Journal of English Language and Culture *Vol. 15 (No. 2) : 113 – 127. Th. 2025*
Versi Online: http://journal.ubm.ac.id/index.php/english-language-culture *p-ISSN: 2087-8346*
Hasil Penelitian *e-ISSN: 2597-8896*

AN ITEM ANALYSIS ON ENGLISH PROFICIENCY TEST OF NECESA 2.0 AT UNIVERSITAS PGRI WIRANEGARA

Hawa Innafiil Jannah [1)]*, Barotun Mabaroh [2)], and Dewi Masitho Istiqomah [3)]

[1)] Universitas PGRI Wiranegara
[2)] Universitas PGRI Wiranegara
[3)] Universitas PGRI Wiranegara

Received on 27 August 2024 / Approved 28 April 2025

Abstract

English is used widely as an international language. The mastery of English skills has a standardization. One way to analyze (either a person qualified or not) is using a test. A good test should be addressed to its target participants. Competition can be used as the language testing for English learners. English Student Association is one of the organizers that held an English competition every year to celebrate their anniversary in February till March, and English Proficiency Contest (EPC) is one of the competitions that contested in this event. In NECESA 2.0, EPC only contested for Senior High School level. From this background, the researcher intended to analyze the content validity, reliability, item distractor, item difficulty, and discriminating power on English Proficiency Test item of NECESA 2.0.The researcher used quantitative method in this research. This study aimed to collect data by documentation and calculate the results of the five characteristics of a good test using human instrument. The research subjects were 63 participants of English Proficiency Contest of NECESA 2.0. Based on the findings, English Proficiency Contest of NECESA 2.0 already had very good content validity and reliability. The content that has stated in each item is 94% suitable with the English ATP (syllabus). The coefficient value of the reliability of the tests is 0,935. This test also has good item difficulty value. The test-maker was successful make a good proportion in levelling the difficulty of the test. The results of the discriminating power analysis of the participants' answers to the EPC questions can be seen that most of the differentiating power (52%) or 52 items on the EPC are sufficient or in satisfactory category. The item distractor of English Proficiency Contest of NECESA 2.0 still had many unfunctional distractor. The very poor distractors really need to be revised.

Keywords: characteristics of a good test; item analysis; English proficiency test

Abstrak

Bahasa Inggris digunakan secara luas sebagai bahasa internasional. Penguasaan kemampuan bahasa Inggris memiliki standarisasi. Salah satu cara untuk menganalisa (apakah seseorang memenuhi syarat atau tidak) adalah dengan menggunakan tes. Tes yang baik harus ditujukan kepada target pesertanya. Kompetisi dapat digunakan sebagai salah satu bentuk tes kemampuan berbahasa Inggris bagi para pembelajar bahasa Inggris. Himpunan Mahasiswa Bahasa Inggris merupakan salah satu organisasi yang mengadakan kompetisi bahasa Inggris setiap tahunnya dalam rangka merayakan hari jadi mereka di bulan Februari hingga Maret, dan English Proficiency Contest (EPC) merupakan salah satu kompetisi yang dilombakan dalam program ini. Pada NECESA 2.0, EPC hanya dilombakan untuk tingkat Sekolah Menengah Atas (SMA). Dari latar belakang tersebut, peneliti bermaksud untuk menganalisis validitas isi, reliabilitas, distraktor butir soal, tingkat kesukaran butir soal, dan daya pembeda butir soal Tes Kemampuan Bahasa Inggris NECESA 2.0. Peneliti menggunakan metode kuantitatif dalam penelitian ini. Penelitian ini bertujuan untuk mengumpulkan data dengan cara dokumentasi dan menghitung hasil dari kelima karakteristik tes yang baik dengan menggunakan human instrument. Subjek penelitian ini adalah 63 peserta English Proficiency Contest NECESA 2.0. Berdasarkan hasil penelitian, tes kemampuan bahasa Inggris NECESA 2.0 telah memiliki validitas dan reliabilitas yang sangat baik. Konten yang dinyatakan dalam setiap butir soal 94% sesuai dengan ATP Bahasa Inggris (silabus). Nilai koefisien reliabilitas tes ini adalah 0,935. Tes ini juga memiliki nilai tingkat kesukaran butir soal yang baik. Pembuat tes berhasil membuat proporsi yang baik

*Author(s) Correspondence:
E-mail: hawainnafiil@gmail.com

dalam meratakan tingkat kesulitan tes. Hasil analisis daya pembeda dari jawaban peserta terhadap soal-soal EPC dapat diketahui bahwa sebagian besar (52%) atau 52 butir soal EPC memiliki daya pembeda yang cukup atau dalam kategori memuaskan. Distraktor butir soal English Proficiency Contest NECESA 2.0 masih memiliki banyak distraktor yang tidak berfungsi. Distraktor yang sangat buruk benar-benar perlu direvisi.
Kata Kunci: *karakteristik tes yang baik, analisis butir soal, tes kemampuan bahasa Inggris*

INTRODUCTION

Language is a fundamental tool for communication, enabling interaction between individuals and groups. English is a global lingua franca widely used for international communication. The mastery of English skills has a standardization. The standardization of English is a sign that the language is stable and consistent. One common way to evaluate whether an individual has qualified or not is through testing. A test is a tool typically used to measure one's mastery of studied material. It produces scores that inform important decisions in various contexts, such as education, employment, immigration, or certification (Chapelle & Voss, 2013).

Language tests can be designed for different purposes, such as academic admissions, employment, immigration, certification, and competitions. In competition settings, language testing is often used to assess learners' English proficiency. Participating in competitions can be a prestigious way to measure one's language skills, offering a broader and more challenging platform for evaluation. Such tests reveal how much competence the learners have achieved in their language-related knowledge, skills, and abilities. The choice of a test format depended on the learning objectives, the subject matter, and the skills to be assessed. A good test should be addressed to its target participants. For instance, an English test designed for high school students cannot be given to middle school students. The design should also consider the specific goals of the test-takers and the institution conducting the evaluation.

One example is the English Proficiency Contest (EPC), organized annually by the English Student Association (ESA) at Universitas PGRI Wiranegara. This contest is part of a larger event known as the National English Championship of English Student Association (NECESA), which takes place from February to March to celebrate ESA's anniversary. In NECESA 2.0, the English Proficiency Contest (EPC) was held at the national level and targeted Senior High School students only. English Proficiency Contest (EPC) is a proficiency test that measures contestants' knowledge and quality of learning about English at each level. Contestants should complete 100 multiple-choice questions created by the committee within a stipulated time. Therefore, seeing the scale and prestige of the event, the test must be carefully and correctly designed to meet academic and professional standards.

Accordingly, test makers must be knowledgeable and skillful in creating high-quality test-items. One effective way to evaluate the quality of a test is through item analysis. Since the test quality is crucial, there are some processes and steps teachers or test-makers can follow to do an item analysis. This process includes checking the test's validity and reliability. Validity is considered the most important and fundamental requirement for any test (Jayanti et al., 2019), while reliability ensures that the test results remain consistent wherever it takes. Furthermore, item analysis also examines the difficulty level, discrimination power, and effectiveness of distractors (Ika Pradanti & Sarosa, 2018).

From this background, the researcher intended to analyze the quality of the English Proficiency Contest, the written test form by NECESA 2.0, which was held by the English Student Association. Since this program was held at the national level, the English Proficiency Test must demonstrate high-quality, standardized content. This research was expected to be beneficial for teachers, test makers, and future researchers. It provides a reference for improving test items and evaluating test quality, offers guidance for creating standardized questions in future NECESA

*Author(s) Correspondence:
E-mail: hawainnafiil@gmail.com

events, and serves as inspiration and a reference for future studies in item analysis. This study aimed to know the quality of the test items in terms of the content validity, reliability, item difficulty, discrimination power, and distractor effectiveness of the English Proficiency Contest of NECESA 2.0 for the Senior High School level.

LITERATURE REVIEW

Definition of The Test

A test is defined as a systematic procedure for observing and describing one or more characteristics using a numerical scale or categorical system (Khariri, 2020). Test is an important thing to done to measure the competency of the students. Test made about in the end of the learning activity. A test was held to know or get a test score as the result of measuring each competency or skill. Tests are used to measure educational abilities and achievements, tests are used to assess personality characteristics, and tests are developed for the measurement of social attitudes (McDonald, 1999).

Language testing was a part of educational assessment. The use of language tests is widespread in various contexts in some fields, including education, employment surrounding, international migration, language planning, and economic policy (Fulcher, 2013). Tests are designed and administered, among other things, to measure proficiency, to classify students into one of several levels of a course, or to diagnose students' strengths and weaknesses based on specific language categories (Brown & Abeyvickrama, 2018). The success of English teaching-learning activities will affect students' language proficiency and ability (Maharani et al., 2020). Hence, that is crucial to make a good item test and content also.

The Characteristics of Good Item Test

Criteria for a good test include clear instructions for administration, scoring, and interpretation. It may also be beneficial if it saves time and cost in administering, scoring, and interpreting tests (Swerdlik, 2009). There were technical criteria for an evaluation that professionals use to assess the quality of tests and other measurement procedures. In educational assessment, item analysis was a crucial process used to evaluate the quality of individual test items. This analysis helped ensure that each question effectively measures the intended knowledge or skills. Teachers or researchers often analyzed validity and reliability to measure good item tests. Not only that, but other aspects, like item discrimination, item difficulties, and item distractor, were usually good test criteria. It examined how test-takers respond to each item, particularly how difficult the question was, how well it discriminated between high and low performers, and how the distractors (incorrect answer choices) function.

Validity is a key concept in item analysis. It refers to how well a test measures what it is supposed to measure. A valid test accurately reflects the content and constructs it is designed to assess. Validity means ensuring that the test is measuring what it wants to measure related to accuracy and suitability between the test as a measuring tool and the measured object (Asrul et al., 2014). More specifically, validity means ensuring that decisions based on test results are appropriate (Weigle, 2012).

The conceptualization of validity was further deepened by dividing the concept into different types of validity: face validity, content validity, criterion-related (or empirical) validity, and construct validity (Harrison, 1983; Akbari, 2012). Face validity was the degree to which a test appears to measure what it is meant to measure on the surface. Content validity ensured that the test covers all relevant material. Construct validity checked whether the test truly measures the theoretical concept behind it. Meanwhile, criterion-related validity compared test results with external benchmarks or outcomes to confirm effectiveness.

*Author(s) Correspondence:
E-mail: hawainnafiil@gmail.com

Closely related to validity is reliability, which refers to the consistency of test scores. Reliability was technically defined as the degree to which a test produces consistent results across different administrations to the same or similar groups of test takers (Akbari, 2012). A reliable test should produce consistent results if the same individuals are assessed under consistent conditions. A reliable test will yield stable and consistent results, regardless of who administered it or when it is taken.

There were two ways to determine reliability instruments, namely external reliability and internal reliability (Asrul et al., 2014). This consistency can be measured in several ways, such as test-retest reliability (comparing results from two different times), inter-rater reliability (consistency across different scorers) as the external reliability, and internal reliability referred to how well the items on the test measure the same concept, often measured using Cronbach's alpha and KR Method. According to Suharsimi Arikunto (2013), a test may be reliable but not valid, and vice versa; a valid test is usually reliable. This theory proved that a high validity value will produce a high reliability value, and vice versa; a low validity value will produce a low reliability value.

There was a balance of the level of difficulty of the question that also affects the quality of the question. A good question is a question that is neither too easy nor too difficult (Arikunto, 2018). The Item Difficulty refers to how challenging a particular test item is for the test-takers. It is commonly measured by calculating the proportion of students who answer the item correctly. The item difficulty index (IF) is calculated as the number of correct responses divided by the total number of responses. It ranges from 0 to 1. The difficulty index of the items can be noticed from the test-taker's ability to answer the test items. To compile a test, question items should have a balanced level of difficulty, namely, difficult category questions as much as 25%, medium category 50%, and easy category 25% (Sunarti & Rahmawati, 2014; Rahmaini & Nur Taufiq, 2018). These various criteria tend for items with a difficulty index of less than 0.25 and more than 0.75 to be avoided or not used because such items were too difficult or too easy, so they did not reflect a good measuring instrument.

The discriminating power or discrimination index showed how well a question differentiates between students who understand the material and those who do not. The discriminating power of a question is the ability of a question to differentiate between smart test-takers (high ability) and weak test-takers (low ability) (Arikunto, 2018). A good test item was that high-performing students are more likely to answer correctly than low-performing students. The higher the discriminating index of the question, the more the question can distinguish between smart and less smart students (Rahmaini & Nur Taufiq, 2018). If an item did not discriminate well or discriminated negatively, it may need to be revised or removed.

Within multiple-choice questions, item distractors played an important role. Distractors were the incorrect answer options provided alongside the correct one. Effective distractors should be plausible enough to attract test-takers who are unsure of the correct answer. If a distractor was rarely selected, it might indicate it was too obviously wrong and ineffective. The distractor is said to have a good function if the exception is chosen by at least 5% of the test participants Daryanto (2012); Rahmaini & Nur Taufiq (2018). A distractor is considered good if the number of students who choose the distractor is the same or close to the ideal number (Sary, 2018). Well-designed distractors helped to identify students' misconceptions and provided insights into how well an item functions.

Curriculum in The School

A curriculum involved delivering courses that assisted students in achieving their academic or professional objectives. Commonly, a curriculum involved establishing general learning objectives and listing courses and materials. Some syllabuses were similar to lesson plans and included detailed information about course instruction, discussion questions, and specific activities for learners (Wahyuni, 2016).

*Author(s) Correspondence:
E-mail: hawainnafiil@gmail.com

Learning activities in schools must be guided by the relevant curriculum at that time. Currently, the Merdeka Curriculum was the guideline for implementing schooling in Indonesia. The Merdeka Curriculum was a curriculum that applied learning in the form of projects based on student development so that the values contained in Pancasila could be embedded in each student. In the Merdeka curriculum, teachers can choose teaching tools according to students' developmental needs and interests (Nurjanah et al., 2022).

RESEARCH METHODOLOGY

The researcher used a quantitative method in this research to measure numerical coefficient result of each characteristic. This research aimed to collect data by calculating the results of the five characteristics of a good test. This study also sought to find information that could be used to describe the quality of the English proficiency test of NECESA 2.0. The research subjects were the 63 participants in the English Proficiency Contest of the National English Championship of English Student Association 2.0. The English proficiency test being analyzed in this study consisted of 100 multiple-choice items with five options or alternatives consisting of 1 answer key, four distractors and additional distractor which was ‘others’ as an empty answer.

The instruments the researcher used were documentation as the file record of this research. In collecting the data, the researcher borrowed the test document, the answer key of the English Proficiency Test for Senior High School, and the syllabus from the National English Championship of the English Student Association committee. Furthermore, the researcher used herself as a human instrument to gather and analyze the data. The researcher analyzed the data for 5 weeks, from May 30th until June 27th, 2024. After the data had already been collected, the researcher prepared the data, identified the data, tabulated the data using Microsoft Excel, and calculated the result of each characteristic of EPC. The researcher analyzed the results of the characteristics of the proficiency test.

Content Validity

The researcher analyzed the data on content validity, which refers to the curriculum's learning outcomes and flow of learning objectives, using the Merdeka curriculum and handbooks. When a test measures goals consistent with the given information or lesson content, it is said to have content validity (Arikunto, 2018). The researcher analyzed the test items by comparing them with the curriculum and calculating the comparison percentage of test items to determine the level of validity. The result of the percentage the basic competence of each grade being tested calculate by using this formula

$$\mathbf{C} = A/B \times 100\%$$

Where,
C refers to percentage of content validity (conformity level)
A refers to frequency of item appearance
B refers to total number of items

Therefore, the researcher described the percentage of each level or criteria adopted by Arikunto (Fadhlullah, 2019).

*Author(s) Correspondence:
E-mail: hawainnafiil@gmail.com

Table 1. The criteria of the conformity level

81% - 100%	Very good
61% - 80%	Good
41% - 60%	Fair
21% - 40%	Poor
0% - 39%	Very poor

Reliability

The Kurder-Richardson 20 formula is the formula used by researchers to estimate the reliability of the test. This formula assessed the reliability of a single test performed on multiple subjects. Reliability is technically defined as the degree to which a test produces consistent results across different administrations to the same or similar groups of test takers (Akbari, 2012). The researcher would calculate the total number of participants who chose correct and incorrect answers and the variance of the total score using Microsoft Excel. The researcher would calculate the result using the formula manually and, lastly, analyze the coefficient of reliability based on the criteria and the degree level of reliability.

$$r_{11} = \left(\frac{k}{k-1}\right)\left(\frac{s^2t - \Sigma pq}{s^2t}\right)$$

Where:
$\boldsymbol{r_{11}}$ refers to reliability value
$\boldsymbol{k}$ refers to the item numbers
$\boldsymbol{p}$ refers to the proportion of participant who choose correct answer
$\boldsymbol{q}$ refers to the proportion of participant who choose incorrect answer (q = 1 – p)
$\boldsymbol{\Sigma pq}$ refers to the total result of $\boldsymbol{p}$ multiply $\boldsymbol{q}$
$\boldsymbol{s^2 t}$ refers to variance of the total score

Moreover, the researcher adopted the criteria and the degree level of reliability by Sudijono (Islami, 2019).

Table 2. The criteria and the degree level of reliability

0,81 – 1,00	Very high
0,61 – 0,80	High
0,41 – 0,60	Moderate
0,21 – 0,40	Low
0,00 – 0,20	Very low

Item Difficulties

The researcher conducted a quantitative analysis by calculating the participants' answers using the IF formula after tabulating the correct and incorrect answers using Microsoft Excel. The

*Author(s) Correspondence:
E-mail: hawainnafiil@gmail.com

researcher calculated the item facilities using the formula Brown & Abeyvickrama (2018) as follows.

$$\boldsymbol{IF} = \frac{Number\ of\ Correct\ Response\ (n)}{Total\ number\ of\ response\ (N)}$$

Moreover, the researcher adopted the criteria of item facility by Arikunto (2018).

Table 3. The criteria of item facility

0,00 – 0,30	Difficult
0,31 – 0,70	Fair
0,71 – 1,00	Easy

Discriminating Power

The researcher calculated the discrimination power by discriminating between the upper and lower groups. Arikunto (2018) stated that the first thing to measure the upper and the lower group is differentiating the small group (less than 100) should be divided equally between the upper and the lower group (50% of the upper group and 50% of the lower group) or the big group (more than 100) by calculating 27% of the upper and the lower group.

The researcher used a small group to measure the discriminating power of the test, which means all groups were divided equally. At the outset, the participants' total scores were ranked from the highest to the lowest, from 1 to 63. The students with the top 33 total scores, ranked from 1 to 33, belong to the upper-group students, while the students with the lowest scores, ranked from 34 to 63, belong to the lower-group students. The researcher analyzed the discriminating power using the formula adopted by Arikunto (2018).

$$\boldsymbol{D} = \frac{Ba}{Ja} - \frac{Bb}{Jb} = Pa - Pb$$

Where :
D represents discriminating Power
J represents number of test participants
Ja refers to number of participants in upper group
Jb refers to number of participants in lower group
Ba refers to number of participants in upper group who answered a question correctly
Bb refers to number of participants in lower group who answered a question correctly
Pa refers to proportion of participants in upper group who answered questions correctly
Pb refers to proportion of participants in lower group who answered questions correctly

Moreover, the researcher adopted the criteria of discrimination power classification by Arikunto (2018)

*Author(s) Correspondence:
E-mail: hawainnafiil@gmail.com

Table 4. The criteria of discrimination power classification

0,71 – 1,00	Excellent items
0,41 – 0,70	Good items
0,21 – 0,40	Satisfactory items
0,00 – 0,20	Poor items
Negative	Bad items

Item Distractor

Lastly, the researcher carried out a quantitative analysis by calculating the participants' answers using the ID formula. Distraction efficiency was also an important measure of multiple-choice tasks. Distraction analysis measured how much each incorrect option contributes to the quality of a multiple-choice item. The steps to measure the item distractor of the test were tabulating the participants' answer choices and the distractor. After that, the researcher measured the item distractor using the following formula adopted by Sary (2018).

$$ID = \frac{P}{(N-B)(n-1)} \times 100\%$$

Where,
ID refers to distractor index
P refers to number of students who choose distractors
N refers to number of students who take the test
B refers to number of students who answered correctly in any items
n refers to number of alternatives answer

Furthermore, the result of the percentage item distractor classified based on the degree which state by Sary (2018) in her book and the researcher will analyzing the result quatitatively

Table 5. The percentage item distractor classified based on the degree

76% - 125%	Excellent ID
51% - 75% or 126% - 150%	Good ID
26% - 50% or 151% - 175%	Deficient ID
0% - 25% or 176% - 200%	Poor ID
More than 200%	Very poor ID

FINDINGS AND DISCUSSION

The Content Validity of English Proficiency Contest of NECESA 2.0

The researcher found a very high content validity in this item tests. The distribution of material in each grade was similarly equal. For 10th grade and 12th grade level almost had an

*Author(s) Correspondence:
E-mail: hawainnafiil@gmail.com

equivalent number of test questions. There were 38 items that were suitable with 10th grade material, 23 items suitable with 11th grade material and 33 itemswith 12th grade material. Therefore, the result index for the content validity of English Proficiency Contest of NECESA 2.0 was 94%. However, this test still had 6 items which were unsuitable with English ATP (learning purposes content and material).

Moreover, the percentage of each grade was quiet balance. The test-maker (committee of NECESA 2.0) already made the good content in this item test. These were the result of content validity based on the basic competence and learning purpose of each grade.

Table 6. The content validity based on the basic competence of each grade

No.	The basic competence and learning purpose of each grade.	Number test item	The percentage
1.	10th grade	1, 2, 3, 4, 5, 6, 7, 8, 9, 27, 28, 29, 30, 31, 32, 33, 34, 35, 49, 65, 66, 67, 69, 70, 73, 74, 75, 76, 79, 83, 84, 88, 89, 96, 97, 98, 99, 100	38%
2.	11th grade	15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 39, 44, 45, 59, 60, 61, 62, 63, 64, 71, 78	23%
3.	12th grade	10, 11, 12, 13, 14, 36, 37, 38, 40, 41, 42, 43, 46, 47, 48, 50, 51, 52, 53, 68, 72, 80, 81, 82, 85, 86, 87, 90, 91, 92, 93, 94, 95	33%
Total Percentage			94%

From the finding above, English Proficiency Contest of NECESA 2.0 already had good and high content validity. The content that had stated in each item was suitable with the English ATP (syllabus) that consisted of the learning purposes for teaching and learning English activities of Senior High School level. A test that has content validity will measures certain specific objectives that are parallel to the material or lesson content provided (Arikunto, 2018). The researcher also found and analyzed that text-based question (reading comprehension) had the highest frequency of occurrence in this test. It was suitable with the content of English learning materials in Senior High School level based on the Merdeka Curriculum which focused on the literacy, understanding and comprehension aspects.

The Reliability of English Proficiency Contest of NECESA 2.0

The researcher measured the reliability of the test by using the Kuder Richardson 20 Formula. KR-20 was chosen to determine the internal consistency of the EPC of NECESA 2.0 item tests. The researcher used Microsoft Excel to tabulate the data and calculate the total result of the participants who chose correct and incorrect answers and the variance of the total score, then put it into the formula. The researcher found the coefficient value of the reliability of the tests is 0,935

Based on Table 2, the criteria and degree level of reliability were very high. This result means that the test was reliable and consistent. In the world of education, with reliable measuring instruments, the measurement results will be the same or have similar results even if the examiner is different, the proofreader is different, or the question items are different but measure the same thing and have the same item characteristics (Retnawati, 2017). This reliable test can be used repeatedly with the same students or participants whose measurement results will remain relatively the same.

*Author(s) Correspondence:
E-mail: hawainnafiil@gmail.com

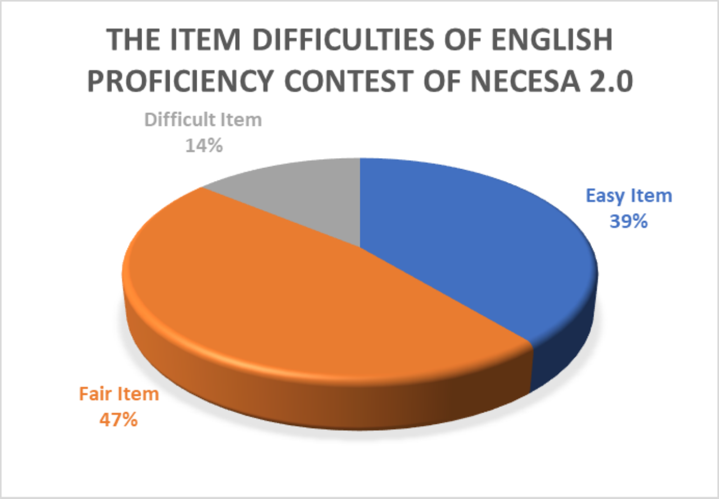
From the validity and reliability results, there was a relationship between the two, where the English Proficiency Contest of NECESA 2.0 has a high validity value and high reliability as well. According to Suharsimi Arikunto (2013), a test may be reliable but not valid, or vice versa, a valid test is usually reliable. This result proves that a high validity value will produce a high-reliability value, and vice versa; a low validity value will produce a low-reliability value.

The Item Difficulties of English Proficiency Contest of NECESA 2.0

Analyzing the difficulty index means examining the test items to identify low, moderate, and high difficulty items (Fauzie et al., 2021). The analysis tabulated the values related to the correct and incorrect answers. Next, the data in the table was calculated using the IF formula. The results obtained in each item test were analyzed and categorized according to the level of difficulty of the questions in Table 3.

The researcher found several difficulty levels in the English Proficiency Contest of NECESA 2.0 item tests. The result were 39 items categorized as easy-level questions, 47 items categorized as fair-level questions, and 14 items categorized as difficult-level questions. The English Proficiency Contest of NECESA 2.0 item difficulty was good and quite balanced in leveling the question. Below is the diagram of the measurement item difficulty of English Proficiency Contest of NECESA 2.0.

Figure 1. The Item Difficulties of English Proficiency Contest of NECESA 2.0



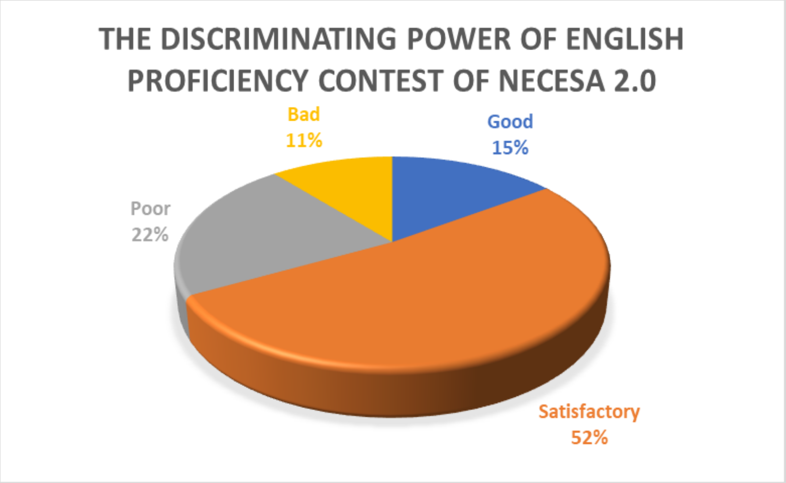
As a follow-up to the item difficulty analysis, Sudijono cited in Fatimah and Alfath (2019) stated that the good category items (fair items) that have a level of difficulty moderate or medium difficulty (47 items) should be quickly recorded in the question bank and can be issued or used again in subsequent tests. For items that are included in the difficult category (14 items), there are three possible follow-ups, namely, discarded or dropped (will not be reissued in the next test), re-examined to find out the cause of the number of students who could not answer the item, and review the use of difficult items. Difficult questions can later be used for rigorous selection tests requiring highly competent skills. So, difficult questions are needed to qualify the best competence. Items that fall into the category of easy items (39 items) can be discarded and not reissued in the following tests; the test-maker can research and track why the item was so easy so that all test-takers can pass the test. Furthermore, there were improvements to be made. It was the same with difficult items; not all easy items have no benefits. Easy items can be used on tests, especially in a selection test that does not require high skills.

The Discriminating Power of English Proficiency Contest of NECESA 2.0

The discriminating power of a question is the ability of a question to differentiate between smart test-takers (high ability) and weak test-takers (low ability) (Arikunto, 2018). For a question that can be answered correctly by both high and low ability test-takers, then the question is not good because it has no distinguishing power. Vice versa, if test-takers who are smart or not smart cannot answer correctly, the question is also not good because it has no distinguishing power (Umi Fatimah & Alfath, 2019).

Discriminating power is calculated by subtracting the proportion of upper group participants who answered correctly from the proportion of lower group participants who answered correctly. After calculated, the researcher found 0 excellent item, 15 number of tests were good items, 52 number of tests were satisfactory items, 22 numbers of tests were poor items, and 11 number of tests were bad items. From the result, English Proficiency Contest of NECESA 2.0 had enough ability to discriminate the participants. Below is the diagram of the measurement discriminating power of English Proficiency Contest of NECESA 2.0

Figure 2. The Discriminating Power of English Proficiency Contest of NECESA 2.0



Based on the finding above, it can be seen that the English Proficiency Contest questions had sufficient quality seen in terms of differentiating power, more than 50% of the total questions were in the satisfactory category, it had enough ability to differentiate upper group and lower group students. As a follow-up to the differentiating power analysis, Sudijono, cited in Fatimah and Alfath (2019), stated items that have good discriminating power (satisfactory and good criteria) should be stored in the question bank. This research found 15 items in the good criteria and 52 items in satisfactory criteria. It mean those questions can be stored in the question bank and can be used again in the next test because the quality is good enough. Then, items with poor discriminating power (22 items) should be corrected and revised so that they can be stored in the question bank to be used for future learning outcomes tests. Items with negative discriminating power (11 items), it should be discarded, and could not be used in future tests because these items have very poor quality.

The Item Distractor of English Proficiency Contest of NECESA 2.0

A distractor is considered good if the number of students who choose the distractor is the same or close to the ideal number (Sary, 2018). The researcher tabulated the values related to the correct answers, four answers to the exceptions, and one 'other' answer. Furthermore, the data in the table is calculated using the ID formula. The results of the distractor index on each of the six answer

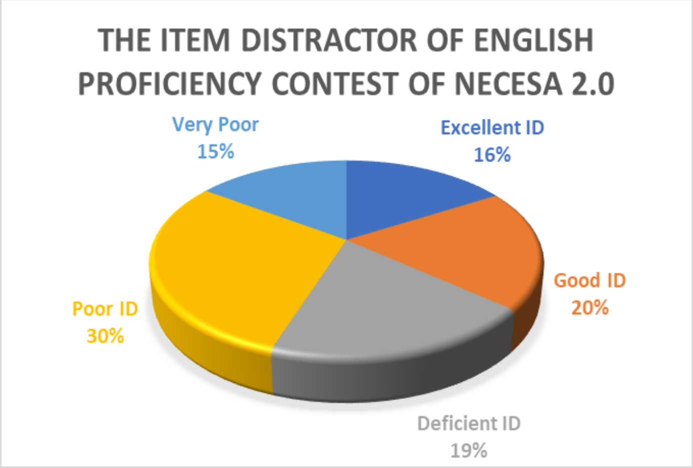
*Author(s) Correspondence:
E-mail: hawainnafiil@gmail.com

choices obtained in each question item were analyzed and categorized according to the percentage of distractor degree items in Table 5.

The researcher found that the item distractor of the English Proficiency Contest of NECESA 2.0 was quite good but still poor. There were 82 distractors categorized as excellent ID, 99 distractors categorized as good ID, 95 distractors categorized as deficient ID, 148 distractors categorized as poor ID, and 76 distractors categorized as very poor ID. The total number of distractors in the English Proficiency Contest of NECESA 2.0 is 500 distractors.

This item test had five distractors: one correct answer for each item and an additional distractor, ‘others, ' as an empty answer. Based on the result above, it showed that many item numbers still had unfunctional distractors. Below is the diagram of the measurement item distractor of English Proficiency Contest of NECESA 2.0

Figure 3. The Distribution of Item Distractor of English Proficiency Contest of NECESA 2.0



In multiple-choice questions there were alternative answers (options) that are distractors. For good questions, the distractors will be chosen equally by students who answer incorrectly. On the other hand, the question items are not good the distractors will be selected unevenly. A distractor is considered good if the number of students who choose the distractor is the same or close to the ideal number (Sary, 2018).

Based on Arikunto's (1984) statement cited in Fatimah & Alfath (2019), a distractor not selected by the test-takers or participants means the distractor is poor, bad, or too conspicuously misleading. On the other hand, a distractor is said to function well if the distractor had an excellent attraction for test-takers who do not understand the concept or do not master the material. Thus, it can be interpreted that poor distractors (30%) do not work, very poor distractors (15%) were misleading, then poor and very poor distractors needed to be replaced because they were bad, and deficient distractors (19%) needed to be revised due to deficiencies. For a good and excellent distractor, it did not need to be replaced or revised.

CONCLUSION AND SUGGESTIONS

Conclusion

In this section, the researcher presented the conclusion of the content validity, reliability, item difficulty, discriminating power, and item distractor of the English Proficiency Contest (EPC) of the National English Competition of English Student Association 2.0 (NECESA 2.0) for Senior High School at Universitas PGRI Wiranegara. Based on the findings, the researcher concluded that the test had the good quality.

*Author(s) Correspondence:
E-mail: hawainnafiil@gmail.com

First, the content validity of the English Proficiency Contest of NECESA 2.0 was good, and it had high content validity. The content stated in each item is suitable with the English ATP (syllabus) that consists of the learning purposes for teaching and learning English activities at the Senior High School level. Second, the researcher found the coefficient reliability of the English Proficiency Contest of NECESA 2.0 of the tests is 0,935. It indicated that this item test is very reliable. Furthermore, it means the test is reliable and can be used repeatedly with the same students or participants whose measurement results will remain relatively the same. Third, the test also had a high item difficulty value and successfully made a good proportion in leveling the difficulty of the English Proficiency Contest of the NECESA 2.0 test. It can be known that the questions were relatively balanced, as seen from the level of difficulty. Fourth, the discriminating power analysis results of the participants' answers to the EPC questions had enough ability to differentiate between upper-group and lower-group students. This research found 15 items in the good criteria and 52 items in the satisfactory criteria. Lastly, the researcher found the item distractor of the English Proficiency Contest of NECESA 2.0 still had many unfunctional distractors and ineffective indexes. It can be concluded that this test had quite good distractors but still had many poor distractors and needs to be revised or discarded.

Suggestions

A test for a competition should be made as good as possible. If necessary, the test can be tested first and then analyzed to determine the quality of the test. English teachers accompanying the participants of EPC can use this research as a reference and guidance to improve the making of items test and to know the quality of the English Proficiency Test that has competed in NECESA 2.0. The teacher should conduct test item analysis because analyzing the test item could help the teacher know the quality of the test made by the teacher. For the test makers (especially the committee of NECESA 2.0), this research can guide and evaluate the test makers in making good and standardized items for the next NECESA 3.0. This research is beneficial to future researchers. It can be helpful to get new information about item analysis and to add references for future researchers.

REFERENCES

Akbari, R. (2012). Validity in Language Testing. In B. O'Sullivan, C. Coombe, P. Davidson, & S. Stoynoff (Eds.), The Cambridge Guide to Second Language Assessment. Cambridge University Press.

Arikunto, S. (2018). DASAR-DASAR EVALUASI PENDIDIKAN (R. Damayanti, Ed.; 3rd ed.). Bumi Aksara.

Asrul, Ananda, R., & Rosnita. (2014). EVALUASI PEMBELAJARAN. Cita Pustaka Media .

Brown, H. D., & Abeyvickrama, P. (2018). LANGUANGE ASSESMENT PRINCIPLES AND CLASSROOM PRACTICES (3rd ed.). Pearson.

Chapelle, C. A., & Voss, E. (2013). Evaluation of Language Tests Through Validation Research. In The Companion to Language Assessment (pp. 1079–1097). John Wiley & Sons, Inc. https://doi.org/10.1002/9781118411360.wbcla110

Dąbrowska, E. (2015). What exactly is Universal Grammar, and has anyone seen it? In Frontiers in Psychology (Vol. 6). Frontiers Media S.A. https://doi.org/10.3389/fpsyg.2015.00852

*Author(s) Correspondence:
E-mail: hawainnafiil@gmail.com

Ekasari, E. N. (2017). Item Analysis of the English Proficiency Contest for SMP at ESA's Anniversary at STKIP PGRI Pasuruan. University of PGRI Wiranegara .

Fadhlullah. (2019). An Item Analysis of The English Proficiency Contest For Junior High School At The 28th Anniversary of English Student Association. University of PGRI Wiranegara.

Farhady, H. (2012). Principle of Language Assessment. In C. Coombe, P. Davidson, B. O'Sullivan, & S. Stoynoff (Eds.), The Cambridge Guide to Second Language Assessment. Cambridge University Press.

Fauzie, M., Pada, A. U. T., & Supriatno, S. (2021). Analysis of the difficulty index of item bank according to cognitive aspects during the Covid-19 pandemic. Jurnal Penelitian Dan Evaluasi Pendidikan, 25(2). https://doi.org/10.21831/pep.v25i2.42603

Fulcher, G. (2013). Practical Languange Testing (Chatham & Kent, Eds.; Vol. 2). Routledge. https://books.google.co.id/books?id=qbAuAgAAQBAJ&lpg=PP1&ots=7jHJILc3Ag&dq=fulcher%202013&lr&hl=id&pg=PR3#v=onepage&q=fulcher%202013&f=false

Ika Pradanti, S., & Sarosa, T. (2018). An Item Analysis of English Summative Test for The First Semester of The Third Grade Junior High School Students in Surakarta.

Islami, Z. F. (2019). Item Analysis on The English Final Test for Eleventh Grade Students in SMAN 1 Pasuruan. University of PGRI Wiranegara.

Jayanti, D., Husna, N., & Hidayat, D. N. (2019). The Validity and Reliability Analysis of English National Final Examination for Junior High School. 3(2). https://doi.org/10.29408/veles.v3i2.1551.g929

Khariri, A. (2020). 20 Item Analysis of English Summative Test at The Eleventh Grade Students of SMA Negeri 9 Kota Jambi City Academic Year. In Journal Of English Language teaching (Vol. 28, Issue 1).

Lestari, J. R. P. (2020). An Item Analysis On English Proficiency Contest For Junior High School at Universitas PGRI Wiranegara Pasuruan. University of PGRI Wiranegara.

Maharani, A. V., Hidayanto, N., Putro, P. S., Vidya, A., Pancoro, H., & Putro, S. (2020). Item Analysis of English Final Semester Test. In Item Analysis of English Final Semester Test Indonesian Journal of EFL and Linguistics (Vol. 5, Issue 2). www.indonesian-efl-journal.org

McDonald, R. P. (1999). Test Theory: A Unified Treatment (1st ed.). Psychology Press.

Nurjanah, K., Saadah, H., Id, K. A., & Id, H. A. (2022). IMPLEMENTASI PROJEK PENGUATAN PROFIL PELAJAR PANCASILADENGAN TEMA SUARA DEMOKRASI DI SMK SETIA KARYA.

Rahmaini, A., & Nur Taufiq, A. (2018). ANALISIS BUTIR SOAL PENDIDIKAN AGAMA ISLAM DI SMK N 1 SEDAYU TAHUN AJARAN 2017/2018 (Analisis Tingkat Kesukaran, Daya Pembeda dan Fungsi Distraktor pada Soal Pilihan Ganda Kelas XI). Jurnal MUDARRISUNA, 8(1).

Retnawati, H. (2017). Reliabilitas Instrumen Penelitian.

Sary, Y. N. E. (2018). BUKU MATA AJAR EVALUASI PENDIDIKAN (1st ed.). DEEPUBLISH.

Swerdlik, C. (2009). Psychological Testing and Assessment: An Introduction to Tests and Measurement 7th Edition (7th ed.). McGraw-Hill Companies .s

*Author(s) Correspondence:
E-mail: hawainnafiil@gmail.com

Umi Fatimah, L., & Alfath, K. (2019). ANALISIS KESUKARAN SOAL, DAYA PEMBEDA DAN FUNGSI DISTRAKTOR. Jurnal Komunikasi Dan Pendidikan Islam, 8(2).

Wahyuni, S. (2016). CURRICULUM DEVELOPMENT IN INDONESIAN CONTEXT The Historical Perspectives and the Implementation.

Weigle, S. C. (2012). Assessing Writing. In B. O'Sullivan, C. Coombe, P. Davidson, & S. Stoynoff (Eds.), The Cambridge Guide to Second Language Assessment. Cambridge University Press.

*Author(s) Correspondence:
E-mail: hawainnafiil@gmail.com