# Item Analysis of Preparation Test for English National Examination

**Madiana Laela, Dewi Rochsantiningsih, Martono**

**English Education Department**
**Teacher Training and Education Faculty**
**Sebelas Maret University of Surakarta**

Email: laylmadina@gmail.com

**Abstract:** This research aims to reveal the quality of English national examination preparation test in terms of qualitative and quantitative aspects. Qualitative aspect includes content validity, technical item quality and cognitive domain learning outcome while quantitative aspect include reliability, difficulty level, item discrimination, and distractor effectiveness. Sample was taken from 3 out of 10 schools in Pati district using simple random sampling. This research employs both qualitative and quantitative analysis in which expert judgement is used to analyze content validity and technical item quality while ITEMAN is used for quantitative analysis. The result showed that the test has good content validity, 99.06% items appropriate with competence being measured, good technical item quality and most items (81.13%)are categorized as cognitive domain learning outcome C2 (Understand). Moreover, the test has high reliability index (> 0.8), fair difficulty, and good discrimination. However, 35.85% items have ineffective distractors.

***Keywords:*** *item analysis, English national examination preparation test*

## INTRODUCTION

National examination as graduation determiner creates such a scary atmosphere for students since students have to pass the exam in order to be able to graduate from school. They are very concerned that they cannot perform well in the test that they will not be able to graduate from school. As a result, some students were depressed. One case reported that a student of SMK Rajapolah, Tasikmalaya, West Java was depressed. Jaenal Mutaqin, the school principal stated, "The cause (mental illness) is the fright, fright of facing national examination" (*"Paranoid UN bikin siswa gangguan jiwa*", 2012). The reason of the problem is that bunch of material tested in national examination.

English, one of subjects tested in national examination, is considered difficult by some students. They feeling difficult may be due to the role of English as an EFL in Indonesia. Students do not get used to reading English text and listen to English dialogs contribute to their feeling difficult, whereas those are tested in English national examination. The other problem that students feel difficult is related to unfamiliar vocabulary.

Administering preparation test is believed to one of many ways to help students prepare national examination. Students can get used to type and form of the question that are going to be asked and

have a chance to find out their readiness facing national examination. From the preparation test result, teacher will be able to find students‟ weaknesses. Preparation test, therefore, should best represent national examination. The content has to be appropriate with national examination‟s test blueprint. When the test is appropriate with the test blueprint, it can measure student‟s ability in each subject otherwise it can give disadvantage to student.

For the purpose of helping senior high school students in Pati well- prepared in facing national examination, education official and Team Teacher (*MGMP*) in Pati administers preparation test in which Team Teacher is mandated to construct the test. With the purpose of preparation test, the test should be appropriate with the material being tested and has good quality. A procedure that can be conducted is by doing item analysis. Shakil (2008) conveys, “Item analysis is a process which examines student responses to individual test items (questions) in order to assess the quality of those items and of the test as a whole” (p. 4). By doing item analysis, the quality of the test items can be discovered. In doing item analysis, there are two analyses that can be utilized: quantitative and qualitative analyses. It is suggested to use both of them in order to assess comprehensively and not to only use one or the other (Kubiszyn and Borich, 2013, 233).

Qualitative analysis is related to analysis towards the appropriateness of the test items with the competence being measured (content validity) and technical item quality. Moreover, cognitive domain learning outcome is also analyzed in this research. Brown (2004) emphasizes that content validity happens if the test represents the ability that is intended to measure and the students should perform the ability being measured (p. 22). Therefore, the more the items in the test are appropriate with competence being measured, the better it fulfills the content validity requirement(s). Technical item quality deals with items writing. Well-written item will make students feel easier to interpret the instruction of the test. Cognitive domain learning outcome analysis is related to what kind of cognitive domain employs for each item in the test.

Quantitative analysis deals with numeral analysis of the test including reliability, difficulty level, item discrimination, and distractor effectiveness. Quantitative analysis is conducted by using ITEMAN™. Reliability deals with the ability of the test to be reliable. Gronlund (1988) proposes the coefficient of reliability for standardized tests of aptitude and achievement over occasions within the same year, the coefficient of reliability should be 0.8 until 0.9 (p. 96). The test having reliability index from 0.8 to 0.9 will be considered as the test having high reliability.

Difficulty level refers to proportion of the test-takers who answer the item correctly (Anastasi and Urbina, 1997, p. 173).The difficulty index ranges from 0.0 until 1.0. The higher the index, the easier the item in the test is. Thorndike and Hagen propose the value to interpret difficulty index. Difficulty index (P) less than 0.30 is considered as too difficult. Items having P 0.30-0.70 is categorized as fair and P with more than 0.70 values belongs to too easy items (Anas 2011, p. 37).

Item discrimination refers to the ability of the item to discriminate the

ability of the test-takers. Brown (2004) emphasizes, "Item discrimination (ID) is the extent to which an item differentiates between high and low-ability test-takers" (p. 59). The value of discrimination index ranges from -1.00 to +1.00. Item should have DI more than 0.19 in order to be accepted, (Ebel and Frisbie, 1991, p.233).

Another analysis conducted in quantitative analysis is distractor effectiveness analysis. According to Cohen et.al (2007), "Distractors are the stuff of multiple choice items, where incorrect alternatives are offered, and students have to select the correct alternatives" (p. 418). Anas (2011) points out that at least 5% of the test-takers must choose the distracters, so the distractor can be considered as well-functioned (p. 411). In addition, Ronalds, Livingston, Willson (2010) argues that the items can be said as "good" if all distractors work and the distractor are chosen by low ability test-takers (p. 158). Therefore, distractors can be said as well-functioned if at least 5% of the test-takers choose it. If the items have all distractors that work in the test, the items can be said as good.

This research focuses to answer the following problem statements: 1) how is English national examination preparation test for senior high schools in Pati in the academic year 2013/2014 viewed from qualitative aspect including content validity, technical item quality, and cognitive domain learning outcome? and 2) how is English national examination preparation test for senior high schools in Pati in the academic year 2013/2014 viewed from quantitative aspect including reliability, difficulty level, item discrimination and distractor effectiveness by using ITEMAN?

## RESEARCH METHODS

The preparation test for English national examination was held in March 22, 2014. This research employs descriptive method in which the approach is content analysis. Descriptive method was chosen since it attempts to truthfully and accurately describe phenomena (Anderson, 1998, p. 260). The research procedure is the following: 1) Collecting a set of test including the question and answer sheet from three schools, test blueprint and answer key from team teacher (MGMP), 2) Qualitative analysis from expert judgement including the appropriateness of the questions with test blueprint and cognitive domain employed in the question, 3) Quantitative analysis by using ITEMAN, and 4) Interpreting the analysis result, and reporting the result.

The population of this study is senior high schools in Pati district in which 3 out of 10 schools were chosen as the sample. To obtain the data, the technique used is documentary. The documents were a set English national examination preparation test for senior high schools in Pati in the academic year 2013/2014, the answer key, and the test blueprint which were collected from Teacher Teams as the test-makers. Students‟ answer sheets were collected from school.

There are two kinds of techniques used to analyze the data in this research: qualitative and quantitative item analysis. In qualitative analysis, content validity of the test, technical item quality, and cognitive domain learning outcome are analyzed. Content validity and technical item quality are analyzed by expert that masters content area(s) covered by the test (Reynolds, Livingston and Willson, 2010, p. 129-130). Content validity is examined by seeing whether the items are

appropriate with competence measured (standard competence and basic competence). Technical item quality is examined by seeing the test construction based on card from National Education Department. Cognitive domain learning outcome analysis intends to find out the information about what kind of learning outcome each test items intended to measure based on Bloom Taxonomy‟s revision.

Quantitative analysis deals with numeral analysis of the test including reliability, difficulty level, item discrimination, and distractor effectiveness. Quantitative analysis is conducted by using ITEMAN™. The software was developed by Assessment Corporation which includes the analysis of item with classical test theory. ITEMAN™ analyzes and provides information including difficulty level, item discrimination, distractor effectiveness for each item, and also reliability of test (Assessment System Corporation, 2006, p. 1-1).The result of reliability index can be known from the score of alpha, and difficulty index (P) can be known by seeing the column of proportion correct. Discrimination index is calculated using point-biserial correlation. It can be known in the column of point biser. Distractor effectiveness can be found out in the column of proportion endorsing.

## RESEARCH FINDINGS AND DISCUSSION

The qualitative analysis of the test includes the analysis of content validity, technical item quality, and cognitive domain learning outcome. Content validity result analysis showed that most items in the test (99.06%) match with the competence being measured. There isone item in the test that is not appropriate with the competence being measured (item number 9 in listening skill categry item).The item that is not appropriate should be rejected since it fails to measure the intended competence. With high percentage of the items being appropriate with competence measured, it can be concluded that the test has good content validity.

The result of technical item quality analysis showed that from 106 items, 8 items have trouble in material aspect, 2 items in construct aspect, and 2 in language/grammatical aspect. 8 items that have trouble in material aspect include 1 item in items‟ appropriateness with competence being measured, 2 items in appropriateness with competency (high urgency, relevancy, continuity, and usefulness), and 5 items in alternatives homogeneity and logicality. Item that does not match with the competency aspect is item number 8 listening skill category. Items that have problem in homogeneity is items number: 18 and 36 (reading skill category type A), 36 and 45 (reading skill category type B), and 45 (reading skill category type C).

For construct aspect, 1 item has problem in stem formulation (item number 17 reading skill category type B) and the other one in the arrangement of number/time form of alternatives (item number 12 listening skill category). In language/grammatical aspect, 2 items have problem in grammaticality (item number 9 listening skill category and item number 45 reading skill category type D).Technical item quality result analysis implies that the test mostly well-constructed and the items that have problems can be revised.

The last aspect investigated for qualitative aspect is cognitive domain learning outcome. The result shows that there are 4.71% items categorized as C1 (remember), 81.13% items as C2 (understand), and 14.15% items as C3 (apply) while there are no items in C4 (analyze), C5 (evaluate), and C6 (create). C1 was found in listening items since the items are asked the students to give correct response in a dialog. In other words, students are asked to recall the expressions that are appropriate in a given situation.

The items categorized C2 on the other hand ask the test-takers not to just remember from the knowledge they have learnt but more to determine the meaning of instructional messages. C2 in this test is mostly found in reading skill category since the items are measured the ability of the students to interpret the reading passages. Meanwhile, the items categorized as C3 are writing skill items. C3 measures higher test-takers ability. It asks the test-takers to use a procedure in a given situation. Items categorized as C3 are the items in writing skill. Items with C3 especially can be found in items with cloze text.

Quantitative aspects include reliability, difficulty level, item discrimination, and detractor effectiveness. There are 5 test forms in English national examination preparation test, and reliability is measured for each test form. 5 test forms are constructed by the test-makers so that students sitting next to each other will not get the same test. The test forms are called *Paket 1* (P1), *Paket 2* (P2), *Paket 3* (P3), *Paket 4* (P4), and *Paket 5* (P5).The resultshows that all of test forms have reliability index ≥ 0.8. The highest is owned by test form P4 with reliability index 0.920 meanwhile the lowest is test form P2 with 8,88. Therefore, it can be noticed that the reliability of English national examination preparation test in Pati in the academic year 2013/2014 is high. With its high number of reliability index, it can be noticed that the test is dependable.

Based on the result of difficulty level analysis, from 106 items, 68 items (64.15%) are considered as item with fair difficulty, 21(19.81%) items are categorized as easy items, and 17 items (16.04%) belong to difficult items. It can be concluded that in terms of difficulty level, the items are good since they are neither too easy nor difficult (Suharsimi, 2005).

Difficult items were mostly found in reading and writing skill items. The competences measured for reading skill items are related to finding the meaning of a word in the text, implicit information, main idea in the text, communicative purpose of the text, general idea, specific information, and the meaning of a phrase. For writing skill category, with total 5 items are the items that most students answer wrongly.The competence measured for writing skill category items is the students can complete the cloze text with suitable word or phrase".This result analysis result enables teachers to find out what kind of materials that students feel difficult and give more emphasize about those in class.

From the analysis result, it can be noticed that items that are categorized as C1, C2 in the test tend to be easy and fair items while items belonging to C3 tend to be difficult items. In relation to the result of analysis and the indicator being measured in the test, there are items in reading and writing skill categories that are categorized as difficult items. Therefore,

with the combination between difficulty level analysis and the test blueprint, the difficult items can be taken into consideration for the teacher to emphasize more on those indicators in which students feel difficult in the instructional process.

Based on the result of item discrimination analysis from 106 items, 81(76,42%) items as very good items, 13 items (12,26%) as good items, 4 items (3,77%) as marginal items, and 8 items (7,55%) as poor items. Items categorized as very good and good items can be accepted and used without any revision for the latter test. Meanwhile, marginal items can be improved by a revision in order that it can be used for the latter test. Poor items in the test must be revised or can also be rejected.

The poor items in the test can be caused by unwell-written items. Besides, high difficulty index also influences the item discrimination since the students will feel difficult to do the test. Hence, the chance for the students to do guessing is high. As a result, the items will fail to discriminate between the high and ability test- takers. To sum up, with few numbers of poor items, it can be concluded that the items can discriminate between high and low- ability test-takers.

Based on the result of distractor effectiveness analysis, from 106 items in the test, there are only 28 items from all skills category that have all distractors that work. The number shows that the items do poorly related to distractor effectiveness. The quality of the distractors in the test can also be seen in the proportion of the test-takers who chose the alternatives. It can be noticed from the result of the analysis that most of the items in the test do poorly in relation to distractors effectiveness.

The items that are categorized as poor item in the test have the same objective with the items that are categorized as difficult ones. The test-takers who felt difficult could not do the items correctly. The test-takers then might do guessing. When many test-takers with high and low- ability did guessing, then the item cannot properly distinguish between high- and low- ability ones.

The poor items in the test can be caused by the item not written well. When the items in the test do not have clear instruction, the students feel difficult in interpreting what the item intends to measure. The clear stem can make students feel easy to read the question. When the stem is no twell-written, student or test-takers will face difficulty to do the items. Besides, high difficulty index also influences the item discrimination since the students will feel difficult to do the test. Hence, the chance for the students to do guessing is high. As a result, the items will fail to discriminate between the high and ability test- takers.

## CONCLUSION, IMPLICATION AND SUGGESTION

From the research findings and the discussion above, the conclusion can be drawn as follows: 1) the result of qualitative analysis shows that test has good validity since most of the items in the test are appropriate with the indicator being measured. 99.06% item is appropriate with the competence being measured Technical item quality analysis shows that from 106 items,8 items have problem in material aspect, 2 items in construct aspect, and 2 items in language, grammatical aspect. Cognitive domain learning outcome analysis shows that 4.71% items categorized as C1

(remember), 81.13% items as C2 (understand), and 14.15% items as C3 (apply) C1, 86 items as C2, and 15 items as C3, 2)Based on the result of data analysis using ITEMAN, the test reliability of the five test forms are categorized as high since all test form have reliability index more than 0.8. In relation to difficulty level analysis from all skill categories, it can be said that the test has fair difficulty. It was found that most of the items (63.21%) categorized as fair items. It was found that most of the items are considered as very good items (76.42%). Distractor effectiveness analysis result showed that there are only 35.85% items in the test that have all distracters that work.

Based on the result of qualitative and quantitative analysis, the quality of the test can be found out. The test items that are accepted can be used to develop the latter test. In addition, the revised and rejected items can be the basis to do revision and improvement of the test itself. The result can be utilized to make item bank in preparation test for English national examination.

In this research, the suggestions will be directed to test-makers, English teacher, and education official. For the test- makers, the suggestions are proposed in relation to test construction including test items‟ appropriateness with competence being measured, stem formulation, and alternatives of the test. Well-constructed will ease the students to interpret the instruction in the test. As a result, it can give more precise result related to the ability of the test as an instrument to measure students‟ ability.

For teacher, the analysis result of difficulty level analysis should be taken into consideration to help the students more prepared. In this case, the finding of difficult items can be used by the teacher to emphasize the material more in the classroom. The suggestion for education official includes the suggestion to administer the test more than one since the official can make evaluation gradually within the same year about the students‟ preparation in facing national examination. Further, it also will advantage and encourage the test-makers to do item analysis about the test quality.

However, there are still some weaknesses, lacks found in this research. The limitation of the research is related to content validity and technical item quality by expert judgement. It can be said that the result drawn has not been well-presented. It is due to time constraint so that there is no much time to for the writer to discuss with the expert. Therefore, this research is still far from perfect. However, the result of analysis reports adequate information of preparation test for English national examination for senior high schools in Pati in the academic year 2013/2014.

**REFERENCES**


Anastasi, Anne & Urbina, Susanna. (1997). *Psychological Testing (seventh edition)*. Upper Saddle River: Prentice-Hall, Inc

Arikunto, Suharsimi. (2005). *Dasar-dasar Evaluasi Pendidikan.* Jakarta: PT Bumi Aksara

ASC. (2006). *User"s Manual for the ITEMAN™ Conventional Item Analysis Program*. Minnesota: Suite

Brown, H. Douglas. (2004). *Language Assessment: Principle and*

*Classroom Practices*. New York: Pearson Education, Inc.

Cohen, Louis, Manion Lawrence, Morrison, Keith, & Wyse, Dominic.(2010). *A Guide to Teaching Practice*. Madison Avenue: Routledge

Didiet. (2012 Apr. 18). Kaget soal UN melenceng dari Tryout. *INDOPOS*. Retrieved from http://www.indopos.co.id/2012/04/kaget-soal-un-melenceng-dari-tryout.html

Ebel, Robert E. & Frisbie, David A. (1991). *Essentials of Educational Measurement (fifth edition)*. New Delhi: Prentice-Hall of India

Gronlund, Norman E. (1981). *Measurement and Evaluation in Teaching (fourth edition)*. New York: Collier-McMillan

Kubiszyn, Tom & Borich, Gary D. (2013). *Educational Testing and Measurement: Classroom Application and Practice (tenth edition)*. Hoboken: John Wiley &Sons, Inc.

Paranoid UN Bikin siswa gangguan jiwa. (2012, Apr. 16). *ANTARA News*. Retrieved from http://antaranews.com/berita/306406/paranoid-un-bikin-siswa-gangguan-jiwa

Reynolds, Cecil R., Livingston, Ronald B., & Willson, Victor. (2010). *Measurement and Assesment in Education (second edition)*. Upper Saddle River: Pearson Education

Shakil, Mohammad. (2008). Assessing Student Performance Using Test Item Analysis and its Relevance to theState Exit Final Exams of MAT0024 Classes - An Action Research Project

Sudijono, Anas. (2011). *Pengantar Evaluasi Pendidikan*. Jakarta: Rajawali Pers