Beta: Jurnal Tadris Matematika, 12(1) 2019: 1-13
DOI 10.20414/betajtm.v12i1.266

The validity and inter-rater reliability of project assessment in mathematics learning

Kusaeri, Sutini, Suparto, Faiqotul Wardah[1]

Abstrak: Subjektivitas dan kurang konsistennya guru/*rater* dalam proses penyekoran merupakan kritik yang umum ditujukan pada penilaian proyek dalam pembelajaran matematika. Oleh karena itu, artikel ini menyajikan hasil uji validitas konstruk dan reliabilitas *inter-rater* instrumen penilaian projek. Instrumen yang dilengkapi rubrik tersebut digunakan untuk menilai tugas proyek siswa kelas VIII SMP pada materi relasi dan fungsi. Tugas diambil dari buku matematika yang digunakan di sekolah. Instrumen diujicobakan pada 10 *raters*/guru dan 94 siswa di tiga sekolah berbeda di Kota Surabaya dan Kabupaten Gresik. Data dikumpulkan melalui lembar penilaian projek yang dilengkapi rubrik penilaian sebagai pedoman guru dalam melakukan penskoran. Validitas konstruk dianalisis dengan menggunakan *Confirmatory Factor Analysis,* sedangkan reliabilitas *inter-rater* dianalisis dengan *Intraclass Correlation Coefficient.* Hasil uji validitas menunjukkan bahwa instrumen penilaian yang digunakan tidak valid secara konstruk. Ketidakvalidan ditandai dengan perbedaan banyaknya faktor hasil konstruksi awal dengan hasil uji empiris. Dari sisi reliabilitas *inter-rater,* instrumen penilaian proyek yang digunakan reliabel. Temuan ini mengindikasikan perlunya dilakukan pengujian instrumen penilaian non tes pada buku matematika, sehingga aspek-aspek dalam lembar penilaian menjadi valid dan reliabel.

Kata Kunci: *Validitas*, *Reliabilitas*, *Inter-rater*, *Penilaian proyek*, *Tugas proyek*

Abstract: A common criticism of project assessment is the subjectivity and inconsistency of raters in scoring. In the present article, we provide the result of validity and inter-rater reliability test of the project assessment instrument. The instrument with a rubric was used to assess students' project task in grade eight for function and relation topic. The task was adopted from mathematics textbooks used in the schools. The instrument has been tested to 10 raters/teachers and 94 grade eight students from three schools (in Surabaya and Gresik). Data were collected through the project assessment sheet along with its rubric as the scoring guidance for the teachers. Construct validity was analyzed through confirmatory factor analysis, while a reliability test was conducted by using inter-rater reliability method with *the Intraclass Correlation Coefficient.* The result of the validity test showed that the instrument did not fulfill the criteria of construct validity. It is indicated by the different number of factors between the initial construction and the empirical test result. In term of inter-rater reliability, the instrument is highly reliable. The findings indicate the need for testing the non-test assessment instrument provided on mathematics textbooks, so the aspects of its assessment sheet fulfill the valid and reliable criteria.

Keywords: *Validity*, *Reliability*, *Inter-rater*, *Project assessment*, *Project task*

A. Introduction

The approach in mathematics assessment continues to change (Dochy, Gijbels, & Segers, 2006; Jonsson & Svingby, 2007; Murtiyasa, 2015), along with the necessity of using various assessment techniques to capture students' mathematics abilities and potentials in the 21st-century context. It is

[1] Universitas Islam Negeri (UIN) Sunan Ampel Surabaya, Jl. A. Yani 117 Surabaya – Indonesia, kusaeri@uinsby.ac.id

due to the assessment in mathematics learning relies more on tests. The changes were triggered by two things: the constantly changing of school mathematics curriculum and the development of mathematics learning theories (Clarke, 1996). Changes in mathematics learning assessment have long been the case in China (Chen, 2011), Malaysia (Chan & Abdullah, 2018) and Australia (Watt, 2005), which were fueled by the global trend in the assessment of mathematics learning.

Changes in assessment at the secondary school level also occurred in Korea, following the change of its curriculum implemented in 2009 (Park & Park, 2012), namely the assessment in mathematics learning should use various ways so that it can capture students' attitude, potentials, and skills in mathematics. Indonesia experiences a similar condition since the enactment of the Curriculum 2013. The demands of the assessment area are more complex. It is not merely focusing on extracting students' knowledge (Kusaeri, 2014). The assessment is led to capture students' skills and attitudes (Sugiyanto, Kartowagiran & Jailani, 2015). In the aspect of knowledge, it is more directed at exploring the potential of students' critical thinking and problem-solving abilities (Kusaeri & Aditomo, 2019). In other words, the implementation of Curriculum 2013 in Indonesia requires an authentic assessment for learning (Sukmasari & Rosana, 2017).

Authentic assessment is an assessment that requires students to perform real-world tasks and shows the essence of knowledge and skill implementation (Mueller, 2005). Retnawati (2015) explains that with authentic assessment, student competency is assessed comprehensively, either knowledge, skills, and attitude or a combination of them. Through authentic assessment, more complete data about the portrait of students' abilities based on the learning sequence they have experienced can be gathered (O'Neill, Huntley, & Race, 2007). On the other hand, mathematical topics have different characteristics (Clement, Sarama, & DiBiase, 2004; Ginsburg, Lee & Pappas, 2016). Not only about numbers but mathematics also about patterns, algebra, geometry or statistics (Kusaeri, 2017), so that not all topics can be assessed in the same way. Thus, the assessment techniques used must also be different (Kane, 2001). One of the assessment techniques that can be given to students authentically is a project task.

Project tasks enable students to learn actively based on the inquiry because students are required to carry out a series of activities to solve the contextual mathematical problems. In solving the problems, a series of activities (planning, implementing, and reporting the project result) must be undergone by students, so they should apply their potential and knowledge. When carrying out the project, students are required to actively seek or collect the required data, and they usually do it in a group. Therefore, students are encouraged to use higher order thinking through problem-solving skills when they work in a project task (Sukmasari & Rosana, 2017; Kusaeri, Hamdani & Suprananto, 2019). By using such tasks, students can learn to solve mathematical problems individually or in a group. It underlies that project task will help students develop their collaboration, communication, and problem-solving abilities (Shariff, Johan, & Jamil, 2013). Those abilities are the keys needed in global life (Kan & Bulut, 2014).

A problem arises when the instrument to assess project tasks is still limited (Pettersen & Braeken, 2019; Riscaputantri & Wening, 2018). In addition, the criticism that often heaves in sight is the effect of subjectivity relating to the scoring process of project assessment which is likely to emerge higher than in the written test. In the written test, it can be easily scored, that is true or false, while in the assessment of the project cannot be done in such ways. When carrying out the project assessment, some irrelevant variables such as moods and the surrounding condition often influence the process. In other words, the teacher's ability to understand and apply the assessment rubric as well as teacher subjectivity level is very influential in giving the assessment. For this reason, Kan and Bulut (2014) suggest two things that can be done to minimize the subjectivity effect on project assessment. First, developing a clear scoring mechanism which contains a description of the project

task to be done by students and teachers. The description contains the criteria for skills and knowledge that will be assessed, set out in a rubric. Second, providing training for teachers or assessors on how to use the rubric to make decisions about project tasks given to the students.

Referring to Kan and Bulut (2014), the effectiveness of the project assessment depends on the quality and coordination between the teacher and the rubric. In other words, the role of rubrics in assessment context is very essential. The rubrics will be a very important tool in transferring what students have done or produced in the project into the form of assessment (Jonsson & Svingby, 2007). Therefore, in order to obtain a valid and reliable result, the rubric must provide sufficient information to help teachers assess the project tasks that have been successfully completed by students (Stuhlmann, Daniel, Dellinger, Denny & Powers, 1999).

Some studies (e.g., Smit, Bachmann, Blum, Birri, & Hess, 2017; Wulan, 2008) show that raters or teachers are often inconsistent in using rubrics. It is due to teachers' lack of experiences in using the rubric as well as the quality of the rubric (Kan & Bulut, 2014). Besides, the inconsistency of using the rubric also occurs because of a lack of understanding of rubric's constructs or aspects. Teachers try using a wide range when scoring result of the project. However, with the same training and teaching experience, assessors can evaluate students' tasks differently (Lumley, 1998; Schafer, Swanson, Bene & Newberry, 2001). The differences in the assessment are due to the way the assessor understands and applies the assessment rubric as well as their subjectivity level in giving an assessment (Eckes, 2008).

The aforementioned researches (e.g., Smith *et al*., 2004; Jonsson & Svingby, 2007; Wulan, 2008) and the experts' opinions (e.g., Kan & Bulut, 2014) indicate the importance of instruments on project assessment along with good rubrics. Thus, teachers who use project assessment can provide relatively similar scores. However, researches that focus on how the construction of a valid project assessment instruments for mathematics learning and its rubric which has high inter-rater reliability, has not been found so far in Indonesia. In fact, when conducting a project assessment, most of it is done only by a teacher (Putra, 2012). Therefore, the issue regarding the inconsistency of teachers in understanding the rubric has implications for the striking differences in the results of the scores given on the assessment sheet (Andrade & Du, 2005). As an effect, the assessment results received by students become bias. This condition is a serious issue and encourages the empirical testing of the validity and the reliability of project assessment instruments used in assessing students' project tasks.

Prior researches put attention on the validity and reliability test of performance assessment instruments (Kan & Bulut, 2014; Avcu & Avcu, 2015; Bashooir & Supahar, 2018). They have not tested the validity and reliability of project assessment instruments. On the other hand, we have not found many studies which employ Confirmatory Factor Analysis and inter-rater reliability test. Thus, for filling the gap in the literature, this article aims to address the results of validity and inter-rater reliability test for project assessment instrument which is adopted from the mathematics textbooks.

B. Methods

Research subject

The participants in this study were 94 lower secondary school students from three different schools in Surabaya and Gresik, East Java. In the first school, the participants consisted of 28 students formed into 5 groups with 3 mathematics teachers. In the second school, the participants consisted of 34 students divided into 5 groups with 4 mathematics teachers. Whereas in the third

school, the participants consisted of 32 students made into 5 groups with 3 mathematics teachers. The five groups in each school were asked to do a project task and the teachers gave an assessment of the projects result carried out by the students.

The selection of students involved in this study was entirely determined by the teachers in each school. Meanwhile, teachers were selected based on their teaching experience and their expertise in assessment. The teachers' experiences in teaching range from 5 to 25 years. Further, the teachers' expertise in assessment was based on their experience of participating in training related to assessment in the curriculum 2013. Project assessment was carried out in the learning process for three meetings conducted by the 10 mathematics teachers (raters).

At the first meeting, students were asked to make plans for the project along with its date and day. They also discussed task sharing among the group members. Furthermore, raters evaluated the project plans made by each group. At the second meeting, students were asked to process and to analyze data obtained through observation. At the end of the activity, the raters gave a conclusion about the relevance of the topics to the project being worked on. At the third meeting, students were asked to present the results of the project. Then, the raters assessed students' presentations, results of data processing, analyzing and drawing conclusion process and the systematics of students' written reports. The raters conducted an assessment using the assessment rubric guidelines that had been prepared with a range of assessment scores for each criterion from 1 to 4. The sequence of students and teachers' activities are depicted in Diagram 1.

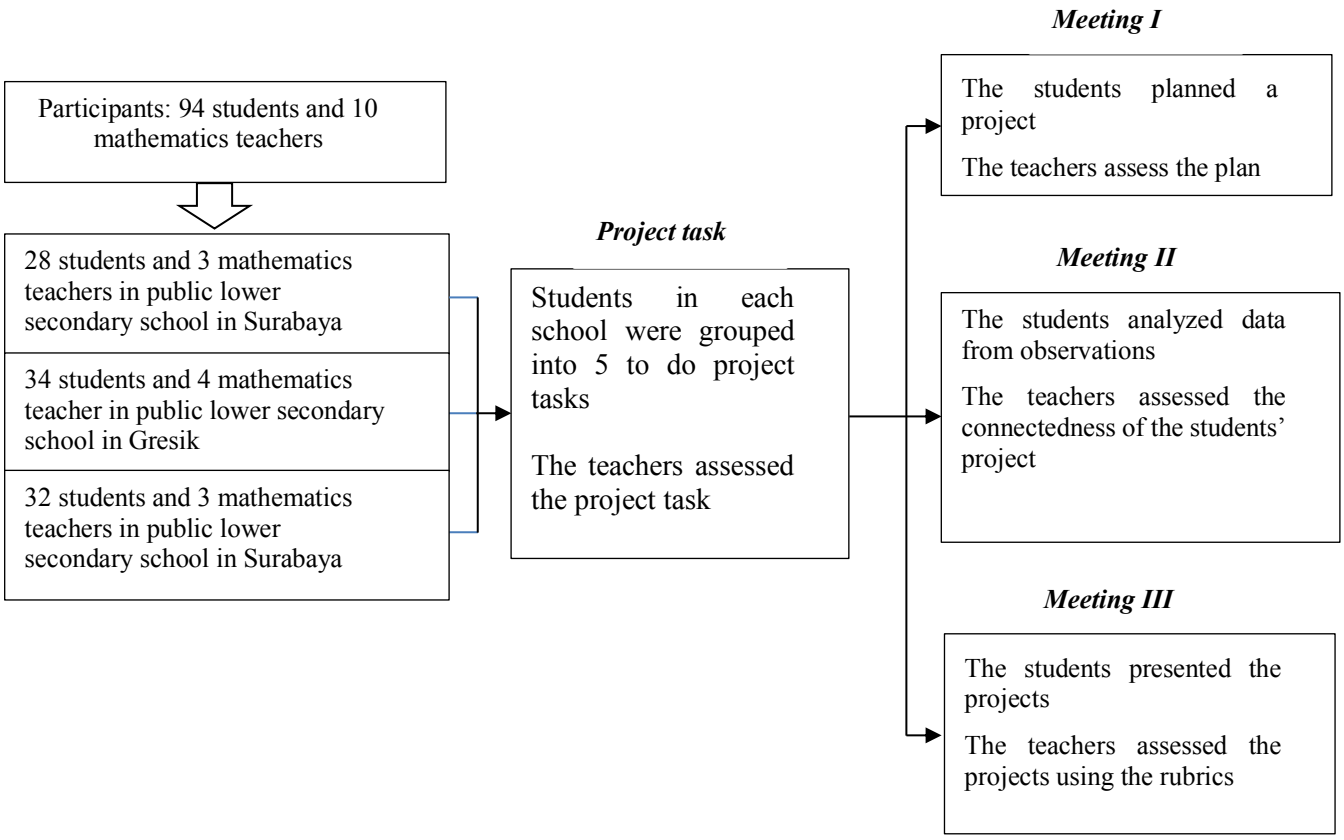

Diagram 1. The students and teachers' sequence of activities in project assessment

Data collection

Data in this study were collected through a project assessment sheet. The sheet comprises the criteria or aspects that will be assessed in the project task and the score that must be given by the

raters or teachers. The sheet was used to assess project tasks given to students. To guide the teacher in assessing the project, a rubric was developed. The rubric was used to assure the objectivity of the assessment. The rubric contained criteria or aspects that are assessed along with its descriptions for a score of 1 to 4.

Project tasks were adapted from the textbook (Curriculum 2013) published by the Ministry of Education and Culture on the topic of relations and functions. Furthermore, in this study, the task was modified by adding goals and indicators of the problems. The language in the project assessment sheet was also clarified for each aspect or assessment criteria, while the rubric was improved by clarifying descriptors on each assessment criterion. In this case, the content of the tasks remains similar to the original one in the textbooks.

The instructions given on the sheet of project task consists of three stages, i.e., (a) planning the project, seeking information on how to determine telephone rates and task sharing to group members. (b) doing data processing, analyzing data that has been obtained, linking the results of observations with the topic of relations and functions and present the results of observations in the form of diagrams (bar charts, tables, lines, and sequential pairs of pairs). And (c) making a project report from the observations and present it to the class.

Furthermore, students completed project tasks, and each of the 10 teachers was asked to assess based on ten predetermined criteria. The ten criteria included: planning the stages of project, tasks sharing among the group members, determining the tools and materials needed, time of project implementation, quantity of data sources, data processing, data analysis, drawing conclusions (relationship between relations and functions in daily life), the format of report and presentation. The maximum score in this assessment was 40 because all criteria have the same score, which was 4.

The teachers assessed those 10 criteria on three occasions. First, the teachers were asked to rate four criteria in the planning and preparation stages of the project, including the plan of project implementation, division of tasks to group members, tools and materials needed to carry out project tasks and time allocation of the project. Second, the teachers were asked to rate four criteria in the project implementation stage, including the amount of data obtained, data processing, data analysis, and drawing a conclusion. Third, the teachers were asked to rate two criteria in the stages of the project report, namely systematic writing of the report and presentation.

Data Analysis Technique

Validity

The results of the assessment data conducted by ten teachers were processed in a table. The first column contains criteria in the rubric, the second until the 11th column contained the evaluation results of each criterion by the raters. Table 1 shows the sample of teachers' scores.

Table 1. Sample of teachers' scores

No	Criteria	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10
1	Plan a project	4	3	3	3	4	3	3	3	3	5
2	…										
.	…										
10	…										

Furthermore, a Confirmatory Factor Analysis (CFA) was carried out, namely variable selection using the SPSS version 21. The procedures for carrying out factor analysis were: the selection of the variables, formation of factors, interpreting analysis result, and factor naming. In the process of analyzing the variable selection, KMO-MSA (Kaiser-Meyer-Olkin Measure of Sampling

Adequacy) was required. The provisions of the KMO-MSA refer to Priyatno (2014), and Hair, Black, Babin, and Anderson (2006). At the last stage of the construct validity test, the criteria used was if the initial construction criteria or factors are the same as the results of the empirical test, it can be said that the existing project assessment instruments meet the construct validity. Otherwise, the instrument is not constructively valid.

Reliability

Assessment results from the 10 raters on student project tasks were included in the table. The first column comprises group order of the students (group 1 to group 5), the second column is filled up with variables or criteria for assessment (criteria 1 to 9), and the third until twelveth column contains the results of the rater's assessment on each variable from each group. Then an inter-rater reliability test or the ICC (Intraclass Correlation Coefficient) was conducted by using the SPSS version 21. Interpretation of the inter-rater reliability coefficient was based on criteria, i.e., the reliability value is < 0.40 (less), 0.40 - 0.59 (low), 0.60 - 0.74 (good) and 0.75 - 1.00 (very good) (Cicchetti, Bronen, Spencer, Haut, Berg, Oliver & Tyrer, 2006)

C. Findings and Discussion

Validity

Validity is the accuracy of a measuring instrument in carrying out its measuring functions (Gay, Mills & Airasian, 2000). The validity of the project assessment instrument in this study will show the extent to which the instrument is able to measure the competency of students' skills in project task. Table 2 shows the description of instruments tested for validity.

Table 2. Project assessment instruments to measure students' competency skills

No.	Criteria or factors	Item	Number of items
1	Preparation and Planning stage	1, 2, 3, 4	4
2	Implementation stage	5, 6, 7, 8	4
3	Reporting stage	9, 10	2
Total			10

The statistical test used in this study was CFA with the help of SPSS version 21. CFA is a method used to determine the construct validity that has been done by previous researchers related to the field of psychology and education (Laher, 2010). Construct validity is one type of validity that is suitable to test the validity of a non-test instrument. The procedure for carrying out factor analysis is (1) selection of variables; (2) factor formation; (3) interpret the results of the analysis; and (4) factor naming.

In the analyzing process of variable selection, the computational results showed that the Kaiser-Meyer-Olkin value Measure of Sampling Adequacy (KMO-MSA) was 0.645 and the significance was 0,000. Given the value of KMO-MSA above 0.500, it is included in the good category. From the Bartlett test for Test of Sphericity, the Chi-Square was 125.063 in the degree of freedom 45 with significance at 0.000 < 0.05. This means that the correlation matrix formed was not an identity matrix and finally, factor analysis can be done (Coakes & Steed, 2007).

The next process was a factor formation analysis that aimed to simplify a set of initial variables. The results of factor formation analysis in the Total Variance Explained table show that the characteristic values (eigen value) of all factors were above 1 (> 1) (in Table 3). As recommended by Hair *et al*., (2006), eigen values more than 1 are accepted for factor formation criteria. This shows the number of factors that formed the project assessment instrument is 4.

Table 3. The result of KMO-MSA and Bartlett's test of sphericity

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.645
Bartlett's Test of Sphericity	Approx. Chi-Square	125.063
	Df	45
	Sig.	.000

In addition, there was a factor load variance found which can explain the quality variance existence of the conducted project by the students. The first factor contributed 26,676% of the variance, the second factor explained 21,434% of the variance, the third factor explained 13,265% of the variance, while the fourth factor explained 10.401% of the total variance (Table 4).

Table 4. Extraction results of project assessment instruments

Factors	Eigen values	Percentage of Variance	Comulative Percentages
I	2.668	26.676	26.676
II	2.143	21.434	48.110
III	1.326	13.265	61.375
IV	1.040	10.401	71.775

Scree plot diagram (Figure 1) shows the decreasing tendency of eigen value. This diagram can also be used to determine subjectively the number of factors used. It also appears that in the fifth factor, the eigen value is below 1. This fact indicates that the four factors as described earlier are enough to summarize the nine existing variables.

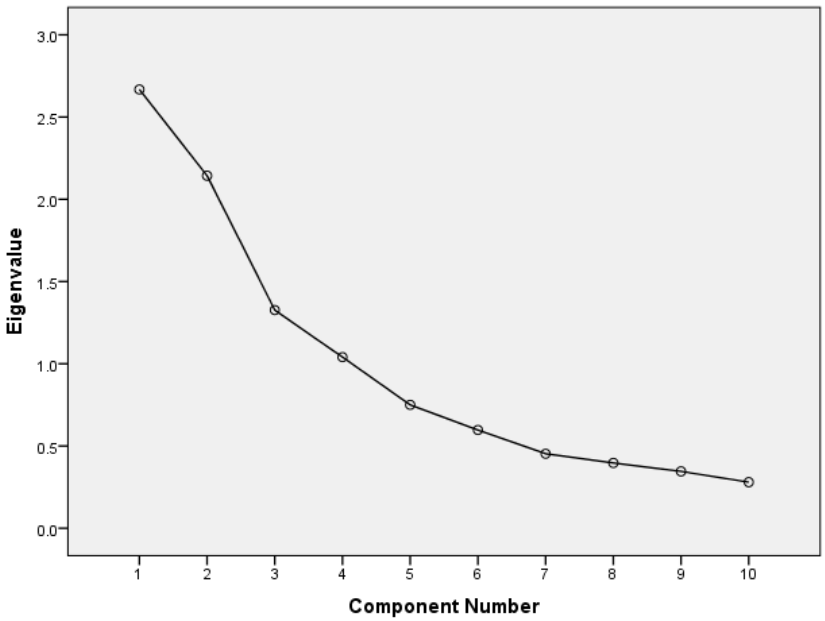

Figure 1. Scree plot eigenvalues for component numbers

After analyzing factor formation, the next step was the interpretation of factor analysis results. Table 5 contains a Rotated Component Matrix that can be used to determine which factor is suitable for a variable. According to Hair et al. (1998), an item is received on a component, if it has a factor

loading more than or equal to 0.500. Thus, it can be explained that items 1 and 2 are both accepted in component 1, because they have a factor loading value of 0.787 and 0.780, respectively.

Furthermore, item 3 has a factor loading value above 0.500 on two components, namely component 1 and component 2. That is, item 3 can be received in component 1 or component 2. However, the factor loading value has a greater effect on component 2 so item 3 included in component 2. In the same way, item 4 and 5 belong to component 2, item 6 and item 7 both are included in component 3. Item 8, 9 and 10 include component 4.

Table 5. Rotated component matrix in project assessment instrument

Items	**Components**			
	1	2	3	4
1	.787		.218	.182
2	.780		-.116	-.254
3	.511	.630	-.334	.110
4	.104	.850	.190	
5		.708		.572
6	.124	.109	.811	
7		.172	.827	.113
8	.508	-.429	.118	.517
9	.173			.817
10	.238	-.104	.438	.536

After a factor was formed with the items under study, the last stage was to give a name to the four factors formed based on the characteristics of its members. In the end, factor 1 consisted of two items, namely item 1 and item 2. Item 1 is related to the planned stages of project implementation and item 2 is related to the division of tasks to group members. Looking at the characteristics of the two items found in factor 1, then the project planning stage is a suite name for it. Factor 2 consists of three items, namely item 3 (determining the tools and materials needed), item 4 (project processing time) and item 5 (quantity of data). Judging from the items in factor 2, the exact name for factor 2 is the project preparation stage. Factor 3 consists of 2 items, namely item 6 (data processing) and item 7 (data analysis), so factor 3 is relevant with the name of the project implementation stage. There are also 3 items contained in factor 4, namely item 8 (conclusion), item 9 (systematic report writing) and item 10 (presentation). Noting the characteristics of the three items, then factor 4 is named the final stage or reporting the project.

Therefore, the composition of items in each factor in Table 2 changes. Next, the final composition of each factor and items contained therein are presented in Table 6. Referring to the number of factors or criteria in Table 2 and Table 6, there seems to be a difference in the number of initial construction factors (3 factors) with the results of the empirical test (become 4 factors). Thus, it can be said that the project assessment instrument used is invalid in terms of the validity of the construct.

Table 6. Project assessment instruments after testing were conducted

No.	**Factors or Criteria**	**Item**	**Number of Items**
1	Project planning stage	1, 2	2
2	Project preparation stage	3, 4, 5	3
3	Project implementation stage	6, 7	2
4	Final stage or reporting the project	8, 9, 10	3
Total			10

The invalidity of the project assessment instrument is influenced by many things. One argument that can be put forward is that the process of instrument construction is not through theoretical review since the instrument was taken directly from the book published by the Ministry of Education and Culture for the implementation of Curriculum 2013 with improvements as needed, such as language. In fact, in determining factors or criteria in the development of affective domain assessment instruments (including skills), it is important to be careful in considering the necessary theories (McCoach, Gable, & Madura, 2013). With a good understanding of the theory, it can certainly produce valid operational definitions on each factor or criterion.

Related to this, Ihsan (2015) asserts that defining constructs to be used as criteria or factors in an instrument of assessment must be careful. According to him, compiling operational definitions is a stage that is very difficult to do. Arranging operational definitions needs to start from a clear and well-understood theory. A theory that is less clear will lead to errors of intent from what we want to measure.

Those facts indicate that in the process of developing a project assessment instrument must be careful, especially when developing operational definitions which are further developed into factors or criteria to be assessed. A good operational definition, according to Riscaputantri and Weing (2018), is certainly formed from a solid theoretical building. A solid theory certainly requires adequate and comprehensive reading. This is the weak point of the project assessment instruments tested in this study.

Reliability

After conducting the construct validity test, the next step was the inter-rater reliability test. Reliability test was carried out after the project assessment instrument was adjusted to the last condition, namely the aspects or items contained in the instrument have been arranged and grouped into 4 factors as Table 5.

For this purpose, the rater or the teacher involved in the research were given the same perception at the beginning of the activity related to how to use the project assessment instrument and its rubric. Including the meaning of each aspect in the rubric. In this way, it was expected that the same understanding among the raters occurred and when using it to assess the results of student work the scores are not far adrift.

Furthermore, the level of inter-rater reliability (ten teachers) can be explained from the results of the calculation of the inter-rater reliability coefficient using the Interclass Correlation Coefficient (ICC). A summary of the ICC calculation results by using SPSS version 21 is presented in Table 7.

Table 7. Intraclass correlation coefficient

	Intraclass Correlation
Single Measures	0.672
Average Measures	0.953

Table 7 shows that 10 existing aspects in the assessment instrument, the mean value between rater is 0.953. While the reliability value for each rater is 0.672. Referring to the opinion of George and Mallery (2003), the closer to 1.00 the higher the reliability or internal consistency of items in the instrument. Thus, it can be concluded that the inter-rater reliability of the project assessment instruments tested in this study belongs to the high category.

The statistical results above are certainly still in general and need to be explored further on how the variance between rater in each aspect or item. The results of this study are very important in order to see aspects of the rubric which still make a significant difference in interpretation among the rater. Hence, it can be used as a basis for improving the instrument at the next stage.

Table 8 presents a case processing summary that can be used as a basis in examining rater behavior when using project assessment instruments with a rubric guide. From Table 8 it can be seen that the results of the assessment of the rater, 20 data are excluded. The excluded data means that from the assessment results of the raters, these data have a high difference score given by one rater compare to the other, from score 1 to score 4 on an item or certain aspect.

Table 8. Case processing summary

		N	%
Cases	Valid	50	71.4
	Excluded[a]	20	28.6
	Total	70	100.0

Descriptions on rubrics that have a range of high scores between rater, occur in item 2 (a division of tasks to group members), item 3 (determine the tools and materials needed), item 4 (project processing time) and item 9 (systematics of writing the report). The four aspects have a wide range of assessment scores between the raters because they have a different understanding of the description of the rubric used. The different understanding between raters, one of which is caused by unclear and too long descriptors given.

Referring to Putra (2012), this indicates that the rubric descriptor has not been practically used since generally the raters only have limited time to assess students' project tasks. In addition, the rater was also burdened with the amount of work, reports, presentation of project results from each class when conducting the assessment. Based on this phenomenon, the descriptor of the rubric should help the rater when assessing student project assignments quicker and more accurate.

Drawing from the findings and discussion, we note some important cases, i.e., (1) to develop a valid instrument of project assessment, we need to decide the constructs which become the factor or criteria to be assessed in an attentive way and (2) the descriptors in the rubric should be clear and short. It intends to overcome the raters' difficulty and promote mutual understanding. A different understanding will possibly lead to the weakness of reliability of project assessment instrument. In this case, the current study significantly contributes to curriculum developers especially the authors of the mathematics textbooks.

D. Conclusion

The validity test of the project assessment instrument shows that the instrument used is not constructively valid. The invalidity is characterized by the difference in the number of factors. It changes from 3 factors in the initial construction to become 4 factors after the empirical test. It is conjectured that the development of the instrument did not equip with a relevant theoretical review. A representative theoretical review of the instrument will be very contributive to the validity. However, in terms of inter-rater reliability, the project assessment instruments used are reliable and included in the high category. Several weaknesses emerged during this research process such as the criteria which are the object of assessment and set forth in the project assessment sheet, are not made based on an in-depth theoretical review. Besides, no further validity testing of the new project assessment sheet has been carried out. Responding to the weaknesses, the following suggestions

are raised: (a) when developing assessments criteria or aspects, it should be derived from in-depth theoretical studies. With a deep and strong theory, a valid operational definition will be produced on each criterion or aspect; (b) when the new project assessment sheet is obtained, the rotation results of several aspects should be tested for further validity. In this way, the validity of the new instrument will be known; and (c) the teacher or education practitioner should formulate a short and clear descriptor on the assessment rubric. With a descriptor that is too long often makes the teacher confused, and in the end, the teacher will give an incorrect assessment.

References

Andrade, H. L. & Du, Y. (2005). Student perspectives on rubric-referenced assessment. *Practical Assessment Research & Evaluation, 10*(3), 1-11.

Avcu, R.& Avcu, S. (2015). Turkish adaptation of utley geometry attitude scale: A validity and reliability study. *Eurasian Journal of Educational Research*, 58, 1-24.

Bashooir, K. & Supahar, S. (2018). Validitas dan reliabilitas instrumen asesmen kinerja literasi sains pelajaran Fisika berbasis STEM. *Jurnal Penelitian dan Evaluasi Pendidikan, 22*(2), 219-230.

Chan, H.Z. & Abdullah, M.L.Y. (2018). Validity and reliability of the mathematics self-efficacy questionnaire (MSEQ) on primary school students. *Pertanika: Journal of Social Sciences & Humanities, 26*(4), 2161 – 2177.

Chen, Q. (2011). *Assessment policy change in relation to English language teaching and learning in China: Study of perspectives from two universities* (Unpublised Dissertation). Queensland: Queensland University of Technology.

Cicchetti, D., Bronen, R., Spencer, S., Haut, S., Berg, A., Oliver, P., & Tyrer, P. (2006). Rating scales, scales of measurement, issues of reliability resolving some critical issues for clinicians and research. *The Journal of Nervous and Mental Disease, 194,* 557–564.

Clarke, D. (1996). Assessment. In A.J. Bishop, K. Clements, C. Keitel, J. Kilpatrick and C. Laborde (eds.) *International Handbook of Mathematics Education*, (pp. 327–370). Kluwer, Dordrecht.

Clement, D.H., Sarama, J. & DiBiase, A.M. (2004). *Engaging young children in mathematics: Standards for early childhood mathematics education.* Mahwah, NJ: Lawrence Erlbaum Associates

Coakes, S. J., & Steed, L. G. (2007). SPSS: *Analysis without anguish: Version 14.0 for Windows.* Brisbane: John Wiley & Sons Australia Ltd.

Dochy, F., Gijbels, D., & Segers, M. (2006). Learning and the emerging new assessment culture. In L. Verschaffel, F. Dochy, M. Boekaerts, & S. Vosniadou (Eds.), *Instructional psychology: Past, present, and future trends* (pp. 191-206). Oxford, Amsterdam: Elsevier.

Eckes, T. (2008). Rater Types in writing performance assessments: A classification approach to rater variability. *Language Testing, 25*, 155-158.

Gay, I.R., Mills, G.E & Airasian, P.W. (2000). *Educational Research Competencies for Analysis and Applications* (10th ed). USA: Pearson Education.

George, D & Mallery, P. (2003). *SPSS for Windows Step by Step*. Canada: Canadian University Collage.

Ginsburg, H.P., Lee, Y.S & Pappas, S. (2016). A research-inspired and computer-guided clinical interview for mathematics assessment: Introduction, reliability and validity. *ZDM Mathematics Education, 48,* 1003-1018.

Hair, J.F., Black, W.C., Babin, B.J. & Anderson, R.E. (2006). *Multivariate Data Analysis,* New Jersey: Pearson Education.

Ihsan, H. (2015). Validitas isi alat ukur penelitian konsep dan panduan penilaianya. *Pedagogia*, *13*(2), 266-273.

Jonsson, A. & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, *2*, 130–144.

Laher, S. (2010). Using exploratory factor analysis in personality research: Best-practice recommendations. *SA Journal of Industrial Psychology*, *36*(1), 1-7.

Lumley, T. (1998) Perceptions of language-trained raters and occupational experts in a test of occupational English language proficiency. *English for Specific Purposes, 17*(4), 347-67.

Kane, M.T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*(4), 319-342.

Kan, A., & Bulut, O. (2014). Crossed random-effect modeling: examining the effects of teacher experience and rubric use in performance assessments. *Eurasian Journal of Educational Research*, *57*, 1-28.

Kusaeri, K., Hamdani, A.S., & Suprananto, S. (2019). Student readiness and challenge in completing higher order thinking skill test type for mathematics. *Infinity Journal, 8*(1), 75-86.

Kusaeri, K. & Aditomo, A. (2019). Pedagogical beliefs about critical thinking among Indonesian mathematics pre-service teachers. *International Journal of Instruction*, *12*(1), 573-590.

Kusaeri, K. (2017). *Historiografi matematika: Rujukan paling otoritatif tentang sejarah perkembangan matematika*. Yogyakarta: Matematika.

Kusaeri, K. (2014). *Acuan & teknik penilaian proses & hasil belajar dalam kurikulum 2013*. Yogyakarta: Ar-Ruzz Media.

McCoach, D. B., Gable, R. K., & Madura, J. P. (2013). *Instrument development in the affective domain*. New York: Springer New York.

Mueller, J. (2005). The authentic assessment toolbox: Enhancing student learning through online faculty development. *Journal of Online Learning and Teaching*, *1*(1), 1-7.

Murtiyasa, B. (2015). Tantangan pembelajaran matematika era global. *Prosiding Seminar Nasional Matematika dan Pendidikan Matematika UMS 2015*. Surakarta: Jurusan Pendidikan Matematika Universitas Muhammadiyah Surakarta.

O'Neill, G., Huntley, R.S., & Race, P. (2007). *Case of good practices inassessment on student learning in higher education.* Dublin: AISH

Park, J.S & Park H.P. (2012). The changes of assessment at middle school level in Korea. *ZDM Mathematics Education, 44*, 201–209.

Pettersen, A & Braeken, J. (2019). Mathematical competency demands of assessment items: a search for empirical evidence. *International Journal of Science and Mathematics Education, 17*(2), 405–425.

Priyatno, D. (2014). *SPSS 22 pengolah data terpraktis*. Yogyakarta: Andi.

Putra, H.S. (2012). Pengembangan rubrik penilaian untuk digunakan guru dalam menilai hasil tulisan siswa SMA (Tesis). Jakarta: Program Studi Linguistik Fakultas Ilmu Pengetahuan Budaya Universitas Indonesia.

Retnawati, H. (2015). Hambatan guru matematika sekolah menengah pertama dalam menerapkan kurikulum baru. *Cakrawala Pendidikan*, *34*(3), 390-403.

Riscaputantri, A. & Wening, S. (2018). Pengembangan penilaian afektif siswa kelas IV Sekolah Dasar di kabupaten Klaten. *Jurnal Penelitian dan Evaluasi Pendidikan, 22*(2), 231-242.

Schafer, W., Swanson, G., Bene, N., & Newberry, G. (2001). Effects of teacher knowledge of rubrics on student achievement in four content areas. *Applied Measurement in Education*, *14*(2), 151-170.

Shariff, S. M., Johan, Z. J., & Jamil, N. A. (2013). Assessment of project management skills and learning outcomes in students' projects. *Procedia Social and Behavioral Sciences, 90*, 745-754.

Smit, R., Bachmann, P., Blum, V., Birri, T., & Hess, K. (2017). Effects of a rubric for mathematical reasoning on teaching and learning in primary school. *Instructional Science*, 45(5), 603-622.

Stuhlmann, J., Daniel, C., Dellinger, A., Denny, R. K., & Powers, T. (1999). A generalizability study of the effects of training on teachers' abilities to rate children's writing using a rubric. *Journal of Reading Psychology*, *20*, 107–127.

Sugiyanto, Kartowagiran, B. & Jailani. (2015). Pengembangan model evaluasi proses pembelajaran matematika di SMP berdasarkan Kurikulum 2013. *Jurnal Penelitian dan Evaluasi Pendidikan, 19*(1), 82-95.

Sukmasari, V.P & Rosana, D. (2017). Pengembangan penilaian proyek pembelajaran IPA berbasis *discovery learning* untuk mengukur keterampilan pemecahan masalah. *Jurnal Inovasi Pendidikan IPA*, *3*(1), 101-110.

Watt, H. (2005). Attitudes to the use of alternative assessment methods in mathematics: A study with secondary mathematics teachers in Sydney, Australia. *Educational Studies in Mathematics*, *58*(1), 21–44.

Wulan, A.R. (2008). Skenario baru bagi implementasi asesmen kinerja pada pembelajaran sains di Indonesia. *Jurnal Mimbar Pendidikan, 32*(3): 1-11.