E-ISSN 2541-4070
JPSE
P-ISSN 2477-829X

ARTICLE HISTORY: *Received: 2022-07-18,*
Revised: 2022-11-10,
Accepted: 2022-12-25.
https://doi.org/10.37729/jpse.v8i2.2164

Factors affecting the reliability of senior high schools' english teacher-made test in Kebumen

Abdul Ngafif[1]*, Semi Sukarni[1], Ismawati Ike Nugraheni[1], Nelu Sahidah[1], Karin Dinda Pithaloka[1], Setiya Cahyaningsih[1]
[1]Universitas Muhammadiyah Purworejo
e-mail: abdulngafif@gmail.com*

ABSTRAK

This study come up with the problem during the informal interview with the senior high school teachers. The researchers found that the teachers never check out whether the test they made reliable or not. They just find that the students' score are high when the test done online and the score was low when the test done offline. This study aims to analyze the reliability and the factors affecting the reliability of senior high schools' English teacher-made test in Kebumen regency. This research belongs to qualitative research, and it used descriptive case study. The instruments used by the researchers to get the data were documentation, close-ended questionnaires, and open-ended interviews. From the research, the reliability of the teacher-made tests are 0.839 (MAN 1 Kebumen), 0.747 (MAN 2 Kebumen), and 0.649 (MAN 3 Kebumen), so the test are reliable. Viewed from the questionnaires' results, the factors affecting the reliability of the test are student-related reliability, rater reliability, test administration reliability, and test reliability. From the interviews' results, the factors influencing the tests' reliability are the mastery of learning material, the preparation of learning before getting the test, students' physical condition, the clarity of instructions of the test, clarity of questions in the test, the atmosphere and time allocation for doing the test.

Kata kunci: English teacher-made test, reliability, senior high school

PENDAHULUAN

Assessment (testing) tools are the way to determine whether students can achieve the learning objective. Assessment or test can determine the success of the learning program (Rixon, as cited in Apsari & Haryudin, 2017). It can also inform students about their mastery of learning materials (Wenno et al., 2021). Assessment is the process of collecting information about learners' achievements to make a good decision (Ulfah et al., 2020). The reasons for testing the students are to assign grades to the students, determine students' knowledge for suitable remediation, and

identify the ineffective instruction. Thus, it can help educators enhance the teaching-learning activity (Reiser & Dick, as cited in Apsari & Haryudin, 2017).

Assessment gives information about instructional decisions, detects learners' strengths and weaknesses in the classroom instruction, and gives them feedback. The assessment also provides immediate feedback to enhance teachers' teaching practices. Teachers should use tests to determine their grades (Tosuncuoglu, 2018).

The test has a crucial role in education at higher levels, especially in university. Types of tests are formative test and summative test. A formative test is the evaluation and analysis of daily learners' learning activities. It helps educator guide learners' learning (Qu & Zhang, 2013). A summative test is conducted after all learning materials are learned and finished at the end of the semester. In summative tests, the educators can detect what learners can remember about the material to give a mark (Qu & Zhang, as cited in Sugianto, 2017). The teacher needs to conduct the summative test to measure students' ability to master all materials in one semester.

In a senior high school, the formative test is conducted for all subjects, including English. Based on the 2013 Curriculum from Indonesian Ministry of Education, English is one of the compulsary subject should be taken by the students. The existence of English subject can be found in 2013 Curriculum revised 2017 published by the Indonesian Ministry of Education. From the 2013 Curriculum also, it is known that English subject has 3 hours per week for classs X, 4 hours a week for class XI, and 4 hours a week for class XII. In this research, the researcher intends to analyze factors influencing the reliability of senior high schools' teacher-made english test in Kebumen regency.

Cosnsidering the background of the study, the researchers have two purposes to be solved through the research. Those purposes are to analyze the reliability of senior high schools teacher-made test in Kebumen regency and to analyze the factors affecting the reliability of senior high schools teacher-made test in Kebumen regency.

Tests have a crucial role in the education for measurement and evaluation processes. The test is one of the vital elements in the learning process. It is a type of evaluation that is a measurement instrument of the learning process. It is an instrument to assess learners' skill or knowledge to take the educational decision (Sugianto, 2017). It measures learners' language proficiency (Hughes, as cited in

Setiabudi et al., 2019). The function of the test is to assess an individual's ability, knowledge, and performance (Brown, as cited in Furwana, 2019). Tests can evaluate the individuals' skills or knowledge in a given standard. In educational practice, tests determine the learners' ability to complete specific tasks (Adom et al., 2020). The test can be designed with the result score provided for learners to detect their ability (Azmi, 2020).

A test must fulfill the characteristics of a good test. It must be valid and reliable (Furwana, 2019). A good test is valid, reliable, objective, practical, and economical (Djiwandono, as cited in Sugianto, 2017; Gyll, & Ragland, as cited in Wenno et al., 2021). By giving a good test, learners have a chance to get a good quality in learning, and its result can enhance the teaching and learning process and determine their grades (Furwana, 2019).

Reliability is the measurement consistency and the scores' stability (Harris, as cited in Sultana, 2015). It means that the same measurement yields the same results (Moser and Kalton, as cited in Taherdoost, 2016). If the teacher conducts the same tests on the same candidates on different occasions, and the results are similar, it is reliable (Arifin, 2018; Heaton, as cited in Sultana, 2015). The principles of reliability are consistency of score, clear instructions, and clear questions (Tosuncuoglu, 2018). The interval between the two tests' administration must not be too long or too short to enhance the reliability (Hughes, as cited in Öz & Özturan, 2018). Reliability is a vital test quality. When the test consistency is achieved, the validity of the test is attained (Linn & Gronlund, as cited in Rosaroso, 2015). To calculate the reliability of certain test, a statistical computation is needed. Zimmerman & Zumbo (2015) define reliability ratio of true-score variance and observed-score variance, where observed-score variance is sum of true and error components. Hinton et.al cited in Taherdoost, (2016) classify the reliability into excellent reliability (0.90 and above), high reliability (0.70-0.90), moderate reliability (0.50-0.70) and low reliability (0.50 and below).

James (2013) and Ary et al. (2010) state that there are three reliability coefficients. The first is test-retest reliability in which the results' consistency from the same samples at different times (Ary et al., 2010; SÜRÜCÜ & MASLAKÇI, 2020). Furhtermore, it is a measure of consistency of the same samples. Expectedly, a reliable instrument must yield similar data (James, 2013). The second is alternating forms or parallel forms of reliability, which is measures reliability using two forms of

an instrument. It has the same domain, the same number of items, the same test specifications, the similar difficulty, and different questions (Ary et al., 2010). The scores are then correlated to measure the coefficient of reliability (James, 2013). The third is reliability as internal consistency in which it tests the homogeneity of items in an instrument (James, 2013). The internal consistency of a test is determined from a single test administration (Rosaroso, 2015).

As cited in Tosuncuoglu (2018), Heaton states that five factors affecting the reliability of the test are first, the extent of the sample of the material selected for testing, means that when the test has more test items, the test will be more reliable. The second is fluctuations in test administration, means test reliability is adversely affected if the test's conditions tend to fluctuate from one administration to another. The third is personal factors, means personal factors are related to a physical and psychological condition such as poor health, fatigue, lack of interest or motivation, anxiety, and sadness. The fourth is test instructions, means when the test instructions are clear, the results of tests will be more reliable. The last is fluctuations in scoring, means subjectivity in scoring may introduce inconsistencies in scores and produce unreliable measurements.

The other theory of things affecting the result of reliability is proposed by Ary et al. (2010) who states that there are six factors affecting reliability. Moreover, she explains in detail that the first factor is the length of the test, in which when the items in the test are greater, the true scores are more representative. Then, the second is group heterogeneity, in which the reliability coefficient will be higher when the learners who take the test are heterogeneous. The third is the individuals' ability, in which when a learner has higher ability, the test will be reliable, but it will not be reliable when a learner has lower ability. The difficulty level of the test also influences the test's reliability. The fourth is the specific technique used for reliability estimation, in which the alternate forms with time lapse technique gives a lower estimation of reliability than either test–retest or split-half procedures. The fifth is the nature of variable being measured, in which academic achievement tests have very high reliability (coefficients of 0.90 or higher). Aptitude tests have lower reliability (0.80 or lower). Personality tests have moderate reliability (0.60 to 0.70). Then, the last is the scoring objectivity, in which inconsistency of scoring reduces the test's reliability.

The other point of view related to things affecting the result of reliability test is brought by Brown (2004) who says that reliability of a test is influenced by four factors. The first factor is student-related reliability, which means the most common learner-related reliability is due to physical or psychology factors for example when a student taking two tests is tired, the test can be unreliable. Then, the second factor is rater reliability, means that inter-rater reliability occurs when two or more scores yield inconsistent scores of the same test because of lack of attention to scoring criteria, inexperience, and inattention. The third is test administration reliability, means when the test is well-administered, the test will be reliable. Then, the fourth is test reliability, means the nature of the test can cause measurement errors.

Dealing with the construction of test itself, test is divided into standardized test and teacher made test. Standardized test is constructed by eligible people, in Indonesia it is called Badan Standar Nasional Pendidikan (BSNP), while teacher-made test, as its name, is a test that is constructed by the teacher (Lebagi et al., 2017). Teacher-made tests are usually criterion referenced tests that are designed to assess student mastery of a specific body of knowledge (Kinyua & Okunya, n.d.). Moreover, Popham as cited in Lebagi et al., 2017 points out that a standardized test is a test, either norm-referenced or criterion-referenced, that is administered, scored, and interpreted in a standard manner.

Additionally, Arifin (2016) argues that teacher-made test is a test constructed by teacher who is going to utilize the test itself and it aims to measure students' mastery on material taught. Commonly, it is administered in daily test, formative test and summative test. By underlying the name of test itself, the writer can conclude that teacher-made test is a test that is constructed by the teacher and will be administered to measure students' mastery after being taught in particular period.

In this research, the researchers take three previous studies. The first study conducted by Setiawaty et al. (2017) was about the validity and reliability of the Indonesian language multiple-choice test in the final examination. The result shows that the reliability index calculation is 0.3657 ≤ 0.6. It means that this instrument cannot be used. The second previous study conducted by Jayanti et al. (2019) was about the validity and reliability of the English national examination. The results showed that the test fulfilled the criteria of validity and reliability. The third previous study conducted by Setiabudi et al. (2019) was about the validity and reliability of a

teacher-made test in a senior high school. The analysis results showed that the test was valid and reliable, but both were in the intermediate category.

The similarity of this research and the previous research is that all research analyses the reliability of the test. The difference between this research and previous research are that the first research analysed the validity and reliability of the Indonesian language multiple-choice test in the final examination, the second research analysed the validity and reliability of the English national examination, the third research analysed the validity and reliability of a teacher-made test in senior high school, while this research analyses the factors affecting the reliability of a test.

METODE PENELITIAN

Seeing the characteristics of the research, this research belongs to qualitative research (Creswell, 2012). Moreover, the research uses a descriptive case study means it is a story about a real world situation facing people or groups and how they addressed it. Furthermore, it is aimed to analyze the sequence of interpersonal events after a certain amount of time has passed. In this research, the aims is to present detailed information on a specific phenomenon to get a deep understanding of the case (Heigham & Croker, 2009). In this study, the researchers presented detailed information about the factors affecting the reliability of a English teacher-made test test in Kebumen. The object of the research was 50 multiple choice test made by the English teacher in MAN 1, MAN 2, and MAN 3 Kebumen. The subject of the research was the students of the eleventh grade from MAN 1, MAN 2, and MAN 3 Kebumen. The researchers chose those schools as the object of the research in order to limit the scope of the research to certain school only. Then, the researchers also want to find out the reliability of the English teacher-made test in a limited area so that the result of the research will be more valid.

Here, the researchers use three instruments namely test, close-ended questionnaire, and in depth interview. The form of the test was in a multiple choice test about the final exam of the semester. The researchers was asking for permition from the teachers to take the data from the test made by the teachers themselves. After getting the permition, then the researchers invoke the teachers to give the paper work of the students to be analyzed then.

In analyzing the reliability of Englsih teacher-made test, the researchers read the result of the students' paper test, then put the score in SPSS, count the reliability result using Cronbach's Alpha, categorizing the result of reliability, explaining the result of reliability, then drawing the conclusion.

The researchers used close-ended questionnaire and in depth interview to recognize the factors affecting the result of reliability test in three of Madrasah Aliyah Negeri, Kebumen Regency. In giving the close ended questionnaire, the researchers asked permition to the English teachers and headmasters of each schools to get the data by using questionnaire and interview. Before giving the questionnaire to the students, the researchers explains them on how to fulfill the questionnaire directly in the classroom, then giving them the link of the questionnaire (google form), asking the students to fulfill the questionnaire, and downloading the result of questionnaire.

In analyzing the questionnaire, the researchers read the result of questionnaire, identifying it, categorizing and explaining factors affecting reliability of the test, and drawing conclusion on it. In interviewing the students, after the researchers got the permition from the teachers and headmasters, then preparing the questions, explaining the procedure of the interview, then interviewing the students while recording it, and keep the recording of the interview to be analyzed then. In analyzing the interview, the researchers see and hear the recording of the interview, identifying it, categorizing and explaining factors affecting reliability test, then drawing conclusions.

FINDINGS AND DISCUSSION

In this section, the researchers will present the findings of the factors affecting the reliability of a test. In this research, the researchers take the data of the final exam score from MAN 1, MAN 2, and MAN 3 Kebumen to be analyzed its reliability of the test. The researchers also gave close-ended questionnaires and conducted one-on-one interviews to analyse the factors affecting the reliability of a test.

Results of the Reliability of a Test

Table 1. The number of students

Case Processing Summary

		N	%
Cases	Valid	29	100.0
	Excluded[a]	0	.0
	Total	29	100.0

a. Listwise deletion based on all variables in the procedure.

Table 2. The result of Cronbach's Alpha

Reliability Statistics

Cronbach's Alpha	N of Items
.839	50

From the table 1, it gives information that the number of students doing the test from MAN 1 Kebumen are 29 students. Because there is no empty data, means that all respondes or students gave their answers to the questions given. Then, from the output of SPSS computation available on table 2, it can be known that the result of Cronbach's Alpha is 0.839 with the total number of items (questions) 50.

Sujarweni (2014) states that the test is said to be Reliable if the Cronbach alpha value > 0.6. In this test, the Cronbach alpha value is 0.839 > 0.6 then. This test is said to be reliable and can be used as a test in research.

Table 3. Cronbach's Alpha if item deleted

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
No_1	48.72	237.135	.291	.836
No_2	48.86	230.909	.505	.831
No_3	48.38	234.530	.399	.833
No_4	48.86	236.052	.332	.835
No_5	48.45	239.185	.234	.837
No_6	48.52	233.973	.402	.833
No_7	49.07	229.638	.588	.829
No_8	48.52	238.687	.245	.837
No_9	49.07	241.209	.178	.838
No_10	47.83	244.791	.139	.838
No_11	48.59	248.608	-.073	.844
No_12	48.45	238.613	.253	.837
No_13	48.62	235.815	.342	.835
No_14	49.10	237.382	.325	.835
No_15	48.62	233.244	.428	.833
No_16	47.83	241.362	.353	.836
No_17	48.24	235.333	.399	.833
No_18	48.38	242.101	.141	.839
No_19	49.07	247.209	-.027	.843
No_20	49.41	252.323	-.248	.845
No_21	48.38	238.815	.252	.837
No_22	48.41	237.751	.291	.836
No_23	49.14	231.195	.552	.830
No_24	48.45	248.185	-.060	.844
No_25	48.79	243.170	.098	.840
No_26	48.52	237.973	.269	.836
No_27	48.79	235.599	.343	.835
No_28	48.24	235.333	.399	.833
No_29	48.62	234.244	.395	.833
No_30	48.28	244.564	.067	.841
No_31	48.83	228.791	.585	.829
No_32	49.00	230.929	.525	.830
No_33	49.03	235.534	.379	.834
No_34	49.10	237.953	.304	.835
No_35	49.34	247.377	-.028	.842
No_36	48.28	240.564	.210	.838
No_37	48.66	240.234	.191	.838
No_38	49.17	240.219	.234	.837
No_39	48.83	231.648	.487	.831
No_40	48.28	241.278	.184	.838
No_41	48.97	250.963	-.151	.846
No_42	48.69	240.436	.189	.838
No_43	49.03	234.820	.404	.833
No_44	49.10	234.953	.414	.833
No_45	48.76	236.047	.334	.835
No_46	48.62	234.387	.390	.833
No_47	48.83	232.219	.467	.832
No_48	49.38	238.672	.369	.835
No_49	48.21	239.884	.246	.837
No_50	48.48	232.259	.472	.832

From the table above, it gives ilustration about the score of students of MAN 1 in doing the test. In the table of Cronbach's Alpha if item deleted, it is seen that the value of of Cronbach's Alpha for the 50 items are more than 0.6 which means that the total 50 questions of final exam made by the teacher of MAN 1 is reliable.

MAN 2 Kebumen

Table 4. The number of students

Case Processing Summary

		N	%
Cases	Valid	35	100.0
	Excluded[a]	0	.0
	Total	35	100.0

a. Listwise deletion based on all variables in the procedure.

Table 5. The result of Cronbach's Alpha

Reliability Statistics

Cronbach's Alpha	N of Items
.747	50

From the table 4, it gives information that the number of students doing the test from MAN 2 Kebumen are 35 students. Because there is no empty data, means that all respondes or students gave their answers to the questions given. Then, from the output of SPSS computation available on table 5, it can be known that the result of Cronbach's Alpha is 0.747 with the total number of items (questions) 50.

Sujarweni (2014) states that the test is said to be Reliable if the Cronbach alpha value > 0.6. In this test, the Cronbach alpha value is 0.747 > 0.6 then. This test is said to be reliable and can be used as a test in research.

Table 6. Cronbach Alpha if item deleted

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
No 1	26.11	34.339	-.107	.757
No_2	25.71	32.034	.362	.736
No 3	25.63	32.064	.441	.735
No_4	25.77	32.358	.268	.740
No_5	25.86	31.420	.419	.733
No 6	25.86	32.479	.223	.742
No_7	26.06	30.997	.485	.729
No_8	25.77	34.417	-.123	.757
No 9	26.03	33.440	.046	.751
No_10	25.57	33.017	.265	.742
No_11	25.71	32.622	.238	.742
No 12	25.91	31.022	.481	.730
No_13	25.74	32.844	.182	.744
No_14	26.43	33.311	.219	.744
No 15	26.14	33.950	-.039	.754
No_16	25.51	33.610	.166	.745
No_17	25.74	33.197	.111	.747
No 18	25.74	32.726	.205	.743
No_19	26.37	33.358	.134	.745
No_20	26.37	33.123	.198	.743
No 21	25.86	33.361	.064	.750
No_22	25.66	33.055	.173	.744
No_23	26.23	34.064	-.057	.754
No_24	25.89	34.692	-.167	.760
No_25	25.86	32.420	.233	.742
No_26	25.83	31.499	.413	.733
No_27	25.51	33.963	-.015	.748
No_28	25.49	33.963	.000	.747
No_29	25.63	31.770	.517	.732
No_30	25.91	31.492	.394	.734
No_31	26.03	30.793	.520	.727
No_32	25.97	31.617	.366	.735
No_33	26.37	33.299	.150	.745
No 34	26.31	33.104	.162	.745
No_35	26.37	33.534	.087	.747
No_36	25.91	30.551	.569	.725
No_37	26.00	32.235	.255	.741
No_38	26.37	33.770	.024	.749
No_39	25.86	31.303	.441	.732
No_40	26.20	32.518	.236	.742
No_41	26.29	33.681	.025	.750
No_42	25.83	32.323	.257	.741
No_43	26.37	32.829	.279	.741
No_44	26.20	32.929	.157	.745
No_45	26.23	33.417	.068	.749
No_46	25.86	32.538	.212	.743
No_47	26.06	30.585	.563	.725
No_48	26.29	34.445	-.136	.756
No_49	25.60	33.953	-.025	.750
No_50	25.77	33.299	.086	.748

From the table above, it gives ilustration about the score of students of MAN 2 in doing the test. In the table of Cronbach's Alpha if item deleted, it is seen that the value of of Cronbach's Alpha for the 50 items are more than 0.6 which means that the total 50 questions of final exam made by the teacher of MAN 2 is reliable.

MAN 3 Kebumen.

Table 7. The number of students

Case Processing Summary

		N	%
Cases	Valid	30	100.0
	Excluded[a]	0	.0
	Total	30	100.0

a. Listwise deletion based on all variables in the procedure.

Table 8. The result of Cronbach's Alpha

Reliability Statistics

Cronbach's Alpha	N of Items
.649	50

From the table 7, it gives information that the number of students doing the test from MAN 3 Kebumen are 30 students. Because there is no empty data, means that all respondes or students gave their answers to the questions given. Then, from the output of SPSS computation available on table 8, it can be known that the result of Cronbach's Alpha is 0.649 with the total number of items (questions) 50.

Sujarweni (2014) states that the test is said to be Reliable if the Cronbach alpha value > 0.6. In this test, the Cronbach alpha value is 0.649 > 0.6 then. This test is said to be reliable and can be used as a test in research.

Table 9. Cronbach Alpha if item deleted

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
No_1	27.90	18.438	.033	.654
No_2	27.83	17.316	.302	.631
No_3	27.87	17.016	.389	.624
No_4	27.87	17.706	.211	.639
No_5	27.83	17.109	.354	.627
No_6	27.23	18.530	.142	.646
No_7	27.23	18.530	.142	.646
No_8	27.23	18.530	.142	.646
No_9	27.23	18.530	.142	.646
No_10	27.23	18.530	.142	.646
No_11	27.53	17.499	.264	.635
No_12	27.93	18.133	.118	.647
No_13	27.47	17.844	.195	.641
No_14	27.90	19.472	-.220	.673
No_15	27.87	18.602	-.011	.658
No_16	27.27	18.271	.208	.642
No_17	27.57	18.944	-.093	.665
No_18	27.57	17.151	.344	.628
No_19	28.17	18.902	-.094	.653
No_20	28.07	18.754	-.029	.655
No_21	27.60	17.076	.355	.626
No_22	27.47	18.740	-.040	.659
No_23	28.10	18.990	-.112	.658
No_24	27.77	18.875	-.078	.664
No_25	27.67	18.782	-.057	.663
No_26	27.83	18.695	-.035	.660
No_27	28.10	17.472	.479	.626
No_28	27.33	17.885	.267	.637
No_29	27.60	16.593	.479	.615
No_30	27.37	18.378	.081	.649
No_31	28.03	17.206	.457	.623
No_32	28.03	16.999	.526	.618
No_33	28.07	18.892	-.075	.657
No_34	28.10	17.472	.479	.626
No_35	28.13	18.464	.118	.646
No_36	27.73	17.789	.173	.643
No_37	27.87	17.913	.159	.644
No_38	28.17	18.764	-.007	.650
No_39	27.83	17.592	.232	.638
No_40	27.90	19.748	-.285	.678
No_41	27.23	18.185	.365	.638
No_42	27.23	18.185	.365	.638
No_43	27.23	18.185	.365	.638
No_44	27.23	18.185	.365	.638
No_45	27.23	18.185	.365	.638
No_46	27.23	18.185	.365	.638
No_47	27.23	18.185	.365	.638
No_48	27.20	18.786	.000	.649
No_49	27.23	18.185	.365	.638
No_50	27.23	18.185	.365	.638

From the table above, it gives ilustration about the score of students of MAN 3 Kebumen in doing the test. In the table of Cronbach's Alpha if item deleted, it is seen that the value of of Cronbach's Alpha for the 50 items are more than 0.6 which means that the total 50 questions of final exam made by the teacher of MAN 3 Kebumen is reliable.

The summary of reliability test

As it has been explained previously, that the researchers will examine the result of students' work in doing the final test. Here, the test were made by the English teachers coming from MAN 1, MAN 2, and MAN 3. After doing the computation using SPSS 16, the summary of the reliability test of teacher-made test coming from 3 schools in Kebumen Regency can be seen as follows:

Table 10. The summary of reliability test

No	School	Result	Conclusion
1	MAN 1 Kebumen	0.839	Reliable
2	MAN 2 Kebumen	0.747	Reliable
3	MAN 3 Kebumen	0.649	Reliable

Results of Close-Ended Questionnaires

In order to answer the second question of the research, the researchers used quesionnaire and interview. These two instruments were used to fit out each other's result in order to get the more valid data.

Table 11. The Factors Affecting the Reliability of a Test

No.	Statement	Strongly Disagree (%)	Disagree (%)	Agree (%)	Strongly Agree (%)
	Student-Related Reliability				
1	When I am sick, I cannot do the test well.	0	13.3	53.3	33.3
2	When I am tired, I cannot do the test well.	3.3	16.7	63.3	16.7
3	When I am sad, I cannot do the test well.	10	33.3	43.3	13.3
4	When I am anxious, I cannot do the test well.	0	20	66.7	13.3
5	When I am angry, I cannot do the test well.	6.7	43.3	36.7	13.3
	Rater Reliability				
6	When my teacher gives unclear instructions in the test, it will influence my score	3.3	3.3	56.7	36.7
7	When my teacher treats students differently, it will influence my score.	3.3	20	56.7	20
8	When my teacher dislikes me, it will influence my score.	3.3	0	40	56.7
9	When my teacher gives unclear scoring criteria, it will influence my score.	3.3	16.7	66.7	13.3
10	When my teacher gives ambiguous questions, it will influence my score.	0	10	63.3	26.7
	Test Administration Reliability				
11	I cannot do the test well when the test item is too much.	0	33.3	50	16.7
12	I cannot do the test well when the weather is too hot.	40	0	46.7	0
13	I cannot do the test well when the internet access is not good.	3.3	0	40	56.7
14	I cannot do the test well when there is too much noise.	0	3.3	60	36.7
15	I cannot do the test well when the place is not comfortable.	3.3	3.3	63.3	30
	Test Reliability				
16	I cannot do the test well when the test is too long.	3.3	30	50	16.7

17	I cannot do the test well when the test has more than one answer.	3.3	36.7	60	0
18	I cannot do the test well in a very limited time.	0	10	46.7	43.3
19	I cannot do the test well when the font size is too small.	10	23.3	53.3	13.3
20	I cannot do the test well when the background of google form is dark (online test) or when the quality of printer ink is low (paper-based test).	13.3	20	50	16.7

In the first statement, no students strongly disagree, 13.3% of students disagree, 53.3% of students agree, and 33.3% of students strongly agree. In the second statement, 33.3% of students strongly disagree, 16.7% of students disagree, 63.3% of students agree, and 16.7% of students strongly agree. In the third statement, 10% of students strongly disagree, 33.3% of students disagree, 43.3% of students agree, and 13.3% of students strongly agree. In the fourth statement, 10% of students strongly disagree, 33.3% of students disagree, 43.3% of students agree, and 13.3% of students strongly agree. In the fourth statement, no students strongly disagree, 20% of students disagree, 66.7% of students agree, and 13.3% of students strongly agree. In the fifth statement, 6.7% of students strongly disagree, 43.3% of students disagree, 36.7% of students agree, and 13.3% of students strongly agree. In the sixth statement, 33.3% of students strongly disagree, 33.3% of students disagree, 56.7% of students agree, and 36.7% of students strongly agree. In the seventh statement, 33.3% of students strongly disagree, 20% disagree, 56.7% of students agree, and 20% strongly agree. In the eighth statement, 33.3% of students strongly disagree, no students disagree, 40% of students agree, and 56.7% of students strongly agree. In the ninth statement, 3.3% of students strongly disagree, 16.7% of students disagree, 66.7% of students agree, and 13.3% of students strongly agree. In the tenth statement, no students strongly disagree, 10% of students state disagree, 63.3% of students agree, and 26.7% strongly agree.

In the eleventh statement, no students strongly disagree, 33.3% of students disagree, 50% of students agree, and 16.7% of students strongly agree. In the twelfth statement, 40% of students strongly disagree, no students disagree, 46.7% of students agree, and no students strongly agree. In the thirteenth statement, 3.3% of students strongly disagree, no students disagree, 40% of students agree, and 56.7% of students strongly agree. In the fourteenth statement, no students strongly disagree, 3.3% of students disagree, 60% of students agree, and 36.7% strongly

agree. In the fifteenth statement, 3.3% of students strongly disagree, 33.3% of students disagree, 63.3% of students agree, and 30% strongly agree. In the sixteenth statement, 3.3% of students strongly disagree, 30% of students disagree, 50% of students agree, and 16.7% of students strongly agree. In the seventeenth statement, 3.3% of students strongly disagree, 36.7% of students disagree, 60% of students agree, and no students strongly agree. In the eighteenth statement, no students strongly disagree, 10% of students disagree, 46.7% of students agree, and 43.3% of students strongly agree. In the nineteenth statement, 10% of students strongly disagree, 23.3% of students disagree, 53.3% of students agree, and 13.3% of students strongly agree. In the twentieth statement, 13.3% of students strongly disagree, 20% of students disagree, 50% of students agree, and 16.7% of students strongly agree.

Based on the questionnaires' results, the factors affecting the reliability of a test are student-related reliability, rater reliability, test administration reliability, and test reliability.

Results of Interview

Here, the researchers present the results of one-on-one interview to sixteen students of MAN 1, MAN 2, MAN 3 Kebumen. From the interview, the researchers found some facts. Related to the score, the students who got high scores on the test stated that they had learned and understood the material well. Students who got medium scores on the test stated that they did not understand one or some parts of the material. Students who got low scores on the test stated that they did not learn the material and did not understand the material. The students who got high and medium scores stated that they had enough time to do the test, so they got good scores. However, the students who got low scores stated that the time was insufficient. There was an influence of the time of the test on students' tests' results.

Then, related to the difficulties when taking the test, the students' difficulty when taking the test was because they did not master the learning material taught by their teachers. Furthermore, they said that they needed to study well to get good scores. Moreover, the students said that the effects of studying before the tests were that they could understand the material, get good scores, and feel more confident. Related to the students' healthy, the students' physical condition was good. All of them were healthy.

Related to clarity of the test instruction, the students stated that the clarity of the test instructions would significantly affect the test results. If the test instructions were not clear, they could not do the test and get bad scores. Furthermore, the students stated that the effects of the clarity of the instructions were that they could understand the test, do the test more quickly, and get better scores' results.

About the scoring criteria and test's atmosphere, the students stated that the scoring criteria were good and very clear. They stated that the unclear scoring criteria would affect their scores because of the inconsistency of scoring. Moreover, the students stated that the atmosphere when students did the test was calm and comfortable. The students also stated that there were no confusing questions in the test. The confusing test items would influence the scores obtained.

Seeing the result of those analysis above, the researchers then realized that the test made by teachers of English subject from MAN 1, MAN 2, and MAN 3 belongs to reliable. This findings are almost similar with the result of the research conducted by Setiabudi et al. (2019) in which the result showed that the test was valid and reliable. The difference is Setiabudi's result shows that the category of reliability belongs to intermediate category, meanwhile this research shows that the category of reliability belongs to high category. Some factors affecting the result of questionnaire based on questionnaire and interview are the test itself, the students' preparation, and the atmosphere during the test done. This findings has the similarity with the research conducted by Setiawaty et al. (2017) and Jayanti et al. (2019) which found the same thing as this research found.

CONCLUSION

After doing a series of steps in the research from analyzing the result of students's paper test, analyzing the questionnaire of the students, then analyzing the result of interview, the researchers come to conclusion. Based on the research results, the reliability of the tests are 0.839 (MAN 1 Kebumen), 0.747 (MAN 2 Kebumen), and 0.649 (MAN 3 Kebumen), so the test made by the English teacher of each school are reliable. Viewed from the questionnaires' results, the factors affecting the reliability of the test are student-related reliability, rater reliability, test administration reliability, and test reliability. Viewed from the interviews' results, the

factors affecting the reliability of the test are the mastery of learning material, the preparation of learning before getting the test, students' physical condition, the clarity of instructions of the test, clarity of questions in the test, the atmosphere and time allocation for doing the test.

SUGGESTION

Based on the result of the research above, although all the teacher-made test of MAN 1, MAN 2, and MAN 3 are reliable, the researchers then will have community service which is conducting a Language Assessment and Evaluation training to make the quality of the teacher-made test better. Then, the researchers also will conduct similar research covering larger area.

REFERENCES

Adom, D., Mensah, J. A., & Dake, D. A. (2020). Test, measurement, and evaluation: Understanding and use of the concepts in education. *International Journal of Evaluation and Research in Education*, *9*(1), 109–119. https://doi.org/10.11591/ijere.v9i1.20457

Apsari, Y., & Haryudin, A. (2017). The Analysis Of English Lecturers' Classroom-Based Reading Assessments To Improve Students' Reading Comprehension. *ELTIN Journal*, *5*(1), 35–44.

Arifin, M. A. (2018). Validity, Reliability and Practicality Of The First Certification in English (FCE) and The Business Language Testing Service (Bulats). *Journal of Language Teaching and Learning, Linguistics and Literature*, 6(2), 80–95.

Ary, D., Jacobs, L. C., Sorensen, C., & Razavieh, A. (2010). Introduction to Research in Education (8th ed.). Wadsworth Cengage Learning.

Azmi, U. (2020). Developing web-based reading tests for the students of English language education. *Journal of Applied Linguistics, Translation, and Literature*, 1(2), 92–104.

Brown, H. D. (2004). Language Assessment: Principles and Classroom Practices. Longman.

Creswell, J. W. (2012). Educational Research: Planning, Conducting, and Evaluating *Quantitative and Qualitative Research* (Fourth). Pearson Education.

Furwana, D. (2019). Validity and Reliability of Teacher-Made English Summative Test at Second Grade of Vocational High School 2 Palopo. *Language Circle: Journal Lof Language and Literature,* 13(2). http://journal.unnes.ac.id

Heigham, J., & Croker, R. A. (2009). Qualitative Research in Applied Linguistics: A Practical Introduction (First). Palgrave Macmillan.

James, C. (2013). Errors in Language Learning and Use: Exploring Error Analysis. Routledge.

Jayanti, D., Husna, N., & Hidayat, D. N. (2019). The Validity and Reliability Analysis of English National Final Examination for Junior High School. Voices of English Language Education Society, 3(2), 127–135.

Kinyua, K., & Okunya, L. O. (n.d.). Validity and reliability of teacher-made tests: Case study of year 11 physics in Nyahururu District of Kenya. 11.

Lebagi, D., Sumardi, S., & Sudjoko, S. (2017). The Quality of Teacher-Made Test in EFL Classroom at the Elementary School and its Washback In The Learning. *Journal of English Education,* 2(2), 97–104. https://doi.org/10.31327/jee.v2i2.289

Öz, H., & Özturan, T. (2018). Computer-based and paper-based testing: Does the test administration mode influence the reliability and validity of achievement tests? *Journal of Language and Linguistic Studies*, *14*(1), 67–85.

Qu, W., & Zhang, C. (2013). The Analysis of Summative Assessment and Formative Assessment and Their Roles in College English Assessment System. *Journal of Language Teaching and Research*, 4(2), 335–339. https://doi.org/10.4304/jltr.4.2.335-339

Rosaroso, R. C. (2015). Using Reliability Measures in Test Validation. European Scientific Journal, 11(18), 369–377.

Setiabudi, A., Mulyadi, & Puspita, H. (2019). An Analysis of Validity and Reliability of A Teacher-Made Test (Case Study at XI Grade of SMA N 6 Bengkulu). *Journal of English Education and Teaching*, 3(4), 522–532.

Setiawaty, R., Sulistyorini, T. B., Margono, & Rahmawati, L. E. (2017). Validity Test and Reliability of Indonesian Language Multiple Choice in Final Term Examination. *The 1st International Seminar on Language, Literature and Education*, *2018*, 43–50. https://doi.org/10.18502/kss.v3i9.2609

Sugianto, A. (2017). Validity and Reliability Of English Summative Test For Senior High School. *Indonesian EFL Journal: Journal of ELT, Linguistics, and Literature*, 3(2), 22–38.

Sultana, R. (2015). Reliability of the Currently Administered Language Tests in Bangladesh: A Case Study. *Journal of Literature, Languages and Linguistics,* 17, 76–85.

Sürücü, L., & Maslakçi, A. (2020). Validity and Reliability in Quantitative Research. Business & Management Studies: *An International Journal*, 8(3), 2694–2726. http://dx.doi.org/10.15295/bmij.v8i3.1540

Taherdoost, H. (2016). Validity and Reliability of the Research Instrument; How to Test the Validation of a Questionnaire/Survey in a Research. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3205040

Tosuncuoglu, I. (2018). Importance of Assessment in ELT. *Journal of Education and Training Studies*, 6(9), 163–167. https://doi.org/10.11114/jets.v6i9.3443

Ulfah, A. A., Kartono, & Susilaningsih, E. (2020). Validity of Content and Reliability of Inter-Rater Instruments Assessing Ability of Problem Solving. *Journal of Educational Research and Evaluation*, *9*(1), 1–7.

Wenno, I. H., Tuhurima, D., & Manoppo, Y. (2021). How to Create a Good Test. Jurnal Pendidikan Profesi Guru Indonesia, 1(1), 11–20.

Zimmerman, D. W., & Zumbo, B. D. (2015). Resolving the Issue of How Reliability is Related to Statistical Power: Adhering to Mathematical Definitions. *Journal of Modern Applied Statistical Methods*, 14(2), 9–26. https://doi.org/10.22237/jmasm/1446350640