JOURNAL OF CULTURAL ANALYSIS AND SOCIAL CHANGE

Journal of Cultural Analysis and Social Change,
2025, 10(4), 3987-3993
ISSN: 2589-1316

# Development of a Reading Comprehension Test for EFL Learners

Suwarto Suwarto*[1], Zakiyah Zakiyah[2], Arini Hidayah[3], Ninik Sudarwati[4]

[1] *Universitas Veteran Bangun Nusantara, Sukoharjo, INDONESIA, Email: suwartowarto@yahoo.com*
[2] *Doctorate Program in ELT, Faculty of Letters, Universitas Negeri Malang, INDONESIA*
[3] *Universitas Surakarta, Surakarta, INDONESIA*
[4] *STKIP PGRI Jombang, INDONESIA*

***Corresponding Author:** suwartowarto@yahoo.com

**Citation:** Suwarto, S., Zakiyah, Z., Hidayah, A., & Sudarwati, N. (2025). Development of a Reading Comprehension Test for EFL Learners, *Journal of Cultural Analysis and Social Change*, 10(4), 3987-3993. https://doi.org/10.64753/jcasc.v10i4.3715

**Published:** December 27, 2025

## ABSTRACT

Reading comprehension is a fundamental skill in language learning, essential for understanding, interpreting, and critically analyzing written texts. This study focuses on the development and evaluation of a reading comprehension test for eleventh-grade students, specifically targeting analytic and hortatory exposition texts. The objectives of this study are to describe the item difficulty, item discrimination, determine the test's reliability, and identify the number of items included in the test. The study employed a quantitative descriptive design, involving 200 students from a senior high school in Sukoharjo, Central Java, Indonesia. The initial test consisted of 30 multiple-choice items with five answer options each, validated by experts and analyzed using the QUEST program based on the Rasch model. Results: (1) Item Difficulty: Analysis of item difficulty showed that 18 items were categorized as easy ($P > 0.70$) and 11 items were categorized as medium ($0.30 < P < 0.70$). There were no difficult items ($P < 0.30$). This distribution indicates that the test provided a balanced range of items suitable for assessing students' reading comprehension. (2) Item Discrimination: The item discrimination analysis using Point Biserial values revealed that 28 out of 29 items (items 1–26, 28, and 29) were classified as "very good" (0.40–1.00), while one item (item 27) was classified as "reasonably good" (0.30–0.39). There were no items in the categories of "acceptable with revision," "poor discrimination," or "negative discrimination." These results indicate that all items effectively distinguished between high- and low-performing students and were therefore considered acceptable for inclusion in the test. (3) Reliability: The test demonstrated a high level of reliability, with a reliability coefficient of 0.960, indicating that the test was highly consistent and dependable for measuring students' reading comprehension skills. (4) Overall Conclusion: The 29-item reading comprehension test is a reliable, valid, and well-constructed instrument, with appropriate item difficulty and strong discrimination power. The test is suitable for assessing eleventh-grade students' reading comprehension skills, and all items were retained for the final test.

**Keywords:** Reading Comprehension, Item Difficulty, Item Discrimination, Reliability, EFL Learners, Test Development.

## INTRODUCTION

Reading comprehension is a fundamental skill in language learning, enabling students to understand, interpret, and critically analyze written texts. In the context of the eleventh-grade English curriculum, students are expected to master various text types, including analytic exposition and hortatory exposition texts. Assessing students' competence in these text types requires a reliable and valid instrument to accurately measure their reading abilities.

The reading comprehension test used in this study serves as a crucial research instrument, specifically designed to evaluate eleventh-grade students' understanding of analytic and hortatory exposition texts as outlined in the

syllabus. The test consists of 30 multiple-choice items, each with five answer options, carefully constructed to assess a range of reading skills while minimizing the probability of correct answers being selected by chance (Brown, 2004). The multiple-choice format with five options not only enhances the reliability and validity of the test but also ensures that students' performance reflects genuine understanding rather than random guessing.

Furthermore, the test aligns with curriculum standards, making it a relevant tool for evaluating students' proficiency in mastering the targeted text types. By implementing this carefully designed instrument, the study aims to obtain valid and reliable data on students' reading comprehension achievements.

The objectives of this study are: (1) to describe the items difficulty in the reading comprehension test, (2) to describe the items discrimination in the reading comprehension test, (3) to determine the reliability of the reading comprehension test, and (4) to identify the number of items included in the reading comprehension test.

## RESEARCH METHODS

The research design of this study is quantitative and descriptive. The research object is the reading comprehension test. Data were collected from the responses of 200 eleventh-grade students from SMA in Sukoharjo, Central Java, Indonesia, based on all students' answer sheets. The materials include the answer key of the reading comprehension test and a set of test items consisting of 30 multiple-choice questions. Data analysis was conducted using the Quest program.

The reading comprehension test serves as a vital research instrument, specifically designed to assess eleventh-grade students' competence of analytic exposition text and hortatory exposition text as outlined in their syllabus. This test consisted of 30 multiple-choice items, each with five answer options, designed by the researcher to assess various comprehension skills that can be seen in the table 1. The used of a multiple-choice formed with five answer options was strategically designed to reduce the probability of random guessing, thereby enhancing the reliability and validity of assessments in measuring students' reading comprehension achievements (Brown, 2004). By offering a wider range of choices, this formed at minimizes the chances of correct responses being selected by mere chance, ensuring that students demonstrate a genuine understanding of the material. The test aligns with the curriculum standards for eleventh-grade students, making it a relevant tool for evaluating their proficiency in understanding the specific text types they were expected to master.

**Table 1.** The Blueprint Reading Comprehension Test

| No | Indicator | Description | Items |
|---|---|---|---|
| 1 | Identifying the main idea | Students skim the text for the main idea. | 1,15,20 |
| 2 | Identifying explicit information | Students scan specific details or facts directly stated in the text. | 2, 10,16, 29 |
| 3 | Inferring implicit information | Students deduce information or meaning that is implied but not explicitly stated in the text. | 3, 11,21, 30 |
| 4 | Inferring vocabulary in context –implied meaning | Students interpret the meaning of words or phrases based on the context of the sentence or passage. | 4, 12,22 |
| 5 | Identifying the author's purpose or intention | Students determine the reason or goal behind the text, such as to inform, persuade, or entertain. | 5, 13,23, 27 |
| 6 | Distinguishing facts from opinions | Students differentiate factual statements from subjective opinions in the text. | 6, 14, 28 |
| 7 | Evaluating arguments and evidence | Students assess the quality, relevance, and sufficiency of arguments and evidence presented in the text. | 7,17,24 |
| 8 | Suggesting a new title for a text | Students create new titles that are creative, engaging, and relevant to the text content. . | 8,18,25 |
| 9 | Summarizing the text | Students condense the text into a brief summary capturing key points and ideas. | 9,19,26 |

(Nation & Macalister, 2020; Grabe & Stoller, 2022)

Developing a reading comprehension test in English requires systematic steps to ensure that the test was not only relevant but also covers all aspects of the reading ability to be measure. One important aspect was content validity, namely the extent to which the test items represent actual reading skills. In this case, the test had been designed to cover various indicators, such as identifying main ideas, understanding vocabulary context and drawing conclusions. To ensure content validity, each item had been validated by an expert in the field of English language education with teaching experience in the Reading course. This validation process includes an assessment of the material aspect (the suitability of the test content to the curriculum and the student's context), the construction

 

aspect (the coherence and clarity of the question formed at), and the language aspect (the used of simple, clear language, and in accordance with the student's ability level). The validation results showed that the test had met the criteria as a good instrument and was worthy of being used in measuring students' reading comprehension abilities. Thus, this test had a strong foundation as a valid and comprehensive measuring tool (Hughes, 2003 & Brown, 2004).

The reliability aspect was also a crucial element in ensuring the quality of the reading comprehension test (Brown, 2004). Reliability ensured that test results were consistent and stable, even when administered at different times or in different contexts (Brown, 2004). One common method to measure reliability was using Cronbach's Alpha, which evaluated the internal consistency of test items. Cronbach's Alpha assessed how well the items in a test measured the same construct, such as reading comprehension skills. A high Cronbach's Alpha value (generally above 0.700) indicates that the test items were reliable measurements (Rudyatmi & Rusllowati, 2017). For example, in a reading comprehension test, items targeting skills like understanding the main idea, inferring meaning, and analyzing text structure should consistently reflect students' skill. By calculating Cronbach's Alpha, researchers can ensure the test provided accurate and dependable results.

Concerning the Alpha Cronbach coefficient ($\alpha$):

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum s^2{}_i}{s^2{}_x}\right)$$

Note:

$\alpha = Alpha - Cronbach$

$k$ = the number of test items.

$\sum s^2{}_i$ = the number of all variant item of test

$s^2{}_x$ = variant total score of tests

(Suwarto, 2021;2023)

For the analysis of item quality in English reading tests, two main aspects used were item difficulty and item discrimination (Brown, 2004). Item difficulty measured the extent to which an item can be answered correctly by test takers, which was usually expressed in the formed of a proportion index. Tis index ranges from 0.00 to 1.00, where lower values indicate difficult items, while higher values indicate easy items. Items with an ideal item difficulty were generally in the range of 0.30 to 0.70. To find the item difficulty of each test item, the following formula:

$$p = \sum B \diagup N$$

Note:

p = proportion of correct

$\sum$ B= the number of correct answers

N = the number of respondents

(Suwarto, 2021;2023)

The item difficulty can be classified into three that were easy, medium, and difficult. According to Suwarto et al. (2023), the category of item difficulty was as follows:

**Table 2.** The Category of the Item Difficulty

| **P = The item difficulty** | **Category** |
|---|---|
| P > 0,700 | Easy |
| 0,300 ≤ p ≤ 0,700 | Medium |
| P < 0,300 | Difficult |

Meanwhile, item discrimination indicates the ability of an item to differentiate between participants with high and low abilities. The item discrimination value ranges from -1.00 to 1.00, where positive values approaching 1.00 indicate that the item had good discriminatory power, while negative values indicate that the item was ineffective or misleading (Ebel & Frisbie, 1991). Point biserial correlation formula was a formula to find out the item discrimination of each test item. The formula that can be used to calculate the item discrimination index as follows:

$$r_{pbi} = \frac{M_p - M_t}{S_T}\sqrt{\frac{p}{q}}$$

Note:

 

$r_{pbi}$ = point biserial correlation coefficient
$M_p$ = the mean criterion score for those who answer the item correctly
$M_t$ = the mean criterion of total score
$S_T$ = standard deviation of total score
p = proportion of correct
q = proportion of false (q = 1 – p) (Suwarto, 2021;2023)

The item discrimination was divided into four, which were bad, acceptable, good and very good. The bad item was removed. However, the acceptable item should be changed for good and very good items. They will be stored in the test bank (Suwarto, et al., 2023). The reading comprehension test were piloted on 200 students. This number was purposefully chosen to meet the minimum data requirements for conducting item analysis using the QUEST program, which is based on the Rasch model. The QUEST program requires an adequate number of respondents to ensure stable and accurate estimations of item parameters, such as item difficulty, item discrimination, and test reliability. According to Linacre (1994), a minimum of 100 respondents is sufficient for exploratory purposes, while a sample size of at least 200–250 is recommended for more robust and high-stakes measurement contexts.

**Table 3.** The Category of the Item Discrimination

| **Item Discrimination** | **Category** |
|---|---|
| 0.40-1.00 | Very good |
| 0.30-0.39 | Reasonably good |
| 0.20-0.29 | Acceptable with revision |
| 0.00-0.19 | Poor discrimination |
| Negative $R_{pbis}$ | Low performing students got the correct answers more with that of high performing students |

Several empirical studies have used similar or slightly larger sample sizes when applying QUEST. For example, Futri et al. (2022) analyzed mathematics test items using QUEST with 398 students and reported stable item parameters and model fit. Ernawati et al. (2024) successfully applied QUEST to analyze test items answered by 187 elementary school students, confirming its applicability for moderate-sized samples. Therefore, piloting the instruments on 200 students in this study follows psychometric recommendations and aligns with established practices in previous QUEST-based research.

## RESULTS AND DISCUSSION

### The Reading Comprehension (30 items)

The reliability of the reading comprehension test was 0.950. It could be seen in internal consistency of the last page of the output of QUEST program. It meant that the test was reliable. As stated Rudyatmi & Rusllowati (2017) that a test was said reliable if the reliability test index was exceeded 0.700.

The item difficulty can be seen in output QUEST item in bar percent (%). The result of item of difficulty based on category was presented in table formed as follows:

**Table 4.** The Result of the Item Difficulty (30 items)

| **Category** | **Item** | **Total** |
|---|---|---|
| Easy<br>(P > 0,700) | 1,2,5,6,8,9,10,13,14,17,18,19,23,24,25,26,27,30 | 18 |
| Medium<br>(0,300 ≤ P ≤ 0,700) | 3,4,7,11,12,15,16,20,21,22,28,29 | 12 |
| Difficult<br>(P < 0,300) | - | - |
| Total | | 30 |

Based on the table 4, it could be concluded that there were 18 items that belonged to easy category in the item difficulty of the test. There were 12 items that belonged to medium category.

 

**Table 5.** The Result of the Item Discrimination (30 items)

| Item Discrimination Category | Item | Total | Note |
|---|---|---|---|
| 0.40 – 1.00: Very good | 2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,29,30 | 28 | Acceptable |
| 0.30 – 0.39: Reasonably good | 28 | 1 | Acceptable |
| 0.20 – 0.29: Acceptable with revision | - | - | - |
| 0.00 – 0.19: Poor discrimination | 1 | 1 | Drop |
| Negative Rpbis: Low-performing students answered correctly more often than high-performing students | - | - | - |
| Total | | 30 | |

The item discrimination can be seen in output QUEST. Discrimination index can be seen in Point Biserial (Pt-Biserial). The result of item discrimination based on category was presented in table 5. Based on the table 5, 28 out of 30 items (items 2–10, 11–27, 29, and 30) fell into the "very good" category, with discrimination values ranging from 0.40 to 1.00. These items were considered effective in distinguishing between high- and low-performing students and were therefore retained for use in the test. One item (item 28) was categorized as "reasonably good" (0.30–0.39) and also retained. Meanwhile, one item (item 1) showed a low discrimination value (0.00–0.19), falling into the "poor discrimination" category. As a result, this item was dropped from the test. No items were found in the "acceptable with revision" or "negative discrimination" categories. In conclusion, a total of 29 items were used in the final test, while only 1 item was excluded based on its poor discrimination performance. The retained items were then reanalyzed using the Quest program, and the results were as follows.

**The Reading Comprehension (29 items)**

The reliability of the reading comprehension test was 0.960. It could be seen in internal consistency of the last page of the output of QUEST program. It meant that the test was reliable. As stated, Rudyatmi & Rusllowati (2017) that a test was said reliable if the reliability test index was up to 0.700. The item difficulty can be seen in output QUEST item in bar percent (%). The result of item of difficulty based on category was presented in table formed as Table 6.

Based on the table 6, it could be concluded that there were 18 items that belonged to easy category in the item difficulty of the test. There were 11 items that belonged to medium category.

**Table 6.** The Result of the Item Difficulty (29 items)

| Category | Item | Total |
|---|---|---|
| Easy (P > 0,700) | 1,3,4,5,7,8,9,12,13,16,17,18,22,23,24,25,26,29 | 18 |
| Medium (0,300 ≤ P ≤ 0,700) | 2,6,10,11,14,15,19,20,21,27,28 | 11 |
| Difficult (P < 0,300) | - | |
| Total | | 29 |

The item discrimination can be seen in output QUEST. Discrimination index can be seen in Point Biserial (Pt-Biserial). The result of item discrimination based on category was presented in table 7.

 

**Table 7.** The Result of the Item Discrimination (29 items)

| Item Discrimination Category | Item | Total | Note |
|---|---|---|---|
| 0.40 – 1.00: Very good | 1,2,3,4,5,6,7,8,9,10,11,12, 13,14,15,16,17,18,19,20, 21,22,23,24,25,26,28,29 | 28 | Acceptable |
| 0.30 – 0.39: Reasonably good | 27 | 1 | Acceptable |
| 0.20 – 0.29: Acceptable with revision | - | - | - |
| 0.00 – 0.19: Poor discrimination | - | - | - |
| Negative $r_{pbis}$: Low-performing students answered correctly more often than high-performing students | - | - | - |
| Total | | 29 | |

In total,

Based on the table 7, 28 out of 29 items (items 1–26, 28, and 29) were categorized as "very good," with discrimination values ranging from 0.40 to 1.00. This indicates that the majority of the items were highly effective in distinguishing between high- and low-performing students, and thus were considered acceptable for use in the test. One item (item 27) fell into the "reasonably good" category (0.30–0.39), which was also considered acceptable. There were no items categorized as "acceptable with revision" (0.20–0.29), "poor discrimination" (0.00–0.19), or with negative point-biserial values.all 29 items demonstrated acceptable discrimination power and were retained. The reading comprehension pretest blueprint was then revised as follows:

**Table 8.** The Blueprint Reading Comprehension Test

| No | Indicator | Description | Items |
|---|---|---|---|
| 1 | Identifying the main idea | Students skim the text for the main idea. | 14, 19 |
| 2 | Identifying explicit information | Students scan specific details or facts directly stated in the text. | 1, 9,15, 28 |
| 3 | Inferring implicit information | Students deduce information or meaning that is implied but not explicitly stated in the text. | 2, 10,20, 29 |
| 4 | Inferring vocabulary in context –implied meaning | Students interpret the meaning of words or phrases based on the context of the sentence or passage. | 3, 11,21 |
| 5 | Identifying the author's purpose or intention | Students determine the reason or goal behind the text, such as to inform, persuade, or entertain. | 4, 12,22, 26 |
| 6 | Distinguishing facts from opinions | Students differentiate factual statements from subjective opinions in the text. | 5, 13, 27 |
| 7 | Evaluating arguments and evidence | Students assess the quality, relevance, and sufficiency of arguments and evidence presented in the text. | 6,16,23 |
| 8 | Suggesting a new title for a text | Students create new titles that are creative, engaging, and relevant to the text content. . | 7,17,24 |
| 9 | Summarizing the text | Students condense the text into a brief summary capturing key points and ideas. | 8,18,25 |

(Nation & Macalister, 2020; Grabe & Stoller, 2022)

## CONCLUSION AND SUGGESTIONS

### Conclusion

Based on the analysis of the 29-item reading comprehension test, the following conclusions can be drawn: (1) Reliability: The test demonstrated a high level of reliability, with a reliability coefficient of 0.960, indicating that the test was highly consistent and dependable for measuring students' reading comprehension skills. This value exceeds the commonly accepted threshold of 0.700 for a reliable test (Rudyatmi & Rusllowati, 2017). (2) Item Difficulty: Analysis of item difficulty showed that 18 items were categorized as easy ($P > 0.70$) and 11 items were categorized as medium ($0.30 < P < 0.70$). There were no difficult items ($P < 0.30$). This distribution indicates that the test provided a balanced range of items suitable for assessing students' reading comprehension. (3) Item Discrimination: The item discrimination analysis using Point Biserial values revealed that 28 out of 29 items (items

 

1–26, 28, and 29) were classified as "very good" (0.40–1.00), while one item (item 27) was classified as "reasonably good" (0.30–0.39). There were no items in the categories of "acceptable with revision," "poor discrimination," or "negative discrimination." These results indicate that all items effectively distinguished between high- and low-performing students and were therefore considered acceptable for inclusion in the test. (4) Overall Conclusion: The 29-item reading comprehension test is a reliable, valid, and well-constructed instrument, with appropriate item difficulty and strong discrimination power. The test is suitable for assessing eleventh-grade students' reading comprehension skills, and all items were retained for the final test.

**Suggestions**

Based on the conclusions of this study, the following suggestions are proposed: (1) For Teachers: Teachers are recommended to use the 29-item reading comprehension test as a reliable and valid instrument to assess eleventh-grade students' reading comprehension skills. The test can also be adapted for classroom practice, remedial teaching, or formative assessment to help identify students' strengths and weaknesses. (2) For Researchers: Future researchers are encouraged to further develop and expand the reading comprehension test by including a wider variety of item types, such as short-answer or essay questions, to measure higher-order thinking skills. Additionally, testing the instrument on a larger and more diverse sample could provide more generalizable results. (3) For Curriculum Developers: The results of this study suggest that the reading comprehension test aligns well with the curriculum standards. Curriculum developers may consider incorporating similar validated tests into standardized assessment tools to ensure the measurement of students' comprehension skills is both reliable and valid. (4) For Test Improvement: Although the current test demonstrates high reliability and good item discrimination, occasional review and revision of items are recommended to maintain the test's quality over time, particularly if it is used repeatedly across different student populations.

## REFERENCES

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

Ebel, R. L., & Frisbie, D. A. (1991). Essentials of educational measurement (5th ed.). Prentice-Hall.

Ernawati, E., Habibah, R. Y., Syarifah, N., Firmansyah, F., & Attamimi, H. R. (2024). Item analysis test of science, Indonesian language, and mathematics using the Rasch model in elementary schools. Jurnal Penelitian dan Evaluasi Pendidikan, 28(2), 195–209.

Futri, V. I., Rosnawati, R., & Rahim, A. (2022). Rasch Model Study on Mathematics Examination Test Using Item Response Theory Approach. International Journal on Emerging Mathematics Education (IJEME), 6(1). https://doi.org/10.12928/ijeme.v6i1.21761

Grabe, W., & Stoller, F. L. (2022). Principles for reading instruction. In Handbook of practical second language teaching and learning (pp. 357-369). Routledge.

Hughes, A. (2003). Testing for language teachers (2nd ed.). Cambridge University Press.

Linacre, J. M. (1994). Many-facet Rasch measurement. MESA Press.

Nation, I. S. P., & Macalister, J. (2020). Teaching ESL/EFL reading and writing (2nd ed.). Routledge.

Rudyatmi, Ely & Rusllowati, A. (2017). Evaluasi Pembelajaran. Semarang: Facultyof Mathematics and Science Unnes.

Suwarto, S. (2021). The Characteristics of Indonesia Second-semester Final Test for Eighth-grade Students. Turkish Online Journal of Qualitative Inquiry, 12(9), 356-370.

Suwarto, S., Suyahman, S., Meidawati, S., Zakiyah, Z., & Arini, H. (2023). The COVID-19 Pandemic and The Characteristic Comparison of English Achievement Tests. Перспективы науки и образования, (2 (62)), 307-329.

 