Scope of English Language Teaching, Literature and Linguistics
Program Studi Pendidikan Bahasa Inggris FKIP
Universitas Muslim Maros
Vol. 3 No. 2, December 2020, pp. 55-64
https://ejournals.umma.ac.id/index.php/seltics
seltics@umma.ac.id, p-ISSN: 2623-2642, e-ISSN: 2655-5417

DEVELOPING COGNITIVE ENGLISH TEST ON RECOUNT TEXT

Nana Erna
Universitas Sawerigading Makassar
nanaerna.enna@gmail.com

ABSTRACT

This study is development research. The purpose of the study is to develop a cognitive test that is valid and reliable, with a test that has the level of difficulty, differentiator, and effectiveness of distracter on the recount text material for the first semester students of the English Education Department at Sawerigading University of Makassar. The desired product of the study is a valid and reliable cognitive test that has the level of difficulty, differentiator, and effectiveness of distracter. The development process of the product employed formative research. The validity of the data obtained was analyzed systematically and categorized based on the set standard. The reliability was calculated by employing the ITEMAN program; whereas, the level of difficulty, differentiator, and distracter was analyzed by employing ANATES version 4 analysis. The trial was conducted at Sawerigading University of Makassar. The result obtained from the trial was a cognitive test with 23 questions in multiple-choice test out of 25 questions made in the first prototype which was valid and reliable. The test has the level of difficulty in the category of easy by 26%, medium by 52%, and difficult by 22%. The differentiator in all questions which consisted of 23 questions was at the range of 0.40. It has 0.1 categorized as good. The effectiveness of the distracter was 87% or 20 items with well-functioned distracter.

***Keywords**:Cognitive Test, Analysis of Item Questions, Recount Text Material*

INTRODUCTION

The written test was an assessment technique that was often used to assess the students' learning achievement. It can obtain information that describes the ability of the students therefore the preparation of the test at the end of the semester should also receive serious attention so that test results can reflect the real ability of the students. According to Sax (cited in Arifin, 2011), a test may be defined as a task or series of the task used to obtain systematic observation presumed to be representative of educational or psychological traits or attributes. Especially for the multiple-

choice test, it must be valid and real. Valid means established if an instrument provides a measure of what it purports to measure, while real means the stability or consistency of the test score or other evaluation result from one measurement to another. A test is called reliable when a student's score on it compared to scores of his classmates is similar to another test in the same information. The reliability of the test scores is typically reported through reliability test exactly through statistic procedure.

The reliability can be measured by three criteria firstly, stability which means the constancy of a test to measure the same phenomenon at different times. Secondly, the dependability that shows the steadiness of the test. Thirdly, the predictability that shows the ability of the test for predicting the results on the measurement of the next symptoms, and it will improve the reliability of the test. Ahmad (2010) said that the reliability of the test is suitability between two efforts that are conducted to measure the same thing through a similar method. Besides valid and reliable, the test must also have a good differentiator and level of difficulty. Purwanto (2013) explained that the differentiator is the ability of test items in knowing the students' learning outcomes, and distinguish who have high ability and low ability. Differentiator should be kept positive and as high as possible. The items of questions have a high positive differentiator. It means that the items of questions can distinguish the top group students and the lower group students. The top group students are the students who are classified as proficient or achieve a total score of high learning outcomes. While the lower groups are the students who obtain a lower total score learning outcomes. Further explained by Suparji (2010) that the distribution of difficulty test items is the instrument must be able to distinguish the group of good students and the group of less intelligent students.

According to Sukiman (2012) the level of difficulty (difficulty index) in an assessment is using the approach of a normal reference assessment, for both easy and difficult questions, tends to produce a low level of reliability. This is because the test results are simple with a limited distribution of test scores that are difficult to answer well. For a simple test, the score will be at the top end of the scale. For both tests (easy and difficult), the difference between learners tends to be very small and cannot be trusted.

From, the aspect of the process, a phenomenon observed in many lecturers did not design good questions, especially for English lecturers. It was supported by the previous research conducted by Pardiyono (2007). In this case, he found that there were some problems faced by the lecturer in designing a test. First, they gave a direct question to the students tested without analysis. Second, items of the question made tend to be in the form of low-level cognition.

Third, they had not known how to design a good question. Fourth they have not fully done the analysis yet, material based on the student textbook. Fifth, they did not consider the level of difficulty of the questions. Based on the phenomenon the lecturer gives a hard or easy test that made them difficult to distinguish the real ability of the students. Besides, it was difficult for the lecturer to diagnose the learning difficulties of the students. Finally, there was no feedback or improvement in the teaching and learning process.

Regarding these problems, the lecturer should perform an analysis of the test and it was a step that was taken to determine the degree of quality of the test, both tests in whole or items of the test. The tests used by the teacher had better quality in many respects. Tests were prepared by the principles and procedures of the preparation of the test. The principles of a good test were validity, reliability, objectivity, practicality, and economist (Arikunto, 2002). Then, there were procedures for the test. First, the researcher determined the level of difficulty of the questions. Second, she determined the differentiator questions. Last, she determined the pattern of the answer to the questions.

Related to the solution, the researcher was interested in developing cognitive English tests, especially for recount text material. She chose this material because most of the students' textbook explained about recount text. The material developed was a daily test. The researcher hoped that this research can solve the problems stated previously by designing questions, especially for recount text material. It was expected to meet the criteria of a good test, so it could show the real ability of the students.

METHOD

This research was Research and Development (R & D). The researcher applied the type of formative research (Tessmer, in Rahayu, T., Purwoko & Zulkardi. 2008), that was more appropriate to this research because the type of formative research consists of some analysis and trials of the test. It makes a good test. The type of formative research (Tessmer, in Rahayu, T., Purwoko & Zulkardi. 2008) consists of several steps. They are 1) *Self Evaluation,* 2) *Prototyping,* and 3) *Product*.

Self-evaluation was the first step of research development. At this step, the preliminary analysis included the analysis of students, curriculum, and assessment instruments that were developed. In prototyping, the researcher made a multiple-choice test based on the material and the objectives, then, it was validated by the experts. The researcher tried out the questions in a small group (selected 20 students randomly). Further, the result of the test was used to revise the questions before conducting a tryout in the field. After revision. The next step was a field test. In this step, the items of questions were tried out on the subject of the research. in the field test, it was the items of

questions that met the criteria and quality of the test. Finally, the researcher analyzed the test by using ITEMAN and Anates version 4.

FINDINGS AND DISCUSSION

The result of the research was finished based on steps of R&D that had been done. Three steps had been done to produce a good product with formative research by Tessmer, in Rahayu, Purwoko & Zulkardi (2008). The data about developing test had been analyzed.

Self-evaluation was the first step of research development. In this step, the preliminary analysis included the analysis of students, curriculum, and assessment instruments that were developed. There were 60 students. In the analysis of the students, the researcher analyzed thinking skills, the background of the student knowledge, the language used by the students. Based on the result analysis, the researcher found that the students had varying levels of intelligence consisting of the students with higher, medium, and low intelligence.

Table 1. The Category of the Students' Intelligence Can Be Shown

No	Score	Total students	Category intelligent
1	85-100	29	Higher Intelligent
2	80-65	22	Medium Intelligent
3	60-10	9	Low Intelligent

The researcher designed a test to develop cognitive tests. In this case, it dealt with designing a blueprint, dealing with the taxonomy table, and determining the assessment instrument. The results of these steps produced cognitive tests which consisted of 25 number multiple-choice questions. The researcher designed it as the first prototype which would be validated by the experts. Then, suggestions from the expert were used to revise the questions. The results of the assessment and advice of two experts on cognitive tests on recount text material were developed. The level of validity that was calculated based on the formula according to Gregory content validity and Martuza Lawshe (in Ruslan, 2009) obtained a value of 0,875 table*s* 2.

Table 2. Result Analysis Assessment from Two Experts

		Irrelevant score (1-2)	Relevant Score (3-4)
Expert II	Irrelevant Score (1-2)	0	2
	Relevant Score (3-4)	0	14

The assessment was given by the two experts above. The validity of the content can be calculated as follows:

Content validation: $=\frac{14}{(0+2+0+14)}=\frac{14}{16}=$ 0,875

It can be said that the relevance of indicators, types of problems, and the dimensions of knowledge on the classification table about the cognitive assessment that was made is valid. it shows that the test items

worthy tested in a small group (20 people). The test results of the small group of content validity by using correlation of coefficient analysis point biserial on ITEMAN software are presented in Table 3.

Table 3. Recapitulation of Result Analysis Validity of Items of Questions

Category	Items of Questions	
	Number of items	**Total**
Valid	1, 2, 3, 4, 6, 7, 8, 9, 10, 11, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25	23 (92%)
Invalid	5, 12	2 (8%)

The test results of the reliability of the questions were declared valid by using a coefficient of reliability Alpha Cronbach analysis of ITEMAN software. It can be argued that the reliability value obtained from the multiple-choice questions was 0,879. The second prototype cognitive tests that had been tested for validity and reliability in the small group were the third prototype, then, it was tested at field test. At this stage, the cognitive tests in the third prototype were tested in English Education Department (40 students). Then the cognitive tests were tested and analyzed the level of difficulty, differentiator, and distracter for each item of questions. Categorizing the level of difficulty is appropriate with the provisions established that if the index level of difficulty from 0.00 to 0.30 classified difficult questions, from 0.31 to 0.70, 0.71 classified medium classified, and then 1.00 classified easy questions. The results of the calculations dealing with the level of difficult questions can be seen in Table 4.

Table 4. Recap of Result Analysis of The Level Difficulty Question

Category	**Items of Questions**	
	Number of questions	**Total**
Easy	2, 9, 12, 15, 18, 19	6 (26%)
Medium	1, 3, 4, 5, 13, 14, 16, 17, 20, 21, 22, 23	12 (52%)
Difficult	6, 7, 8, 10, 11	5 (22%)

The results analysis of differentiator questions coefficients 0.40 to 1.0 were good; 0.30 to 0.39 were acceptable; 0.20 to 0.29 need revisions and -1.00 to 0.19 poor. The results of the differentiator calculation can be seen in Table 5.

Table 5. Recap Result Analysis of Differentiator

Category	**Items of Question**	
	Number of questions	**Total**
Good	All questions	23 (100%)
Need Revisions	-	0
Bad	-	0

The results analysis of the effectiveness of distracters has been done by counting the number of students who chose the answers for each question. The calculation results of the distracter can be seen in Table 6.

Table 6. Recap Analysis Results of Distracters

Category	Items of Questions	
	Number of Questions	Total
Effective	1, 2, 4, 5, 6, 7, 8, 9, 10, 13, 14, 15, 16, 17, 18, 19, 20, 22, 23, 24	20 (87%)
Not Effective	3, 12, 21	3 (13%)

Research development is intended to produce Cognitive Tests on recount text material. It is valid and reliable with test questions that have high levels of difficulty, differentiator, and effectiveness of detractors. This study used a model of the type of formative research study (Tessmer, in Rahayu, T., Purwoko & Zulkardi. 2008). The results analysis from the validator on cognitive tests in the first prototype was made to declare valid criteria. It was found that content validity was 0.94. Previously, several revisions were made following the suggestions given by the experts. The validity of the content included aspects of the material, construction, and languages based on the analysis of the blueprint preparation. According to Nasution (2007) in preparing a test to measure the students' learning outcomes, content validity is the most important, because it can measure the entire material that has been taught. The validity of the content based on the expert of material judgment provided a high validity value of cognitive tests. This is consistent with Matondang (2009) who found that the validity of the content shows how far the question, task, or item in a test or instrument can represent the overall and proportional of the test samples. This means that the test is valid if the test questions reflect the entire content or material that should be tested or controlled proportionally. The results of field trials in small group students to determine the validity of the content from each item had been done by analyzing the results of the provision of cognitive tests using analysis correlation coefficient point biserial. The result analysis of validity showed that there still an item of question is not valid. According to Matondang (2009), a test is valid for a particular purpose or particular decision-making. The test may not be valid for other purposes of making a decision. In the cognitive tests, the decision can be taken on an item of question that is invalid, then the researcher did not use the question again in the next trial because a valid question already met the indicators of achievement of competencies in learning.

Based on the results, validity is one of the requirements to obtain a good question. It was similar to the result of research conducted by Anwar (2006). Anwar stated that those who meet the quality requirements are valid questions. The validity of an item is the suitability or accuracy of a test to measure something to be measured. The result of the reliability testing of cognitive tests empirically by using Alpha Cronbach coefficients was 0,879. The value of the coefficient of reliability of the

test is interpreted by using a standard. According to Sukiman (2012) giving an interpretation of the instrument reliability coefficient (r) is generally used standard when r ≥ 0.70. It means that the instruments are reliable. Based on the results, it can be said the cognitive tests that are developed can be used because it has high reliability. The statement is also confirmed by Mahaputri, et al (2013) which states that the test developed is of good quality and meets the standards of reliability.

Based on the analysis of the level of difficulty questions, cognitive tests that had been developed, it had a difficulty level in the category of easy, medium, and difficult. It is strengthened with the opinion Nasution (2007), which is considered a useful item that has a level of difficulty in the medium category. These results were consistent with the views expressed by Arikunto (2010) that a good question is a question that is not too easy or too difficult. The questions that are too easy do not stimulate the students to increase their solving efforts. Otherwise, the question that is too hard will cause students to become hopeless and do not have the spirit to try again because out of their reach. Similar to the above opinion, according to Sukiman (2012) that for the kind of formative tests, the proportion of the difficulty level of easy categories is 25%, 50% for the medium category, and 25% for the difficult category.

According to Sukiman (2012), no further than the analysis of the results of the level of difficulty these items are as follows: Record the good items in the question bank book, for difficult questions, there are two possibilities, namely: discard or re-examine what makes the question difficult, maybe the sentences are not good or the instructions are unclear, and so on, then used again after being corrected; or use (such as for a selection test).

The analysis is a differentiating assessment of items that are intended to determine the ability of students to distinguish students who are in the capable category from those who are unable (Uno, 2012). According to Nasution (2007), some things need to be considered and one of them is paying attention to differentiator items. Items are considered good if the key or the answer assumed to be true has high power positive difference and the distracters have a differentiator negatively which is very different from the other options. According to Mansyur (2009), the higher differentiator is better if it can distinguish a group of participants who have high ability from a group of students who have low ability.

A distracter can be said well functions if those distracters have great appeal for the test participants who do not understand the material. The effectiveness of distracter analysis or analysis of the pattern of responses is done by calculating the test participants who chose each alternative answer on each item (Uno, 2012). Distracter functions well if it is chosen by more than 5% of test-takers (p> 5%) if there are four choices and 3% for the five answer choices

(Depdiknas, 2004). This is consistent with the research conducted by Widyantoro (2009) that the question has a bad distracter because there are some distracters on any question that has not been chosen by 5% of the test participants. According to Purwanto (2013), distracters are said to function most effectively if no student answers incorrectly. A good item's question's quality can provide appropriate information about where the students do not understand the material that has been taught. One of the characteristics of a good question is that the questions can distinguish each student's ability. The higher the students' ability to understands the material, the higher the chance to answer the questions correctly. The lower the students' ability to understand the material, the smaller the chance to answer the questions correctly (Safari, 2003).

The analysis of the test is one of the activities that need to be done to improve the quality of a test, both the overall quality of testing and the quality of each item that is part of the test. Surapranata (in Mansyur, 2009) states that the function of the analysis is to improve the quality of the question, namely, whether a question (1) can be accepted because it has been supported by adequate statistical data, (2) has some weaknesses, or (3) is not used at all because it proved empirically not functioning at all.

The item analysis is a systematic procedure. It primarily can be done for an objective test. The analysis of items, among others, aims to hold the identification of the questions whether it is a good or bad question. Then, through the analysis of a question, information can be obtained about the poor quality of the items of question or a "guidance" to make improvements (Arikunto, 1999).

CONCLUSION AND SUGGESTIONS

Based on the analysis and discussion of the research that has been done, and is associated with the formulation of the problem, it can be concluded that some key points relating to the development of cognitive tests on the recount text material as follows:

1. The development of cognitive tests on the recount text material was developed based on the results of the validation sheet by the expert with the validity, the value of 0.87, and a coefficient of results from correlation analysis by point biserial. The content validity of each item acquired cognitive tests were declared valid as much as 23 or 92% items of 25 items of questions. Reliability testing results of cognitive tests empirically by using Alpha Cronbach coefficient values obtained for the multiple-choice test was 0.879 that met with reliable criteria.
2. Development of cognitive tests on the recount text material was developed based on the item analysis. The cognitive test on the recount text material characteristics met the test items covering the difficulty level of multiple-choice tests that had a problem with the level of difficulty in the

category easy 6 items or 26%, in the medium category was 52% or 12 items and 22% or 5 items for difficult category. As for differentiators, the test consisted of 23 numbers of differentiator features in the range of 0.40 to 0.1 that was well categorized. The effectiveness of distracter 87% or 20 items that have well-functioning.

Based on the results obtained in this study, several suggestions are made as follows:

1. A question that has been developed can be used as a reference for English teachers, especially the teachers who want to test the cognitive abilities of the students.
2. It is expected to become a question bank for the university or the school
3. Further researchers can conduct further research dealing with this topic, meanwhile, they have to make it more specific.

REFERENCES

Ahmad. (2010). *Pengembangan Instrument Evaluasi Hasil Belajar Fisika Berbasis Literasi Di sekolah Rintisan Bertaraf Internasional SMAN 15 Makassar*. Tesis Tidak Diterbitkan. Pasca Sarjana UNM.

Anwar, Syafri. (2006). *Peningkatan Kinerja Guru dalam Membuat Soal Objektif melalui Umpan Balik*. Jurnal Pendidikan don Kebudayaan. No. 059, Tahun Ke-12, Maret 2006. Vol.12. Hal. 238

Arifin, Z. (2011). *Evaluasi Pembelajaran Prinsip, Teknik, Prosedur*. Bandung: PT. Remaja Rosdakarya.

Arikunto, S. (2002). *Dasar-dasar Evaluasi Pendidikan*. Jakarta: Bumi Aksara

Arikunto, S. (2010). *Dasar-Dasar Evaluasi Pendidikan (Edisi Revisi)*. Jakarta: PT. Bumi Aksara.

Mahaputri, Ni Luh Putu., Nyoman Dantes, I Wayan Sadia. *Pengembangan tes prestasi belajar berbasis taksonomi Anderson dan krathwohl pada kompetensi dasar fisika Smk kelas x semester ganjil se-kota singaraja*. e-Journal Program Pascasarjana Universitas Pendidikan Ganesha Program Studi Penelitian dan Evaluasi Pendidikan (Volum 3 Tahun 2013)

Mansyur, Rasyid, H. & Suratno. (2009). *Asesmen Pembelajaran di Sekolah*. Yogyakarta: Multi Pressindo.

Matondang, Zulkifly. (2009). *Validitas dan reliabilitas suatu instrumen penelitian*. Jurnal Tabularasa PPS UNIMED Vol.6 No.1, Juni 2009. Hal 87:97

Nasution, Noehi. (2007). *Materi pokok evaluasi pengajaran*. Jakarta: Universitas terbuka

Pardiyono. (2007). *12 writing clues for better writing competence*. Yogyakarta: Andi Yogyakarta

Purwanto. (2013). *Evaluasi Hasil Belajar*. Yogyakarta: Pustaka Pelajar.

Rahayu, T., Purwoko & Zulkardi. (2008). Pengembangan Instrumen Penilaian dalam Pendidikan Matematika Realistik Indonesia (PMRI) di SMPN 17 Palembang. *Jurnal Pendidikan Matematika, (Online),* Vol. 2. No. 2.

Ruslan. (2009). *Penilaian Kinerja Dosen Berdasarkan Kepuasan Mahasiswa dan Pengaruhnya terhadap Perilaku Pasca Kuliah (Studi di FMIPA Universitas Negeri Makassar)*. Jakarta: Pustaka Yaspindo.

Sukiman. (2012). *Pengembangan Sistem Evaluasi*. Yogyakarta: Insan Madani.

Sumarna (2009). *Analisis, Validitas, Reliabilitas dan Interpretasi Hasil Tes*. Bandung: PT Remaja Rosdakarya Offset

Suparji, (2010). *Kualitas Butir Soal Buatan Guru-Guru SMA Mata Pelajaran Matematika dan IPA Di Kabupaten SUMENEP.* Jurnal Pendidikan Volume 11 no. 1. UNESA. Hal 57

Uno, Hamzah B., Koni, satria. (2012). *Assessment Pembelajaran*. Jakarta: Bumi Aksara.