Developing Conservation-based English Proficiency Test as Implementation of Strengthening Universitas Negeri Semarang's Vision

Amir Sisbiyanto
English Department
Universitas Negeri Semarang
Semarang, Indonesia
amier_sis@yahoo.com

Mohamad Ikhwan Rosyidi
English Department
Universitas Negeri Semarang
Semarang, Indonesia
mirosyidi@mail.unnes.ac.id

Lukman Hardi
Language and Education Training Center
Universitas Negeri Semarang
Semarang, Indonesia
lukmanhardi@mail.unnes.ac.id

Abstract - the aim of this research is to investigate the validity and reliability of English proficiency test items as a measure of students' English achievement for requirement to have final examination and to explain conservation context of test items as manifestation of UNNES vision. This research applied research and development method of educational research (see Gall et. al., 2003; Sukmadinata, 2016; Sugiyono, 2017). It used descriptive, evaluative, and experimental methods. It gave a description of condition in UNNES, especially students' English proficiency, first prototype draft which would be tested, and result of comparison to different English test prototype of UNNES test takers The making of test items which reflects on the conservation-based materials gives significant improvement of the test takers' score of UNNES.

Keywords— English proficiency test, Computer-based, research and development, Universitas Negeri Semarang

I. INTRODUCTION

World Bank released data that fresh graduates of University entered labor market without any skills. It encouraged them to fulfill job opportunity which did not require particular skill (The World Bank Office Jakarta, 2014). It indicates irony in developing skilled labor in Indonesia. University as one of pioneers to create skillful graduates has been challenged to solve it. Universitas Negeri Semarang has challenged this condition by making a policy in producing particular skill, especially language skill, for its graduates. Rector of UNNES in 2015 responded the challenge by issuing Rector's Decree relating to an obligatory requirement for students to exit UNNES. The decree required students to take English Proficiency Test before they take final examination of their undergraduate thesis, thesis, and dissertation (Decree of UNNES Rector Number 465/P/2015). In a consequence, students of UNNES have to take the test as prerequisite before taking the exam.

Assessment is an ongoing process to measure something. It can bring students' response, comments, and tries out a new word or a new structure (Brown, 2003). Test is a subset of assessment, and it can be a useful device to measure student's performance (Brown, 2003). A test then is a method of measuring students' ability, knowledge, or performance in a given domain (Brown, 2003). It has at least three components. They are method, measurement, and students. A method is a set of techniques, procedures, and items (Hornby, 1995: Brown, 2003). It is a method since it uses some particular techniques, it entails steps to do the analysis, and it has also some items used to measure students' performance. The word measure is generally to find size in standard unit (Hornby, 1995). Measurement in language assessment becomes a means of offering test-takers some kinds of result (Brown, 2003). Students' performance is seen as process or product they produce in classroom activities. The product or the process must be linked to students' language competence and knowledge (Brown, 2003). In another word, a test is one of realization of assessment. It has techniques to measure students' language competence and knowledge.

The requirement of taking English proficiency test for students of UNNES leads to produce regular new tests. Language and education training center as an institution having responsibility to escort this UNNES policy has been trying to create this test. Every year, the centers produce new test items. To develop a test that would be acceptable, it has to have the qualities of validity and reliability. The experience in the current project has shown that these test qualities are not achievable when considerable time, resources, administrative support, and budget are not made available (Akmar, Abidin, & Jamil, 2015). The development of test items creates a problem. The problem is about in what reasoning the centers need to make test items. The productions lead to questions about the validity and reliability of preceded test items which are significant or not. The validity and reliability of preceded test item and the recent on give reasoning for policy making relating to make new test item. Result of validity and reliability has no doubt indicated the difficulty of students who are test-takers to do test. The difficulty of doing the test delivers a problem to them. The question raised is whether they are guided to do the test item in cultural context of UNNES as conservation university or not. Another question can be whether the context of test item does not include some particular contexts of conservation as one of UNNES vision. Those questions above crystalize on problems in accordance to validity and reliability of test items and conservation context as manifestation of UNNES vision.

Test development starts with the concern that a test can be shown to produce scores that are an accurate reflection of UNNES students' ability in a particular area, such as reading for specific ideas, writing a research proposal, breadth of vocabulary knowledge, or speaking in a class presentation (Akmar et al., 2015). Testing also has an ethical dimension, which many in the testing field have referred to as consequential validity (Akmar et al., 2015). A good case in point embracing this perspective is the assessment series such as good practice in the testing or assessment of a specific language ability or skill, including reading, listening, grammar (Fan & Jin, 2013). "Theory-based validity" concerns the test takers; elements include language and content knowledge that test takers possess and the processes or strategies that they utilize when performing the test task. These are important considerations regarding the test takers, which test developers need to be mindful of when creating the test (Akmar et al., 2015).

The studies show different investigations. This study examined the comparability of reading and writing tasks of two English language proficiency tests—the General English Proficiency Test-A (GEPT-A) developed by Language Training Center, Taipei and the Internet-Based Test of English as a Foreign Language (iBT) developed by Educational Testing Service, Princeton. The results of the text analysis showed the reading passages on the two tests are comparable in many ways but differ in several key regards. The task analysis revealed that the construct coverage, item scope, and task formats of the two tests are clearly distinct. Analysis of test performance showed that scores on the GEPT-A and iBT are highly inter-correlated with each other. Exploratory and confirmatory factor analyses of the test score data indicated that the two tests appeared to be measuring reading and writing ability but emphasize different aspects of the reading construct (Kunnan & Carr, 2017).

Another study also examines the use of qualitative data analysis to improve students' English language performance while taking an ESP course, which gives an insight into what teacher needs for a successful English language course and lesson planning, and that means having an overlook into gray areas of the English language that largely cause problems for students. At present, however, most language teaching and testing around the world provide the kind of tripartite service for students: teach language courses, conduct exams and produce learning and test practice materials. The shortcomings of a diagnostic test, however, can be large groups of students (as is the case at the University of Niš Medical School) so that the kind of service advocated here could be provided for all of them which leaves options open to advanced students only, that is, those who need English for professional reasons, and approximate to native speaker level (Baki, 2012).

Those studies become the foundation of this research. Therefore, the objectives of this research are to investigate the validity and reliability of English proficiency test items as a measure of students' English achievement for requirement to have final examination and to explain conservation context of test items as manifestation of UNNES vision. Although validity and reliability of test items having no evidence limited, there is no previous published studies about it (Baldauf, JR, 1978) in UNNES.

II. METHOD

In the beginning stages of the test development process, the participants were randomly sampled from Undergraduate students currently enrolled in Universitas Negeri Semarang (UNNES). When the students has completed the test and obtains a score, they was placed on a band, in line with the levels of the TOEFL score. Their score was analyzed by investigating the validity and reliability of the test.

Having this result, the researchers develop new test items reflected on conservation insight regarding UNNES as conservation university. The conservation-based test items them are tested to participants as randomly sampled. They were still Undergraduate students currently enrolled in Universitas Negeri Semarang (UNNES). The test is resulted in the score to be investigated as the validity and reliability.

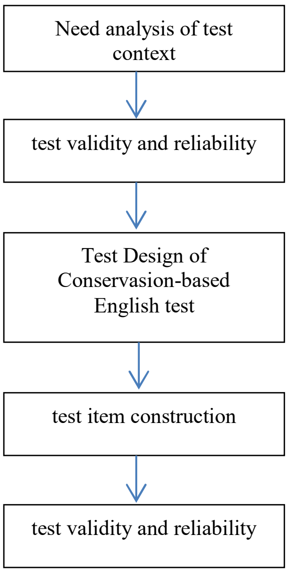
Need analysis of test context
test validity and reliability
Test Design of Conservasion-based English test
test item construction
test validity and reliability

Figure
1. Diagram of Test Development Style
The diagram is adapted from Test development style (Akmar et al., 2015)

III. FINDING AND DISCUSSION

First phase of investigating the need analysis of the test is by giving students of UNNES existed test items. There three sections here. They are listening, grammar, and reading. The skills are adapted from Test of English as Foreign Language (TOEFL). There are 140 test items tested. The participants are

57 students taken randomly from various non-English students for different study programs in UNNES. The aim of testing those participants is investigate that the test item is appropriate for a research. The test was taken for about two hours. It was located in Language Laboratory of Language and Education Training Centers of UNNES. Result of the test was:

Table 1. Final score of the first phase test takers

First Phase Test

No	Code	Score
1	T-01	27
2	T-02	37
3	T-03	19
4	T-04	26
5	T-05	35
6	T-06	37
7	T-07	33
8	T-08	27
9	T-09	38
10	T-10	34
11	T-11	39
12	T-12	28
13	T-13	35
14	T-14	31
15	T-15	37
16	T-16	33
17	T-17	25
18	T-18	32
19	T-19	36
20	T-20	39
21	T-21	36
22	T-22	41
23	T-23	33
24	T-24	32
25	T-25	27
26	T-26	22
27	T-27	24
28	T-28	32
29	T-29	27
30	T-30	83
31	T-31	34
32	T-32	33
33	T-33	31
34	T-34	30
35	T-35	27
36	T-36	30
37	T-37	29
38	T-38	39
39	T-39	31
40	T-40	43
41	T-41	31
42	T-42	30
43	T-43	42
44	T-44	37
45	T-45	34
46	T-46	32
47	T-47	31
48	T-48	34
49	T-49	34
50	T-50	34
51	T-51	28
52	T-52	38
53	T-53	31
54	T-54	36
55	T-55	34
56	T-56	31
57	T-57	33
	TOTAL	1865
	MEAN	32,719298

The table above shows the final score of the participants. The total score is 1865. The mean of the scores is 32,719298. It shows that most participants got scores around 30-40. If the total score for one test item is 140, it will give picture that most participants are difficult to do the test items. The difficulty of doing the test item can be seen from the following example of a test item.

The city of Beverly Hills is surrounded on ________ the city of Los Angeles.

a. its sides
b. the sides are
c. it is the side of
d. all sides by

(cited from Phillips, 2001)

The participants are difficult to choose which one is correct since the test item choices are difficult to be differentiated. They are also difficult to contextualize the choices into correct order.

Besides grammar, the participants are also difficult to do reading test item. Below is one example.

Reading Passage

The locations of stars in the sky is relative to one another do not appear to the naked eye to change, and as a result stars are often considered to be fixed in position. Many unaware stargazers falsely assume that each star has its own permanent home in the nighttime sky….

The expression “naked eye” in line 1 most probably refers to ….

a. a telescope
b. a scientific method for observing stars
c. unassisted vision
d. a camera with a powerful lens

(cited from Phillips, 2001)

The example test item above requires an expansion of vocabulary. It is difficult to be done since the participants did not understand the context of thing questioned. The mistake the participants chose indicates misunderstanding of the connected idea contextually. It relates to their proficiency of English.

After describing the test item example in first phase test, the researchers explores validity and reliability. Validity is the extent to which inferences made from assessment results are appropriate, meaningful, and useful in terms of the purpose of assessment (Brown, 2000). Validity shows a test instrument has standard or criteria whether it is valid or not. A test instrument is categorized valid if it has high standard validity. The researcher uses content validity to measure the test item.

Content validity or content evidence validity (Brown, 2000) gives measurement for sampling subject matter about which conclusion are drawn, and it requires the test taker to perform the behaviour that is being measured. The researcher uses this validity for the test item is categorised to direct test (Brown, 2000). The researcher uses Person Product Moment (Brown, 2000) to measure test item validity. It is categorized valid if r_{xy} for each test item is higher than r_{table} ($r_{xy} > r_{table}$) (Creswell, 2012). The result shows that 57 participants of taking test and 5% significance level has reached r_{table} 0,874. Below is test item validity table.

Table 2: Test Item Validity of Second Phase Test Items

Status	Item number	Total
Valid	1, 2, 3, 11, 12, 13, 16, 17, 19, 20, 31, 32, 33, 38, 39, 40, 42, 48, 49, 52, 60, 63, 65, 72, 83, 84, 86, 88, 90, 92, 100, 103, 105, 111, 114, 117, 119, 125, 126, 127, 128.	41
Not Valid	4, 5, 6, 7, 8, 9, 10, 14, 15, 18, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 34, 35, 36, 37, 41, 43, 44, 45, 46, 47, 50, 51, 53, 54, 55, 56, 57, 58, 59, 61, 62, 64, 66, 67, 68, 69, 70, 71, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 85, 87, 89, 91, 93, 94, 95, 96, 97, 98, 99, 101, 102, 104, 106, 107, 108, 109, 110, 112, 113, 115, 116, 118, 120, 121, 122, 123, 124, 129, 130	99

The above table shows that 41 test items are categorized valid, and 99 test items are evidently categorized not valid. The non-valid categorization is because of mistakes made by 57 test takers or participants. The participants cannot answer 99 test items since they could not catch the 'idea' of the test. They misunderstood the instruction directed within the test items.

Based on the finding above, the researchers reconstruct the test items. The 99 test items which are not valid are reconstructed based on conservation perspective. The conservation perspective is included to put in test takers or participants' mind something that is not far for daily mind-set of them. They are engaged on conservation-based curriculum, courses, and activities since conservation becomes a 'spirit' of UNNES academia. Some of the test items replaced with conservation-based items are:

________ Java Man, who lived before the first Ice Age, is the first manlike animal.

A. it is generally believed that
B. Generally believed it is
C. Believed generally is
D. That is generally believed

(Adapted from Phillips, 2001)

Of all the cereals, rice is the one food ________for more people than any of the other grain crops.

A. it provides
B. that providing
C. provides
D. that provides

(Adapted from Phillips, 2001)

Those test items show something relating to conservation value in UNNES. They give pre-understanding to the test takers since the items are familiar to them. Evidences of the increasing validity and reliability can be seen from the table below.

Table 3: Second Phase Test

No	Code	Score
1	T-01	28
2	T-02	47
3	T-03	37
4	T-04	48
5	T-05	22
6	T-06	40
7	T-07	48
8	T-08	37
9	T-09	35
10	T-10	56
11	T-11	45
12	T-12	45
13	T-13	41
14	T-14	37
15	T-15	38
16	T-16	43
17	T-17	43
18	T-18	31
19	T-19	29
20	T-20	45
21	T-21	37
22	T-22	36
23	T-23	36
24	T-24	38
25	T-25	43
26	T-26	45
27	T-27	55
28	T-28	42
29	T-29	50
30	T-30	49
31	T-31	38
32	T-32	40
33	T-33	41
34	T-34	35
35	T-35	44
36	T-36	41
37	T-37	36
38	T-38	44
39	T-39	42
40	T-40	39

Table 3: Second Phase Test, cont

41	T-41	42
42	T-42	37
43	T-43	30
44	T-44	42
45	T-45	52
46	T-46	31
47	T-47	40
48	T-48	41
49	T-49	35
50	T-50	58
51	T-51	43
52	T-52	44
53	T-53	48
54	T-54	35
55	T-55	34
56	T-56	34
57	T-57	50
	TOTAL	2275
	MEAN	39,9123

The table shows the increasing number of the correct items or total score done by the test takers. It indicates that the test items given are more acceptable to be understood since the content of the items is not far from the test takers' background knowledge. The mean of the score is also increasing, from 32,719298 to 39,9123. The increasing number of the mean also gives significant result of the test takers' understanding.

Dealing with normality test, the test is conducted to find out whether the data is distributed normally or not. In the context of testing *paired t test*, data are distributed normally. It can be seen the following table.

Table 4: The Results of Normality Test

One-Sample Kolmogorov-Smirnov Test

		Posttest
N		57
Normal Parameters[a,b]	Mean	40,7719
	Std. Deviation	6,95141
Most Extreme Differences	Absolute	,079
	Positive	,079
	Negative	-,063
Kolmogorov-Smirnov Z		,593
Asymp. Sig. (2-tailed)		,874

a. Test distribution is Normal.

b. Calculated from data.

The table above shows testing criteria. It is written the data above that if score $ig. > \alpha$, the data distributes normally. The score Sig is 0,874. It is more α score.

Besides that, the researchers also test *Paired T Test*. It can be seen from the table below.

Table 5: The Results of Paired T Test

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	First phase test	33,2982	57	8,27208	1,09566
	Second Phase test	40,7719	57	6,95141	,92074

The table above shows the mean of 57 test takers is different. The data is for 57 samples. The average variable of first phase test is 33,2982, and The average variable of first phase test is 40,7719. It gives description that the test items from second phase test are more valid and reliable than the first phase test items.

Table 6: The Results of Paired Samples T - Test

Paired Samples Test

		Paired Differences					t	df	Sig. (1-tailed)
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper			
Pair 1	First phase test – Second phase test	-7,47368	10,60846	1,40513	-10,28849	-4,65888	-5,319	56	,00000095

The table above describes the score $Sig.$ is 0,00000095. Since $Sig.$ is smaller to α, it can be inferred that changing content of the test items into conservation-based materials is significant.

IV. CONCLUSION

UNNES policy about testing students for requirement before having final examination has been implemented by making test items for English Proficiency test. The making of test items which reflects on the conservation-based materials gives significant improvement of the test takers' score of UNNES. This research basically gives opportunity to develop especially in testing some different test items provided in Language and Education Training Centers of UNNES.

References

[1] Akmar, S., Abidin, Z., & Jamil, A. (2015). Toward an English Proficiency Test for Postgraduates in Malaysia, 1–10. http://doi.org/10.1177/2158244015597725

[2] Baki, N. (2012). SIGNIFICANCE OF CREATING A CUSTOM DIAGNOSTIC ENGLISH LANGUAGE TEST IN ENGLISH FOR SPECIFIC PURPOSES COURSE AT, 31–35. http://doi.org/10.5633/amm.2012.0105

[3] Baldauf, JR, R. B. (1978). THE VALIDITY OF THE MICHIGAN TEST OF ENGLISH Michigan Test of English Language Proficiency (MTELP) (Division. *Educational and Psychological Measurement*, *38*, 429–432.

[4] Brown, H. D. (2000). Principles of Language Learning and Teaching. New York: Pearson Education.

[5] Creswell, J. W. (2012). *Educational Research: Planning, Conducting and Evaluating* Quantitative *and Qualitative Research*. (P. A. Smith, Ed.) (4th Editio). Boston: Pearson Education, Inc. Retrieved from www.pearsonhighered.com

[6] Fan, J., & Jin, Y. (2013). A survey of English language testing practice in China : the case of six examination boards, 1–16.

[7] Gall, M.D., Gall, J.P. & Borg, W.R. 2003. *Educational Research*. Boston: Pearson Education.

[8] Kunnan, A. J., & Carr, N. (2017). A comparability study between the General English Proficiency Test- Advanced and the Internet-Based Test of English as a Foreign Language. http://doi.org/10.1186/s40468-017-0048-x

[9] Phillips, Deborah. 2001. *Longman Complete Course for the TOEFL test*: preparation *for the computer and paper test*. New York: Pearson Education.

[10] Sugiyono.2017. *Metode Penelitian Pendidikan (Pendekatan Kuantitatif, Kualitatif, dan R&D)*. Bandung: Alfabeta

[11] Sukmadinata, Nana Syaoqih. 2016. *Metode Penelitian Pendidikan*. Cet. 16. Bandung: PT Remaja Rosdakarya.