Interrater Reliability: Comparison of essay’s tests and scoring rubrics

Lussy Dwiutami Wahyuni¹, Gumgum Gumela², Herdiyan Maulana³
Universitas Negeri Jakarta, Jalan Rawamangun Muka, Jakarta, Indonesia
E-Mail: lussysf@unj.ac.id, ggumelar@unj.ac.id

Abstract. This comparative study aimed to examine the difference of interrater reliability between two different methods of measurement: essay tests and rubrics scoring. Thirty students and thirty science teachers were participated in this study. The interrater reliability was estimated using Fleiss Kappa. While the hypotheses were tested using Mann-Witney U with Exact to increase data validity. The results of this study showed that interrater reliability of restricted response items was higher than context dependent tasks as well as applied to extended response items scored using the analytic rubric and the holistic rubric. While the interrater reliability of extended response items is higher if its compared to background dependent tasks which was scored using the analytic rubric and the holistic rubric.

1. Introduction
Scoring has an important role to understand the learning success rate achieved by student in class. Measurement instrument is needed to determine student’s learning achievement. The instruments that being used commonly have a form in test or non-test, and the use of the instruments depend on indicator characteristic that will be measured.

The suitability between selected scoring type with indicator that are going to be measured implicates to the degree of validity and reliability of information describing student’s learning result. This idea parallel with Cohen & Swerdlik (1996) idea that “... a good test is a reliable and valid, a good test item is reliable and valid. Further, a good test item helps to discriminate test takers” [1]. Following that Azwar (2003) states that unreliable measurement would produce untrusted scores, because the difference in scores that occur between individuals is more determined by the error factor rather than the actual difference factor[2].

Reliability is the indicator of measurement’s consistency. A low reliability score means that there is a high measurement error in the test. On the other hand, a high reliability score means that the test has a low measurement error.

Scholars have identified many methods to estimate reliability. One of which is interrater reliability. This approach is commonly used to measure the consistency of subjective test. Lange (2011) states that Interrater reliability is the extent to which two or more raters (or observers, coders, examiners) agree. It addresses the issue of consistency of the implementation of a rating system” [3]. Thus, interrater reliability analysis produces coefficient score indicating agreement between two or more rater. The higher the coefficient produced, the higher the rater’s agreement consistencies level [4].
Essay test is one of type of test that have a high scoring subjectivity. According to Tuckman “Essay tests, at best, are easy to construct and relatively valid to be used as test of higher cognitive processes. But the reliability of this type of test is hard to estimate” [5]. Parallel with that, Harold Gulliksen also implied that the test is prone to reliability issue due to interpretation differences which are emerged among raters[6]. However, this kind of test is important to measure student’s capabilities in organize, formulate, and response their answers [7]. Further, this test did not have a single answer as it found in the objective test. This condition provides chances for student to use their high level thinking process and likely to minimize unethical (e.g. cheating) behavior during administration of the test.

Essay test generally divided into two types: restricted responses items and extended responses items. Beside the type of those two tests, Nitko have been also adapted multiple-choice discourse question type into restricted essay test which in the formulation of the problem related with discourse that have been put in the question, which become known by the name of context dependent tasks. Restricted essay test can measure student’s ability in explaining cause-and-effect relationship, explaining the application of a concept, presenting relevant argumentation, formulating hypothesis, concluding things, making an assumption, explaining limitation of the data, explaining method and procedure, and other things that related with student’s ability in making some answers. Whereas extended essay test can measure student’s ability in producing, organizing, dan expressing idea; integrating learning experiences in different context; and the other things that related in student’s ability in writing an essay with determined goals[8].

One of the method that can be used to increase essay test reliability is scoring guidelines. Scoring guidelines is usually known as a rubric model. Andrade (in Zainul & Mulyana) define rubric as one of the scoring tools which consist of lists of criteria or what must be calculated [9]. There are number of purposes when using rubric; 1) is to increase the scoring consistency; 2) increase score consistency which obtained by student when being scored by different rater, and 3) provide validity measures in scoring complex idea [10].

In general, rubric is divided in two type: analytic and holistic. Analytic rubric commonly being used as a helping tool in essay test with restricted response items. Analytic rubric scoring requires the scorer to determine list of important elements which have to be scored [11]. Analytic rubric also has a detail and restricted characteristic, so that rater can be easily to use or remember the descriptor which is the criterion of the answer that is being scored by them. Therefore the analytic rubric can differentiate which things that have been measured with points which were obtained previously[12]. Further, essay test with extended response items, commonly use holistic rubric as a helping tool in scoring process. Holistic rubric is one of scoring method that requires the question maker to score the quality of each element in the student’s answers [13][12].

Knoch states that interrater reliability resulting from analytic rubric is higher and better to distinguish aspects which are contained in scoring, because scoring aspect in analytic rubric is more detail than holistic rubric [14].

| Table 1. The Differences in Rubric Quality Judging from 3 Aspect of Measurement Criteria |
|---------------------------------|---------------------------------|------------------|
| **Rating scales** | **Ease of construct** | **Scoring efficiency** | **Reliability** |
| analytic | Moderate | Moderate | Moderate |
| Holistic scoring | High | High | Low |

On the contrary, research by Paul Diederich showed that the outcome that essay test with holistic rubric scale of 6, resulting the interrater agreement level for about 95% [12]. Until now it has not been found the specific research looking for the reliability from any kind of essay test. There is only study about essay test in general or essay as one type of paper. That inversely proportional with research in analytic and holistic rubric reliability.
However, with regard to available studies, the findings are still controversial. There are studies that show the reliability of analytic rubric higher than holistic rubric, while other studies showed opposite results. Cynthia S Wiseman is one of the researchers that worked in comparation in analytic and holistic rubric in scoring about the ability in writing second language (L2/second language). This research that have been published in Iranian Journal of Language Testing vol. 2, No. 1, March 2012 get the result that low scale which are owned by holistic rubric was not being chosen by rater. Whereas, the low scale which are owned by analytic rubric spread evenly. On the contrary, the research from Bunmi S. Malau-Aduli, Sue Mulcahy, Emma Warnecke, Petr Otahal, Peta-Ann Teague, Richard Turner, and Cees Van der Vleuten about the comparation of interrater reliability in holistic and analytic rubric in scoring OSCEs (Objective Structured Clinical Examinations) showed the result of interrater reliability in global ratings (holistic) higher than checklist (analytic).

With that matters, then it is examined whether there is differences between interrater reliability in restricted essay test, discourse based essay test, and extended essay test with using analytic and holistic rubric. The objective of this study is looking for the difference with essay test and rubric scoring method in more clearly, decisively, and empirically manner. Further, it also aimed to investigate the suitability with type of essay test and rubric scoring method in this research.

2. Methodology

This research uses comparative method, which compare the presence one or more variable on two or more different sample or different time. The study design was 1 x 3 as much as two times.

Table 2. Research Design in the Type of Essay Test and Analytic Rubric

| Essay Type               | Rubric                        |
|-------------------------|-------------------------------|
| restricted response items (RRI) | Analytic Rubric (A) RRI-A ($\rho_1$) |
| context dependent tasks (CDT) | Analytic Rubric (A) CDT-A ($\rho_2$) |
| extended response items (ERI) | Analytic Rubric (A)ERI-A ($\rho_3$) |
| restricted response items (RRI) | Holistic Rubric (H)RRI-H ($\rho_4$) |
| context dependent tasks (CDT) | Holistic Rubric (H)CDT-H ($\rho_5$) |
| extended response items (ERI) | Holistic Rubric (H)ERI-H ($\rho_6$) |

Data for this research are obtained from each of rater (30 raters) when they scoring overall subject answers (30 subjects) from 3 type of essay test with using 2 different type of scoring test. In this study, the population divided into two parts; they were subject population and rater population as well as the sample. Subject population is something/individual that will be scored, and rater population is the individual that will give the score/rate [15].

The instrument that were used are essay test and rubric. Essay test consist of 15 questions which divided in 5 restricted essay test or restricted response items (RRI), 5 essay tests based on discourse or context dependent tasks (CDT), and 5 extended essay tests or extended response items (ERI).

The Rubric consist of 2 types: Analytic Rubric and Holistic Rubric. All of the instrument have been validated previously by 25 individuals, which consist of 19 science teacher in Middle School, 1 lecture as material expert, 1 lecture as the measurement expert, and 1 lecture as the language expert. Then, the data that has been gathered will be analyze qualitatively and quantitatively. Quantitatively, interrater reliability was calculated with Fleiss’Kappa, the hypothesis was examined by using nonparametric inferential statistic Mann Whitney U test with Exact assisted using Statistical Product and Service Solutions (SPSS) 16.
3. Result and Discussion

As seen on Table 3 below, question number 1 has high interrater reliability in the type of restricted response items (RRI), with the coefficient score 0.218 although according to Fleiss’ Kappa Benchmark Scale the score in the poor agreement degree. Whereas question number 3 in the type of context dependent tasks (CDT) has a low score, with coefficient score 0.003 (poor).

| Table 3. Interrater Reliability Fleiss Kappa based on Scored Type with using Analytic Rubric |
|-----------------------------------------------|
| Type of Test | Scoring Method With Analytic Rubric |
| RRI | 0.218 | 0.213 | 0.045 | 0.053 | 0.138 |
| CDT | 0.037 | 0.028 | 0.003 | 0.008 | 0.023 |
| ERI | 0.030 | 0.021 | 0.014 | 0.041 | 0.038 |

regarding to Table 4, question number 1 has higher interrater reliability in the type of restricted response items (RRI), with the coefficient score 0.317 although according to Fleiss’ Kappa Benchmark Scale the score in the poor agreement degree. Whereas question number 3 in the type of context dependent tasks (CDT) has a low score, with coefficient score 0.009 (poor).

| Table 4. Interrater Reliability Fleiss Kappa based on Scored Type with using Holistic Rubric |
|-----------------------------------------------|
| Type of Test | Scoring Method With Holistic Rubric |
| RRI | 0.317 | 0.271 | 0.121 | 0.090 | 0.108 |
| CDT | 0.042 | 0.035 | 0.009 | 0.056 | 0.043 |
| ERI | 0.022 | 0.048 | 0.022 | 0.050 | 0.061 |

From overall the data, then hypothesis testing shows the result in 4 hypothesis which significantly different with the comparison.

| Table 5. Mann-Whitney U test with Exact |
|-----------------------------------------------|
| Hypothesis | Sig. P(U) | Hypothesis Result | Testing | Interpretation |
| RRI-A > CDT-A | 0.004 | H0 rejected | Success |
| RRI-A > ERI-A | 0.004 | H0 rejected | Success |
| CDT-A > ERI-A | 0.155 | H0 accepted | Fail |
| RRI-H > CDT-H | 0.004 | H0 rejected | Success |
| RRI-H > ERI-H | 0.004 | H0 rejected | Success |
| CDT-H > ERI-H | 0.325 | H0 accepted | Fail |

The finding proven by significant score which is obtained is under the acceptance of confidence result parameter (p < 0.05), the rest of the result has been known that the hypothesis that submitted fail to be rejected or the empirical data does not support the hypothesis. This finding evident by the significant score which is obtained is above the acceptance of confidence result parameter (p > 0.05). From all of the hypothesis constellation that being made, only 4 hypotheses that were tested while the rest are not tested. Many factors that make it so, one of them as mentioned by Gavin[15] that holistic rubric generally more consistent and the descriptor have a good correlation with the final increase in
scoring. As quoted by Gavin [15], “Ratings that are guided by explicit holistic criteria are generally more consistent compare to [16]. Holistic rating scales with rich descriptions of each progressive stage of development appear to correlate well with final grades (e.g., Biggs & Collis, 1982; Kember, McKay, Sinclair, & Wong, 2008). However, explicit analytic criteria are also advocated [17], though these require considerably more effort. Whatever the rubric, errors do arise as different individuals apply the scoring criteria [18]. Hence, the scoring rubric alone is insufficient.

This finding was reinforced by paper from Bunmi S. Malau-Aduli, Sue Mulcahy, Emma Warnecke, Petr Otahal, Peta-Ann Teague, Richard Turner, Cees Van der Vleuten, which is published in Journal of Creative Education Vol. 3, Special Issues, 2012[19] with the title of “Interrater Reliability: Comparison of Checklist and Global Scoring for OSCEs” showing that the result of reliability of interrater global ratings (holistic) higher than checklist score (analytic). Although there is limited literature or research finding showing that analytic rubric is more reliable than holisticrubric. One of those as quoted by Nakamura[19] about the comparation between analytic rubric and holistic rubric, that show the analytic rubric reliability higher than holistic rubric, but when viewed in terms of practicality, analytic rubric tends to spend much time and expensive, whereas holistic rubric relatively faster and easier.

Interrater reliability which is estimated based on the consensus will get the result range between 0 – 33% or 0,33 if maximized or poor when being converted to Fleiss’ Kappa Benchmark Scale index. Therefore, seems a bit difficult to increase essay test quality if based on the consensus only. This was answered by Stemler statement that states there is 3 categories that can be used to estimate interrater reliability, which are: 1) consensus, the purpose is to know appropriately the agreement level that happened in rater that do not related in scoring. The statistic that can be used to calculate is the agreement percentage, Cohen’s Kappa, dan Jaccard’s J; 2) consistency, the objective is to know which rater is more consistent in rating/using rubric. Next to that, to obtain the score summary from each of the test participant. Statistics that being used to calculate is Pearson’s r, Spearman’s rho, dan Cronbach’s alpha; 3) to measure, the objective is to obtain the information as much as possible from each rather for further suitable information analysis (fit) with model. Statistic that can be used for that information is principal components analysis, generalizability theory, and multifaced. Those three categories from interrater reliability can be used so that the quality and information become more detail and comprehensive.

4. Conclusion

The finding of this study indicated that the interrater reliability in restricted essay test was higher compare to restricted essay test based on discourse and extended essay test that being scored by using analytic rubric or holistic rubric. While in the other hand, interrater reliability in extended essay test was higher compare to restricted essay test based on discourse which is scored by analytic rubric and holistic rubric.

5. References
[1] M. E. Swerdlik and R. J. Cohen, *Psychological testing and assessment: An introduction to tests and measurement*. Boston: McGraw-Hill, 2005.
[2] S. Azwar, *Penyusunan Skala Psikologi*. 2003.
[3] Lange, “Inter-rater Reliability,” *Springer*, 2011.
[4] B. Sumintono and W. Widhiarso, *Aplikasi model rasch untuk penelitian ilmu-ilmu sosial (Application of rasch model in social sciences research)*. 2014.
[5] B. W. Tuckman, “The Essay Test: A Look at the Advantages and Disadvantages,” *NASPP Bull.*, vol. 77, no. 555, pp. 20–26, 1993.
[6] H. Gulliksen, “The content reliability of a test,” *Psychometrika*, vol. 1, no. 3, pp. 189–194, Sep. 1936.
[7] M. Muslich, *Authentic assessment: Penilaian berbasis kelas dan kompetensi*. 2011.
[8] S.-G. Baek, “Measurement and assessment in teaching (8th edition),” *Asia Pacific Educ. Rev.*, ...
[9] A. Zainul, *Tes dan Asesmen di SD*. Universitas Terbuka, 2005.

[10] A. Jonsson and G. Svingby, “The use of scoring rubrics: Reliability, validity and educational consequences,” *Educ. Res. Rev.*, vol. 2, pp. 130–144, 2007.

[11] S. Surapranata, *Analisis, Validitas, Reliabilitas dan Interpretasi Hasil Tes*. Bandung: Remaja Rosdakarya, 2004.

[12] K. L. Gwet, *Handbook of Inter-rater reliability*. 2012.

[13] S. Surapranata, *Analisis, validitas, reliabilitas dan interpretasi hasil tes*. .

[14] C. S. Wiseman, “A Comparison of the Performance of Analytic vs. Holistic Scoring Rubrics to Assess L2 Writing,” *Iran. J. Lang. Test.*, vol. 2, no. 1, pp. 59–92, 2012.

[15] G. T. L. Brown, “The reliability of essay scores: The necessity of rubrics and moderation,” in *Tertiary assessment and higher education student outcomes: Policy, practice and research*, no. March, 2009, pp. 40–48.

[16] R. Williams, J. Sanford, P. W. Stratford, and A. Newman, “Grading written essays: A reliability study,” *Phys. Ther.*, vol. 71, no. 9, pp. 679–686, 1991.

[17] P. G. Caryl, “Psychology Examiners Re-examined: A 5-year perspective,” *Stud. High. Educ.*, vol. 24, no. 1, pp. 61–74, 1999.

[18] B. Sherifat Malau-Aduli, E. Warnecke, P.-A. Teague, and R. C. Turner, “Inter-Rater Reliability: Comparison of Checklist and Global Scoring for OSCEs Omega-3 PUFA and lamb View project Improving on the SBA question format in medical education View project,” 2017.

[19] Y. Nakamura, “A comparison of holistic and analytic scoring methods in the assessment of writing,” in *The Interface Between Interlanguage, Pragmatics and Assessment: Proceedings of the 3rd Annual JALT Pan-SIG Conference*, 2004, pp. 45–52.