used as a basis of what is measured. Kizlik (2012: 3) says the test is a method used to determine the ability of students to complete certain tasks or show mastery of skills or knowledge about the content.

From some opinions mentioned before, it can be concluded that the test is a tool used to make measurements about certain aspects and in order to obtain good measurement results then the test instruments should be well prepared.

An instrument can be said to be good if the instrument is valid, reliable, and usable. One way that can be used is to use a rater. Widhiharso (2010) says that involving rater will ensure that the items we make are relevant to what we measure and represent the overall domain. Hariansyah (2013) also said that involving rater can improve the quality of measuring instruments developed. Determining who the rater is, someone should have the same background while many rater may vary. The use of rater in providing an assessment of an instrument can improve the quality of the instrument. The reliability coefficient obtained by using the rater is also called the interrater reliability coefficient.

On reliability of inter-rater, it is to test the consistency of the raters. According to Azwar, (2015: 88) if the rating is done by several raters then the reliability value of the rating result is more consistent among the raters (inter-rater reliability). Multon (2010: 1) states that inter-rater of reliability is currently possible to use more than two rater.

McQuillian (2001: 49) said the reliability of inter-rater also called inter-assessor reliability refers to the extent to which two or more rater agree and express in the form of correlation between raters. Crocker & Algina (2008: 143) say that in assessing an instrument will be more interesting if using consistency of more than two assessors. According to Multon (2010: 2) who says that if the subject is judged by many different raters while the same valued object is recommended using the Fleiss Kappa method.

C. Methodology

1. *Design*

This research is a quantitative research with comparative method. This research was done to know the difference of inter-rater reliability coefficient between 18 raters and 12 raters. After that, the results were analyzed statistically. The research method used in the form of experimental method. Experiments in this study were conducted after the raters gave an assessment on the mathematical test instrument. The design or treatment design used is as follows:

Table 1. Research design

Method	Number of Rater	
Fleiss Kappa	12 raters	18 raters

2. *Setting*

This research was conducted from June to September 2016 at FMIPA Sriwijaya University, FKIP Sriwijaya University, FKIP University PGRI Palembang, SMP Negeri 2, SMP Negeri 18, SMP

Negeri 43, and SMP PGRI 11 Palembang. This research was done in different places, this was because the raters were 20 of lecturers of mathematics FMIPA, 20 lecturers of mathematics education and 20 mathematics teachers.

3. *Instrument*

The instrument used was a class IX mathematics test instrument in the form of multiple-choice as much as 40 questions with 4 options of answers. This math test instrument was a teacher-made instrument.

4. *Procedure*

In this study, all rater were asked to provide an assessment on the test instrument. Rater who numbered 60 respondents were spread in various places that were 20 raters of lecturers of mathematics FMIPA in Sriwijaya university, 7 raters from lecturers of mathematics education at FKIP Sriwijaya university, 13 raters from lecturers of mathematics education FKIP PGRI Palembang, 6 raters from mathematics teacher SMP Negeri 17 Palembang, 5 raters from mathematics teacher of SMP Negeri 2 Palembang, 4 raters from mathematics teacher of SMP PGRI 11 Palembang, and 5 raters were from mathematics teacher of SMP Negeri 43 Palembang.

The results of the assessment of 60 raters was grouped into 3 i.e. 20 assessments of the raters derived from mathematics lecturers FMIPA, 20 assessments of lecturers from mathematics education and 20 assessments of junior mathematics teachers. After that, they were randomly taken as 4 grades from teachers, 4 grades from lecturers of mathematics education, and 4 grades from lecturers of mathematics FMIPA, so obtained 12 assessments from 12 raters. Then, the 12 raters were analyzed using Fleiss Kappa method to obtain one value of inter-rater reliability coefficient. In other word, it was done as many as 20 repetitions so that obtained 20 data coefficients of inter-rater reliability analyzed using Fleiss Kappa method for 12 raters.

In the same way, it was also done for the 18 rater assessments, each of which was 6 ratings from each group and repeated 20 times so that obtained 20 scores of inter-rater reliability coefficient analyzed using Fleiss kappa method for 18 raters.

5. *Technique of Data Collection*

Data were collected by asking the raters to see the suitability of some aspects. The conformity of some aspects was the suitability of KD / indicators with question indicators, the conformity of the indicator about the item, the questions were formulated briefly and the workmanship of the item was clearly written, the accuracy of the use of standard Indonesian language, and the problem of using communicative language was easy to understand and not giving rise to multiple interpretations. After that, the raters were asked to give an assessment of the score on the instrument of class IX mathematics test with the provision of score 1 if only two criterias appear, score 2 if only 3 criterias appear, score 3 if only 4 criterias appear, and score 4 if more than four criterias appear. Furthermore, the scores obtained were analyzed using Fleiss Kappa method.

6. *Technique of Data Analysis*

Before the data were analyzed, prerequisite tests were tested for normality and homogeneity. Normality test and homogeneity were done with assisted program SPSS version 20. After that, for the test difference was applied a t-test that was also an assisted program SPSS version 20.

D. Finding and Discussion

1. *Findings*

The results obtained in this study are 2 groups of data, namely the data group assessed by 18 raters as many as 20 data and data group assessed by 12 raters also as many as 20 data. The following data is presented in tabular form on Table 2.

Table 2. Data of Inter-rater Reliability Coefficients

No.	FK 12	FK 18	No.	FK 12	FK 18
1	0.0643	0.0771	11	0.0754	0.0756
2	0.0616	**0.0614**	12	0.0614	0.0813
3	0.0657	0.0794	13	0.0674	0.0786
4	0.0595	0.0714	14	0.0855	0.0717

5	0.0643	0.0717	15	0.0680	0.0759
6	0.0684	0.0757	16	0.0746	0.0786
7	0.0680	0.0773	17	0.0744	0.0839
8	0.0717	0.0721	18	0.0692	0.0802
9	0.0695	0.0781	19	**0.0921**	0.0734
10	0.0629	0.0862	**20**	**0.0811**	**0.0730**

Data obtained from 18 raters is 1.5226 and has an average value of inter-rater reliability coefficient of 0.07613. The highest inter-rater reliability coefficient for 18 raters is 0.0862 and the lowest is 0.0614. So, the range is 0.0248. As for data from 12 raters obtained the amount of reliability coefficient inter-rater for 1.4050 and the average value of inter rater reliability coefficient of 0.07025. The highest interconnect reliability coefficient for 12 raters is 0.0921 and the lowest is 0.0595 so that the range is 0.0326. To see the distribution of data from both groups is presented in the boxplot on figure 1.

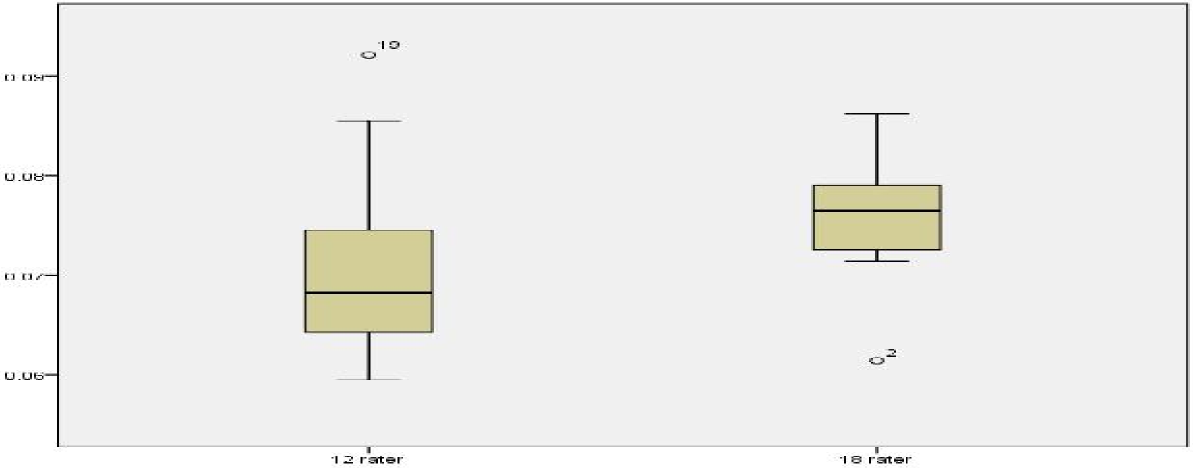

Figure.1 Initial reliability coefficient data

From Figure 1, it can be seen that there are two groups of data data that is data group for 18 raters and for 12 raters. In the boxplot, drawings of both groups there are several similarities and differences. The medians of these two distinct data groups are visible from the lines on the box or box. The median in the 18-rater data group was higher than the median on a 12-rater data cluster. The Median in the 18 rater's data group is 0.0758 while the median in the 12-rater group is 0.0682. The length of the box for 2 groups of data is also not the same. Box for 12 rater data is longer than the box for 18 rater data. The length of this box is determined by the larger quartile range this box indicates that the data is spreading. Quartile range for data of 18 rater equal to 0.0071 while for data of 12 rater equal to 0,0102. Comparison of data symmetry, if the data is symmetrical, the median line will be in the middle of the box and the whisker at the top and bottom will have the same length. If the data is not symmetrical (skew), the median will not be in the middle of the box and one of the whiskers is longer than the other.

Based on this, it appears that the data for 12 raters and 18 raters are not symmetrical but extends right. For data 12 raters, there is one data that is data number 19 which is at the top of Whisker indicating that data is data outlier, whereas for data 18 raters on data number 2 is at bottom of Whisker so this data also said outlier.

Filzmoser (2005) satets that data outlier can allow for bias on estimated parameters, by eliminating outlier data will not eliminate the information to be measured. Referring to the opinion of Filzmoser above, then the outlier data is removed from the analysis so that the existing data of each group there are as many as 19 data. The following data are presented in tabular form on Table 3:

Table 3. Data of Inter-rater Reliability Coefficients

No.	FK 12	FK 18	No.	FK 12	FK 18
1	0.0643	0.0771	11	0.0754	0.0813
2	0.0616	0.0794	12	0.0614	0.0786
3	0.0657	0.0714	13	0.0674	0.0717
4	0.0595	0.0717	14	0.0855	0.0759
5	0.0643	0.0757	15	0.0680	0.0786
6	0.0684	0.0773	16	0.0746	0.0839

7	0.0680	0.0721	17	0.0744	0.0802
8	0.0717	0.0781	18	0.0692	0.0734
9	0.0695	0.0862	20	0.0811	0.0730
10	0.0629	0.0756			

Data obtained from 18 raters is amounted to 1.4612 and has an average value of inter rater reliability coefficient of 0.0769. The highest inter-rater reliability coefficient for 18 raters is 0.0862 and the lowest is 0.0714 so the range is 0.0148. As for data from 12 raters obtained the amount of coefficient reliability inter rater of 1.3129 and the average value of inter rater reliability coefficient of 0.0691. The highest inter-reliability coefficient for 12 raters is 0.0855 and the lowest is 0.0595 so obtained a range of 0.0260. Below are two data groups in the boxplot on Figure 2.

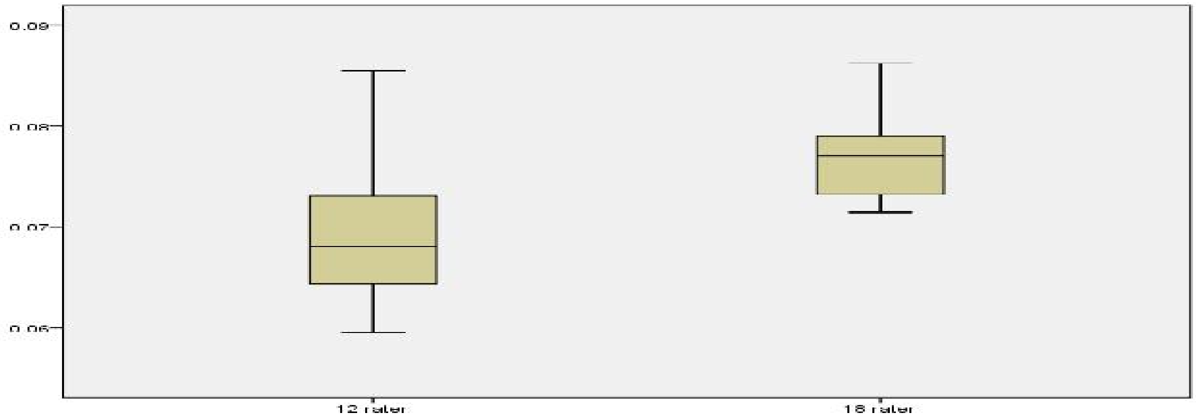

Figure 2. Data Coefficient of Reliability

From Figure 2, it can be seen that the median in the 18-rater data group is higher than the 12-rater data group. The Median for the 18 rater data group is 0.0771 while for 12 rater is 0.0680. The length of this box is determined by the larger quartile range this box indicates that the data is spreading. Quartile range for data of 18 rater equal to 0,0064 while for data of 12 rater equal to 0,0101 so it can be said that box for 12 rater data is longer than box for data 18 rater. The two groups of data are not symmetrical but extending this right is seen from whiskers longer up. In the Figure 2 there is no longer data that outlier, both in the group 18 raters and the 12 raters so that existing data can be analyzed.

Prior to t-test, normality and homogeneity tests should be performed. Normality data test is done by using SPSS version 20. The result is as follows:

Table 4. Normality Test

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
12 rater	.161	19	.200*	.940	19	.260
18 rater	.114	19	.200*	.947	19	.353

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

From table 4, it can be seen that on 12 rater sig. data = 0,200> 0,05 so it can be said as normal distributed data. For data of 18 rater sig. = 0,200> 0,05 then data is normal distribution. From the table, it can be concluded that both groups of data are normally distributed. Furthermore, conducting the data homogeneity test. This homogeneity test is also done by using SPSS version 20. The results can be seen in table 5 below:

Table 5. Homogeneity Test

Levene's Test of Equality of Error Variances[a]			
Dependent Variable: FK			
F	df1	df2	Sig.
2.488	1	36	.123

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept + VAR00002

From table 5, it can be seen that F = 2,488; Df1 = 1; Df2 = 36 and sig. or p-value = 0.123> 0.05 then the data is said to be homogeneous. From prerequisite test result, it can be concluded that the data is normal and homogeneous so it is continued with t test to see the difference of interrater reliability coefficient about math test analyzed by Fleiss kappa method between 18 raters and 12 raters. The statistical t-test is also done with the assistance of SPSS version 20, and its results can be seen in table 6 below:

Table 6. Group of Statistic

Group Statistics					
	Combination	N	Mean	Std. Deviation	Std. Error Mean
FK	1	19	.069100	.0067803	.0015555
	2	19	.076905	.0042056	.0009648

From table 6 above, it can be seen that the mean for 12 rater = 0.069 whereas for 18 rater = 0.0769. Mathematically, it is clear that the inter-rater reliability coefficient for 18 raters is higher than 12 rater. However, this difference should also be reviewed statistically, for that can be seen in the Independent sample test table in table 7 below:

Table 7. Independent Sample Test

Independent Samples Test										
		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
FK	Equal variances assumed	2.488	.123	-4.264	36	.000	-.0078053	.0018304	-.0115176	-.0040930
	Equal variances not assumed			-4.264	30.065	.000	-.0078053	.0018304	-.0115432	-.0040674

In Table 7, note on the Equal variance assmed line, this is because the data is normal and homogeneous. From table 7 above it can be seen that t = -4.264; Df = 36 and sig. (2 tailed) = 0,000 / 2 = 0,000 <0.05 which means that the interrater reliability coefficient of mathematical test analyzed by Fleiss Kappa method for 18 rater is higher than 12 rater.

2. *Discussion*

From the calculation results, it is found that the mean of the inter-rater reliability coefficient for 18 raters is 0.0769 while for 12 raters is 0.0691. It can also be seen in figure 2, shown that the data group for 18 raters is higher than 12 raters. Median data group of 18 raters is higher than 12 raters. From these data, mathematically, it is said that the reliability coefficient of inter-rater of 18 raters is higher than 12 raters.

Statistically, it can be seen from table 6 and 7 above that t = -4.264; Df = 36 and sig. (2 tailed) = 0,000 / 2 = 0,000 <0.05 which means that the inter-rater reliability coefficient of mathematical test analyzed using Fleiss Kappa method for 18 raters is higher than 12 raters. This is in line with the opinion of Azwar (2015: 88) which says that the more rater used the inter-rater reliability coefficient will be better or more accurate it is.

Menurut Naga (2013: 225), says that the Spearman-Brown prophecy which states if the test is extended then the reliability coefficient is higher with the requirement of equal parity. Interrater reliability is meant in this research is consistency rater. So, the item position is replaced with the position of the person (rater), so the more rater used, the inter-rater reliability coefficient will increase. The equivalent rater here is that all rater have the same background of mathematics.

Dari pendapat widhiarso dan Hariansyah (2013) says that involving rater can improve the quality of measuring instruments. This study used 18 raters and 12 raters. According to Dragon (2013: 225) and Azwar (2015: 88) that the inter-rater reliability coefficient will be higher if the rater used has more and more background. In this study, the raters have the same background of mathematics, so it can be said that the coefficient of inter-rater reliability of mathematical tests analyzed using Fleiss Kappa method for 18 raters is higher than 12 raters.

E. Conclusion

From the results of data analysis, it can be concluded that there are differences in the reliability coefficient of the interrater mathematical test instrument analyzed using Fleiss kappa method that was assessed by 18 raters and 12 raters. The inter-rater reliability coefficient of the mathematical test instrument analyzed using Fleiss kappa method and rated by 18 raters is higher than that assessed by 12 raters.

References

Ahmad, Nurul Q. (2015). *Pengaruh Pendekatan Pembelajaran dan Belief tentang IPA Terhadap Kemampuan Penalaran IPA*. Jurnal Pencerahan Volume 9, Nomor 1 (Maret 2015) ISSN: 1693–7775.

Azwar, S. (2015). *Reliabiltas dan Validitas*. Yogyakarta: Pustaka Pelajar.

Crocker, L. & Algina, J. (2008). *Introduction to Classical and Modern Test Theory.* Printed in The United of Amerika.

Djanuarsih, Eri. (2012). *Validitas dan Reliabilitas Butir Soal. E-Jurnal Dinas Pendidikan Kota Surabaya*; Volume 1. No.1. ISSN: 2337-3253.

van Daalen, E. et al. (2009). Inter-rater reliability and stability of diagnoses of autism spectrum disorder in children identified through screening at a very young age. *Eur Child Adolesc Psychiatry 18:663–674, DOI 10.1007/s00787-009-0025-8. 2009.*

Hansson, Eva E., et al. (2014). Inter-rater Reliability and Agreement of Rubrics for Assessment of Scientific Writing. *Education 2014, 4(1): 12-17 DOI: 10.5923/j.edu.20140401.03.* 2014*.*

Fleming, A., & Judith, A. (2004). Comparasion of two Methods of Determining Interrater Reliability. *AssessmmentFor Effective Intervention*.

Hariansyah. (2013). Reliabilitas Inter Rater. Accessed on https://hanriansyahjaya*.wordpress. Com /2013/06/22/reabilitas-inter-reter/*2013.

Kerlinger, Fred N. (1973). *Foundations of Behavioral Research*. *Second Edition* New York: Holt, Rinehart and Winston, Inc.

Khaliq, T. (2011). *Reliability of Results Produced Through Objectively Structured Assessment of Technical Skills (OSATS) for Endotracheal Intubation (ETI).*

Kizlik, B. (2012). *Measurement, Assessment, and Evaluation in Education*.

Kusaeri, K. (2014). *Acuan dan Tehnik Penilaian Proses dan Hasil Belajar dalam Kurikulum 2013*. Yogyakarta: Ar-Ruzz Media.

Multon, Karen D. (2010). *"Interrater Reliability", Encyclopedia Of Research Design*.

Naga, Dali S. (2013). *Teori Sekor pada Pengukuran Mental Edisi Kedua*. Jakarta: PT. Nagarani Citrayasa.

Nitko, Anthony J. (1996). *Educational Assessment of Sudent, 2nd Edition.* New Jersey Prentice Hall, Inc. A Simon & Schuster Company, Englewood Cliffs.

Filzmoser, P. (2005). *"Indentification of Multivariate Outliers: A Performance Study," Australian Journal of Statistiks, Vol. 35, No. 2.*

Afrizal, S., Hakiem, N., & Sensuse, Dana I. (2015). Analisis Kesiapan Implementasi E-Goverment Pada Direktorat Jenderal Penyelenggaaraan Haji Dan Umrah Kementerian Agama Republik Indonesia. *Journal of Information Systems, Volume 11, Issue 2.*

McQuillian, S. (2001)*. Inter-Rater Reliability Testing For Utilization Management Staff.* USA.

Tsuchiya, Kenji J. dkk. (2013). Reliability and Validity of Autism Diagnostic Interview-Revised. *Japanese Version, J Autism Dev Disord 43:643–662 DOI 10.1007/s10803-012-1606-9.2013*.

Widhiharso, W. (2010). *Melibatkan Rater dalam Pengembangan AlatUkur.* Fakultas Psikologi Universitas Gadjah Mada.

Widoyoko, S. Eko P. (2015). *Evaluasi Program pembelajaran*. Yogyakarta: Pustaka Pelajar.