

**A VALIDATION STUDY ON NATIONAL ENGLISH EXAMINATION OF
JUNIOR HIGH SCHOOL IN INDONESIA**

by

TEDDY FIKTORIUS
NIM F52212025

A research paper presented to Tanjungpura University
in partial fulfillment of the requirements for the degree of
Master of Education



**MASTERS STUDY PROGRAM OF ENGLISH LANGUAGE EDUCATION
TEACHER TRAINING AND EDUCATION FACULTY
TANJUNGPURA UNIVERSITY
PONTIANAK
2014**

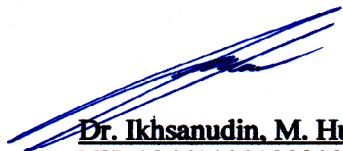
**A VALIDATION STUDY ON NATIONAL ENGLISH EXAMINATION OF
JUNIOR HIGH SCHOOL IN INDONESIA**

Jurisdiction responsibility

TEDDY FIKTORIUS
NIM F52212025

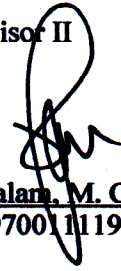
Approved by

Supervisor I




Dr. Ikhsanudin, M. Hum.
NIP 196611051992031003

Supervisor II



Urai Salam, M. CALL., Ph.D.
NIP 197001111998031001

Acknowledged by
Dean, Teacher Training and Education Faculty
Tanjungpura University



Dr. Aswandi
NIP 195805131986031002

A VALIDATION STUDY ON NATIONAL ENGLISH EXAMINATION OF JUNIOR HIGH SCHOOL IN INDONESIA

Teddy Fiktorius, Ikhsanudin, Urai Salam

Masters Study Program of English Language Education, Teacher Training and Education Faculty, Tanjungpura University Pontianak

fiktoriusteddy@yahoo.com

Abstract: The research purpose is to provide feedback about the test quality from a local context that leads to the improvement of National English Examination in Indonesia. Simple random sampling is applied to obtain 1 packet of test items out of 42 packets. Next, the sample of participants as the examinees, 3 junior high schools, is obtained through stratified purposive sampling based on the school accreditation ranks. Lastly, for the sample as the content validity analysts, 3 English subject teachers are purposively selected based on the set criteria. The findings show that the test has fulfilled the criteria of having content validity. However, the test score reliability coefficient calculated with KR-21 is .65 which is categorised *unreliable*. The test developers need to consider revising items with very low or very high item difficulty and very low item discrimination. Finally, a further action needs to be taken to revise the implausible distractors.

Keywords: National English Examination, validity, reliability, item difficulty, item discrimination, distractor

Abstrak: Penelitian ini bertujuan untuk menghasilkan umpan balik tentang kualitas tes dari konteks lokal yang mengarah kepada peningkatan kualitas Ujian Nasional Bahasa Inggris di Indonesia. *Simple random sampling* digunakan untuk mendapatkan 1 dari 42 paket soal ujian. Selanjutnya, sampel peserta ujian diperoleh melalui *stratified purposive sampling* berdasarkan peringkat akreditasi sekolah. Kemudian, untuk sampel analis validitas isi, 3 orang guru bahasa Inggris terpilih menurut tujuan berdasarkan kriteria yang ditentukan. Hasil penelitian menyimpulkan bahwa tes tersebut memiliki validitas isi. Tetapi nilai reliabilitas yang dihitung menggunakan KR-21 adalah 0,65 sehingga dikategorikan *tidak memiliki reliabilitas*. Para pengembang tes perlu mempertimbangkan revisi butir soal dengan tingkat kesukaran yang sangat rendah atau sangat tinggi dan butir soal dengan daya pembeda yang sangat rendah. Akhirnya, tindak lanjut perlu diambil untuk merevisi pengecoh yang tidak berfungsi dengan baik.

Kata kunci: Ujian Nasional Bahasa Inggris, validitas, reliabilitas, tingkat kesukaran, daya pembeda, pengecoh

The quality of any assessment in any educational setting results from the “quality of the instruments” that test administrators use as a basis of decision making (Anderson & Morgan, 2008, p. xi). In the context of National Examination in Indonesia, the Ministry of Education and Culture of the Republic of Indonesia uses the results for making a criterion-referenced decision. Miller,

Linn, & Gronlund (2009) define it as a decision that evaluates a student's performance based on a concisely defined area of knowledge of learning task. In other words, students with scores above the standard are considered to have demonstrated an expected level of ability and therefore pass the examination or vice versa. Considering these concerns, the test items used as the basis of decision making must meet the criteria of having high quality.

However, it has been publicly assumed that the examination lacks some authenticity; given that authenticity is defined as the extent of connection between the characteristics of test tasks and real world tasks (Bachman, 2005; Bachman & Palmer, 1996; Brown, 2003). The table of specifications listed in BSNP Regulation No. 0019/P/BSNP/XI/2012 (2012) shows that the examination only covers a number of reading materials and a small portion of writing materials in a multiple-choice test format. This means that the authenticity issue is related to the curriculum content that draws upon the guidelines set by the Ministry. As the aims of teaching and learning process in the English curriculum are to develop not only the reading skill but also listening, speaking, and writing skills, English tests that omit the assessment of those skills are claimed to lack authenticity.

Holding firmly onto the arguments, teachers start questioning the examination quality. Particularly, they question, "Are current standardized tests of language proficiency accurate and reliable?" (Brown, 2003, p. ix). Therefore, a validation analysis should be conducted to impose quality assurance (Bachman, 2005; Permendikbud No. 66, 2013). It is an ongoing process that should be continuously conducted to build a larger base of evidence. Moreover, validity evidence is always incomplete, it is important to perform a justification of the test use and to direct the research required to obtain a better insight of what the test scores mean and how they can be used in decision making (Weir, 2005).

Basing on these viewpoints, it is interesting to carry out an inquiry to answer the question: To what extent is the test valid and reliable? The research focus is on the content validity, reliability, item difficulty, item discrimination, and effectiveness of each distractor of 50 multiple-choice test items of Junior High School National English Examination in the academic year 2012/2013.

METHODS

In this research, some qualitative evidence and quantitative data are collected, analysed, and interpreted. The research is descriptive and exploratory in nature. Descriptive statistics allow the researcher to describe data and examine relationships between variables that provide information about conditions, situations, and events that occur in the present (Brown, 2011). Additionally, Brown (2011, p. 192) asserts that an exploratory study is conducted to scrutinise "relationships and correlations" of any data that are collected and analysed.

Population refers to the entire group of people, events, or things of interest that the researcher wishes to investigate (McMillan, 1996). In this research, the population of things of interest is 42 packets of the test items of Junior High School National English Examination. Then, the population of people of interest is defined as the students who sit in the ninth grade in the academic year 2013/2014 and the English subject teachers.

Simple random sampling is applied, in which every test item packet shares the same opportunity to be chosen as the sample (Cohen, Manion, & Morrison, 2007). The randomly selected test item packet is BHS_ING_SP_74_SPERTIGA Front 1. Next, stratified purposive sampling is applied to obtain a sample size of 169 ninth graders as the examinees from 3 purposively selected junior high schools out of 49 schools situated in Kota Pontianak, Kalimantan Barat in the academic year 2013/2014 accredited by Badan Akreditasi Nasional (National Accreditation Board). Lastly, taken for the basis of judgment, 3 English subject teachers are purposively selected to be the content validity analysts.

To collect some qualitative data, a validity form for classifying the test items into the content of table of specifications is used. An agreement with some concepts particularly on the categorisation of items into the suitability domains is reached based on the consensus of the majority. Meanwhile, to obtain some quantitative data, documentary study technique is employed by collecting the students' answer sheets and the test question paper of Junior High School National English Examination in the academic year 2012/2013 that is readministered on 17 February 2014 and its table of specifications.

The content validity analysis is concerned with whether or not the content is sufficiently representative and comprehensive for the test to be a valid measure of what it is supposed to measure that can be best examined with a table of specifications. Then, the test score reliability is measured through KR-21.

$$r_{21} = \frac{k}{k-1} \left[1 - \frac{M(k-M)}{kS^2} \right]$$

where, r_{21} = reliability coefficient of the whole test

k = number of items in the test

S^2 = variance of scores

M = mean of the scores

(Rudner & Schafer, 2002, p. 18).

The interpretation of the KR-21 reliability estimate (r_{11}) is as follows. If r_{11} equals to or higher than .70, the test is considered to be reliable. Whereas, if r_{11} is lower than .70, the test is considered to be unreliable (Braun, Kanjee, Bettinger, & Kremer, 2006).

Next, the formula for item difficulty is as follows.

$$IF = \frac{\sum C_r}{N}$$

where, IF = item facility (level of difficulty)

$\sum C_r$ = the sum of correct responses

N = the number of examinees

(Weir, 2005, p. 202).

Thorndike and Hagen (as cited in Anas Sudijono, 2008) assert that a proportion of correct answers less than .30 is classified *too difficult* while a proportion of correct answers that exceeds .70 is labelled *too easy*. In other words,

any given items that have the difficulty indices ranging from .30 to .70 are acceptable and therefore are considered to be good items.

Furthermore, the first step of computing item discrimination is to separate the highest and the lowest scoring groups on the basis of the total score of the test (Brown, 2003). The decision to employ the number of students in each of the two groups is based on the optimal size of each group, which is 27.00% of the total sample. The formula of item discrimination is as follows.

$$D = (RU - RL) / (5T)$$

where, D = item discriminating power

RU = number of students in the upper group who get the item right

RL = number of students in the lower group who get the item right

$5T$ = one half the total number of students included in the item analysis

(Miller, et al., 2009, p. 357).

According to Anas Sudijono (2008, p. 389), the following are the classifications and interpretations of discriminability indices. The item discrimination indices that are below .20 are classified *poor*, those lying from .20 to .40 are labelled *satisfactory*, the classification *good* is addressed to those between .41 to .70, the classification *excellent* is labelled to those above .71, and item discrimination indices with negative signs are labelled with a negative sign.

Lastly, the computation of how well a distractor works is done by applying the computation of 5.00% of the total number of test takers through a frequency table that presents the number and per cent of test takers who select given distractors (Anderson & Morgan, 2008).

RESULTS AND DISCUSSION

Results

First, it is concluded that the test has fulfilled the criteria of having the content validity. Nonetheless, a further analysis is done to provide a better understanding. The analysis on the distribution of the indicator domains reveals that there are two domains, namely communicative purposes of certain texts (in the reading section of the test) and words arrangement (in the writing section) that are not represented by any item in the test although they are stated in the table of specifications as listed in BSNP Regulation No. 0019/P/BSNP/XI/2012. Next, the reliability coefficient of the test scores is .65 which is categorised *unreliable*.

Then, the analysis on the difficulty level shows that the item difficulty indices range from .29 to .93. It indicates a negative result of item difficulty range with the majority of the items (72.00%) or 36 items (items no. 1-5, 7-15, 18-19, 22-26, 31-32, 35-36, 38-41, 43-49) are classified *too easy*, 26.00% of the items (13 items: items no. 6, 16-17, 20-21, 27-30, 33-34, 42, and 50) is categorised *moderate*, and 2.00% (item no. 37) is *too difficult*. After that, the items that discriminate reasonably well between proficiency levels are items no. 35 and 44 with the item discrimination values .46 and .50. Next, there are 32 items (64.00%) with the item discrimination values ranging from .20 to .39. (items no. 1-2, 6, 8-11, 13-14, 18, 21-26, 28-29, 31-34, 39-43, 45-46, and 48-50) that discriminate moderately between the highest and lowest scoring groups. Then, 14 items (28.00%) are classified to have *low* discriminating power with the discrimination

values lying from .07 to .17. Those items are items no. 3-5, 7, 12, 15-17, 19, 30, 36-38, and 47. In addition, items no. 20 and 27 have *negative* item discrimination values indicating that these items fail to discriminate between the stronger and weaker examinees. Lastly, none of the test items has the discrimination ability that is categorised *excellent* with the value ranging from .71 to 1.00.

Finally, by looking at the distractor performances, it is concluded that the 50 items comprise 98 plausible distractors, 52 implausible distractors, and 50 answer keys. Items no. 5-6, 8, 13-14, 16, 20-21, 26-30, 33-35, 37, 39-42, 44, and 50 contain the distractors of incorrect answers that are chosen by at least 5.00% of the total number of examinees. Next, Items no. 4, 11, 17, 22, 25, 31, and 32 contain one of the distractors in each item that does not function well. In addition, items no. 2-3, 10, 15, 18-19, 23-24, 36, 43, and 45-49 are items with two of the distractors in each item that do not obtain the minimum effectiveness index. Furthermore, all of the distractors in items no. 1, 7, 9, 12, and 38 do not function well. Next, the zero effectiveness index is found in items no. 2 and 47, distractors of which fail to attract any responses from the examinees. Finally, item no. 37 contains a distractor whose effectiveness index is higher than the answer key.

Discussion

Content Validity

The content validity analysis shows a very positive result. Nevertheless, more attention should be paid to empirical evidence that there are clearly a domain in the reading section and 3 domains in the writing section that are certainly listed in BSNP Regulation No. 0019/P/BSNP/XI/2012 but do not appear in the test. From this conclusion, the teachers should question why the domains should be listed in the table of specifications but, in fact, do not appear in the test.

Besides, it is also crucial to remember that the test content represents only a general language construct that consists of the reading and writing skills. What is more crucial to discuss is that the writing skill is assessed through the multiple-choice format. Language teachers should question how a student's productive skill (writing skill) can be assessed without asking the student to really write. Subsequently, National English Examination in the academic year 2012/2013 is perceived to lack authenticity. This implies that the test results do not have any capacity to be accepted as a perfect reflection of students' English proficiency.

Reliability

The reliability coefficient of the test scores is .65 which is categorised *unreliable*. Several strategies that can increase the reliability relevant to Junior High School National English Examination are as follows.

1. Increase the number of items and the types of questions in the test.
It is beneficial to have more items and the types of questions to increase the range of scores that subsequently increases the test score reliability.
2. Omit test items which do not perform well in an item analysis.
A statistical item analysis provides information about whether an item is of high quality or not.
3. Use items that permit scoring to be as objective as possible.

In the context of National English Examination in Indonesia, the test also covers some multiple-choice items in the writing section. To test the writing skill, it would not be suitable to use multiple-choice items that test writing concepts without having the students actually write a composition.

Item Difficulty

Having known the difficulty level, the following actions might be taken. First, the test items that are neither *too easy* nor *too difficult*, items no. 6, 16-17, 20-21, 27-30, 33-34, 42, and 50 can be stored in an item bank. These items can be reused as good quality items in the future test administration. Second, there are three possibilities of follow-ups for the items that are classified *too easy*.

1. Those items are discarded.
2. The test developers are expected to investigate why these items can be easily answered correctly by the examinees.
3. The items with low difficulty level can also be stored in an item bank without any revisions. This type of items can be reused in any entrance tests in which taking the test becomes just a formality.

Finally, there are also three possibilities of follow-ups for the items that are labelled *too difficult*.

1. Those items are left out.
2. The test developers are expected to investigate what factors that become the reasons why most of the examinees get into difficulties.
3. The items labelled *too difficult* might also be reused in any tight entrance tests that are strictly administered to filter high-scoring participants.

Item Discrimination

The item discrimination computation results have a negative effect on the validity of the test. If an item cannot discriminate well, the result of the test can give misleading and incorrect information, which in turn may mislead the decision making process. The reason for *low* discrimination values in this case is that the items generally seem not to be *too difficult* for the examinees. The majority of examinees in both the highest and lowest scoring groups answer the items correctly, which does not differentiate well between the various levels of ability.

The follow-ups towards the analysis are stated as follows.

1. The items that have *good* discriminating power are stored in an item bank. They can be reused in the next test administration with the same materials.
2. The items with *satisfactory* discriminating power can be improved.
3. There are two actions that can be possibly taken towards the items with the classification of having *low* discriminating power.
 - It needs a further analysis for modification or revision before the test items are reused in the next test administration.
 - Such items are simply eliminated because they won't be reused in any future test administration.
4. The items with *negative (very bad)* item discriminating can be omitted because they fail to discriminate between weak and strong examinees.

Effectiveness of Each Distractor

The test developers should make sure that all the distractors are plausible. If one distractor is obviously ridiculous, that distractor is not helping to test and discriminate between the highest and lowest scoring groups of examinees. Furthermore, the incorrect distractors that are more prominent than the correct distractors need to be reviewed as the quality of the distractors influences examinees' performances on the test items.

The actions to be taken as the follow-ups towards the results of the analysis on the effectiveness of each distractor are as follows.

1. The distractors with the minimum value of 5.00% of being chosen by the examinees have functioned well and can be reused in the future test.
2. The distractors with the effectiveness indices below 5.00% and thus fail to function well can be either revised or eliminated.

CONCLUSION

The study concludes that the test has fulfilled the criteria of having content validity. However, the test score reliability coefficient calculated with KR-21 is .65 which is categorised *unreliable*. Through item analysis, 42 items (84%) are found to be problematic items. The test developers need to consider revising items with very low or very high item difficulty and very low item discrimination. Finally, a further action needs to be taken to revise the implausible distractors.

REFERENCES

- Anas Sudijono. (2008). *Pengantar Evaluasi Pendidikan*. Jakarta: Raja Grafindo Persada.
- Anderson, P. & Morgan, G. (2008). *Developing Tests and Questionnaires for a National Assessment of Educational Achievement* (Vol. 1). Washington, DC: The World Bank.
- Bachman, L. F. (2005). Building and Supporting a Case for Test Use. *Language Assessment Quarterly*, 2(1), 1-34.
- Bachman, L. F., & Palmer, A. S. (1996). *Language Testing in Practice*. Oxford: Oxford University Press.
- Badan Standar Nasional Pendidikan. (2012). *Peraturan Badan Standar Nasional Pendidikan Nomor 0019/P/BSNP/XI/2012 tentang Kisi-kisi Ujian Nasional untuk Satuan Pendidikan Dasar dan Menengah Tahun Pelajaran 2012/2013*. Jakarta: Badan Standar Nasional Pendidikan.
- Braun, H., Kanjee, A., Bettinger, E., & Kremer, M. (2006). *Improving Education through Assessment, Innovation, and Evaluation*. Cambridge: American Academy of Arts and Sciences.
- Brown, H. D. (2003). *Language Assessment: Principles and Classroom Practices*. California: Longman.

- Brown, J. D. (2011). Quantitative Research in Second Language Studies. In Eli Hinkel (Ed.), *Handbook of Research in Second Language Teaching and Learning* (Vol. II, pp. 190-206). London and New York: Routledge/Falmer Taylor & Francis E-Library.
- Cohen, L., Manion, L., & Morrison, K. (2007). *Research Methods in Education* (6th ed.). London and New York: Routledge/Falmer Taylor & Francis E-Library.
- McMillan, J. H. (1996). *Educational Research Fundamentals for the Consumer* (2nd ed.). New York, NY: Harper Collins.
- Miller, M. D., Linn, R.L., & Gronlund, N. E. (2009). *Measurement and Assessment in Teaching* (10th ed.). Upper Saddle River, NJ: Pearson Education, Inc.
- Rudner, L. W., & Schafer, W. D. (2002). *What Teachers Need to Know about Assessment*. Washington, DC: National Education Association.
- Weir, C. J. (2005). *Language Testing and Validation: An Evidence-Based Approach*. New York, NY: Palgrave Macmillan.