# Education Quarterly Reviews

**Wiyasa, Putu Irmayanti, Laksana, I Ketut Darma, and Indrawati, Ni Luh Ketut Mas. (2019)**, Evaluating Quality of Teacher-Developed English Test in Vocational High School: Content Validity and Item Analysis. In: *Education Quarterly Reviews*, Vol.2, No.2, 344-356.

The online version of this article can be found at:
**https://www.asianinstituteofresearch.org/**

The Asian Institute of Research *Education Quarterly Reviews* is a peer-reviewed International Journal. The journal covers scholarly articles in the fields of education, linguistics, literature, educational theory, research, and methodologies, curriculum, elementary and secondary education, higher education, foreign language education, teaching and learning, teacher education, education of special groups, and other fields of study related to education. As the journal is Open Access, it ensures high visibility and the increase of citations for all research articles published. The *Education Quarterly Reviews* aims to facilitate scholarly work on recent theoretical and practical aspects of Education.

**ASIAN INSTITUTE OF RESEARCH**
Connecting Scholars Worldwide

# Evaluating Quality of Teacher-Developed English Test in Vocational High School: Content Validity and Item Analysis

Putu Irmayanti Wiyasa[1], I Ketut Darma Laksana[2], Ni Luh Ketut Mas Indrawati[3]

[1] Magister of Applied Linguistics Program, Universitas Udayana, Denpasar, Indonesia
[2] Universitas Udayana, Denpasar, Indonesia
[3] Universitas Udayana, Denpasar, Indonesia

Corresponding Author: Putu Irmayanti Wiyasa, Magister of Applied Linguistics Program, Universitas Udayana, Denpasar, Indonesia Tel: +6287 862 388 556. E-mail: irmayantiwiyasa@yahoo.com

**Abstract**

Teacher-developed test, in the form of multiple-choice questions has widely used for measuring the students' final learning process. Content validity and items analysis is used to evaluate the quality of the test that is developed by the teacher. The study focused on content validity that analyzed the content of the test and basic competency which have been asserted on the syllabus. The content validity analysis using Gregory Formula calculation. In addition, the items test includes validity, reliability, index of difficulty (I), discrimination index (DI) and distractor efficiency (DE) is measured by using respective formula that processing in Microsoft Excel program. This mixed-method study was conducted among 211 students of grade X, XI, and XII of Senior Vocational High School, Sekolah Menengah Kejuruan Negeri 2 Denpasar. Total test items that analyzed in this study were 180 items. The result of this study showed that according to the basic competency and also the topic that occurred on the test, there are 6 out of 180 test items which are not in line with the basic competency on the syllabus of K13. In addition, the items test analysis resulted that in terms of validity, reliability, index of difficulty (I), discrimination index (DI) and distractor efficiency (DE) the tests are categorized as moderate.

**Keywords:** Content Validity, Test Items Analysis, Validity, Reliability, Index of Difficulty (I), Discrimination Index (DI) and Distractor Efficiency (DE)

## 1. Introduction

### 1.1 Introduce the Problem

In 2016, the President of the Republic of Indonesia issued Presidential Instruction (Inpres) Number 9 of 2016 concerning Revitalization of Vocational High Schools (SMK). The Presidential Instruction aims at reversing the pyramid of qualifications of Indonesian workers who are educated in Elementary School (SD) and Middle School (Vocational High School) to become educated and skilled workforce by taking education in Vocational Schools. The Presidential Instruction also answers the challenges of the Asean Economic Community (MEA) program in which vocational graduates have the potential to face global competition so as to produce superior and competitive graduates, it is expected that the quality of education at the SMK level should be increased by reforming the SMK development roadmap; perfecting and harmonizing the Vocational curriculum with competencies in accordance with graduate users (link and match). In addition, the Ministry of Education and

Culture has the duty to be able to increase the number and competence of vocational educators and education staff; enhance cooperation with ministries or institutions, regional government, business world and industry; and increasing access to vocational school certification and vocational accreditation; and form a vocational development working group.

From the revitalization of vocational schools that has been proclaimed, the quality of education can then be measured from educational evaluations in accordance with the Education Assessment Standards stipulated in the Minister of Education and Culture of the Republic of Indonesia Number 23 of 2016. Evaluation plays an important role in determining the success of education. Good evaluation and assessment give a big impact towards the learning process (Popham, 2009) and influence the education policy (Mardapi, 2008). The results of the evaluation are closely related to the assessment method and instruments used to take measurements. The accuracy of the selection of assessment methods and instruments are greatly influenced the objectivity and validity of the quality of education.

By looking into the importance of the evaluation role on the quality of education, the evaluation must be carried out periodically to measure students' abilities on a regular basis. However, the implementation of evaluation remains a big problem in the field. Problems related to the evaluation of learning outcomes, for instance the validity of evaluation instruments (tests), assessment methods that are used, and students' answers that are not well analyzed after being tested, have been found in several previous related studies related and also during the pre-observation in several vocational high schools in Denpasar.

According to the phenomena, this study aims at improving the quality of evaluation instrument used in an English language class in Vocational High School. Appropriate evaluation instruments can measure the quality of education well so that the analysis of the quality of evaluation instruments at SMK Negeri 2 Denpasar in the academic year 2018/2019 is undeniably important to be conducted. The quality analysis of the instrument includes the relevance of the test with basic competency on syllabus, validity, reliability, difficulty level, discrimination index, and efficiency of the distractors. By paying attention to the quality of test items henceforth the function of the test namely for education (identify the problem in order to create better learning in the future), policy maker (the institution could evaluate the policy that has been implemented either it works well or not), diagnostic (diagnose the students' difficulty in learning and help them to overcome the problem), administrative (enhance the teacher's ability in preparing, conducting, and evaluating the learning process) (Rohmawati, 2015).

*1.2 Importance of the Problem*
The focus of this study is evaluating the tests that have been developed by the teacher from the content validity and items analysis perspective. There are three rationales of this study important to be conducted. First, for the teacher, this study would bring a reflection of the tests that have been created and administered to the students. Teacher somehow preparing and conducting the lesson well, however, they take less attention towards the evaluation process (according to the pre-observation result). Nevertheless, the teacher needs to be realized that evaluation brings essentials impact towards the learning continuity. Second, for the students, by developing good evaluation instrument, it thoroughly helps the students mapping their learning achievement. They feel appreciated when they did the test that is in accordance with what they have learnt. Additionally, they also can evaluate their learning if there is any particular material or topic that they have not comprehend well yet. Third, from the researcher perspective, this study supports the government plan for revitalizing and giving more attention towards education in vocational high school. Furthermore, this study would bring complete insight of good evaluation instrument, according to Harmer (2007) theory. Several previous studies conducted similar studies in terms of index of difficulty, index of discrimination, and efficiency of distractors. However, this study evaluates the test holistically from the content validity and items analysis that includes the analysis of the test validity, reliability, index of difficulty, index of discrimination, and efficiency of the distractors.

*1.3 Relevant Scholarship*
Several former studies related to evaluation, particularly test items analysis had been conducted and used as a reference in this study are as follows.

The first study conducted by Quaigrain and Arhin in 2017 with a study entitled "Using reliability and item analysis to evaluate a teacher-developed test in educational measurement and evaluation." That study aims at improving the quality of test and avoiding misleading items occurred in a test. The study conducted in Education at Cape Coast Polytechnic with taken answers from 247 first-year students.  The study focused on item and test quality by analyzing the difficulty index (p-value) and discrimination index (DI) with distractor efficiency (DE). The result of their study showed that the internal consistency reliability of the test was 0.77 using Kuder–Richardson 20 coefficient (KR-20). The mean score was 29.23, with a standard deviation of 6.36. Mean difficulty index (p) value and DI were 58.46% (SD 21.23%) and 0.22 (SD 0.17), respectively. DI was noted to be a maximum at a p-value range between 40 and 60%. Mean DE was 55.04% (SD 24.09%). Items having average difficulty and high discriminating power with functional distractors should be integrated into future tests to improve the quality of the assessment. Using DI, it was observed that 30 (60%) of the test items fell into a good category.

Comparing the previous and current study, this study is not merely conducted the test analysis of difficulty index (p-value) and discrimination index (DI) with distractor efficiency (DE) but also includes the content validity that analysed the accordance of basic competence and the content in the test, items validity as well as the reliability. In addition, in this study, the sample was taken from test items that administered in vocational high school level. The items test were gathered from English summative test for grade X, XI, and XII with two different types of test for each grade (Type A and B). Therefore, this current study brings more various result than the previous one.

The second related study entitled Item Analysis of a Multiple-Choice Exam by Toksöz and Ertunç in 2017. Similar with Quaigrain and Arhin (2017), this study focused on the analysis of test items from the difficulty index, discrimination index (DI) and distractor efficiency (DE). The study was conducted at the University Level with the answer gathered from 453 participants who are in the language preparation class. Data was collected from the responses given by the participants to 50 multiple choice items. The multiple choice part included three main sections: vocabulary, grammar, and reading. The result of this study revealed that most of the items are at the moderate level in terms of item difficulty. Besides, the results show that 28% of the items have a low item discrimination value. Finally, the frequency results were analyzed in terms of distractor efficiency, and it is found that there some ineffective distractors that should be revised subsequently.

The current study has some similarities as well as differences towards the study of Toksöz and Ertunç (2017). First of all, the current study not only analyzed the difficulty index, discrimination index (DI) and distractor efficiency (DE) but also the content validity, validity, and reliability of the test. Second, the previous study conducted at the level of University students as same as Quaigrain and Arhin (2017), but this study were taken the responses from the vocational high school students. Nevertheless, the result of this study also found that the test items have low discrimination value, moderate item difficulty, and there some distractors given as an option on the test that did not work effectively.

Next, Setiyana (2016) conducted a study related to the quality of the summative English test at the Meulaboh I. MAN boarding school. This study focused on the analysis of validity, reliability, level of difficulty of the problem, discrimination level, and distractor efficiency. The data collection technique used is a checklist and document analysis. The results of the study on the English language test at the Meulaboh MAN boarding school for question consistency (reliability) and the difficulty level of the questions obtained results above 70%. Additionally, for discrimination level, and distractor efficiency, it is also revealed a good result.

Current research with research conducted by Setiyana (2016) has similarities about analyzing validity, reliability, level of difficulty of the problem, discrimination level, and distractor efficiency. In this study, the object of the study was a summative test for classes X, XI, and XII at Vocational High School 2 Denpasar, and this study analyzed the relevance of the test with the basic competency in the syllabus and curriculum.

*1.4 Research Design*

The research problem of this study is evaluating the quality of the test developed by the teachers. The research design in this study is mixed-method research or a combination of quantitative and qualitative research. Combination research aims at gaining a full understanding to solve problems or phenomena that had been occurred. In addition, this study used to verify and complete the findings and discussions of the study (Dornyei, 2007). In this study, quantitative methods were used to calculate the data presented in the form of numbers. From the results of data quantification, it is supported by qualitative data. Qualitative descriptions are used to describe the suitability of the basic competencies in the syllabus with the questions on the test (content validity), explanation of the results of the items by conducting expert judgment, and the description of the results of integrated interviews. The grand theory of this study is designing an evaluation instrument from Harmer (2007). After forming the research study and determining the research design, then it is started by gathering the data from document analysis and interview. Then, analyzing the data by using the statistic program in Microsoft Excel and also describing the data qualitatively. Next, presented the data in the form of a percentage, numbers, as well as data description. The last is drawing a conclusion as well as giving a recommendation for future studies.

## 2. Method

The research problem of this study is evaluating the quality of test developed by the teachers. The test quality that is intended in this study includes the content validity, items test analysis (validity, reliability, index of difficulty, index of discrimination, and efficiency of distractors (Harmer, 2007). The tests that developed by teachers in this study refer to the English summative test created by teachers at vocational high school *SMK Negeri 2 Denpasar* for grade X, XI, and XII. The research design in this study is mixed-method research or a combination of quantitative and qualitative research. Combination research aims at gaining a full understanding to solve problems or phenomena that had been occurred. In this study, quantitative methods were used to calculate the data presented in the form of numbers. From the results of data quantification, it is supported by qualitative data. Qualitative descriptions are used to describe the suitability of the basic competencies in the syllabus with the questions on the test (content validity), explanation of the results of the items by conducting expert judgment, and the description of the results of integrated interviews. The grand theory of this study is designing an evaluation instrument from Harmer (2007). After forming the research study and determining the research design, then it is started by gathering the data from document analysis and interview. Then, analyzing the data by using the statistic program in Microsoft Excel and also describing the data qualitatively. Next, presented the data in the form of a percentage, numbers, as well as data description. The last is drawing a conclusion as well as giving a recommendation for future studies.

*2.1 Object of the Study*

The object of this research is an English summative test for students of class X, XI, and XII at SMK Negeri 2 Denpasar. In total, there are 180 items test that being analyzed. This is because vocational education is still an interesting focus of study especially the evaluation and implementation component so that it can be expected that vocational students are ready to compete with a better quality evaluation and learning system. In addition, in order to meet the demands of national education goals, namely the implementation of national examinations and the SMK revitalization program, summative test quality analysis research can be used as a reference to training students to be ready to face national examinations with satisfactory results.

*2.2 Data of the Study*

The types of data in this study are quantitative and qualitative data. The type of quantitative is an English summative test result that has been answered by students. Whereas for qualitative data is a description or explanation of the results of document studies and the results of interviews. The source of the data from this study is the summative test results of students of class X, XI, and XII in English subjects in the 2018/2019 school year, syllabus, and the results of interviews.

*2.3 Instrument of Evaluation*

In this study, there were two research instruments used, namely checklists and interview guides. Checklist instrument is a series of statements filled out by the evaluated respondent or researcher for collecting data by

putting a matching sign (√) in the place provided (Fadarwati, 2015). In this study, the checklist filled out by researchers and experts based on document studies. Next instrument is a list of integrated interview questions. List of integrated interview questions is a form of question prepared by the interviewer to obtain information from the interviewee at the time of conducting the interview. In this study, the researcher interviewed the teacher as the creator of the test to gather supporting information after quantitative analysis.

*2.4 Procedures for Conducting the Study*

The procedure of conducting research of evaluating the quality of teacher-developed English test in vocational high school from the content validity and item analysis is divided into three parts, namely, planning, implementation, and the final stage of research. Each of the more detailed stages is presented as follows. First, research planning includes the stages of conducting preliminary observations and interviews with English teachers at SMK Negeri 2 Denpasar (informally), requesting permission for research to go to school, and identifying problems and formulating problems. Second, for research implementation includes the stage of preparing research instruments, gathering syllabus and collecting research data, namely summative test questions in the form of multiple-choice English for grade X, XI, and XII in Vocational High School 2 Denpasar academic year 2018/2019. Third, the final phase of research includes analyzing the results of item analysis to determine the validity, reliability, index of difficulty, index of discriminators, and efficiency of distractors of English summative tests at Vocational High School 2 Denpasar. Next, analyzing the relevance of multiple-choice test with the basic competency, drawing conclusions to answer research problems and the last is reporting the results of research, drawing conclusions, and recommendations.

2.4.1 Data Collection

In this study, the data collection methods used were document studies and integrated interviews. The study of documents uses a checklist instrument while the interview method uses an interview guide instrument equipped with a questionnaire. In this study, a document study was conducted by collecting English summative test items for the 2018/2019 academic year to obtain the data. Document studies are also supported by checklist instruments. The first checklist is used to test the validity for the relevance of the test with the basic competence on the syllabus used in the school. The results of the first checklist will also be attached by the researcher as supporting data on the review of the content validity component. The expert then put a check mark (√) related to the relevance of the basic learning competence and the questions. The second checklist used was developed according to the criteria of good evaluation instrument by Harmer (2007). In this study, integrated interviews were used to obtain additional information related to the results of the analysis that had been obtained, either in the form of reasons for a problem that could occur, or the advantages found in the tests that had been made so that they could add recommendations to be given in this study. In the interview method, the researcher interviewed the teacher and gathered other information to support the quantitative data result.

2.4.2 Data Analysis

Data analysis is a method that is used to determine the results of the analysis. The type of data obtained in this study is quantitative and qualitative data, namely data obtained from the results of item analysis (quantitative data), and interview result (qualitative data) of summative English test for grade X, XI, and XII at SMK Negeri 2 Denpasar in the academic year 2018/2019.

2.4.3 Research Design

The research design in this study is mixed-method research or a combination of quantitative and qualitative research. In this study, quantitative methods were used to calculate the data presented in the form of numbers. From the results of data quantification, it is supported by qualitative data. Qualitative descriptions are used to describe the suitability of the basic competencies in the syllabus with the questions on the test (content validity), explanation of the results of the items by conducting expert judgment, and the description of the results of integrated interviews. The grand theory of this study is designing an evaluation instrument from Harmer (2007). After forming the research study and determining the research design, then it is started by gathering the data from document analysis and interview. Then, this research is analyzing the data by using statistic program in Microsoft Excel, and after getting the result, the analysis then supported by qualitative data supporting. Next,

presenting the data in the form of a percentage, numbers, as well as a description. The last is drawing a conclusion as well as giving a recommendation for future studies.

## 3. Results and Discussion

The results of this study discussed in the findings and discussion of the problem formulated in this study. The problem discussed with quantitative analysis (Likert scale calculation and statistics) and qualitative regarding to the relevance of the items with basic competency and item analysis including validity, reliability, index of difficulty, index of discrimination, and the efficiency of distractors.

*3.1 The relevance of basic competencies and items of the test (content validity)*
The relevance between items with learning indicators was carried out by analyzing the basic competencies in the syllabus, the items used, and cognitive domains achieved in the respective item. After that, expert judgment sheets then assessed by experts, practitioners, or someone who has the ability in the assessment field. In this study, two expert judges were used namely lecturers from one of the educational universities with educational backgrounds in the field of assessment.

For grade X test there are a number of topics taught had to be taught in the first semester namely Personal Identity, Congratulating, Showing Intention, Descriptive, Announcement, Present Perfect & Simple Past Tense (Sentences for expressing actions or events in the past and / or those that happened in the past and are still valid today). The test items for grade X are divided into type A and B in which according to document analysis, all the items given in the test already included all the material contained in the topic in the syllabus for the first semester at SMK Negeri 2 Denpasar.

For grade XI test, the topics taught are Suggest & Offer, Opinions, Invitation, Personal Letters, Passive Voice, Conditional, and Factual Report. For the test of grade XI type A and B already cover the entire contents of the material contained in the syllabus.

For grade XII, the topics taught were Offering Help, Surprising News, Asking for Attention, Caption (accompanying text images), Application Letter (letters applying for jobs), and Factual Report. Overall, the questions given include the material provided. But there are a number of questions that need to be concerned namely, on questions number 24 and 25 are questions with the topic of Personal Letter, whereas on the basic competencies in the syllabus are not included in the topic of the first semester for XII grade. In addition to numbers 24 and 25, there are also questions number 29 and 30, which contain material about Personal Identity in which the material does not appear in the syllabus for grade XII. In addition to the test of type A problem, the type B also found a number of questions whose topics were not included in the syllabus, namely questions number 21, 22, and 23 where the questions contained material about Procedure Text and for questions number 9, were not available in the questions so in total there are only 29 questions available on the test.

*3.2 Items Analysis*
The results of the analysis of the items included tests of validity, reliability, level of difficulty, index of discrimination, and index of difficulty (Harmer, 2007) and Brown (2004). In this study, dichotomous analysis of test items was used because the types of tests used had true or false values or 0 or 1. Multiple choice tests merely have the right or wrong answers, and there is no range of scores between right and wrong answers. Overall, the items analyzed were 180 items divided into 60 questions at each level with packages A and B for the respective grade. The summary of the analysis is presented as the following table.

Table 1. Analysis of Test Items Result

| Class/ Type | Items Analysis | | | | |
|---|---|---|---|---|---|
| | Validity | Reliability | Index of Difficulty | Index of Discrimination | Distractor Efficiency |
| X / A | 15 items | 0.828 High | • 18 (Easy) • 8 (Moderate) • 4 (Difficult) | • 23 items (*poor*) • 7 items (*satisfactory*) | • 14 items (ineffective) • 13 items (moderate) • 3 items (effective) |
| X / B | 10 items | 0.708 High | • 15 (Easy) • 10 (Moderate) • 5 (Difficult) | • 28 items (*poor*) • 2 items (*satisfactory*) | • 11 items (ineffective) • 7 items (moderate) • 9 items (effective) |
| XI / A | 15 items | 0.741 High | • 18 (Easy) • 9 (Moderate) • 3 (Difficult) | • 25 items (*poor*) • 5 items (*satisfactory*) | • 16 items (ineffective) • 12 items (moderate) • 2 items (effective) |
| XI / B | 15 items | 0.868 High | • 18 (Easy) • 10 (Moderate) • 2 (Difficult) | • 24 items (*poor*) • 6 items (*satisfactory*) | • 4 items (ineffective) • 17 items (moderate) • 2 items (effective) |
| XII / A | 9 items | 0.836 High | • 24 (Easy) • 5 (Moderate) • 1 (Difficult) | • 28 items (*poor*) • 2 items (*satisfactory*) | • 17 items (ineffective) • 11 items (moderate) • 2 items (effective) |
| XII / B | 15 items | 0.792 High | • 15 (Easy) • 12 (Moderate) • 3 (Difficult) | • 24 items (*poor*) • 6 items (*satisfactory*) | • 11 items (ineffective) • 12 items (moderate) • 7 items (effective) |

### 3.2.1 *Items Analysis of Grade X*

*Validity*

The results of the items test validity showed the test accuracy to measure students' abilities. The results of this validity test are obtained from the calculation of student scores compared to the total value of all correct answers. Quantitative calculations on content validity are also supported by data from expert judgment result. For the type A test, there are 15 valid questions and 15 invalid questions. Invalid questions constitute half of the total number of questions. The questions that are classified as invalid are questions number 1, 4, 6, 8, 10, 11, 12, 13 18, 19, 20, 21, 25, 26 and 29 meanwhile valid questions are number 2, 3, 5, 7, 9, 14, 15, 16, 17, 22, 23, 24, 27, and 28. For the type B resulted in 10 questions were valid, and 20 questions were invalid. Items with valid category are questions number 1, 3, 4, 5, 12, 14, 15, 16, 28, & 29 whereas for questions number 2, 6, 7, 8, 9, 10, 11, 13, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, & 30 are belonged to invalid category.

*Reliability*

The reliability test results are carried out after the validity test. The items that categorized as valid items then tested for reliability. Based on the results of the validity test of grade X type A, there are 15 valid items, and in type B, there are 10 valid items. The results of the reliability test that belonged to dichotomy test are calculated using the KR - 20 formula. The KR - 20 formula is used because the level of difficulty of the items is heterogeneous that each question has a different level of difficulty. The high-reliability test results show that the level of trustworthiness or good consistency of the question even though it is used in different groups of students. The result of the reliability test of grade X type B showed that the reliability level is 0828, which means that a very high level of reliability. On the other hands, for the type B, the results are 0.708 belonged to a high level of reliability.

*Index of Difficulty*

Index of difficulty is the result of calculation from the numbers of students who answer the questions correctly divided by the total number of students whom taking the test. The test results of the index of difficulty type A there are 18 questions are classified as easy, 8 questions are moderate, and 4 questions belong to a difficult category. The results indicated that the questions used are dominated by easy category whereas most students are able to answer all the questions correctly. However, the category of moderate and difficult questions does not have a balanced portion with the category of easy questions. On the other side, the test for grade X type B consisted of 5 questions with difficult category, 15 with easy category, and 10 questions with a moderate category.

*Index of Discrimination*

Index of discrimination is an analysis that analyzed how the questions enable to distinguish groups of students who are capable and less capable at the end of the assessment. For the test of grade X type A, 23 items are categorized as poor category which means that the questions are not able to distinguish the ability of students because most students answer are similar whether it is all true or all wrong meanwhile the other 7 questions belong to the satisfactory category which shows that the questions have a moderate level to distinguish the students' ability. For type B, the test is not good enough to classify students' abilities between the upper, middle, and lower groups. A total of 28 questions are included in the poor or low category where the questions can be answered correctly both with the upper-class students or students with the lower groups meanwhile the other 2 questions belong to the satisfactory category where the questions are able to be answered correctly by the students who are classified in the upper ability group of students.

*Efficiency of Distractors*

The result of the distractor effectiveness test is used to check the quality of the answer choices given besides the answer key. For the grade X type A, questions no 1, 2, 5, 8, 9, 12, 13, 19, 20, 21, 24, 25, 26, & 29, the distractors do not work effectively because it is chosen by less than 5% of the total number of students. For questions number 3, 4, 6, 7, 10, 11, 14, 15, 16, 17, 18, 22, & 30, the choices are categorized as moderate because one of the options is chosen by more than 5% of a total number of students. Next, for questions no. 23, 27, and 28, the distractor is effective is because two options are chosen by more than 5% of the students. For type B, there are 11 questions where the three choices given are not functioned properly as it was not chosen by more than 5% of students in total. In addition, there are 7 questions where 1 option is chosen by more than 5% of the students. Lastly, the distractors of 9 questions are categorized as affective because two options are chosen by more than 5% of students.

3.2.2 *Items Analysis of Grade XI*

*Validity*

For the test of grade XI type A, there are 15 questions which are included in the valid category, and other 15 items are the invalid. The items that valid are the questions number 1, 6, 7, 11, 12, 13, 18, 19, 20, 21, 23, 28, and 30 and the invalid are no. 2, 3, 4, 5, 8, 14, 15, 16, 17, 22, 24, 25, 26, 27, and 29.  For the type B questions number 3, 6, 8, 9, 10, 11, 12, 13, 15, 17, 18, 22, 23, 28, 29 & 30 belong to the category of valid questions. On the other hands, questions number 1, 2, 4, 5, 7, 14, 16, 19, 20, 21, 24, 25, 26, 27 are classified as invalid.

*Reliability*

For class XI packages A and B, there is the same number of valid question packages, which are 15 questions. For the UTS class XI package A problem has a very high level of reliability as well as the results of 0.836.
For the UTS class XI package B, the results show that the question is included in the very high-reliability category with the results of the quantitative calculation of 0.868.

*Index of Difficulty*

For the test of grade XI type A, the results of the analysis show that there are 18 easy questions, 9 moderate questions, and 3 difficult questions. The results of the index difficulty of grade XI test have similar results as class X that dominated by easy questions. For the type B test, 18 easy questions, 10 moderate questions, and 2 difficult questions.

*Index of Discrimination*

For type, A test, 25 questions are classified as poor, and 5 questions are in the satisfactory category. The poor or low category questions indicate that the related questions can be answered both by students with upper group abilities or students with lower group abilities. For type B, there are 24 questions that belong to a poor category, and 6 questions are classified as satisfactory. This is showed that the test is dominated by questions which are not good enough to distinguish groups of student abilities.

*Efficiency of Distractor*

The result of efficiency of distractor of test type A, questions no. 1, 2, 3, 4, 5, 7, 8, 9, 10, 11, 17, 18, 20, 26, 28, & 30, the choices given besides the answer key is not well-functioned as it is chosen less than 5% of the total number of students. Furthermore, for questions no. 6, 9, 12, 13, 14, 16, 19, 21, 22, 23, 24, & 29, there are two choices of answers besides the answer key hence the distractor level is moderate. For questions number 15 and 27, the three choices given as a distractor had functioned properly, because there two choices are chosen by more than 5% of students.

3.2.3 *Items Analysis of Grade XII*
*Validity*
For the test for grade XII type A there are 21 invalid questions and 9 others are valid. Questions that are classified as invalid are questions number 1, 2, 3, 4, 5, 6, 7, 8, 9, 12, 13, 14, 15, 16, 23, 24, 25, 26, 28, 29, and 30 For questions that are classified as valid are questions number 10, 11, 17, 18, 19, 20, 21, 22, & 27. For the type B, 15 questions are classified as invalid and 15 other questions are valid. Questions that are classified as invalid are questions no. 2, 3, 6, 8, 9, 14, 17, 19, 21, 22, 23, 24, 25, 28, & 30. As for the questions with valid categories are questions number 1, 4, 5, 7, 10, 11, 12, 13, 15, 16, 18, 20, 26, 27, 29, & 30.

*Reliability*
For the reliability result of the test grade XII type A, there are 9 questions classified as valid, and for type B, there are 15 questions valid. Therefore, the reliability of the type A test is 0.741 categorized as high. Moreover, for type B, the results of the reliability test indicate that the reliability level is 0.792 classified as a high level of reliability.

*Index of Difficulty*
For the grade XII type A test, questions with easy categories are 24, moderate categories are 5, and difficult category only 1 question. Henceforth, questions on the test are dominated by easy category. This result showed that the quality of the questions is unable to distinguish or classify the abilities of each student in the class. Nevertheless, for type B, 15 questions are easy, 12 questions are moderate, and 3 questions are difficult.

*Index of Discrimination*
For the test of grade XII type A, the results of the analysis showed that 28 questions belong to the poor category with 2 items having a minus (-) result which means that there are questions answered correctly by groups of lower ability students but answered incorrectly by upper ability student group. In addition, there are 2 questions categorized as satisfactory. For grade XII type B, 24 questions are classified as poor or not good enough to differentiate students in the class because this right answer is given equally by groups of students with upper and lower abilities. In addition, 6 questions are classified as satisfactory because it is relevant between the number of correct answers and the number of students with high ability to answer correctly.

*Efficiency of Distractor*
For grade XII type A, the distractors for questions number 1, 2, 3, 4, 5, 7, 8, 10, 11, 12, 17, 18, 20, 22, 26, 28 & 30 are not work effectively as the three distractors given are chosen less than 5% of students. Next, questions number 6, 9, 13, 14, 16, 19, 21, 22, 23, 24, 25, & 29 have distractors that the level of the efficiency is moderate. Lastly, 2 questions, number 15 & 27 have good function of the distractor because it is chosen by 5% more of a total number of students.

**Conclusion**

The result of this evaluative study divided into content validity and items analysis of English summative test for grade X, XI, and XII in vocational high school that developed by the teachers. The research design is a mixed method as the data gathered in this study required calculation and statistics, and for enriching and supporting the quantitative data, the researcher also conducted the qualitative study. The result showed that from the content analysis, the test of grade X and XI already in relevance with the syllabus that is developed by the Indonesian government. However, for the test of grade XII, there are 6 questions that are not in line with the material

required on the basic competence. For the item analysis result, according to validity, reliability, index of difficulty, index of discrimination, and efficiency of the distractor, the test for grade X, XI, dan XII are categorized as moderate.

## Acknowledgments

## References

Brown, H. D., (2004). *Language assessment: Principles and classroom practices*. White Plains, NY: Pearson Education

Dornyei, Z. (2007). *Research methods in applied linguistics: Quantitative, qualitative, and mixed methodologies*. New York: Oxford University Press.

Fadarwati, I. P. (2015). Penggunaan Teknik Evaluasi Non-tes dan Hambatannya pada Penilaian Pembelajaran PKn SD di DABIN IV Kecamatan Tonjong Kabupaten Brebes [Skripis]. Semarang (ID): Universitas Negeri Semarang.

Selamat, Suzanah, and Kaur Gurnam Sidhu. 2011. "Student Perceptions of Metacognitive Strategy Use in Lecture Listening Comprehension." *Language Education in Asia* 185-198. Accessed November 23, 2015. http://dx.doi.org/10.5746/LEiA/11/V2/12/A02/Selamat_Sidhu.

Suchitra, Nair, Yew Li Koo, and Abu Bakar Kesumawati. 2014. "Exploring the Listening Processes of Pre-university ESL Students." *Procedia Social and Behavioral Sciences* 118: 475-472. www.sciencedirect.com.

Robinson, Martin, Evelina D Galaczi, Coreen Docherty, Anthony King, and Hanan Khalifa. 2014. "Supporting national education reform: The Cambridge Malaysia Baseline Project." *Cambridge English: Research Notes*, November: 40-44. Accessed November 8, 2016. www.cambridgeenglish.org/research-notes.

2010-2015. "Official Portal Malaysian Examination Council." *STPM and MUET Examination Reports*. Accessed May 1, 2016. portal.mpm.edu.my.

Mehrabian, A. 1981. *Silent Messages: Implicit Communication of Emotions and Attitudes*. Wadsworth: Belmont.

McDevitt, T.M., E.P. Sheenan, and N McMenamin. 1991. "Self-reports of academic listening activities by traditional and non-traditional college students." *College Student Journal* 25: 478-489.

Brown, J.D. 2005. *Testing in language programs: A comprehensive guide to English language assessment*. New York: Mc Graw-Hill.

Vivien, Berry, and Barry O'Sullivan. 2014. "Assessing General Language Proficiency." *Introducing Language Assessment*. London: British Council Digital Services. Accessed November 15, 2016. https://www.britishcouncil.org/exam/aptis/research/project/assessment-literacy/general-language-proficiency.

Taylor, Lynda, and Ardeshir Geranpayeh. 2011. "Assessing listening for academic purposes: Defining and operationalising the test construct." *Journal of English for Academic Purposes* (Elsevier Ltd) (10): 89-101. doi:10.1016/j.jeap.2011.03.002.

Buck, Gary. 2001. *Assessing Listening*. Cambridge Language Assessment Series. Edited by J.Charles Anderson and Lyle F Bachman. Cambridge: Cambridge University Press.

English Language Standards and Quality Council. 2015. *English Language Education Reform in Malaysia: The Roadmap 2015-2025*. Ministry of Education Malaysia.

Weir. 2005. *Language Testing and Validation An Evidence-Based Approach*. Hampshire: Palgrave Macmillan.

North, Brian; Ortega, Angeles; Sheehan, Susan;. 2010. *A Core Inventory for General English*. British Councils / EQUALS. Accessed February https://englishagenda.britishcouncil.org, 2017.

Mehrens, William A, and Irvin J Lehmann. 1991. *Measurement and Evaluation in Education and Psychology*. 4th. Belmont, CA: Wadsworth Thomson Learning.

Field, John. 2008. *Listening in the Language Classroom*. Cambridge: Cambridge University Press.

Vandergriff, Larry. 2007. "Listening: theory and practice in modern foreign language competence." *Subject Centre for Languages, Linguistics and Area Studies Guide to Good Practice*. February 22. Accessed January 3, 2016. http://www.llas.ac.uk/resources/gpg/67#toc_1.

Mendelsohn, David. 2001. "Listening Comprehension: We've come a long way but..." *Research Symposium*, Fall: 33-37. Accessed February 16, 2016. http://www.teslontario.net/publications/researchsympossium/ResearchSympossium2001.pdf.

Rost, Michael. 2002. *Teaching & Researching Listening.* London: Longman.

D.Wolvin, Andrew, ed. n.d. *Listening and Human Communication in the 21st Century.* 2010: Wiley-Blackwell.

2018. "DepEd vision, mission, core values and goals." *Department of Education.* October 30. www.deped.gov.ph.

2018. "K to 12 Curriculum Guide in Science." *DepEd Tambayan.* October 30. https://depedtambayan.blogspot.com.

2018. "Creating Learning Objectives." *Convergence Training.* October 30. https://www.convergencetraining.com.

Lucas, Maria Rita D., and Brenda B. Corpuz. 2014. *Facilitating Learning : Metacognitive Process.* Quezon City, Metro Manila: Lorimar Publishing, Inc.

Mizbani, Maryam, and Azizeh Chalak. 2017. "Analyzing Listening and Speaking Activities of Iranian EFL Textbook Prospect 3 Through Bloom's Revised Taxonomy." *Advances in Language and Literacy Studies* 1-6.

Zareian, Gholamreza, Mohammad Davoudi, Zahra Heshmatifar, and Javad Rahimi. 2015. "An Evaluation of Questions in Two ESP Coursebooks Based on Bloom's New Taxonomy of Cognitive Learning Domain." *International Journal of Education and Research* 1-14.

Mizbani, Maryam, and Azizeh Chalak. 2017. "Analyzing Reading and Writing Activities of Iranian EFL Textbook Prospect 3 Based on Bloom's Revised Taxonomy." *Journal of Applied Linguistics and Language Research* 1-15.

Assaly, Ibtihal R., and Oqlah M Smadi. 2015. "Using Bloom's Taxonomy to Evaluate the Cognitive Levels of Master Class Textbook's Questions." *English Language Teaching* 1-11.

Olimat, Mohammad. 2015. "Analyzing Action Pack Textbooks' Questions according to Revised Taxonomy." *Journal of Education and Practice* 1-8.

Al-hasanat, Hasan AbdRabbeh Ali. 2016. "Analyzing Assessment Questions in an Arabic Textbook (Communication skills) for Eight Grade in Jordan According to Bloom's Taxonomy of Levels of Knowledge Aims." *World Journal of Education* 1-12.

Rahpeyma, Ayoub. 2015. "The Analysis of Learning Objectives in Iranian Junior High School English Textbooks based on Bloom's Revised Taxonomy." *International Journal of Education and Literacy Studies* 1-12.

Abdi, Heidar, and Mohammad Mohammadi. 2014. "Textbook Evaluation: A Case Study." *Procedia Social and Behavioral Sciences* 1-8.

2018. "Effective Use of Performance OBjectives for Learning and Assessment." *Teacher and Educational Development.* September 26.

Bigcas, Jimayma. 2016. "A Mini Critique of K-12 Basi Education Curriculum." *ResearchGate* 1-3.

Wilson, Leslie Owen. 2018. "Anderson and Krathwol Bloom's Taxonomy Revised." *The Second Principle.* October 30. https://thesecondprinciple.com.

2018. "Revised Bloom's Taxonomy ." October 30. http://www.kurwongbss.qld.edu.au/thinking/Bloom/blooms.htm.

2018. "Revised Bloom's Taxonomy Action Verbs." October 30. https://www.apu.edu.

Cherry, Kendra. 2018. "The 4 stages of Cognitive Development by Jean Piaget." *very well mind.* November 12. https://www.verywellmind.com.

n.d. *National Book Development Board.* Accessed November 21, 2018. books.philippines.gov.ph.

Abistado, Jonna M., Meliza P. Valdoz, Marites D. Aquino, and Mary Anne T. Bascara. 2014. *Science Links 9.* Quezon City: REX Printing Company Inc.

Alvarez, Liza A., Dave G. Angeles, Hernan L. Apurada, Ma. Pillar P. Carmona, Oliver A. Lahorra, Judith f. Marcaida, Ma. Regaele A. Olarte, et al. 2014. *Science Learner's Module 9.* Pasig city: FEP Printing Corporation.

Deauna, Melecio C. 1996. *Elementary Statistics for Basic Education.* Quezon City: Phoenix Publishing House, Inc.

Harmer, Jeremy. 2007. *The practice of English language teaching.* England: Pearson Education Limited.

Selamat, Suzanah, dan Kaur Gurnam Sidhu. 2011. "Student Perceptions of Metacognitive Strategy Use in Lecture Listening Comprehension." *Language Education in Asia* 185-198. Diakses November 23, 2015. http://dx.doi.org/10.5746/LEiA/11/V2/12/A02/Selamat_Sidhu.

Suchitra, Nair, Yew Li Koo, dan Abu Bakar Kesumawati. 2014. "Exploring the Listening Processes of Pre-university ESL Students." *Procedia Social and Behavioral Sciences* 118: 475-472. www.sciencedirect.com.

Robinson, Martin, Evelina D Galaczi, Coreen Docherty, Anthony King, dan Hanan Khalifa. 2014. "Supporting national education reform: The Cambridge Malaysia Baseline Project." *Cambridge English: Research Notes*, November: 40-44. Diakses November 8, 2016. www.cambridgeenglish.org/research-notes.

2010-2015. "Official Portal Malaysian Examination Council." *STPM and MUET Examination Reports.* Diakses May 1, 2016. portal.mpm.edu.my.

Mehrabian, A. 1981. *Silent Messages: Implicit Communication of Emotions and Attitudes.* Wadsworth: Belmont.

McDevitt, T.M., E.P. Sheenan, dan N McMenamin. 1991. "Self-reports of academic listening activities by traditional and non-traditional college students." *College Student Journal* 25: 478-489.

Brown, J.D. 2005. *Testing in language programns: A comprehensive guide to English language assessment.* New York: Mc Graw-Hill.

Vivien, Berry, dan Barry O'Sullivan. 2014. "Assessing General Language Proficiency." *Introducing Language Assessment.* London: British Council Digital Services. Diakses November 15, 2016. https://www.britishcouncil.org/exam/aptis/research/project/assessment-literacy/general-language-proficiency.

Taylor, Lynda, dan Ardeshir Geranpayeh. 2011. "Assessing listening for academic purposes: Defining and operationalising the test construct." *Journal of English for Academic Purposes* (Elsevier Ltd) (10): 89-101. doi:10.1016/j.jeap.2011.03.002.

Buck, Gary. 2001. *Assessing Listening.* Cambridge Language Assessment Series. Disunting oleh J.Charles Anderson dan Lyle F Bachman. Cambdrige: Cambdrige University Press.

English Language Standards and Quality Council. 2015. *English Language Education Reform in Malaysia: The Roadmap 2015-2025.* Ministry of Education Malaysia.

Weir. 2005. *Language Testing and Validation An Evidence-Based Approach.* Hampshire: Palgrave Macmillan.

North, Brian; Ortega, Angeles; Sheehan, Susan;. 2010. *A Core Inventory for General English.* British Councils / EQUALS. Diakses February https://englishagenda.britishcouncil.org, 2017.

Mehrens, William A, dan Irvin J Lehmann. 1991. *Measurement and Evaluation in Education and Psychology.* 4th. Belmont, CA: Wadsworth Thomson Learning.

Field, John. 2008. *Listening in the Language Classroom.* Cambridge: Cambridge University Press.

Vandergriff, Larry. 2007. "Listening: theory and practice in modern foreign language competence." *Subject Centre for Languages, Linguistics and Area Studies Guide to Good Practice.* 22 February. Diakses January 3, 2016. http://www.llas.ac.uk/resources/gpg/67#toc_1.

Mendelsohn, David. 2001. "Listening Comprehension: We've come a long way but..." *Research Sympossium*, Fall: 33-37. Diakses February 16, 2016. http://www.teslontario.net/publications/researchsympossium/ResearchSympossium2001.pdf.

Rost, Michael. 2002. *Teaching & Researching Listening.* London: Longman.

D.Wolvin, Andrew, penyunt. t.thn. *Listening and Human Communication in the 21st Century.* 2010: Wiley-Blackwell.

2018. "DepEd vision, mission, core values and goals." *Department of Education.* 30 October. www.deped.gov.ph.

2018. "K to 12 Curriculum Guide in Science." *DepEd Tambayan.* 30 October. https://depedtambayan.blogspot.com.

2018. "Creating Learning Objectives." *Convergence Training.* 30 October. https://www.convergencetraining.com.

Lucas, Maria Rita D., dan Brenda B. Corpuz. 2014. *Facilitating Learning : Metacognitive Process.* Quezon City, Metro Manila: Lorimar Publishing, Inc.

Mizbani, Maryam, and Azizeh Chalak. 2017. "Analyzing Listening and Speaking Activities of Iranian EFL Textbook Prospect 3 Through Bloom's Revised Taxonomy." *Advances in Language and Literacy Studies* 1-6.

Zareian, Gholamreza, Mohammad Davoudi, Zahra Heshmatifar, dan Javad Rahimi. 2015. "An Evaluation of Questions in Two ESP Coursebooks Based on Bloom's New Taxonomy of Cognitive Learning Domain." *International Journal of Education and Research* 1-14.

Mizbani, Maryam, dan Azizeh Chalak. 2017. "Analyzing Reading and Writing Activities of Iranian EFL Textbook Prospect 3 Based on Bloom's Revised Taxonomy." *Journal of Applied Linguistic and Language Research* 1-15.

Assaly, Ibtihal R., dan Oqlah M Smadi. 2015. "Usisng Bloom's Taxonomy to Evaluate the Cognitive Levels of Master Class Textbook's Questions." *English Language Teaching* 1-11.

Olimat, Mohammad. 2015. "Analyzing Action Pack Textbooks' Questions according to Revised Taxonomy." *Journal of Education and Practice* 1-8.

Al-hasanat, Hasan AbdRabbeh Ali. 2016. "Analyzing Assessment Questions in an Arabic Textbook (Communication skills) for Eight Grade in Jordan According to Bloom's Taxonomy of Levels of Knowledge Aims." *World Journal of Education* 1-12.

Rahpeyma, Ayoub. 2015. "The Analysis of Learning Objectives in Iranian Junior High School English Text books based on Bloom's Revised Taxonomy." *International Journal of Education and Literacy Studies* 1-12.

Abdi, Heidar, dan Mohammad Mohammadi. 2014. "Textbook Evaluation: A Case Study." *Procedia Social and Behavioral Scciences* 1-8.

2018. "Effective Use of Performance OBjectives for Learning and Assessment." *Teacher and Educational Development.* 26 September.

Bigcas, Jimayma. 2016. "A Mini Critique of K-12 Basi Education Curriculum." *ResearchGate* 1-3.

Wilson, Leslie Owen. 2018. "Anderson and Krathwol Bloom's Taxonomy Revised." *The Second Principle.* 30 October. https://thesecondprinciple.com.

2018. "Revised Bloom's Taxonomy ." 30 October. http://www.kurwongbss.qld.edu.au/thinking/Bloom/blooms.htm.

2018. "Revised Bloom's Taxonomy Action Verbs." 30 October. https://www.apu.edu.

Cherry, Kendra. 2018. "The 4 stages of Cognitive Development by Jean Piaget." *verywell mind.* 12 November. https://www.verywellmind.com.

t.thn. *National Book Development Board.* Diakses November 21, 2018. books.philippines.gov.ph.

Abistado, Jonna M., Meliza P. Valdoz, Marites D. Aquino, dan Mary Anne T. Bascara. 2014. *Science Links 9.* Quezon City: REX Printing Company Inc.

Alvarez, Liza A., Dave G. Angeles, Hernan L. Apurada, Ma. Pillar P. Carmona, Oliver A. Lahorra, Judith f. Marcaida, Ma. Regaele A. Olarte, et al. 2014. *Science Learner's Module 9.* Pasig city: FEP Printing Corporation.

Deauna, Melecio C. 1996. *Elementary Statistics for Basic Education.* Quezon City: Phoenix Publishing House, Inc.

Harmer, Jeremy. 2007. *The practice of English language teaching.* England: Pearson Education Limited.

**Copyrights**