CHAPTER II

THEORETICAL FRAMEWORK

This chapter discusses relevant works of literature on aspects of tests and national standardized school examinations. It addresses the discussions about the definition, the type, and the criteria of a good test. Then, it presents the national standardized school examination issues, including educational assessment in Indonesia and how it is implemented in Indonesia.

A. A glance at the test

1. Definition

Teachers and other educational professionals have tasks to design any tests as part of considered are typical for them despite whether or not the teachers know and have been trained test designing. They will spend lots of their time to make a good test for their respective students. A test is an essential part of the teaching and learning process, one of the assessment methods. As Fulcher (2010:1) argues, tests are vehicles by which society can implement equality of opportunity or learner empowerment. They have a function as a gatekeeper for education. It means that by evaluating students' performance, the success of the education process will be identified, revealing whether or not the process has met a set of established standards or outcomes.

The “test” term is often associated with the terms assessment and evaluation. For some teachers, these three terms, test, assessment, and evaluation,

become difficult questions. Fulcher (2010) asserted that tests are tools used to verify the extent to which an individual has mastered a specific skill(s). On the other hand, assessment is concerned with documenting the target knowledge, skills, and attitudes in that a performance is observed, described, documented, and interpreted for improvement purposes. In contrast, evaluation is a process of making judgments to a performance-based on evidence. Chen and Mathies (2016) also clarified that assessment is learner-centred and process-oriented, which points to distinguish areas where teaching and learning can improve, whereas evaluation is judgmental and arrives at a valuation of performance.

2. Type of the test

a. Formative and Summative Tests

Regarding test designing, teachers have two working mandates at their hands, internal and external mandates. The internal mandate means that the test is regularly built up by teachers or by the school administration. The test is conducted to meet the teachers' needs (or the school). Such tests include placement and evaluation tests; A placement test is used to group learners based on their competence levels. An evaluation test is the one functioning to determine how much students have understood the learning or whether students have difficulties on certain topics during the learning process. In doing so, teachers usually will organize a spoken or written test.

Besides the teachers (and the school), such tests are also important for the pupils, a motor function. For instance, when a teacher announces at the beginning of the lesson that there will be a test after the learning process, students will usually be more motivated to pay more attention to the lesson. Another case is when students know that there will be a test at the end of a month, then the students usually will spare much time just for studying because they want to get a high score on the test. It can be concluded that the test has a positive effect on students since the test can increase students' motivation in learning.

Such a test that plays a role in the teaching-learning process is called a formative test. The formative test takes put whereas students are still preparing to learn something and are utilized to monitor how well that learning is advancing (Carr, 2011: 11). This is similar to Hughes (2011:5) as saying that assessment is formative when teachers use it to test on their students' progress, to work out how far they need mastered what they ought to have learned, then use this information to switch their future teaching plans.

The external mandate is a test initiated by an agent beyond the local context (Fulcher, 2010:2) is fundamentally outlined to measure the capability of learners without reference to the setting in which they are learning. The tests measure the learners' competence after a period of learning, during which learners may be anticipated to have come to a specific standard. The data in the tests do not continuously nourish back into the learning process but functions as an accountability part. These kinds of tests are called summative tests. Carr (2011: 11) also wrote that summative tests are ordinarily given at the end of a unit

course, program, etc., and they give data approximately how much students learned. Hughes (2011:5) says that summative assessment is employed at the end of the term, semester, or year to measure what has been achieved by groups and individuals.

b. Objective and Subjective Tests

Viewed from its scoring, Carr (2011:12) divided the tests into two kinds, objective and subjective tests. An objective test can be scored impartially and thus uses selected-response questions. The examples of the test are multiple-choice questions, true-false questions, matching, and completion. At the same time, a subjective test is the one that includes human judgment to score as in most tests of writing and speaking. It demands students to organize and present an original answer. An example of such a test is an essay test.

1) Multiple-choice item

Among the numerous testing types, the multiple-choice (MC) item remains the most used for many reasons. Multiple-choice is very productive and effective for measuring knowledge and cognitive skills (Haladyna et al., 2019). It is also stated by Siegfried and Wuttke (2019) that multiple-choice (MC) tests are presently being expanded to other subjects and are, in the meantime, one of the foremost broadly utilized test formats in higher education to measure cognitive performance. Wu et al. (2019) also give a reason for using multiple-choice at various educational levels because it has objectivity in the evaluation and is

simple to score. This fact is echoed by Gierl et al. (2017), stating that multiple-choice is the most effective and enduring educational evaluation.

Multiple-choice questions have many advantages since they cover a broad scope of material. They can measure various cognitive levels, can be scored easily, quickly, and have high objectivity, and are appropriate to be used for large-scale examinations whose results must be announced immediately, such as national examination, final school examinations, national standardized school examination, and civil service selection test.

In Indonesia, multiple-choice is the most commonly used format for presenting objective-style questions. According to PUSPENDIK (2017), the multiple-choice question is a question whose answer can be chosen from several possible answers (options) provided. However, this kind of item is not easy to make. Khan et al. (2013) revealed that creating a single best question is challenging and time-consuming, even for those formally trained in multiple-choice question designing. Kelly (1916) in Gierl et al. (2017) underlined three criteria as the first guidelines: (1) all students should interpret the item in the same way; (2) the item should target a single problem so that its answer would be completely right or completely wrong, and not partly right and partly wrong; and the difficulty level of the item should not depend on either obscure words or unintentional cues in the stem. However, the multiple-choice items also have limitations since they are time-consuming in designing the test, challenging to make a homogeneous and functional distractor, and there is an opportunity to guess the answer key. In line with the fact, Wu et al. (2019) are concerned that

multiple-choice tests with the single-best answer informational have some disadvantages: helplessness to speculating and obliviousness to partial knowledge.

Some studies have proved that guessing is one of the limitations of multiple-choice items. Haladyna et al. (2019) stated that guessing came in two forms: random guessing and strategic guessing, or blind guessing and informed guessing (Wu et al.: 2019). Random or blind guessings refer to the same situation in which an examinee selects one of the answers randomly or blindly, even without reading the question. The test takers only depend on luck. He/she has no idea for choosing one option as he/she thinks that all options are equally plausible. It means he has a chance of 25% to tick a correct answer, provided four options available.

Meanwhile, in strategic or informed guessing, the examinee involves partial knowledge for solving his/her multiple-choice questions. A study about guessing done by male and female students conducted by Riener G. and Wagner (2018) revealed that males are more likely to guess and females are less willing to enter a competition and guess in multiple-choice tests.

How are multiple-choice questions drafted? Butler (2017) confirmed that a multiple-choice item comprises a stem (i.e., the context, content, or question the test-taker is required to answer) and a set of potential responses. As there is only a single correct option, the other choices are commonly referred to as lures or distractors. This can be described in more detail. Multiple-choice questions consist of two components, they are:

(1) stem (the subject matter)

(2) options (answer choices, typically labeled A, B, C, etc.)

The options of multiple-choice question contents two parts :

(1) distractors (the incorrect options)

(2) key (the correct option)

The multiple-choice test option for senior high school students' tests in Indonesia is five (A, B, C, D, and E). The use of five options is supported by many scholars (Delgado and Prieto, 1998; Vyas and Supe, 2008, Gierl et al., 2017); five response options, which consist of one correct answer and four distractors, are recommended by most authors of measurement textbooks and are widely used in educational testing. In Indonesia, writing options for senior high school, multiple-choice questions use capital letters (A, B, C, D, E). The purpose of capital letters using in options is for students/test-takers to feel effortless in distinguishing lower-case letters "a" and "d" and "c" and "e." Here is an example of a multiple-choice question:

In one <u>particular</u> village, there was a man who had two beautiful daughters.

What is the synonym of the underlined word above?

A. Specific

B. Element

C. Feature

D. Scoop

E. Point

The sentence:

What is the synonym of the underlined word above?

It is said as a **stem** or subject matter. While

A. Specific

B. Element

C. Feature

D. Scoop

E. Point

They are mentioned as **Options**. From the five options above, one option is a crucial answer. The answer to the question above is A (specific).

In designing good multiple-choice questions, Fulcher (2010: 172) wrote guidelines as follows :

(1) The stem should contain all the information necessary to select the key but should not contain unnecessary material.

(2) The stem should not contain vocabulary unknown to the test-takers unless it is a vocabulary item.

(3) Avoid giving clues to the key in the stem (e.g., using words from the key or writing distractors that are not grammatically consistent with the stem).

(4) Each multiple-choice item should test only one construct (e.g., if it is a vocabulary item, it should not also test grammar, and so on).

(5) The key to any item should not give a clue about the key to another item. This often happens in reading tests where multiple items are based on the same text.

(6) Ensure that the key cannot be selected without reading the stem or other textual material on which the item is based.

(7) Avoid trick items, including ambiguous content, too much information in the stem, and too little difference between some of the options with the possibility of more than one correct answer. While trick items may distract some students, others could get the item correct by using a test-taking strategy.

(8) Avoid negatives such as 'not' and 'except' if possible, as such questions increase cognitive processing and make the item more difficult.

(9) Make sure that only one option can reasonably be keyed.

(10) Randomize the location of the key. If you don't, you will find that (on average) option (C) will tend to be the key more often than other options.

(11) Options should be similar in structure and (most importantly) length. If all else fails in a multiple-choice test, students will select the most extended option.

(12) Avoid options that use 'all of the above or 'none of the above.

(13) Avoid using qualifiers such as 'always' or 'never', which are less likely to be in the key than qualifiers like 'sometimes' or 'probably'.

In designing a good multiple-choice question, Indonesian English teachers as test writers must conform to the principles of designing multiple-choice questions guided by the Education Assessment Center (PUSPENDIK:2019) They are :

a) Material

(1) Questions must conform to the indicator. The indicator is a reference in designing questions. It is a guideline for a test writer. It consists of the basic competence characteristic to be measured, cognitive level/intellectual understanding, material, type of question. Basic competence refers to the minimum ability which has to be mastered by learners, as written in a syllabus. The cognitive level/intellectual understanding denotes cognitive behavior that can explain thinking skills and abilities used in the classroom (and in real life) (Papas, 2015). This cognitive level refers to the revised

Bloom's taxonomy of educational objectives, i.e., knowledge, comprehension, application, analysis, synthesis, evaluation. Puspendik (2017) categorizes the cognitive level into three stages; they are Level 1 (L1), Level 2 (L2), and Level 3 (L3). L1 indicates low ability (knowledge and comprehension), L2 points higher ability (application), L3 shows the level of analysis, synthesis, and evaluation.

(2) The options must be homogeneous and logical in terms of material. It means that all answer choices must be taken from the same material as contained in the subject matter, the writing must be equal, and all answer choices must be functional.

(3) Each question must have only one correct or the correct answer.

b) Construction

(1) The stem must be formulated clearly and explicitly.

(2) The formulation of the stem and its answer options must only make one right and complete statement. There should not be unnecessary information.

(3) The stem does not lead to the key answer.

(4) The stem does not contain a statement that tends to create double negatives.

(5) The length of the option must be relatively the same.

(6) The options do not contain statements :

- All the answers are wrong
- All the answers are true

(7) The options in the form of numbers or times must be arranged chronologically.

(8) Images, graphics, tables, diagrams, and the like must be clear and functioning.

(9) The question is not dependent on the answer to the previous question.

c) Language

(1) The question is written in a standard language following the rules of Bahasa Indonesia, and so with other languages, local languages, and foreign languages.

(2) The question does not use the dialect of the area where the questions are tested.

(3) Each question must use communicative language.

(4) Every option does not repeat unnecessary words or phrases.

Other important things that need to be considered in writing multiple-choice questions are:

(1) Questions should not be offensive to ethnicity, religion, race, and intergroup.

(2) Questions may not be politically charged or contain pornography, promotion of commercial products (advertisements) or agencies (school

names, area names), violence, and other forms that can cause negative effects or things that can benefit or harm certain groups.

2) True-False item

True-false item is an ideal measuring device to assess the students' analytic decision for choosing two options. They may need to choose right from wrong, to stop, or to continue. In English, for example, the selection to use singular or plural construction and so on. True-false items are suitable for concepts with two logical responses and evaluate students' understanding of popular misconceptions. It is suitable for concepts with two logical responses and to evaluate students' understanding of popular misconceptions. The advantages of the true-false item are that students can answer 3-4 questions per minute, and it can test large amounts of content. From the teacher's perspective, writing the true-false item can be done quickly. However, the true-false item also has disadvantages. It needs a large number of items for high reliability. Students have a 50 – 50 chance of testing the correct answer by guessing; it is difficult to discriminate between the students who have mastered from those who fail.

Below is the example of a True-false item

Direction:

For each question below, circle True or False.

- BOOK can be used as a noun or a verb. True False
- The past tense of EAT is ATE True False

3) Matching Test item

This test is very efficient to test knowledge in which events, dates, names, and places are important, and to test, special terms and definitions have to be remembered. A simple matching test item consists of two columns. The first column, traditionally placed on the left side, is a stem or problem to be answered. The second column is placed on the right side; it is responses from which the answers are to be chosen. The students have to read a stem on the left side and find the correct response on the right side.

The matching test item is good for some comprehension levels if it is appropriately constructed. It is valuable in content areas that have many facts. However, it is time-consuming for students and may not be appropriate for higher levels of learning.

The example of the matching test item can be shown below.

Direction:

Find the synonym by matching the words in column A with the words in column B.

A	B
Glad	Terrible
Nice	Peculiar
Wonderful	Pleasant
Awful	Hopeful
Strange	Amusing
Optimistic	Happy
Intelligent	Simple
Easy	Marvellous
Cheap	Clever
Funny	Inexpensive

(source: Watcyn - Jones; 1991 page: 2)

4) Completion items

Completion is also said as Fill-in-the-blank test items. They are useful in assessing the mastery of factual information when a specific word or phrase is important to know. They require students to answer a question by finishing an incomplete statement by filling in a blank with the correct word or phrase.

The completion item is good for recalling and memorizing facts and testing about who, what, where, and when content. It can minimize guessing and encourage students to do more intensive study because they must know the answer versus recognizing the answer. However, the completion item is challenging to assess higher levels of learning because of its limited words of the answers. Sometimes, questions may have more than one correct answer, and scoring is also time-consuming.

Below is an example of the completion or fill-in-the-blank test item.

Direction: by listening to a song, complete the blank spaces of the song lyric below.

Whatever Will Be, Will Be (Que sera, sera)

Doris Day

When I was just (1) … girl

I asked my mother, what will I be

Will I be (2) …, will I be rich

Here's what she said to me.

Que Sera, Sera,

(3) … will be, will be

The future's not ours, (4) …

Que Sera, Sera

What will be, will be.

When I was (5) …, I fell in love

I asked my (6) … what lies ahead

Will we have rainbows, (7) …

Here's what my sweetheart said.

Que Sera, Sera,

Whatever will be, will be

(8) … not ours, to see

Que Sera, Sera

What will be, will be.

Now I have children of (9) …

They ask their mother, what will I be

Will I be (10) …, will I be rich

I tell them tenderly.

Que Sera, Sera,

Whatever will be, will be

The future is not ours, to see

Que Sera, Sera

What will be, will be.…

(Source: Kapanlagi.com)

5) Essay tests

Essay tests present a real-life test to students because he or she is required to organize and communicate opinions and thought rather than respond to multiple-choice and true-false questions. Therefore, they permit students to demonstrate achievement of such higher-level objectives as analyzing and critical thinking. They offer students an opportunity to use their judgment, writing styles, and vocabulary. From the teachers' point of view, the essay tests are easy to construct but time-consuming in scoring.

Here is an example of a reading essay test.

This text is for questions 1 and 2

Jakarta. The roof of Genting State Elementary School in Pasuruan, East Java, collapsed on Tuesday morning, killing a teacher and a second-grade student.

The National Disaster Mitigation Agency (BNPB) said at least 12 others were injured, two of them seriously, in the incident at around

9 a.m. while classes were in session. "The collapse of the school's roof is believed to have been caused by faulty construction when it was built in 2017," BNPB spokesman Agus Wibowo said in a statement." The incident caused five classrooms to collapse, trapping everyone inside," he added.

The teacher has been identified as 19-year-old Sevina Arsy Wijaya and the student as 9-year-old Irza Amira – both female.

An investigation was launched to determine the cause of the incident. Agus said the BNPB's disaster response team had set up tents as temporary classrooms.

(Source: https://jakartaglobe.id/news/elementary-schools-roof-collapses-in-east-java-killing-teacher-student/)

(1) What is the best headline for the news?

(2) After the disaster happened, where did the students study?

3. Criteria of a good test

How to judge that a test is qualified? For answering the question, Carr (2011: 20) used several *qualities of usefulness* proposed by Bachman and Palmer, as shown in table 1 below.

Table 1. Criteria of a Good Test

Quality	Definition
Reliability	Consistency of scoring estimated statistically
Authenticity	The degree to which test tasks resemble Target Language Use tasks
Construct Validity	The degree to which it is appropriate to interpret a test score as an indicator of the construct of interest
Impact	Effects of the test on people and institutions, including wash back – the effect of a test on teaching and learning
Practically	The degree to which there are enough resources to develop and use the test

(Source: Carr;2011 page 20)

Pramono (2014:223-246) proposed quite similar criteria for a qualified test. A test should have the following.

a. Validity means that the test assesses what should be judged using an appropriate assessment tool.

b. Reliability means that the test has consistency; when it is retested, then the score will not be changed dramatically.

c. A test must have a discriminating power for distinguishing students' abilities between high, average, and low levels.

d. Practicality, a good test is practical, no wordy, easy to understand, and easy to administer.

e. Objectivity, a test, does not have a personal element to influence.

f. Economically, a test can be carried out efficiently, quickly, and shortly.

Hughes (2011) emphasizes the points of the criteria, whatever test or testing system, it should consistently provide accurate measures, encompasses a beneficial effect on teaching, and is economical in terms of time and money. Validity and reliability are also discussed as the primary term for determining a good test. As they also stated by Harmer (2007), Bachman and Palmer (1996). A test is said to be valid if it measures accurately what to measure. A test is said to be reliable if it has consistency. It means that when the test is retested in different circumstances, the score will not change dramatically.

Test validity is implemented in studying the impact of the mistakes on the quality of the multiple-choice question items designed by Banyumas English teachers. Since the criteria of a good test is validity, and Harmer (2007) expressed that a test is valid if it tests what it is supposed to test or it refers to how well a test can measure what it is intended to measure. To achieve it, the test designer must consider that the test can be used to make a decision and give information about teaching and learning. The test also has to have a positive effect on students, teachers, and institutions. It means that the test must be helpful for its purpose.

For the reasons, a test must be well designed. The test designers must obey the principles of the test writing. Why is it so? Because according to Haladyna (2004), a test item development is a primary source of evidence in validating a test score interpretation or use.

B. Ujian Sekolah Berstandar Nasional (USBN)/ National Standardized School Examination(NSSE)

Republic of Indonesia's Minister of Education and Culture regulation, Number 23 the Year 2016 concerning Educational Assessment Standards stated that there are three kinds of educational evaluation, namely:

1. educational evaluation by educator
2. educational evaluation by an educational unit
3. educational evaluation by the government

Teachers in their classes do educational evaluation by an educator. Teachers will assess the attitude, knowledge, and skill of their students. It is also stated by Suwartono and Riyani (2019) that the assessment covers cognitive, affective, and psychomotor domains and considers the characteristics and level of the learners. The evaluation aims to monitor and evaluate the learning process, the progress of learning, and the betterment of students learning results. It is done by teachers continuously.

Educational evaluation by an educational unit is an evaluation organized by the educational unit. The meaning of the educational unit, in this case, is a school,

as it is stated in the Minister of Education and Culture regulation, number 4 the year 2018 about Assessment of Learning Outcomes by the Education Unit and Assessment of Learning Outcomes by the Government. The name of the evaluation is *Ujian Sekolah* or School Examination. It is a measurement and evaluation activity of students' competencies which is done by an educational unit. The educational evaluation organized by an educational unit aims to evaluate the attainment of graduate competencies standard or *Standard Kompetensi Lulusan (SKL)*.

Educational evaluation is an evaluation organized by the government. It aims to evaluate graduate competency standards or (*Standard Kompetensi Lulusan/SKL*) nationally. It is not all subjects learned by students are examined. They are only *Bahasa Indonesia*, English, Mathematics, and one major lesson as students optional.

National Standardized School Examination/NSSE or *Ujian Sekolah Berstandar Nasional/USBN* is an activity for measuring students' competencies attainment as recognition of learning achievement and graduation from school (*Permendikbud No. 23 Tahun 2016*). The test script of school examination is 75% designed by the teacher at school, and the remaining 25% is from the Ministry of Education and Culture (*Permendikbud No. 4 Tahun 2018*).

There are many parties involved in sustainability of NSSE/USBN, namely Kemdikbud (*Kementerian Pendidikan dan Kebudayaan*/Minister of Education and Culture), BSNP (*Badan Standar Nasional Pendidikan*/National Education Standards Board), Educational Assessment Center/*Puspendik*, MKKS

(*Musyawarah Kerja Kepala Sekolah*/School Principal Association), MGMP (*Musyawarah Guru Mata Pelajaran*/Teachers Association), and school as the organizer of USBN.

C. The Teachers' steps in designing the NSSE Multiple-choice Test Items

The requirement of being the designer of the NSSE was that the teacher taught the twelve grades of senior high school. There was no particular requirement such as the TOEFL score, the years of teaching experience, the teacher training on the test writing, etc. Therefore, after the selected teachers at the schools had been given a duty from their principal to write questions for school examination, they had to pay attention to the table of specifications given by the school management.

The table of specifications is the most detailed level of test architecture. They are also sometimes called test 'blueprints' (Fulcher, 2010:144). Fives and Barnes (2013) stated that a table of specifications (TOS), sometimes called a test blueprint, can help teachers shape the decision-making process of the construction of the test and enhance the validity of teachers' assessments based on a classroom test. It is written in PUSPENDIK (2017) that a test blueprint is a matrix format containing information that can be used as a guide for writing questions. It can be concluded that forming the test blueprint is an important step that a test designer must do before writing a test.

A good test blueprint has met the following requirements:

1. The test blueprint has represented the content of the curriculum tested.
2. The components of the test blueprint must be detail, clear, and understandable.
3. The indicator of the questions must be well-defined. It means that the indicator can be a guide to write a good question.

Fives and Barnes (2018) wrote that the specification table aims to ensure alignment between the assessment items and the content, skills, or constructs that the assessment intends to assess. It helps test designers focus on response content, ensuring that the test measures what it intends to measure. The school management distributed the table of specifications received by the teachers. The school management received it from the English Community (MGMP), as it was the second hand of the Assessment Centre. The writer of the specification table was the Education Assessment Centre of the Indonesian Education and Culture ministry. Table 2 is the example of the national standardized school examination's table of specifications.

Table 2. An Example of USBN Table of Specification

KISI-KISI SOAL UJIAN SEKOLAH BERSTANDAR NASIONAL

Jenjang Pendidikan : SMA/MA **Alokasi Waktu : 120 Menit**

Mata Pelajaran : BAHASA INGGRIS **Jumlah Soal : 45 soal**

Program : IPA/IPS/BAHASA **Bentuk Soal : 40 PG &**

Kurikulum : 2013 **5 Uraian**

No.	Kompetensi yang diuji	Lingkup Materi	Materi	Level Kognitif	Indikator Soal	Bentuk Soal	Nomor Soal
11	**Reading** Peserta didik dapat menentukan tujuan dari bacaan	Fungsi sosial	*Announ-cement*	L3	Disajikan sebuah teks berbentuk *Annoucement* (tentang kegiatan sekolah: Liburan), peserta didik dapat menentukan tujuan dari bacaan.	Pilihan Ganda	11
12	**Reading** Peserta didik dapat menentukan rincian kegiatan dari bacaan	Struktur Teks	*Announ-cement*	L2	Disajikan sebuah teks berbentuk *Annoucement* (tentang kegiatan sekolah: Liburan), peserta didik dapat menentukan rincian kegiatan dari bacaan.	Pilihan Ganda	12

(Source: Kisi-Kisi USBN SMA)

The table of the specification contains two components. The first component is called as identity component, and the second one is the matrix component. The identity component is placed on top of the matrix component. It consists of Education level (*Jenjang pendidikan*), Time allocation (*alokasi waktu*), subject (*Mata Pelajaran*), number of questions (*Jumlah pertanyaan*), Program, question form (*bentuk soal*), curriculum. While, in the matrix component where it is written in the table, it consists of a number, competency tested (*kompetensi yang diuji*), the scope of material *(lingkup materi*), material (*materi*), cognitive level (*level kognitif*), question indikator (*indikator soal*), question form (*bentuk soal*), and question number (*nomor soal*).

Every selected teacher or teacher who had the duty to write school examination questions must comprehend each sub-components of the identity and matrix component. The identity sub-component is very easy to understand by all teachers since they are only the identity where the teachers work and about the identity of school examination questions teachers will design.

For comprehending each matrix subcomponent, the teachers should figure out that the number in the first column means the number of the information written in the matrix component. The column of competencies tested, is about what language skill and competency are tested. As the example in table 2, it is written that the competency tested in number 11 is:

Reading

Peserta didik dapat menentukan tujuan dari bacaan (students are able to determine the purpose of the text).

It means that the language skill to be tested in number 11 is reading, and the competency is to understand the purpose of the text.

In the column of the scope of material, the example in table 2 is *fungsi social* (social function). For the school examination questions, as it is written in the table of specifications designed by the Assessment Center, there are three scopes of test material, i.e. social function, text structure, and language feature.

The first scope of test material is a social function. Some text aspects that can be written as questions in the social function's scope of material is topic of the text, the aim of writing the text, background/reason of writing the text, audience target, the moral value of the text. The second scope of the test material is text structure. The inner parts of a text that can be designed as questions are about the text's main idea, detailed information, specific information, and implied information. Whereas in the third scope of the test material is about language features. In this scope of the test material, the question items writers can use these language aspects related to the text content to be base of question writing, i.e. word synonym, reference, word meaning, etc.

The next column is about *materi* (material). It contains material taught to students during their six-semester learning in senior high school. In designing multiple-choice question items, this part is as a stimulus of the question. Based on

the table of specifications given by the Assessment center, the material for school examination are announcement, invitation letter, personal letter, application letter, caption, song, descriptive text, recount text, narrative text, exposition text, explanation text, news item text, and procedure text. Each text can be applied for one to five questions. Table 2 gives an example that the material as a question stimulus is an announcement text.

The column next to the material is *level kognitif* (cognitive level). This part, as written by Kemdikbud (2017), describes the student's ability level individually that can be elaborated into three cognitive levels, i.e., first-level (L1), second-level (L2), and third-level (L3). L1 shows that students in this level have the minimum standard in mastering the lesson. At the same time, L2 means that students in this level have an applicative ability. For L3, it portrays students that have an ability of reasoning and logic. The L1, L2, and L3 cognitive levels, if presented in the form of cognitive process dimension and active verb that can be used to formulate indicators based on Bloom's taxonomy, can be described in table 3 as follows.

Table 3. Cognitive Level

L1 (Knowing)		L2 (Applying)	L3 (Reasoning		
C1 Remembering	C2 Understanding	C3 Applying	C4 Analyzing	C5 Evaluating	C6 Creating

(Source: Kemdikbud; 2017, page 7)

Table 2 explains that question number 11 is L3, meaning that the cognitive level of question number 11 must be C3.

The next column is written as *indikator soal* (question indicator). This part guides the question designer more clearly. Since it is guidance, it is an important part that can be a mirror to know the question's validity. The indicator of the question consists of three components. They are subject, Stimulus, and behavior to be measured. The subject of the indicator refers to students who have to answer the question. The stimulus component in English questions is usually a source/reading material, for example, a text, paragraphs, pictures, graphics, photos, tables, etc. they are used as a basis for writing a question. In this component, an English teacher as a test designer has to pay attention much to the length of a text, and he/she must be able not only to distinguish the text type but also to be observant in reading the topic of a text. While in the component of behavior to be measured, competency must be mastered by students. An English teacher as a test writer is required to interpret a student's competency in answering the question as seen in a question indicator as shown in table 2 above, and the snippet of the sentence is as written below:

Disajikan sebuah teks berbentuk Annoucement (tentang kegiatan sekolah: Liburan), peserta didik dapat menentukan tujuan dari bacaan.

It is presented an announcement text about school activity: holiday, student
Stimulus, subject,

can determine the aim of the text.
behavior to be measured

The question indicator tells that for designing a multiple-choice item, and the teacher needs an announcement text about the school activity. The topic of school activity is a holiday. A student as a subject is hoped to be able to determine the aim of the announcement text.

Another example of question indicator taken from the school examination table of specification is written as follows:

It is presented a news item text (150-200 words) about a natural disaster,
Stimulus

Students can determine an answer about the reading topic.
Subject Behavior to be measured

The second indicator example above describes that in the Stimulus component, a question designer must have a news item text about a natural disaster, and the length of the text is about 150 to 200 words. The indicator component of behavior to be measured, the student as the subject of the indicator must answer a question about the reading topic or the topic of a news item text.

The following table of specification column is question form and number of questions. They are not difficult to understand since a question form to be tested a multiple-choice question, and in the column of question number, it has been written based on what number the question will be.

After reading and comprehending the table of specifications, teachers designed a table of specification level mapping, the design aimed to guide teachers as the questions writers in knowing at what level the material was taught

since the material examined in the national standardized school examination was the material that students learned during their study in high school for six semesters. So, it would be very beneficial for teachers as the questions writer to know at what level or grade the material written as a question had been learned by students. Table 4 is the example of a table of specification level mapping written by a teacher.

Table 4. An Example of Table of Specification Level Mapping

FORM PEMETAAN LEVEL

KISI-KISI BAHASA INGGRIS KURIKULUM 2013

No.	Ruang Lingkup	Materi	Kelas/ Semester	Level	Bentuk soal	
					Pilihan ganda	Uraian
11	Fungsi Sosial	*Announcement*	XI/2	L3	√	
12	Struktur Teks	*Announcement*	XI/2	L2	√	
13	Fungsi Sosial	Undangan Resmi	XII/1	L1	√	
14	Struktur Teks	Undangan Resmi	XII/1	L2	√	

(Source: USBN Administration of SMA Negeri 1 Baturraden)

Furthermore, teachers would work with question cards. Table 4 describes a question card form used by teachers to write a question. One question card is for one question. It means that if a teacher writes 30 questions, he/she will write in 30 question cards. The question card part that a teacher can copy and paste without adding new information is in the identity component, which is consist of *Satuan*

Pendidikan (education unit), *Penyusun* (question items writer), *Mata Pelajaran* (Subject), *Alokasi waktu* (time allocation), Program, , *Tahun Pelajaran* (Academic year), *Kurikulum* (Curriculum), while *Kelas/semester* (class/semester) teacher cannot just copy and paste. The explanation of each information is clarified in the next paragraph.

Education unit is the information about where the students as the test-takers study during their 6 semesters. A question writer must write the answer SMA (*Sekolah Menengah Atas*/ Senior High School). While question items writer must be answered by writing what the question writer's name is. In the Subject part, the question writer must fill about the subject tested to students. Time allocation needs an answer about how long the test will be, for example, 120 minutes. For the academic year information, the question writer has to write about in what academic year the test will be applied. The next is information about the curriculum. It means that what curriculum is applied for students as test-takers in teaching-learning before school examination is held. While the information about class/semester, the teacher cannot just copy and paste since it needs information about in what class/semester the material of question items had been learned by students during six semesters learning. Therefore, the teacher needs a table of specification level mapping.

For completing other information in a question card, the teacher needs to move some information from the table of specifications and a table of specification level mapping. In *Kompetensi yang diuji* (Competency tested), the teacher must copy the information from the table of specifications as well as in

Lingkup materi (material scope), *Materi* (material), *Indikator soal* (question indicator), and *Level kognitif* (cognitive level).

Kunci jawaban (key answer) must be completed by the key answer of question item written by the teacher. *Nomor soal* (question number) is in what number the question item will be. While, *Rumusan soal* (question formulation) must be completed not only by text as a question stimulus but also by question designed by the teacher based on the information about competency tested, Material Scope, Material, Question Indicator and Cognitive level that written on the left of question formulation. Furthermore, *Buku Sumber* (Reference) must be filled with the information about the reference of text used as a question stimulus.

Table 5. Question Card Form

<table>
<tr><td colspan="3">KARTU SOAL PILIHAN GANDA

Satuan Pendidikan :
Mata Pelajaran :
Program :
Tahun Pelajaran :

Penyusun :
Alokasi Waktu :
Kelas/Semester :
Kurikulum :</td></tr>
<tr><td rowspan="2">Kompetensi yang diuji:</td><td>Kunci Jawaban :</td><td rowspan="2">Buku Sumber:</td></tr>
<tr><td>Nomor Soal :</td></tr>
<tr><td>Lingkup Materi:</td><td colspan="2" rowspan="4">Rumusan Soal:</td></tr>
<tr><td>Materi :</td></tr>
<tr><td>Indikator Soal:</td></tr>
<tr><td>Level Kognitif:</td></tr>
</table>

(Source: USBN Administration of MGMP Bahasa Inggris SMA Kab. Banyumas)

Table 6 below will illustrate a question card that has been filled with the information needed.

Table 6. Question Card Form (with information)

<table>
<tr><td colspan="3">KARTU SOAL PILIHAN GANDA

Satuan Pendidikan : SMA Harapan Siswa
Mata Pelajaran : Bahasa Inggris
Program : IPA/IPS/Bahasa
Tahun Pelajaran : 2019/2020

Penyusun : Citra Raisa HS
Alokasi Waktu : 120 menit
Kelas/Semester : XI/2
Kurikulum : K-2013</td></tr>
<tr><td>Kompetensi yang diuji:
Reading
Peserta didik dapat menentukan rincian kegiatan dari bacaan</td><td>Kunci Jawaban :
E

Nomor Soal :
13</td><td>Buku Sumber:
http://www.happykids.taipei/home/school-announcements</td></tr>
<tr><td>Lingkup Materi:
Struktur teks

Materi :
Announcement

Indikator Soal:
Disajikan sebuah teks berbentuk Annoucement (tentang kegiatan sekolah: Liburan), peserta didik dapat menentukan rincian kegiatan dari bacaan.

Level Kognitif:
L1</td><td colspan="2">Rumusan Soal:

Please remember that there are no classes this week (Dec. 25-Jan. 1) due to Happy Kids' Christmas Break. This is a time our staff takes some time to be with their families on this very important holiday.

We wish all of our Happy Kids families to have a wonderful Christmas together. We will see all of our Happy kids back in classes on January 2.
Merry Christmas!

When will the students be back in their class?
A. December 25
B. December 26
C. December 27
D. January 1
E. January 2</td></tr>
</table>

(Source: USBN Question card of SMA Negeri 1 Baturraden)

From the facts written above, it seems that writing a question is not a simple work to do, where teachers had to read and analyze the table of specifications, design the table of specification level mapping, write the question cards form with the number of question items that they had to design and then fill the question cards with the information needed. For the national standardized school examination, teachers had to write 30 multiple-choice question items. They had to design 30 question cards and fill the form with the information needed.

After finishing their works with the question cards, teachers should design the multiple-choice questions from the question cards into a good design test book for the test takers. Then teachers had to bind: the table of specifications, the table of specification level mapping, all the question cards, and the test book. The binding was called as the school examination administration. The teacher submitted the school examination administration as the test designer to the school management. The school management submitted the teachers' work to the MGMP to be reviewed by the reviewer team. From the MGMP, the school examination administration was given back to the school management. The school management would ask the teacher to make some changes based on the review given by the MGMP reviewer team. After that, teacher, as the test designer, gave the school examination administration to the school management, then it sent to the school supervisor to have the legalization. If there were some revisions from the school supervisor, the teachers should revise the question cards and submit them back to the school supervisor. This would be done until the school

supervisor gave the legalization. After the school supervisor's legalization was received, the book test would be the test tool for the students as the test takers.

D. Relevant Research

Studies on multiple-choice items have been conducted by many researchers in Indonesia and other countries (Manalu et al., 2019; Hartati & Yogi, 2019; Toksöz & Ertunç, 2017; Arora, 2018). Manalu et al. (2019) highlighted the quality of multiple-choice items in terms of reliability, level difficulty, discriminating power, and the level or the effectiveness of distractor by using ANATES program version 4.0.9. The analysis showed that from 25 questions, 56% of multiple-choice items are valid, and 44% are invalid. From the reliability side, it is said reliable because it equals 0,90. The multiple-choice items that categorized easy are three items (12%), the satisfactory item's category are 7 (28%), difficult items category are 2 (8%), the category of the poor item is 12 items (48%), average items category are 2 items (8%), good item category is 1 item (4%), and excellent items category is 8 (32%).

A similar focus was also found in Hartati and Yogi (2019). Using Heaton's formula, they examined the quality of multiple-choice items in terms of the difficulty level, the discriminating power, and the effectiveness of distractors on a teacher's own-made summative test. The teacher's English summative test document and the students' answer sheets are used as their research instrument. It was showed that the summative test has more easy items than difficult items with

the ratio of 19:25:6. It was gained information that 21.5% of all distractors are dysfunctional.

The other two studies (Toksöz & Ertunç, 2017; Arora, 2018) explored multiple-choice analysis regarding difficulty index, discrimination index, and distractor effectiveness. Toksöz & Ertunç analyzed the multiple-choice items to test grammar, vocabulary, and reading comprehension and administrated as a state university to preparatory class students. There were 453 students' responses to be analyzed in terms of item facility, item discrimination, and distractor efficiency. It revealed that most of the items are at a moderate level in terms of item facility. 28 % of the items have a low item discrimination value. It had been found that some distractors in the exam are significantly ineffective, and they should be revised.

Arora assessed multiple-choice questions from three aspects. They are difficulty index, discrimination index, and distractor effectiveness. The difficulty index is a measure of whether an item was too easy or too hard. The discrimination index measures the items which can be solved by the students and not. In contrast, distractor effectiveness determines whether distractors tend to be marked by the less able students and not by the intelligent students.

The four pieces of research above discussed the quality of multiple-choice items in terms of reliability, level difficulty, discriminating power, and the distractor's level or effectiveness. They used the students' answer sheets as the response to the multiple-choice questions for conducting the studies. They did not pay much attention to the quality of the multiple-choice questions from the point of view of the design.

USBN test has been implemented for two years in Indonesia. It is a new phenomenon in Indonesian education after the National Final Exam (UAN) era, which long haunted many stakeholders in unfortunate areas. Compared to UAN, USBN is more decentralized, and the school has a certain role in determining the passing grade for their students. As part of the school role, teachers there have more space and chances to propose the question items in the exam. Seen from another perspective, the teachers get more responsibility in the evaluation process. They are required to design a good test. With its focus, the study is about an analysis of USBN or National Standardized School Examination. The multiple-choice items test is a new field. Since the test is bottom-up, the school teachers design them. Therefore, this study will be beneficial for all parties involved in the sustainability of USBN.