User Pro

Ok pdf_230218.doc

Upload Here (No Repository)-7

paper 2025

Punjab Agricultural University

Document Details

Submission ID

trn:oid:::1:3375078905

Submission Date

Oct 16, 2025, 1:22 PM GMT+5:30

Download Date

Oct 16, 2025, 1:32 PM GMT+5:30

File Name

Ok_pdf_230218.doc

File Size

2.6 MB

23 Pages

11,588 Words

70,939 Characters

18% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography

Match Groups

44 Not Cited or Quoted 18%
Matches with neither in-text citation nor quotation marks

0 Missing Quotations 0%
Matches that are still very similar to source material

1 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

16% Internet sources
2% Publications
4% Submitted works (Student Papers)

Match Groups

- **44** Not Cited or Quoted 18%
  Matches with neither in-text citation nor quotation marks
- **0** Missing Quotations 0%
  Matches that are still very similar to source material
- **1** Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- **0** Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

16% Internet sources
2% Publications
4% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Internet	
pnojournalarchive.files.wordpress.com		14%
2	Internet	
eprints.univetbantara.ac.id		1%
3	Student papers	
Panipat Institute of Engineering & Technology		<1%
4	Student papers	
Caleb University		<1%
5	Student papers	
NEBOSH		<1%
6	Student papers	
Pope High School		<1%
7	Publication	
Endah Retnowati, Anik Ghufron, Marzuki, Kasiyan, Adi Cilik Pierawan, Ashadi. "Ch...		<1%
8	Publication	
In Sung Hwang, Woo Young Lee. "Block Toeplitz operators with rational symbols (...		<1%
9	Internet	
eprints.gla.ac.uk		<1%

Perspectives of Science & Education

International Scientific Electronic Journal
ISSN 2307-2334 (Online)

Available: https://pnojournal.wordpress.com/2023-2/23-02/
Accepted: 23 November 2022
Published: 30 April 2023

S. SUWARTO, S. SUYAHMAN, M. SUSWANDARI, Z. ZAKIYAH, A. HIDAYAH

The COVID-19 pandemic and the characteristic comparison of English achievement tests

Aim. The objectives of this research were to characterize and contrast the features of English language proficiency tests conducted before and during the COVID-19 pandemic.

Methodology and research methods. Before coronavirus pandemic, there were 287 students; during pandemic, there were 288 pupils; there were also an English teacher and a forum for English teachers. Through documentation and interviews, the information was gathered from eighth-graders at SMP Negeri 2 Semarang in Central Java, Indonesia.

Results. Some aspects of English accomplishment tests made before COVID-19 can be seen. First, the percentages of items in the Easy, Moderate, and Difficult categories are 55%, 37.5%, and 7.5%, respectively. The item discrimination percentages for the Poor, Fair, Good, and Very Good categories are 10%, 30%, 25%, and 35%, respectively. Third, the distractor's effectiveness as a percentage is 53.30% and 46.70%. (effective: ineffective). Finally, the test reliability value is 0.990. The English proficiency test created during COVID-19 exhibits some of the same traits. First, the percentages for Easy, Moderate, and Difficulty categories for item difficulty: 10%, 84%, and 6%. The item discrimination percentages for the Poor, Fair, Good, and Very Good categories are 2%, 4%, 14%, and 80%, respectively. Third, the distractor's percentage efficacy is 99.30%: 0.70% (effective: ineffective). Finally, the test reliability value is 0.960. The foundation of classical test theory (CTT) was the effectiveness of the distractor, item difficulty, and item discrimination. The exams administered during coronavirus pandemic were more normally distributed than the tests administered prior to pandemic based on item difficulty. The tests given during coronavirus pandemic fell more into the very good category than the tests given before pandemic, according to item discrimination. In comparison to tests conducted before to coronavirus pandemic, more tests during pandemic were classified as effective based on the distractor's effectiveness. Both tests were compared based on the data of the collected features. The English achievement exam created during the epidemic was determined to be superior to the test created prior to the outbreak based on CTT. However, the English performance exam created before the epidemic is superior than that created during the pandemic, according to Item Response Theory (IRT). IRT was based on item fit and dependability. Testing for dependability before COVID-19 is more accurate than during pandemic. Before COVID-19, item fit tests were more favorable than during pandemic.

Conclusions. The English proficiency test that was created during the epidemic is superior to the test that was created prior to the pandemic based on CTT. But according to IRT, the English proficiency exam created before the pandemic is superior to that created during the pandemic.

Keywords: COVID-19, coronavirus pandemic, achievement test, characteristics test, item difficulty, item discrimination, distractor effectiveness, validity, reliability

For Reference:
Suwarto, S., Suyahman, S., Suswandari, M., Zakiyah, Z., Hidayah, A. (2023). The COVID-19 pandemic and the characteristic comparison of English achievement tests. *Perspektivy nauki i obrazovania – Perspectives of Science and Education, 62* (2), 307-329. doi: 10.32744/pse.2023.2.18

Introduction

Program for International Student Assessment (PISA) [1], national Australian standardized tests [2], international standardized tests, and the Post-Soviet Central Asian educational context where standardized examinations are used are all examples of standardized testing utilize only established tests [3].

Students will have their language skills evaluated regularly throughout a particular period of English study. They may learn about their level of competence by doing this. The test [4] shows that a teacher can obtain students' English learning outcomes at the end of the teaching-learning process. and progress. Each educational level conducts evaluations at regular intervals during the learning process [5], and assessments and instruction are provided online [6]. the significance of this English test [7]. It suggests that teachers should always give an assessment or a way to assess students' academic development. In order to evaluate students' learning outcomes, teachers should regularly monitor their progress. Testing, measurement [8], classroom the evaluation process [9], the teaching and learning process [10], and teaching analytics [11]. assessing student progress and development at the conclusion of a unit or instructional period, and A teacher's duties include informing stakeholders about the results and achieving success. Giving pupils feedback on their progress is one of a teacher's responsibilities as an educator. and striving to improve the learning atmosphere. Teachers employ tests, one of the most effective methods for assessing the amount and quality of students' learning. [12]. Students must answer questions on standardized exams with multiple options. Additionally, [13] defined a test as a frequent way for collecting behavioral data from a certain field. When performing a test, it is essential to comprehend what it is and what it measures. comes to test creation. In other words, a test is a well-thought-out instrument that, in its entirety, evaluates actual learning results from the real world. reflects the desired behavioral traits [14]. A comprehensive learning objective should contain the following components, according to some: 1) observable actions,

2) the circumstances in which the planned action should occur, and 3) the degree of performance that is considered adequate to show mastery of learning results. when assessing the knowledge and concepts that promote students' cognitive, affective, and psychomotor growth.

To determine the students' level of English competence, the instructor often gives the class a test. After covering each chapter of the subject, teachers are able to They can administer tests or give them at the end of the semester. This sort of test is called an achievement test, which is defined as follows: It is a popular way to measure educational assignments and a vital data source for decision-making. the degree to which students have achieved the intended learning outcomes, as well as their ability to absorb new information during a lesson or throughout a course. Teachers, schools, and educational institutions must evaluate them. The researchers believed that achievement assessments should be used to demonstrate the students' greatest performance. well structured. Educators can utilize the results of achievement assessments to guide choices or provide insights on ways to improve teaching and learning. During official classroom instruction, achievement tests may be administered in the form of formative assessments, summative tests, the National Final Exam, and college entrance exams [15].

A summative test is a task that yields grades or scores that are used to evaluate the performance of the students. When all subjects have been covered, this test is conducted. The English Summative Test [16] is one example of a summative test that is used to categorize awards and grades at the conclusion of a course or program [17]. Formative tests, on the other hand, are used to monitor pupils' academic development and give them feedback to help them do better. Student understanding of their strengths and limitations is improved by formative assessments. Teachers can use the findings to help pupils become more proficient in their weak areas.

To create an achievement test that is valid and trustworthy, a teacher must be familiar with the principles of excellent test development. The facets of its application in the classrooms must also be known by the teachers. They should also be able to score, and most crucially, analyze, the outcomes of these assessments. According to [8], test creators and users should consciously work to improve the validity and reliability of the tests by lowering measurement errors in order to obtain objective data. Well-designed test questions should be employed, and test developers should fit the learning objectives, when assessing what students already know or have learned in their field of study. For a test score to be considered reliable, learning, teaching, and subject understanding must all be in balance. Learning outcomes are a useful approach to maintain high standards and enhance instruction. Li et al. make the point that a practical exam [18] must be precisely specified [19] in order to measure the actual score. A valid test should have high-quality items that adhere to test requirements and provide accurate data with few errors [20]. The excellent test result may help to explain actual learning outcomes.

A good test must satisfy S. Suwarto's [21] definition of a good test, which includes validity, reliability, item difficulty, item discrimination, and effective distractors (for multiple choice questions). To ascertain the degree of validity and reliability of the assessments, it is important to analyze the test items. As a result, the test's quality will be similar to the quality of each item's test result, which in turn affects the test's overall quality. Teachers should concentrate on the test item quality, thus they should perform item analysis to evaluate each item's quality and determine which questions must be updated or removed if they fulfill the criteria.

Numerous research have been conducted, particularly in junior high schools in Indonesia, on the characteristics of English accomplishment assessments created by the English teacher forum/Musyawarah Guru Mata Pelajaran (MGMP) [22]. However, from December 2019 until the present, all teaching and learning activities, including exams, are temporarily conducted at home owing to the Corona Virus Disease 2019 (COVID-19) pandemic. It must be done to reduce physical contact that promotes the spread of the virus. As a result, testing and assessment are done online utilizing laptops or mobile devices [23].

It is thought that one of the best ways to stop the spread of COVID-19 in the educational setting is to employ online media to limit engagement [24]. The teacher offers tests to students or parents via computers or cellphones. The kids then do their homework or tests at home [25]. The COVID-19 pandemic condition prevented researchers from doing a study on this subject. There is a necessity to look into the end product in a pandemic since English success exams are created individually by teachers due to distance restrictions [26]. Researchers' interviews indicate that the English subject test utilized in SMPN 2 Semarang was never administered. According to the syllabus, the English teacher created the test without using a plan. The English teacher forum in sub-rayon 01 East Semarang created the English language learning accomplishment exam at SMPN 2 Semarang before the COVID-19

outbreak, but the test instrument was never put to use on students. The English teacher forum was the only entity to cross-check the instrument.

In order to examine the qualities of the English accomplishment tests developed before and after COVID-19 in terms of validity, reliability, item difficulty, item discrimination, and distractor efficacy based on Classical Test Theory (CTT) and Item Response Theory (IRT). The two assessments were compared in part because no researchers had ever examined tests conducted by the teacher forum under normal circumstances and independently by English teachers during the coronavirus pandemic. For English teachers, educators, test makers, and other parties involved in the test's development, this research is anticipated to offer comments and examples. Additionally, this research was done to serve as a guide for future studies that will be similar to it.

The English achievement test created prior to COVID-19 included 40 multiple-choice items and 5 essay questions, according to interviews and supporting evidence. The test was created by the English forum teachers by first defining the accomplishment test development based on the area, subject, goal, resources, test type, and amount of test elements. Second, a strategy that included precise goals and metrics was created. Third, test objects were built in accordance with a test blueprint. Fourth, test validation was carried out by cross-checking with other forum participants who spoke English. Fifth, editing was used to revise the test after cross-checking. The sixth step was grouping good items into a set of tests. The test equipment was finally printed and shipped to schools.

Only 50 multiple-choice questions were included on the English proficiency test that was created during the pandemic. The teacher initially defined the accomplishment exam's subject, objective, source material, test format, and quantity of test elements. She then created the test by copying and pasting the questions that had previously been created by English forum teachers into the google form.

Based on the aforementioned test development procedure, neither test developer examined the content validity, reliability, item difficulty, item discrimination, or effective distractors of the English accomplishment test. They were therefore unaware of whether or not the test items were thought to be valid indicators of students' true aptitude. It was considered that they were unable to assess the qualities of a good test because of time restrictions and high prices.

Sumadi [27] asserts that the test region, test subject, objective test, test material, type test, and other test items should all be included in the specificity of accomplishment test formulation. Teachers should carefully and appropriately create test items. They must first create a blueprint achievement test with a clear purpose, a clear value, and indicators. Second, they should create test objects in accordance with a blueprint while creating a test. The test must be validated a third time by review, expert opinion, and validation. Fourth, the exam must be revised in light of the validator's recommendations. Fifth, in order to analyze the test characteristics, which include item difficulty, item discrimination, the role of the distractor, and reliability based on CTT, the test items that are deemed to be good are placed in the draft before being tried out with a group of students in accordance with the test subject. Sixth, the test items are chosen depending on the findings of the IRT study. Finally, the items included in the standardized test are those that pass the test. The examination will be printed and provided to pupils or schools.

The researchers narrowed their analysis of multiple-choice exams based on the identification of the aforementioned issue because the English accomplishment test created during COVID-19 did not include any essay questions. It would be simpler to compare the

traits of the English proficiency tests created before and during the coronavirus pandemic as a result. Thus, the following study questions were put forth: (1) What were the features of the English performance test that was created prior to the pandemic? (2) What features distinguish the English proficiency exam created during the pandemic? (3) Were the tests created prior to the epidemic and those created during it different in any way?

Methods

1. Research Design

Methods of analysis and comparison were used in this study. The characteristics of the tests created before and during the epidemic were described and analyzed using the test analysis study. The test's qualities were divided into Very Good, Good, and Poor categories. The status of the test item – acceptance, amendment, or abolition – was then explained. The researchers compared the test's properties using the comparison approach after they had examined the test.

2. Research Site

The characteristics of English Achievement tests created before and during the epidemic were compared in the study. The assessments were created at SMPN 2 Semarang, which is located on Brigjend Katamso Street No. 14 in Karangtempel East Semarang, Semarang City, Central Java, for eighth-grade students in the academic years 2017–2018 and 2020–2021. The study was conducted between September and December of 2021.

3. Research Objectives

The purpose of this study was to examine the traits of tests created both before and during the pandemic. Students' replies on the test answer sheets were used to compile the data. In the academic year 2017–2018, there were 287 student answer sheets, and in the academic year 2020–2021, there were 288 student answer sheets online. A teacher of English and the director of the English teacher forum were both present, and they both learned more about how the English accomplishment test was created.

Data Collection

Through interviews and documentation, data were gathered. The eighth-grade English curriculum, the English achievement test grid, the English achievement test papers, the answer keys, and the student answer sheets were all examined. Validity and reliability of the test were determined by analysis. Distractors, item discrimination, and item difficulty were also examined. The purpose of the interview with English teachers and members of the English teacher forum was to learn more about how English accomplishment assessments are developed. Exams created prior to the pandemic had 40 items, whereas tests created during the pandemic had 50 multiple-choice questions. The English teacher at SMPN 2 Semarang and the head of the English teacher forum in sub rayon 01 of East Semarang Region 01 provided the answer key. The item difficulty, item discrimination, alternatives, and dependability based on CTT were all examined using the answer sheets. To assess the validity, the English course syllabus and template were employed.

Unstructured interviews were undertaken by the researchers as one of the methods for gathering data. This was consistent with the research methodology that was used, which

heavily relies on the researchers' comprehension and the data gathered through observations and interviews. The researchers requested authorization from the administrative team and the school principal to conduct study at SMPN 2 Semarang. The English teacher was also contacted by the researchers to obtain data on the eighth-grade pupils in the academic years 2018 and 2021 as well as information on the school's curriculum. They were questioned about how the COVID-19 pandemic affected the creation of the English accomplishment test and received information on the leader of the English teacher forum in Sub Rayon 01 East Semarang. Then, in order to learn more about the process for creating the English achievement test for the 2018 academic year, the head of the English teacher forum in Sub Rayon 01 East Semarang was interviewed.

4. Data Analysis Technique

Quest was used to analyze the data.

Item Difficulty

The total number of right responses divided by the total number of respondents [28; 29], represents the difficulty of each test item.

Three levels of difficulty—Easy, Moderate, and Difficulty—can be applied to the object. The category of item difficulty is as listed in [21].

Table 1

The Category of the Item Difficulty

P = The item difficulty	Category
P > 0.700	Easy
0.300 < p < 0.700	Moderate
P < 0.300	Difficult

The output file of the Quest software displays item difficulty as a percentage (%) row based on the Quest. The proportion of students' accurate answers is expressed as a percentage (%) of the overall Quest output. When the item difficulty index is near to 0 or 1, it means that the question is either too simple or too complex for students [30].

Item Discrimination

The point biserial correlation formula can be used to determine each test item's item discrimination. The item discrimination index can be calculated using the Point biserial (Pt-Biserial) formula, which can detect item discrimination in Quest output [30]. Since many teachers used the technique, the researchers used a point correlation model to statistically determine the item discrimination [15]. According to Suwarto [31], a point-biserial correlation is a bivariate correlation approach. To apply the approach, variable 1 is discrete data (dichotomy), and variable 2 is continuous data (interval data). By developing a correlation between item scores and total value, this method is primarily used to assess item discrimination. The strength of the relationship between a dichotomous nominal scale and an interval scale is assessed statistically [12]. The item discrimination in the current study was broken down into four categories: Poor, Fair, Good, and Very Good. The subpar products have been removed, and the Fair ones need to be improved too Good or Very Good. They were after that kept in the test bank [21].

Table 2

The Category of Item Discrimination

Item Discrimination	Category
0.40-1.00	Very good
0.30-0.39	Good
0.20-0.29	Fair
0.00-0.19	Poor
Negative r_{pbis}	Low-performing students got the correct answers more than high-performing students

Distractor Analysis

Distractors are considered effective if respondents choose them for at least 5% of responses (0.050), while they are considered ineffectual if respondents choose them for less than 5% of responses [21]. There needs to be an update to the ineffective detractors. New distractions that are more appealing and difficult to choose from should take their place.

Quest

The Quest application is simple to set up on any laptop or computer. Inputting commands into the notepad program, entering student responses into the notepad program, and processing the data on the Quest software are the three basic components of conducting item analysis using the Quest program. All of those files must be kept in a single folder. There are a few steps that must be taken in order to do item analysis using the Quest application. [33].

With the Itanal command on the syntax, the Quest software can carry out classical analysis. Information on item statistics and test statistics is included in classic files. Item statistics represent the attributes of items, such as their degree of difficulty, their capacity for discrimination, and how well they work as distractions. The % figure, which displays the percentage of pupils in each choice, is used to determine the difficulty level. The criteria for the item difficulty level are based on the percentage of the correct response. The discriminating power of the questions as determined by biserial correlation points (r_{pbis}) is the second statistic.

Item Analysis According to Item Response Theory

PROX (normal approximation estimation) is the method used to estimate items and responses. Accurate measurement will arise from a match between the respondent's aptitude and the difficulty index of item (b). When P = 0.5, accuracy is at its highest. All true and false responses are disregarded while performing parameter estimation. Because they are still unknown, respondent and item parameters are estimated simultaneously. Up till the respondents and item parameters remain consistent, the estimation is carried out.

The Rasch model appropriateness of the item as well as the item difficulty index define the quality of the item. The requirements of the item response theory must be met by a good item. Items in this study were evaluated for appropriateness using the infit mean-square value.

Table 3

Fit Item Criteria with Rasch model

Infit Mean square	Judgment
> 1.33	Mismatch
0.77-1.33	Match
< 0.77	Mismatch

The second step is to evaluate the value of the clothing items using the following standards.

Table 4

Criteria for Accepted and Rejected Item

Criteria	Judgment
Outfit t < 2.00	Accepted
Outfit t > 2.00	Rejected
< 0.77	Mismatch

The value of Delta or Threshold (b) is examined in the third step using the following standards.

Table 5

Threshold Category

Threshold	Category
b > 2	Very Difficult
1 < b < 2	Difficult
-1 < b < 1	Moderate
-1 > b > -2	Easy
b < -2	Very Easy

Research results

The Characteristics of English Achievement Test before COVID-19

The head of the English teacher forum in Sub-Rayon 01 revealed during interviews that they first created a blueprint in accordance with the 2013 Curriculum syllabus. Following that, the blueprint was given out to participants in the East Semarang Sub-Rayon 01 English Teacher Forum. All forum participants had access to each indicator from the blueprint, and they were instructed to create questions based on each indicator. They were then given a deadline by which to submit their inquiries to the forum's leader, an English teacher. Before being cross-checked with other members of the English teacher forum, all items were assembled into a single test. When something went wrong, they informed the forum's administrator and worked to repair it collectively. All of the schools in sub-rayon 01 received the test when it was determined that the test items were accurate. Based on the results of the interview, it was decided not to administer this test to students first and to analyze each

item's difficulty, discrimination, and efficacy as a distraction. Additionally, the test's validity and reliability were not examined as a whole.

The item with the lowest item difficulty index is item number 2, while the item with the greatest item difficulty index is item number 10. The English achievement test created by the English teacher forum in sub rayon 01 East Semarang prior to the pandemic's results indicate that item number two is the most challenging. Item number 10 is the test's simplest question. Table 6 below displays the outcomes of the English accomplishment test's item difficulty test based on category.

Table 6

The Item Difficulty Result of the English Achievement Test

Category	Item	Total	Percentage
Easy (0.71 – 1.00)	1,4,5,6,7,9,10,11,12,13,14,15,17,18,24,27,28,30,31,33,39,40	22	55
Moderate (0.31- 0.70)	3,16,19,20,21,22,23,25,29,32,34,35,36,37,38	15	37.5
Difficult (0.00-0.300)	2,8,26	3	7.5
Total		40	100

Based on Table 6, it is determined that 22 items, or 22/40 x 100% = 55%, fall into the Easy group. The percentage for the 15 items in the Moderate group is 15/40 x 100%, or 37.5%. Three things are then included in the Difficult category. This category's item difficulty percentage is 7.5% (or 3/40 x 100%). The Easy category dominated the exam with a dominance of 55%, followed by the Difficult category with a dominance of 7.5% based on the percentage of item difficulty for each category.

The item number 38 has the lowest item discrimination index (0.01), whereas the item number 19 has the greatest item discrimination index (0.52). Table 7 below displays the results of the item difficulty test depending on category.

Table 7

The English Achievement Test's Item Discrimination Findings Before the Pandemic

Category	Item	Total	Percentage
Poor (Pt. Biser < 0.19)	9,27,31,38	4	10
Fair (0.20-0.29)	1,2,5,6,8,10,11,15,23,25,30,37	12	30
Good (0.30-0.39)	3,7,13,17,20,33,35,36,39,40	10	25
Very Good (0.40 < Pt. Biser)	4,12,14,16,18,19,21,22,24,26,28,29,32,34	14	35
Total		40	100

Four items fall into the Poor category in terms of item discrimination, according to Table 7 below. For items in this Poor category, the item discrimination percentage is 4/40 x 100%, or 10%. The Fair category has up to 12 items, with a percentage of 12/40 x 100% = 30%. 10 things are then classified as Good, with a percentage of 10/40 times 100% equaling 25%. Last but not least, there are 14 products in the Very Good category. For items in this category, the item discrimination percentage is 14/40 x 100%, or 35%. Based on the aforementioned item discrimination percentages, it can be deduced that the Very Good category (35%) and the Poor category (10%) are the two categories with the highest and lowest respective item discrimination percentages.

There are 64 effective and 56 ineffective distractions on the English achievement test created before COVID-19. There are 12 things with useful distractions. The test's ineffective distractor proportion is 56/120 times 100%, or 46.70%. The test's effective distractor percentage is 64/120 x 100%, or 53.30 percent. Before the epidemic, the English Teacher Forum conducted an English accomplishment test in East Semarang's Sub-Rayon 01 with a 0.990 reliability rating.

The COVID-19 Characteristics of English Achievement Test

No stages were used in the development of the COVID-19 English accomplishment test. A instructor of English created it. She made the test without using a blueprint, according to the researcher's conversation with her. She created a Google Form to modify the material she taught in class for a particular time period as a result. She merely copied and pasted answers from the earlier test that she and other English teachers on the site had created. She also skipped the opportunity to use it to evaluate test characteristics including item difficulty, item discrimination, and distractions. Overall, the validity and reliability of the test were not examined.

The lowest item difficulty index is 0.13 for item number 2 and the highest item difficulty index is 0.96 for item number 10. Based on the indexes, it is concluded that the most difficult item of the English achievement test made by an English teacher during COVID-19 is item number 2, while the easiest item of the test is item number 10. The result of the item difficulty test is presented in Table 8.

Table 8

Result of Item Difficulty Test on the English Achievement Test

Category	Item	Total	Percentage
Easy (0.71 – 1.00)	1,12,27,31,35	5	10
Moderate (0.31- 0.70)	2,3,4,5,6,7,8,9,10,11,13,14,15,16,17,18,19,20,21,22,23,24,25,26,28,29,30,32,33,34,36,37,38,40,41,43,44,46,47,48,50	42	84
Difficult (0.00-0.300)	42,45,49	3	6
Total		50	100

According to Table 8, five items fall into the easy category, and their percentage is 5/50 x 100%, or 10%. The moderate category includes 42 items, with a proportion of 42/50 x 100% = 84%. Three items, with a proportion of 3/50 x 100% = 6%, are in the difficult group. According to the percentage of difficult items in each category, it can be deduced that the moderate category (84%) is the most prevalent, while the tough category (6%), is the least prevalent.

Item 6 has the lowest item discrimination index (0.14), whereas item 19 has the greatest item discrimination value (item 20). Table 9 displays the outcomes of the item discrimination test. Based on the item discrimination of the test, it is demonstrated that there is one item that falls under the poor group. Items in the poor category have an item discrimination rate of 1/50 x 100%, or 2%. In the fair category, there are two things. This category makes up 4% of the total, or 2/50 times 100%. Additionally, there are seven entries in the good category.

Table 9

The English Achievement Test Developed During the Pandemic: Item Discrimination Results

Category	Item	Total	Percentage
Poor Discrimination (Pt. Biser < 0.19)	6	1	2
Fair (0.20-0.29)	42,49	2	4
Good (0.30-0.39)	1,30,33,34,46,47,50	7	14
Very Good (0.40 < Pt. Biser)	2,3,4,5,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23, 24,25,26,27,28,29,31,32,35,36,37,38,40,41,43,44,45,48	40	80
Total		50	100

For items in the good category, the item discrimination percentage is 7/40 x 100%, or 14%. The category of "very good" has 40 items. For items in the very good category, the item discrimination percentage is 40/50 x 100%, or 80%. Based on the aforementioned item discrimination percentage, it can be deduced that the very good category (80%) and bad category (2%), respectively, are the most and least prevalent categories of item discrimination in this exam.

The English achievement exam created during the pandemic featured 150 useful distractions compared to one ineffective one. There are 49 items that can be effectively distracted. The test's ineffective distractor percentage is 1/150x 100%, or 0.70 percent. The test's effective distractor percentage is 149/150 multiplied by 100%, or 99.30%. The English achievement test created by an English teacher for COVID-19 has a 0.960 reliability rating. These numbers can be seen in the output file for the quest's Summary of Item Estimates.

Tests Developed Before and During the COVID-19 Pandemic: Comparisons

The researchers discovered variances and similarities between the English Achievement exams created before COVID-19 and after COVID-19 after gathering data from both tests.

Table 10

The Variations in Tests Conducted Prior to and During the COVID-19 Pandemic

Test Characteristics	Category	English Achievement Test Made **before** the pandemic	English Achievement Test Made **during** the pandemic
Item Difficulty	Easy	22 (55%)	5 (10%)
	Moderate	15 (37.5%)	42 (84%)
	Difficult	3 (7.5%)	3 (6%)
Item Discrimination	Poor	4 (10%)	1 (2%)
	Fair	12 (30%)	2 (4%)
	Good	10 (25%)	7 (14%)
	Very Good	14 (35%)	40 (80%)
Distractors	Effective distractors	64 (53.30%)	149 (99.30%)
	Ineffective distractors	56 (46.70%)	1 (0.70%)
Reliability	Reliable	0.990	0.960

The procedure of Designing Test	It was made based on the blueprint. It was not tried out. It was not analyzed. It was cross-checked with other members of the English teacher forum. The English teacher forum made it by themselves.	It was made without referring to any blueprint. It was not tried out. It was not analyzed. It was not cross-checked with other English teachers. The English teacher copied and pasted from the previous test made English teacher forum and teachers including herself.

Discussion

The English Achievement Test's Pre-COVID-19 Characteristics

Based on CTT, which emphasizes item complexity, item discrimination, distractors, and dependability [21], the characteristics of the English achievement exam created prior to the pandemic are recognized. According to the item difficulty test, there are 22 easy items, which account for 55% of the total, 15 moderate items, which account for 37.5% of the total, and 3 difficult items, which account for 7.5% of the total. The exam does not have proportional item difficulty because the difficult items are more prevalent than the easy ones based on the results. Test questions should ideally be divided into three difficulty levels: 25% easy, 50% moderate, and 25% challenging [34]. Tests without items of proportional complexity cannot reflect pupils' true talents, claim Roid & Haladyna [35].

Most of the test's questions are simple. According to Brown[12], a well-made item shouldn't be too easy or challenging, and the percentage of each item difficulty category needs to be balanced in order to fully reflect students' talents or scores [16]. According to Djiwandono [36], a test item is ineffective if it can be answered correctly by every test taker or if it cannot be answered by every test taker. A test with lots of simple questions, in
S. Suwarto's opinion [21], is used to evaluate pupils who perform below average. Students who have a mid-level of achievement will take a test with numerous items of moderate complexity. High-achieving pupils will be put to the test on an exam with a lot of challenging questions.

According to those definitions, this test does not fairly represent the talents of all pupils. Madsen [37] further supports the idea that researchers categorize subjects into simple and tough based on the proportion of students who correctly respond to each question. The results of the item difficulty test can be compared to other research, such as item analysis [38] and validity analysis [17] for the English summative exam [39], as well as English summative tests [40]. Despite the fact that the test settings are different, earlier research discovered that the distribution of item difficulty amongst simple, moderate, and tough items is uneven. Cognitive abilities including comprehension, coding, transition, observation, and working memory might have an impact on an item's difficulty. These mental elements may have an impact on students' performance.

According to the Quest program, there are four poor products with a discrimination proportion of 10%, 12 fair items with a 30% discrimination percentage, 10 good items with a 25% discrimination percentage, and 14 very good items with a 35% discrimination percentage. According to these findings, 12 fair things should be updated, whereas 4 poor items should be rejected [32]. The outcome is favorable because 25% of the things are good and predominate, and 35% of them fall into the very good category [32]. This indicates that the majority of the items can be included to the test bank and used to assess students' actual

English proficiency. These factors can reveal information regarding the distinctions between high, mid, and low achievers. This is consistent with [21], which claims that a higher item discrimination score suggests that the item can identify differences between students who have high achievement and those who have low achievement. Although the test settings are somewhat different, the results of this item discrimination test are comparable to those of other studies that have looked at item test characteristics [41], multiple choice questions [42], and education research [43]. They discovered effective item discrimination. According to students who reported that the item discrimination was poor and that they were unable to differentiate between the upper group and lower group after reading [44] item analysis [45] and taking a multiple-choice exam [46] in the meantime, different findings had been found from earlier studies.

Third, the exams contain 64 effective distractors (53.3%) of the 120 distractors and 56 ineffective distractors (46.7%), which should be altered. This study's percentage of effective distractions is nearly identical to that of the Rehman et al. study from 2018. Out of the 120 distractors, they found 31.07% to be useful. On the other hand, [46] discovered more ineffective distractions that no students chose to use during the test. Therefore, the useless distraction was either too simple or unimportant. The claim that all multiple-choice items are not always created to satisfy the testing objectives in terms of giving students with four or more choices is supported by all of the ineffective refutators. The majority of the English achievement test items created prior to COVID-19 can identify high and low performers. Therefore, it can be inferred that effective distractors are produced by large index item discrimination [47]. They added that at least three distractions are recommended for each item. The findings of this study demonstrate that both tests have more potent deterrents, which raises the quality of the things.

The test reliability index is 0.990 from a reliability perspective. It shows how highly trustworthy the test items are. A good test is one that has a high level of reliability [48]. A good test can also be applied to later time testing. The findings of this study also demonstrate how well the English accomplishment test measurement made prior to COVID-19 holds up over time and under identical testing circumstances [15]. Although the test settings are different, this dependable test is nearly identical to earlier studies' reliability tests of 0.651 [49] and 0.631 [50]. Because its value is below the reliability coefficient limit of 0.700, test reliability estimation can be trusted. Group homogeneity, allotted time, and test length are a few variables that affect dependability estimation. Additionally, the proportion of difficult items has an impact on how reliable it is estimated to be [13].

The study of the test item is included in the quantitative analysis of the English proficiency exam that was created prior to the pandemic. There are 12 test items that need to be altered (30%) and 24 acceptable (60%) test items. Four test items were, however, disqualified (10%). The following is a summary of the test items' analysis.

Table 11

English Achievement Test Items Developed Before COVID-19: Analysis

Criteria	Test Items	Total (%)	Percentage
Accepted	3,4,7,12,13,14,16,17,18,19,20,21,22,24,26,28,29,32,33,34,35,36,39,40	24 (60%)	2
Revised	1,2,5,6,8,10,11,15, 23,25,30,37	12 (30%)	4
Rejected	9,27,31,38	4 (10%)	14

24 accepted items have an index between 0.30 and 1.00, according to table 11. These articles were accepted without modification, according to [32]. They fall under the very good and good categories. The remaining 12 items have an index of 0.20 to 0.29. These items are accepted with amendment, according to [32]. They fall under the very good and good categories. Finally, four items that were rejected have indexes below 0.20. These goods should be excluded since they fall under the poor group, as shown by [33]. This outcome is consistent with [51].

The English Achievement Test for COVID-19's Characteristics

The criteria of item difficulty, item discrimination, distractors, and reliability were used to identify the properties of the English accomplishment test produced during COVID-19. The test has five easy things with a percentage of 10%, 42 moderate items with a percentage of 84%, and three difficult items with a percentage of 6%. According to the results, the exam has more moderate difficulty items than easy items, leading the researchers to draw the incorrect conclusion that the test's item difficulty is not proportional [21]. An ideal test would have 25% easy questions, 50% moderate questions, and 25% difficult questions. According to Roid and Haladyna [35], a test that lacks proportional item difficulty cannot reflect students' true proficiency. Additional moderate-level questions are included in the test. According to Brown [12], something that is well-made cannot be overly simple or challenging. The test needs to be fair so that teachers can learn about the pupils' level of proficiency.

In contrast, the moderate item category, where more than half of the students responded correctly, can suggest that students have a solid grasp of the content being tested. The item difficulty test's findings are comparable to earlier research looking at the level of difficulty for summative tests [51], analysis challenges [41], and development tests [18] under various circumstances. According to earlier research, some item categories had more products with a moderate level of difficulty than others. It suggests that the examinations have more carefully crafted questions than poorly crafted questions, but the ratio of easy, moderate, and tough questions is unbalanced. The COVID-19 epidemic, which required kids to work from home so they could ask their friends for the answers and conduct online searches for the answers, is likely to blame for the difficulty of imbalanced items. The item difficulty index may be impacted by these circumstances. Additionally, the students' responses are impacted by the question instructions. When a question is unclear, it is anticipated that pupils will give a false response. Additionally, this will impact the item difficulty index.

Second, this test has good item discrimination. According to the Quest program, there are 40 very good items with a percentage of 80%, seven good items with a percentage of 14%, one poor item with a percentage of 2%, two fair items with a percentage of 4%, and seven fair items. According to these findings, 2 fair items should be updated, whereas 1 subpar item should be rejected [32]. The fact that 80% of the goods fall into the very good category and 14% fall into the good category makes this outcome in some ways positive [32]. The majority of the items can therefore be kept in the question bank and utilized to assess students' actual English proficiency. These tools can also gather data on the distinction between high performers, mid achievers, and poor achievers. According to S. Suwarto [21], the higher item discrimination score suggests that the item makes a distinction between students' high accomplishment and low achievement. Item discrimination index is capable of identifying differences between students. This test's outcomes are comparable to those of other investigations. Although the test settings are different, the test was shown to have an

80% discriminating power [43], good discrimination [41], and a discrimination index of 50% great items. Researchers discovered good item discrimination in earlier trials. In contrast, according to other investigations, the test's discrimination power was weak in 67.5% of cases [45]. The [48] test had a modest item discrimination, however [46] claimed that the item was subpar, thus the items couldn't tell the high achievers from the low achievers.

Third, only one ineffective distractor (0.7%) was produced out of the 150 total distractors used in this investigation. This study's percentage of effective distractions is nearly identical to that of the earlier study by [51], which discovered 80% of effective distractions. Because of this circumstance, the distractor indexes may be impacted by the item discrimination indexes. Because tests with a high item discrimination index have an effective distractors index, the majority of the test items created during COVID-19 can distinguish between high and low achievers [47]. Furthermore, each item has a maximum of three distractors. The study's findings demonstrate that there are more potent distractions, demonstrating the high caliber of the objects. The findings of studies on the English proficiency test created during the pandemic demonstrate that nearly every item has useful detractors. It is therefore presumed that the English teacher who created the test is quite familiar with the traits of the students. Because she works at one of Semarang's most well-liked schools, the teacher is competent.

The reliability value is 0.996 as well. It shows how highly trustworthy the test items are. A good test is one that has a high level of reliability [48]. A good test can also be applied to later time testing. The findings of this study also show how well the English achievement test measurement during the epidemic holds up after being repeated on the topic and under the identical circumstances [15]. This reliable test is nearly identical to the tests created in earlier investigations, including the tests with reliability scores of 0.756 [18], 0.800 [41], and 0.907 [16]. They also created or studied trustworthy tests. Because it was higher than the reliability coefficient limit of 0.700, the test reliability estimation could be trusted. The English achievement exam, on the other hand, was shown to be unreliable in other earlier research [52] because the data analysis did not adhere to the standards of consistency and dependability. The exam could be used in a classroom to evaluate a student's proficiency in English, but it couldn't be used as a component of a bank exam. The test was variable, so it could be used to a changing circumstance to evaluate a student's performance on a midterm or final exam. The fact that the data analysis's findings were imbalanced was another factor in why it wasn't constant. The researchers opted to stop this research because it was time-constrained and another factor.

The analysis of the test item was included in the quantitative analysis of the English achievement exam that was created prior to the pandemic. 47 things have been accepted (94%) and 2 have been altered (4%); one item has been denied (2%). The analysis of the test items' executive summary is shown below.

Table 12

Results of the Analysis of Pandemic-Era English Achievement Test Items

Criteria	Test Items	Total (%)	Percentage
Accepted	1,2,3,4,5,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28, 29,30,31,32,33,34,35,36,37,38,39,40,41,43,44,45,46,47,48,50	47 (94%)	2
Revised	42,49	2 (4%)	4
Rejected	6	1 (2%)	14

47 accepted items had indexes more than 0.30, according to Table 12. These submissions are approved as-is. Next, two items are included with indexes that fall between 0.20 and
0.29. These are acceptable suggestions with changes. The last rejected item had an index of less than 0.09. This thing is removed [32].

The characteristics of English achievement tests conducted before and during the pandemic were compared

The process of creating the test, which involves the stages of blueprint generation and cross-checking, is what sets it apart from the tests made before during the pandemic. The English teacher followed a plan when creating the assessments. Additionally, they double-checked each answer with members of the English teacher community. However, because the English teacher created the assessments using Google Form right away, they lacked a template during the pandemic. It can be said that the teacher forum's method of creating tests is more thorough than the method employed by English teachers.

However, the two exams are comparable in that they do not adhere to the rules for creating good questions. When creating examinations to accurately assess students' English proficiency, teachers should adhere to the proper approach. Making a good exam involves multiple steps, including creating indicators that correspond to the syllabus's fundamental competencies and allocating items to each indication, according to [27]. The questions were created using a blueprint and then made available for testing. The reliability, item discrimination, item difficulty, and distractor effectiveness of the trials were then assessed. As a result, the test's quality may be determined by its creators. Additionally, some components could be changed or removed. The entire process of creating tests is expensive and time-consuming.

Boopathiraj & Chellamani [43] assert that test preparation should include test design, test execution, and results management. The test creators can be directed by the instructional objectives or evaluation objectives to be tested when choosing which types of learning outcomes or degrees of thinking ability to be assessed. A blueprint should be created before any materials are prepared since it outlines the criteria for the objectives to be evaluated, the scope of the content, and the questions to be utilized.

The level of item difficulty differs across tests created before and during the pandemic. The English accomplishment exam created during the epidemic had a more evenly distributed item difficulty distribution (closer to the normal distribution) than the test created prior to the outbreak. The pandemic-era English accomplishment test has 5 easy items (ten percent), 42 intermediate items (84%), and 3 difficult items (six percent). The English proficiency test that was created prior to the pandemic had 22 easy items (55%), 15 moderate items (37.5%), and 3 difficult items (7.5%). There are too many moderate things on the test that was created before to the pandemic, making it unbalanced (dominant). The amount of challenging items on the two tests is where they are similar (3 items). [19] asserts that a good test has three different item difficulty categories: easy, moderate, and difficult, with easy items falling into the 25% category. The test can identify kids with high, mid, or low accomplishment levels based on the balance of tough items. If the English proficiency of every student is accurately assessed, the instructor may determine which subject has not been fully grasped by the students and can improve it in the subsequent learning process utilizing more effective learning resources, teaching strategies, and teaching methodologies.

The distinction between the tests created before and during the epidemic is examined in terms of the discrimination index. There are 40 good items on the test that the English teacher created during the pandemic. The test conducted prior to the pandemic showed 14 positive results, though. This demonstrates that English teachers' creations are more widely appreciated [53]. Additionally, there are a few items that the English teacher only needs to slightly alter. The results produced by the English instructor are nearly identical to the findings of [51]'s analysis of the test items from the English teacher's final semester test. It has been discovered that very nice items predominate (97.5%). The commonality between the two tests, on the other hand, is that both feature a few items that must be eliminated since they fall under the poor category. These must be removed, and new inquiries must be added in their stead [54]. As a result, both the test created by the English teacher forum and the test created by the English teacher must exclude four weak items.

The test created during the pandemic contains 149 useful distractor functions, while the test created during the pandemic has 64 useful distractor functions. This demonstrates that the English instructor who created the exam during the epidemic had a greater understanding of the traits of pupils as seen in the test results for each chapter. As a result, the English teacher's blinded distractors are more effective than those in the exam that the English teacher forum designed. The English Teacher Forum's examination of the distractions is consistent with Sugiarti's study, which looked at the distractions on English multiple-choice tests given to eighth graders. On the test, she discovered a lot of useless distractions (82.5%). In terms of effective distractors, the two tests are very similar in that they both have an identical amount of dominance.

The English proficiency exam created before and during the pandemic had a reliability score of 0.990 and 0.960, respectively. According to [15], if the dependability index is more than 0.700, the test is considered reliable. The reliability index distinguishes the two of them. The English Teacher Forum test's reliability index (0.990) is greater than the English Teacher Forum test's (0.990). (0.960). The English teacher did not follow the proper procedure for creating the test, and the test was created by copying and pasting from the previous tests created in 2016, 2017, and 2018 by the English teacher forum, among other factors that contributed to the test's lower reliability value before the pandemic.

Because the features of the items depend on the group of test-takers who are exposed to them, the analysis based on CTT has a flaw. The statistics for questions in the CTT, such as the difficulty index of the questions, are dependent on the test-takers' demographics. When brilliant students take the test, the questions are regarded as easy (the level of difficulty of the questions increases), and when less intelligent students take the test, the questions are regarded as challenging (the level of difficulty gets lower). Therefore, depending on the exam-takers' skill levels, the question qualities can vary or even change.

The IRT measurement is demonstrated to eliminate the distinction between the test-taker group and the test-item group, thus resolving the CTT measurement issue. Despite the fact that test taker characteristics vary, IRT measurement essentially dictates the features of the items. In other words, despite the fact that test takers' responses varied, the item group's properties remained constant. It follows that even though they choose to respond to various test items, the participants' traits will remain constant. The primary distinction between IRT measurements and CTT measurements is that the IRT score is invariant (unchanged) to both the test item and the test taker [55].

Table 13

Test Threshold Category Developed Prior to COVID-19

Category	Items	Total	Percentage
Very difficult	2,8,26	3	7.5%
Difficult	16,20,23,25,34,35,36,37,38,	9	22.5%
Moderate	3,4,5,11,12,15,18,19,21,22,27,29,32,33	14	35%
Easy	1,6,14,17,24,28,30,31,39,40	10	25%
Very easy	7,9,10,13	4	10%
Total		40	100%

According to Table 13, the threshold percentage for the English proficiency test prior to COVID-19 is 7.5%, 22.5%, 35.5%, 25%, and 10%, respectively.

Table 14

Test Threshold Category Created During COVID-19

Category	Items	Total	Percentage
Very difficult	-	0 (0%)	7.5%
Difficult	6,21,42,45,48,49	6 (12%)	22.5%
Moderate	2,3,4,5,7,8,9,10,11,13,14,15,16,17,18,19,20,22,23,24,25, 26,28,29,30,32,33,34,36,37,38,39,40,41,43,44,46,47,50	39 (78%)	35%
Easy	1,12,27,31,35	5 (10%)	25%
Very easy	-	0 (0%)	10%
Total		50 (100%)	100%

According to Table 14, the English achievement test threshold percentages for the extremely difficult, difficult, moderate, easy, and very easy categories are 0%: 12%: 78%: 10%: 0%. The percentages of the English accomplishment test developed during the pandemic are therefore more balanced than the percentage of the English achievement test developed prior to the pandemic, according to the two tables above. Additionally, the English performance exam levels created prior to the pandemic primarily contain questions with a moderate level of difficulty.

Table 15

The Evaluation of Items Accepted and Rejected Before COVID-19

Category (Criteria)	Test Items	Total (%)	Percentage
Accepted (Outfit t < 2.00)	1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22, 23,24,26,28,29,30,31,32,33,34,35,36,37,39,40	37 (92.5%)	7.5%
Rejected (Outfit t > 2.00)	25,27,38	3 (7.5%)	22.5%
Total		40 (100%)	100%

According to Table 15, there were 92.5% accepted items and 7.5% rejected items that were created before the epidemic. The percentage of test items generated during the pandemic that have been accepted is 80%, while the percentage of test items that have been refused is 20%, according to Table 16. These tables can be used to draw the conclusion

that a bigger percentage of acceptable things are created before the pandemic than are accepted items created during the pandemic. The test created before the pandemic had superior qualities than the test created after the pandemic, according to the number of acceptable and rejected items.

Table 16

The Evaluation of COVID-19's Accepted and Rejected Items

Category (Criteria)	Test Items	Total (%)	Percentage
Accepted (Outfit t < 2.00)	1,2,3,4,5,7,8,9,10,11,12,13,14,16,17,18,19,21,22,23,25,26,27,28,29,30,31,32,33,35,36,37,38,39,40,41,43,44,45,48	40 (80%)	7.5%
Rejected (Outfit t > 2.00)	6,15,20,24,34,42,46,47,49,50	10 (20%)	22.5%
Total		50 (100%)	100%

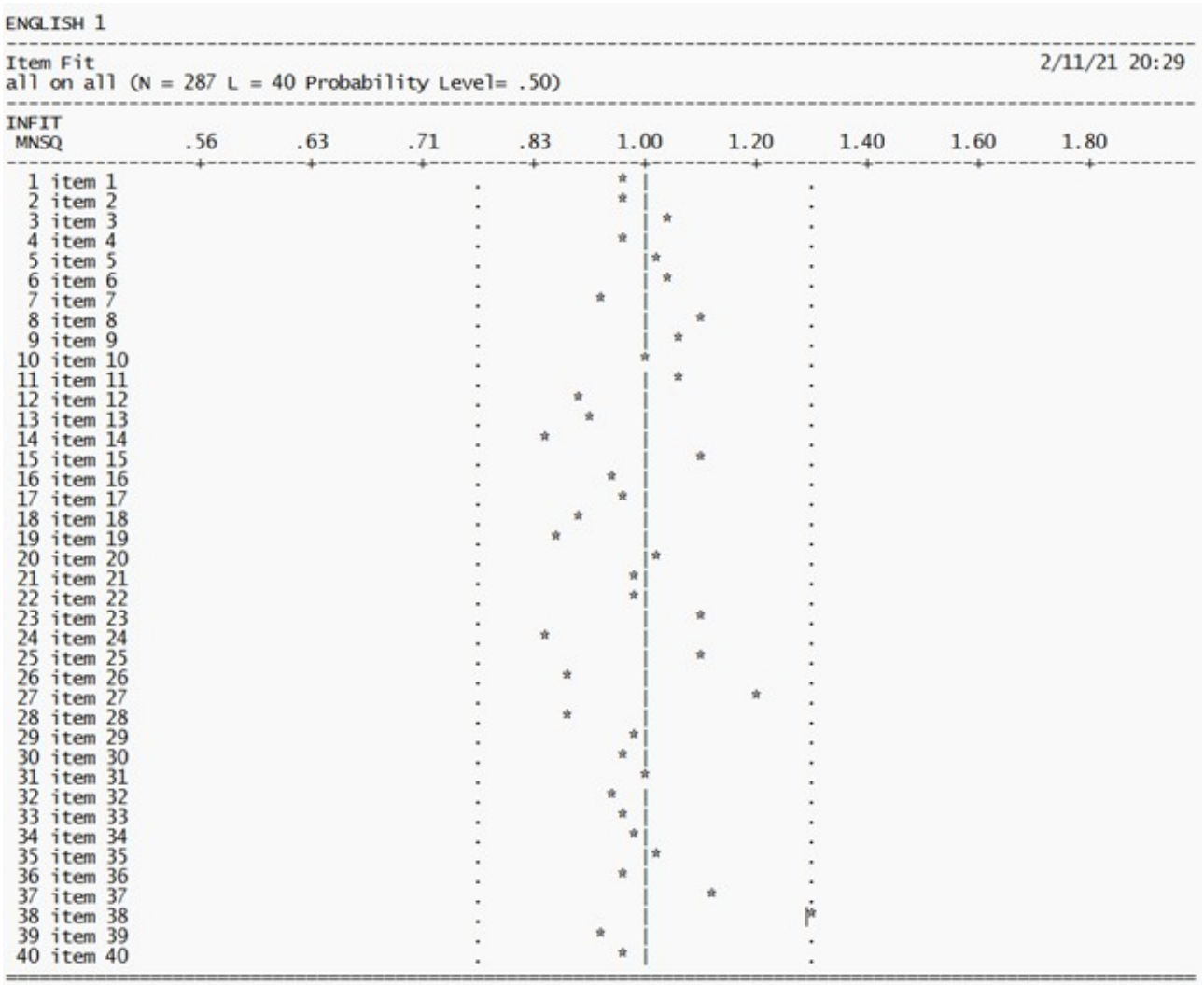

Figure 1 Item Fit Map for the English Achievement Test Developed before COVID-19

The asterisks are between two dotted vertical lines, as can be seen in Figure 1, and there are 40 fit items of the English achievement exam created prior to the pandemic [30]. It shows that all test items created prior to the pandemic (100%) are compatible with the Rasch Model (one-parameter logistic model) with an acceptability range of > 0.77 to 1.30 [33]. Then, according to Figure 2, eight items of the English proficiency test created during the pandemic are not regarded as fit since the asterisks are outside of fit statistics that are

inside the two dotted vertical lines, although there are 42 fit items [30]. The proportion of goods that fit is 42/50 x 100%, or 84%. Based on those two numbers, it can be said that the English teacher forum's characteristics of the English achievement test developed prior to COVID-19 were superior to those of the English teacher's characteristics of the English achievement test developed during COVID-19.

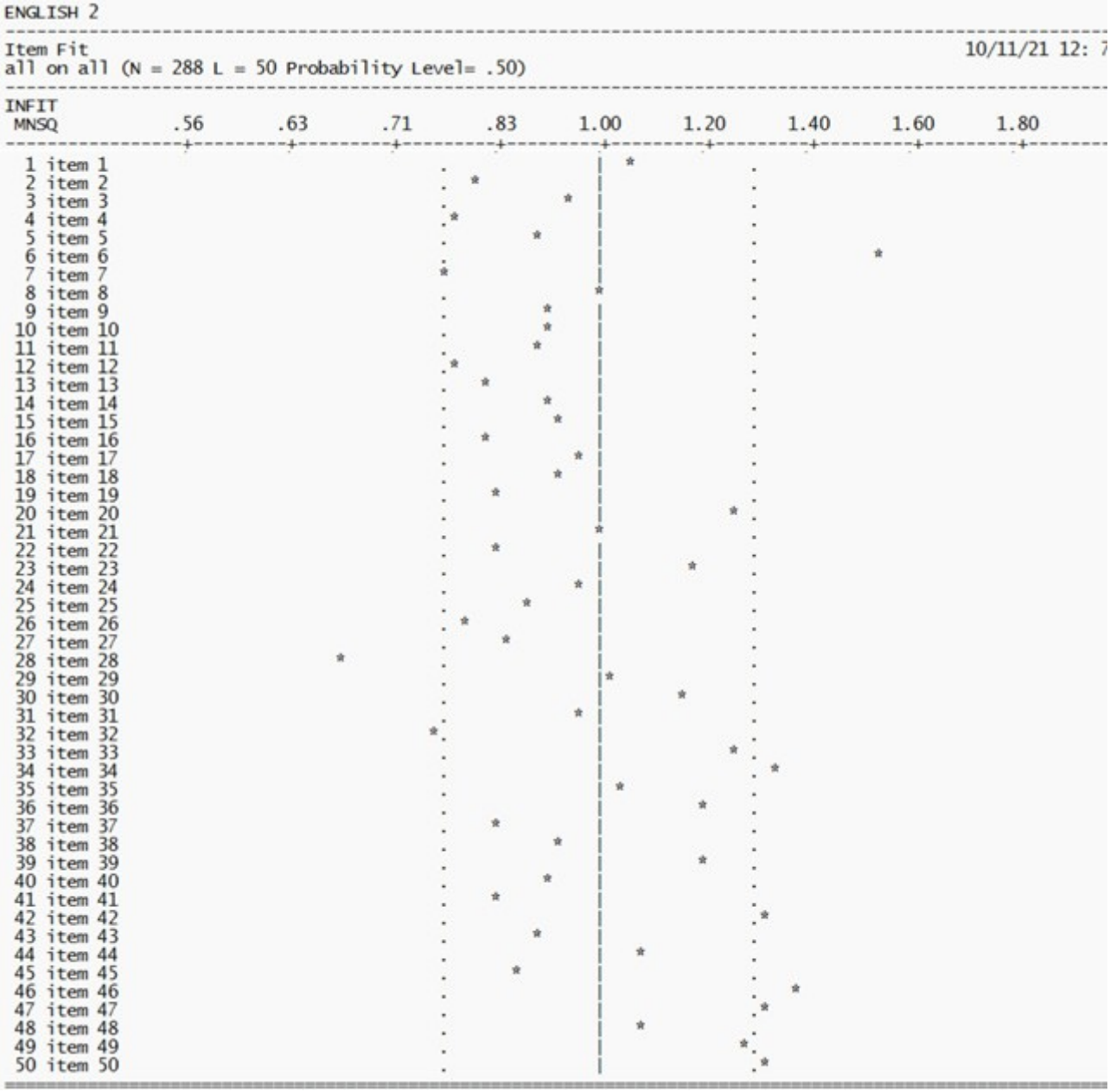

Figure 2 English Achievement Test item fit map created during COVID-19

Conclusion

The characteristics of the English accomplishment test items created before COVID-19 and during COVID-19 for eighth-grade students at SMPN 2 Semarang were elaborated by the researchers based on the research findings and discussions. The properties of the test that was designed are further detailed in light of CTT and IRT. First, there are 22 easy items (55%) 15 intermediate items (37.5%), and 3 difficult items (7.5%) based on item difficulty. Second,

there are 4 subpar things (10%), 12 fair items (30%), 10 good items (25%), and 14 very good items (35%), in terms of item discrimination. Third, in terms of distractor performance, there are 64 effective (53.30%) and 56 ineffective (46.70%) distractions. Finally, the exam is regarded as trustworthy.

Then, based on the same factors, the characteristics of the English accomplishment test created during COVID-19 are elaborated. First, there are 5 easy items (10%), 42 moderate items (84%), and 3 difficult items (6%), according to item difficulty. Second, there is one poor item (2%), two fair things (4%), seven good products (14%), and forty very good items (80%) in terms of item discrimination. Third, 149 effective distractors (99.30%) and one unsuccessful distractor (0.70%) were found using distractor analysis. Finally, the exam is regarded as trustworthy.

It is discovered that the test created before to the epidemic had a more evenly distributed item difficulty. The test maker created 40 extremely good things during the pandemic, compared to just 14 in the test created prior to the outbreak, according to item discrimination. The test created by the English teacher has 149 items, more functional distractions, and 64 distractions created by the English teacher forum. The test created by the English teacher forum has a higher reliability rating than the test created by an English instructor. The test created before to COVID-19 is legitimate based on content validity, however the test created during COVID-19 is invalid. The test created before to COVID-19 was correctly created using the blueprint that had been produced, however the test created during COVID-19 was not created using any blueprint.

1 The threshold percentages for the English achievement test created prior to COVID-19, based on IRT, are 7.5%, 22.5%, 35%, 25%, and 10% for categories that are very tough, difficult, moderately difficult, easy, and very easy. The English achievement test cutoff percentages for the very tough, difficult, moderate, and easy categories are 0%, 12%, 78%, 10%, and 0%, respectively. For accepted and rejected categories, the percentages of test items created prior to COVID-19 are 92.5% and 7.5%, respectively. The percentages of test items created during COVID-19 for accepted and rejected categories are 80% and 20%, respectively. The English achievement exam created prior to COVID-19 has 40 fit items with a percentage of 100% and is based on the Rush Model or the one-parameter logistic model. The English achievement test that was prepared for COVID-19 has 42 fit items (84%) and 8 unfit items (16%). Based on the percentages of approved and rejected test items based on the Rush Model, or one-parameter logistic model, the characteristics of the English achievement test developed prior to COVID-19 are superior to those of the English achievement test developed during COVID-19. Both tests mostly feature a Moderate level of difficulty when it comes to threshold.

REFERENCES

1. Pizmony-Levy O. Big Comparisons, Little Knowledge: Public Engagement with PISA in the United States and Israel. *The Impact of the OECD on Education Worldwide*, vol. 31, Emerald Publishing Limited, 2017, pp. 125–156. doi: 10.1108/S1479-367920160000031008.
2. Froese-Germain B. The OECD, PISA and the Impacts on Educational Policy VIRTUAL RESEARCH CENTRE (VRC). 2010.
3. Shamatov D., Sainazarov K. The impact of standardized testing on education quality in Kyrgyzstan: The case of the Program for International Student Assessment (PISA) 2006. *International Perspectives on Education and Society*, vol. 13, pp. 145–179, 2010. doi: 10.1108/S1479-3679(2010)0000013009.
4. Luthfiyyah R., Aisyah A., Sulistyo G. H. Technology-enhanced formative assessment in higher education: A voice from Indonesian EFL teachers. *EduLite: Journal of English Education, Literature and Culture*, 2021, vol. 6, no. 1, pp. 42–54.

5. Gamage K. A. A., de Silva E. K., and Gunawardhana N. Online Delivery and Assessment during COVID-19: Safeguarding Academic Integrity. *Education Science*, 2020, vol. 10, no. 301, pp. 1–24. doi: 10.3390/educsci10110301.
6. Joshi A., Vinay M., Bhaskar P. Impact of coronavirus pandemic on the Indian education sector: perspectives of teachers on online teaching and assessments. *Interactive Technology and Smart Education*, 2020, vol. 18, no. 2, pp. 205–226. doi: 10.1108/ITSE-06-2020-0087/FULL/PDF.
7. Sultana N. Test review of the English public examination at the secondary level in Bangladesh. *Language Testing in Asia*, 2018, vol. 8, no. 1, pp. 1–9.
8. Adom D., Mensah J. A., Dake D. A. Test, Measurement, and Evaluation: Understanding and Use of the Concepts in Education. *International Journal of Evaluation and Research in Education*, 2022, vol. 9, no. 1, pp. 109–119.
9. Chen P. P., Bonner S. M. A framework for classroom assessment, learning, and self-regulation. *Assessment in Education: Principles, Policy & Practice*, 2019, vol. 27, no. 4, pp. 373–393. doi: 10.1080/0969594X.2019.1619515.
10. Luckin R., Cukurova M. Designing educational technologies in the age of AI: A learning sciences-driven approach. *British Journal of Educational Technology*, 2019, vol. 50, no. 6, pp. 2824–2838. doi: 10.1111/BJET.12861.
11. Ndukwe I. G., Daniel B. K. Teaching analytics, value and tools for teacher data literacy: a systematic and tripartite approach. *Ndukwe and Daniel International Journal of Educational Technology in Higher Education*, vol. 17, p. 22, 2020, doi: 10.1186/s41239-020-00201-6.
12. Brown H. D. Language Assessment Principles and Classroom Practice, 2003.
13. Crocker J., Algina L. Introduction to Classical and Modern Test Theory. 1986. [Online]. Available: http://www.mich.gov/documents/mde/3_Classical_Test_Theory_293437_7.pdf
14. Etsey K. A. Assessing performance in schools: Issues and practice. *IFE PsychologIA: An International Journal*, 2005, vol. 13, no. 1, pp. 123–135.
15. Suwarto D. Pengembangan Tes Diagnostik Dalam Pembelajaran. Yogyakarta, Pustaka Pelajar, 2013.
16. Sugianto A. Validity and reliability of English summative test for senior high school. *Indonesian EFL Journal: Journal of ELT, Linguistics, and Literature*, 2017, vol. 3, no. 2, pp. 22–38.
17. Putri B. D. T. The validity analysis of English Summative test of junior high school. *Journal of Languages and Language Teaching*, 2018, vol. 5, no. 1, pp. 6–11.
18. Mulianah S., Hidayat, W. Pengembangan Tes Berbasis Komputer. *Kuriositas*, 2013, vol. 2, no. 6, pp. 27–43.
19. Suwarto S. Karakteristik Tes Biologi Kelas 7 Semester Gasal. *Jurnal Penelitian Humaniora*, 2016, vol. 17, no. 1, p. 1. doi: 10.23917/humaniora.v17i1.2346.
20. Kalmukov Y., Valova I., MladenovaT. Covid 19-A Major Cause of Digital Transformation in Education or Just an Evaluation Test. *TEM Journal*, 2020, vol. 9, no. 3, pp. 1163–1170. doi: 10.18421/TEM93-42.
21. Suwarto S. The Characteristics of Indonesia Second-semester Final Test for Eighth-grade Students. *Turkish Online Journal of Qualitative Inquiry (TOJQI)*, 2021, vol. 12, no. 9, pp. 356–370. Available: https://www.tojqi.net/index.php/journal/article/view/5499
22. Salwa A. The validity, reliability, level of difficulty and appropriateness of curriculum of the English test. *Diponegoro University,* 2012.
23. Setiawan A., Sunardi B.Gunarhadi.Teaching Language Proficiency: The Implementation of Virtual Multimedia-Based Learning for Indonesian Vocational High School. Journal of Hunan University Natural Sciences, 2021, vol. 48, no. 11, pp. 289–297. Available: http://jonuns.com/index.php/journal/article/view/865
24. Alqurashi E. Predicting student satisfaction and perceived learning within online learning environments. Distance Education, 2018, vol. 40, no. 1, pp. 133–148. doi: 10.1080/01587919.2018.1553562.
25. Bubb S., Jones M. A. Learning from the COVID-19 home-schooling experience: Listening to pupils, parents/carers and teachers. Improving Schools, 2020, vol. 23, no. 3, pp. 209–222. doi: 10.1177/1365480220958797
26. Shim T. E., Lee S. Y. College students' experience of emergency remote teaching due to COVID-19. *Child Youth Serv Rev*, 2020, vol. 119, p. 105578. doi: 10.1016/J.CHILDYOUTH.2020.105578.
27. Sumadi S. Pengembangan alat ukur psikologis. Yogykarta, Andi Offset, 2005.
28. Lababa J. Analisis Butir Soal dengan Teori Tes Klasik: Sebuah Pengantar. *Jurnal Pendidikan Islam Iqra*', 2018, vol. 5, pp. 29–37. doi: 10.30984/jpii.v2i2.538.
29. Kartowagiran B. Pengantar teori tes klasik (ttk)*). *Pengantar teori tes klasik*, 2009, no. April, pp. 1–19.
30. Adam R. J., Khoo S.-T. Acer Quest: The Interactive Test Analysis System. *Australian Council for Educational Research*. 196, pp. 1–96.
31. Suwarto. Statistik Pendidikan. Yogyakarta, Pustaka Pelajar, 2018.
32. Dichoso A. A., Joy M. R. J. Test item analyzer using point-biserial correlation and p-values. *International Journal of Scientific & Technology Research*, 2020, vol. 9, no. 4, pp. 2122–2126.
33. Subali B., Suyata P. Panduan analisis data pengukuran pendidikan untuk memperoleh bukti empirik kesahihan menggunakan program Quest. Yogyakarta, Lembaga Penelitian dan Pengabdian pada Masyarakat UNY, 2011.
34. Kunandar K. Penilaian autentik (Penilaian hasil belajar peserta didik berdasarkan Kurikulum 2013). Jakarta, Rajawali Pers, 2013.
35. Roid G. H., Haladyna T. M. A technology for test-item writing. Academic Press, 1982.
36. Djiwandono S. Tes bahasa pegangan bagi pengajar bahasa. Jakarta: PT Indeks, 2008.
37. Madsen H. S. Techniques in Testing. ERIC, 1983.
38. Huda N., Wahyuni T. S. Analisis butir soal IPA Try Out USBN Tahun Ajaran 2018/2019 dalam kaitannya dengan level kognitif. *Madrasah: Jurnal Pendidikan dan Pembelajaran Dasar*, 2019, vol. 12, no. 1, pp. 29–39.
39. Haryudin A. Validity and Reliability of English Summative Tests at junior High School in West Bandung. *Jurnal Ilmiah UPT P2M STKIP Siliwangi*, 2015, vol. 2, no. 1, pp. 77–90. doi: 10.22460/p2m.v2i1p77-90.167.

40. Masruroh H. Z. An Item Anaalysis on English Summative Test for Second Grade Students of MAN Tulungagung 1 in Academic Year 2013/2014. A Script: State Islamic Institute Tulungagung, 2014.
41. Saputra A. N. S., Retnawati H., Yusron E. Analysis Difficulties and Characteristics of Item Test of on Biology National Standard School Examination. *6th International Seminar on Science Education (ISSE 2020)*, 2021, pp. 8–14.
42. Singh J. P., Kariwal P., Gupta S. B., Shrotriya V. P. Original Article Improving Multiple Choice Questions (MCQs) through item analysis: An assessment of the assessment tool. *Int J Sci Appl Res*, 2014, vol. 1, no. 2, pp. 53–57.
43. Boopathiraj C., Chellamani K. Analysis of Test Items on Difficulty Level and Discrimination Index in the Test for Research in Education. *International Journal of Social Science & Interdisciplinary Research*, 2013, vol. 2, no. 2, pp. 189–193.
44. Manalu D. An Analysis of Students Reading Final Examination by Using Item Analysis Program on Eleventh Grade of SMA Negeri 8 Medan, 2019.
45. Rehman A., Aslam A., Hassan, S. H. Item analysis of multiple-choice questions. *Pakistan Oral & Dental Journal*, 2018, vol. 38, no. 2, pp. 291–293.
46. Toksöz S., Ertunç A. Item analysis of a multiple-choice exam. *Advances in Language and Literary Studies*, 2017, vol. 8, no. 6, pp. 141–146.
47. Kheyami D., Jaradat A., Al-Shibani T., Ali F. A. Item analysis of multiple-choice questions at the department of pediatrics, Arabian Gulf University, Manama, Bahrain. *Sultan Qaboos Univ Med J*, 2018, vol. 18, no. 1, p. e68.
48. Sa'adah N. The Analysis of English Mid-Term Test Items Based on the Criteria of a Good Test At the First Semester of the Eighth Grade Students of MTs. Mathalibul Huda Mlonggo in the Academic Year of 2016 / 2017. *Jurnal Edulingua*, 2017, vol. 4, no. 1, pp. 45–57.
49. Handoko B. L., Pamungkas, H. R. Effect of Independence, Time Budget Pressure, and Auditor Ethics on Audit Quality. *International Journal of Psychosocial Rehabilitation*, 2020, vol. 24, no. 9, pp. 1–6.
50. Yousoof M. Employees Perceptions on Factors Affecting Organizational Climate-An Emperical Study Employees perceptions on factors affecting organziational climate-an emperical study, 2016. Available: https://www.researchgate.net/publication/303921076
51. Maharani A. V., Putro, N. Item analysis of English final semester test. Indonesian Journal of EFL and Linguistics, 2020, vol. 5, no. 2, pp. 491–504.
52. Mahirah R., Ahmad D. Designing Multiple Choice Test of Vocabulary for The First Semester Students at English Education Department of Alauddin State Islamic University of Makassar. *ETERNAL (English, Teaching, Learning, and Research Journal)*, 2016, vol. 2, no. 2, pp. 194–208.
53. Sugianto, A. Item Analysis of English Summative Test: EFL Teacher-made Test. *Indonesian EFL Research & Practice*, 2020, vol. 1 (1), pp. 35-54.
54. Fan J., Frost K., Liu B. Teachers' involvement in high-stakes language assessment reforms: The case of Test for English Majors (TEM) in China. *Studies in Educational Evaluation*, 2020, vol. 66, p. 100898.
55. Hambleton R. K., Swaminathan H., Rogers H. J. Fundamentals of item response theory, vol. 2. Sage, 1991.

Information about the authors

Suwarto Suwarto
(Indonesia, Sukoharjo)
Professor, Doctor, Faculty of Education and Teacher Training of Veteran Bangun Nusantara University, Sukoharjo, Indonesia
E-mail: suwartowarto@yahoo.com
ORCID ID: 0000-0002-7925-8017
Researcher ID: AAT-2165-2021

Suyahman Suyahman
(Indonesia, Sukoharjo)
Doctor, Faculty of Education and Teacher Training of Veteran Bangun Nusantara University, Sukoharjo, Indonesia
E-mail: suyahman.suyahman@yahoo.com
ORCID ID: 0000-0001-7029-3396
Scopus Author ID: 57211791547

Meidawati Suswandari
(Indonesia, Sukoharjo)
Doctor, Faculty of Education and Teacher Training of Veteran Bangun Nusantara University, Sukoharjo, Indonesia
E-mail: moetis_meida@yahoo.co.id
ORCID ID: 0000-0002-5340-9075
Scopus Author ID: 57215844785

Zakiyah Zakiyah
(Indonesia, Malang)
Doctoral Student in English Language Education, State University of Malang, Malang, Indonesia
E-mail: zakiyahpagi@gmail.com
ORCID ID: 0000-0002-2119-7012

Arini Hidayah
(Indonesia, Surakarta)
Lecturer, Surakarta University, Surakarta, Indonesia
E-mail: ariniunsa@gmail.com
ORCID ID: 0000-0002-1640-235X
Researcher ID: AEV-9891-2022