

Investigating Students' English Proficiency Based on TOEFL Scores Using the Rasch Model

Ilham Sandy Pratama¹

¹Universitas Bandar Lampung

Helta Anggia²

²Universitas Bandar Lampung

Email :ilhamsandypratama@gmail.com , helta@ubl.ac.id

Abstract

Learning English is very important in the era of globalisation, especially for non-native speakers who want to master English language skills. In the field of education, English as a Foreign Language (EFL) is becoming increasingly important as it is considered an essential skill. Along with the increasing need for English in educational institutions, such as the Test of English as a Foreign Language (TOEFL) is used to measure students' abilities and identify areas that need improvement. This study aims to investigate students' English proficiency using Rasch Model analysis on the TOEFL ITP test. This study used a quantitative approach with data analysed using Winstep 3.73 software. The results showed that students were successful in answering long logical inference questions and short dialogue restatements on listening comprehension, as well as clause questions on written structures & expressions. On reading comprehension, students successfully answered reference questions, vocabulary in context, unstated details, and implied details. This study also found that all subtests on the TOEFL ITP have good reliability and suitability. In conclusion, Rasch Model analysis is effective in identifying the types of TOEFL questions that affect student success, as well as ensuring that the subtests used are reliable and appropriate for the measurement of English language skills.

Keywords: EFL, English Language Proficiency, Rasch Model Analysis, TOEFL ITP



INTRODUCTION

English as a Foreign Language (EFL) is increasingly important in today's academic environment due to globalization, making mastery of the language an essential asset (Köroğlu, 2021). The widespread use of English as a global lingua franca has contributed to its rise in importance. With an increasing need for English proficiency in educational institutions, standardized exams to assess students' abilities and identify areas for improvement have become increasingly important (Pussegoda et al., 2017). Tests are used as evaluation tools to determine the level of comprehension and skill in the content covered. Notably, the Test of English as a Foreign Language (TOEFL) is one of the most widely acknowledged examinations (Halim et al., 2018). According to ETS (Educational Testing Service), more than 9,000 universities in over 130 countries utilize TOEFL to guide admissions decisions, and around 27 million people worldwide have taken the test (Yuniarti & Pratiwi, 2022). TOEFL is widely used in Indonesia as a prerequisite for higher education, work chances at home and abroad, and university graduation requirements, providing insight into an individual's degree of English language understanding and ability. A study by Gonzalez (2013), the TOEFL serves as a baseline for measuring native speakers' English competence, consequently it must be administered fairly in order to minimize bias based on gender, race, or religion. However, gender dynamics in the exam present a distinct perspective, since discovered that female students outperform male students in language-related skills such as reading and writing. As a result, monitoring against gender bias in exam questions is essential. TOEFL is widely classified as an Internet-Based Test (IBT) and an Institutional Testing Program (ITP), examines hearing, reading, speaking, and writing skills, all of which address different areas of language competency (Özdemir Yilmazer, 2022)

Speaking, writing, listening, and reading are the four fundamental skills that comprise English proficiency. According to Anggia & Habók (2023), reading proficiency is the capacity to comprehend written material. It calls for word recognition, fluency, and a large vocabulary. It provides a gateway to knowledge, encouraging the development of critical thinking abilities and intellectual progress (Huettig & Pickering, 2019). Effective listening is a crucial component of daily communication and calls for the interpretation and active processing of auditory data. Understanding, remembering, and identifying important details are necessary for effective listening. Despite its difficulty, writing competency is the ability to organize thoughts into coherent stories and educational content and to present them in a creative and clear manner (Tenri Ampa & Quraisy, 2018). Conversely, speaking competency is the ability to use vocabulary, pronunciation,



and fluency in the language to communicate ideas succinctly and clearly (Pinatih, 2021). Verbal communication emphasizes the significance of consistency and clarity to prevent misconceptions (Firmansyah et al., 2019). However, speaking fluency is frequently hampered by poor reading comprehension and a lack of practice opportunities.

Considering this complexity, it is essential to evaluate students' English language proficiency. English language learners confront a number of difficulties to the TOEFL exam (Hashemi, 2011). Using the Rasch model, a method known for its validity and reliability, provides information about the classification of items and test performance on an individual basis (Andrich & Marais, 2018). In order to assess English proficiency at Bandar Lampung University, this study will employ the Rasch model, which will offer important context for using this analytical framework. Through illuminating the subtleties of evaluating English language proficiency, this research aims to deepen comprehension and guide instructional strategies, ultimately promoting ongoing enhancements in students' language skills.

The assessment of English as a Foreign Language (EFL) proficiency in various contexts has drawn a lot of attention to the field of language proficiency testing in recent years. There is still a significant lack of literature regarding thorough analyses of differential item functioning (DIF) across various TOEFL subtests, despite the fact that prior research has examined a variety of aspects of language proficiency testing, including gender bias, item analysis, and cognitive variables (Aryadoust, 2015; Dewi et al., 2023). Some research has been done to systematically examine DIF across all TOEFL subtests, despite the attention that some researchers have paid to particular areas of the test, such as reading comprehension or listening comprehension questions. In order to fill this gap, this study carefully examine the kinds of reading comprehension, writing structures and expressions, and listening comprehension questions that help students succeed. Additionally, it will look for instances of differential item functioning in the TOEFL. To give a complete picture of the TOEFL's efficacy as a gauge of English proficiency, this study also attempts to evaluate the validity and suitability of each subtest (Gu et al., 2015). The purpose of this study is to improve our understanding of language proficiency assessment and offer fresh perspectives on the subtleties of differential item functioning across various TOEFL subtests by conducting this multifaceted analysis.

METHODS

This study examines how well Bandar Lampung University students' English language



proficiency is measured by TOEFL questions using a quantitative methodology. Through the analysis of numerical patterns, this approach described by (Creswell, 2014) offers a methodical way to comprehend individual or group problems. According to Kesmodel (2018), a cross-sectional design was employed for data collection, enabling a momentary evaluation of students' abilities. Students majoring in English education at the university made up the study's population; these students were chosen through total population sampling, which included all students who met specific requirements. Students from semesters 1 through 7 represent the research sample, which provides insight into their level of English proficiency at various academic levels.

The research tool consisted of reading comprehension, structure and written expression, listening comprehension, and TOEFL Longman preparation questions. Students took the TOEFL exam twice in order to collect data, and the results of the test were used to determine the correct answers. The data were evaluated for each test item's difficulty level and differentiating ability using the Rasch model and Winstep 3.73 software. According to (Judd et al., 2016), the Rasch Model contributes in evaluating how well the items measure the intended ability and produces accurate data regarding student competency.

Item fit sequences that indicate the presence or absence of mismatches in items or respondents are used to test the measurement validity of the Rasch model. The three psychometric characteristics of each item derived from the Rasch model, according to Boonee et al. (2014), are used as criteria to assess the appropriateness of items that might be misfits or outliers. While the respondent fit order measures how well the response patterns match or deviate from the students' ability levels, the item fit order evaluates how well the measurement's items work. Boonee et al. (2014) list the following criteria for determining item fit and person fit:

- (a) *Mean Square (MNSQ) antara 0.5 dan 1.5*
- (b) *Z standard (ZSTD) antara -0.2 dan +2.0*
- (c) *Point measure correlation antara 0.4 dan 0.85*

RESULTS AND DISCUSSION

Listening comprehension

The listening comprehension analysis reveals the difficulty level of the questions on the right of the item map and the distribution of student abilities on the left. There are three groups for question difficulty—the most difficult, medium, and easiest—as well as three groups for student ability levels—the high, medium, and low categories. The most difficult questions to



respond to are numbers 36 and 18, while the easiest questions to respond to is number 31. The values in problem number 36 were outside of the standard deviation, indicating a significant level of difficulty for the students. In contrast, even though question number 31 was the easiest, not all students answered it correctly because it decreased within the standard deviation (T). Questions with a logit value of less than -2.0 are considered easy, those with a value between -2.0 and 2.0 are considered moderate, and those with a value greater than 2.0 are considered difficult.

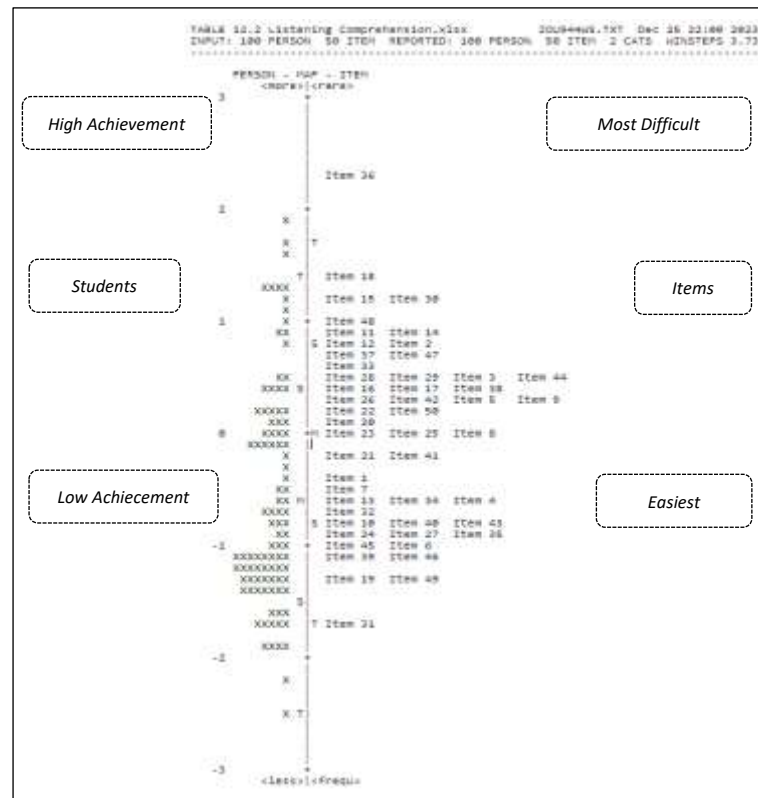


Figure 1. Item map listening comprehension

The Rasch model's measurement validity is examined using item fit analysis on both individuals and items. According to Boone et al. (2014), each item produced by the Rasch model has three psychometric attributes that are used as a criterion for assessing the fit of potentially outlier or misfit items. While the person fit order is used to evaluate whether students' response patterns match or deviate from their ability levels, the item fit order is used to evaluate whether the items function properly in measurement. Following analysis of listening comprehension items,

the following table was produced.

Table 1. Reliability and fit statistics of listening comprehension test

Construct	Male		Female	
	Items	Persons	Items	Persons
Number	50	28	50	72
Mean	11.8	21.0	27.3	19.0
Standard Deviation	4.4	10.7	11.5	8.7
Reliability (Cronbach's alpha)	0.73	0.91	0.90	0.86
Separation	1.64	3.10	2.85	2.51
MNSQ (infit)	0.99	1.00	1.00	1.00
MNSQ (outfit)	1.00	1.00	1.03	1.03
ZSTD (infit)	0.0	0.0	-0.1	0.0
ZSTD (outfit)	0.0	0.0	0.0	0.1
Chi-squared (X^2)	1447.70		3830.61	
df	1322		3479	

Table 1 displays the logit means for item and person scores in males, which were 10.9 and 15.5, respectively. The logit means for person and item scores in females were 29.5 and 16.4, respectively. Positive standard deviation values were present in all scores. Table 1 shows that the MNSQ outfit value for items and person was 1.03 and the average MNSQ infit value for items and person in males was 1.00. On the other hand, the MNSQ outfit value and infit value for both person and item were both 1.00 for females. If MNSQ is between 0.5 and 1.5, it is considered appropriate. ZSTD test results indicated that males had infit values on items of 0.0 and 0.0, while ZSTD outfit values were found on items of 0.2 and individuals on 0.0. In contrast, the infit and outfit ZSTD values for both individuals and objects in females were 0.0. Since this value falls between -2 and 2, all objects and people have rational likelihood values. Values are regarded as reliable if they are > 0.70 and unreliable if they are < 0.70 in terms of reliability (Cronbach's alpha). All of the items and persons in the reliability table have values greater than 0.70 for both males and females, indicating their reliability.



The study also used plot Differential Item Functioning (DIF) to see if there was any bias between men and women in the understanding of listening. Note the following chart showing Differential item functioning listening comprehension.

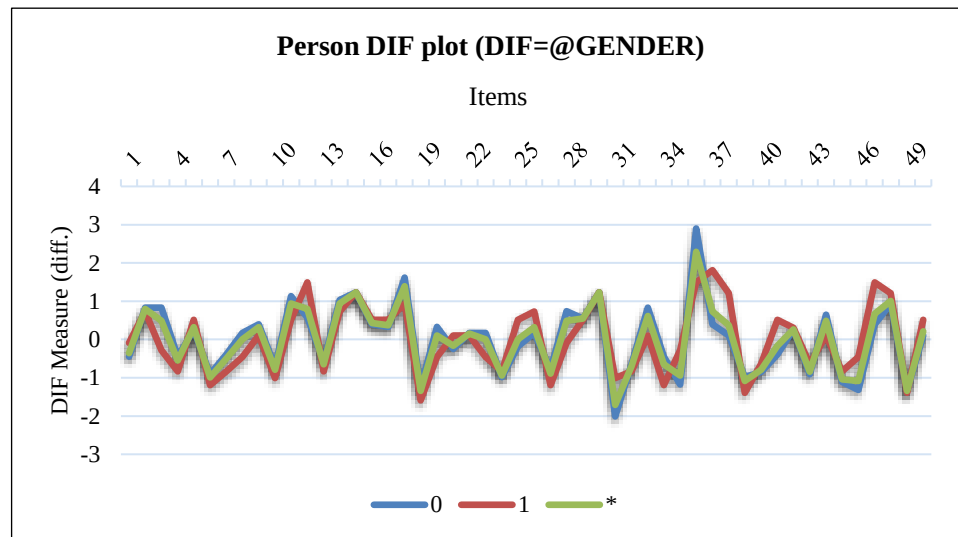


Chart 1. Differential item functioning listening comprehension

The plot presents a significant distance, as the chart demonstrates. It indicates that the items' degrees of difficulty vary amongst the groups; consequently, some groups are better off than others in this situation. Instruments or measuring instruments in the form of bias-free items are necessary for valid measurements. When some individuals with particular characteristics are given preference over other individual characteristics—like gender—an instrument or item is considered biased. On the plot, the average (green line) is represented by number 2, the female group (orange line) by number 1, and the male group (blue line) by number 0. Chart 1 indicates that there is a significant difference between the responses of males and females on items 3 and 35. More disadvantage is shown by a higher chart, and more advantage is shown by a lower chart. The chart indicates that there is an increase in the male group for items 3 and 35, indicating that the female group was more successful in answering the question while the male group struggled. Girls are given preference. In contrast, the chart indicates that the female group is increasing in items 31, 37, 38, and 47, indicating that the male group is ahead in these questions. This analysis shows that there are gender-specific variations in item difficulty, suggesting that some items are

biased. It is important to take note of this in order to confirm the validity and fairness of the assessment of English language proficiency.

Structure and written expression

According to analyses of the items' structures and written expressions, students found item number 15 to be most difficult to complete, while item number 8 was the easiest to answer. The logit value of problem number 8 was -2.16, suggesting that it was extremely easy, whereas the logit value of problem number 15 was +2.20, indicating a high degree of difficulty. The majority of students, even those with high aptitude, frequently provided incorrect answers to problem number 15, rendering it an anomaly. However, question number 8 was ineffective in differentiating students' abilities because nearly all students were able to answer it correctly. When a question's logit value is less than -2.0, it is classified as low difficulty; -2.0 to 2.0 is considered medium difficulty; and > 2.0 indicates high difficulty.

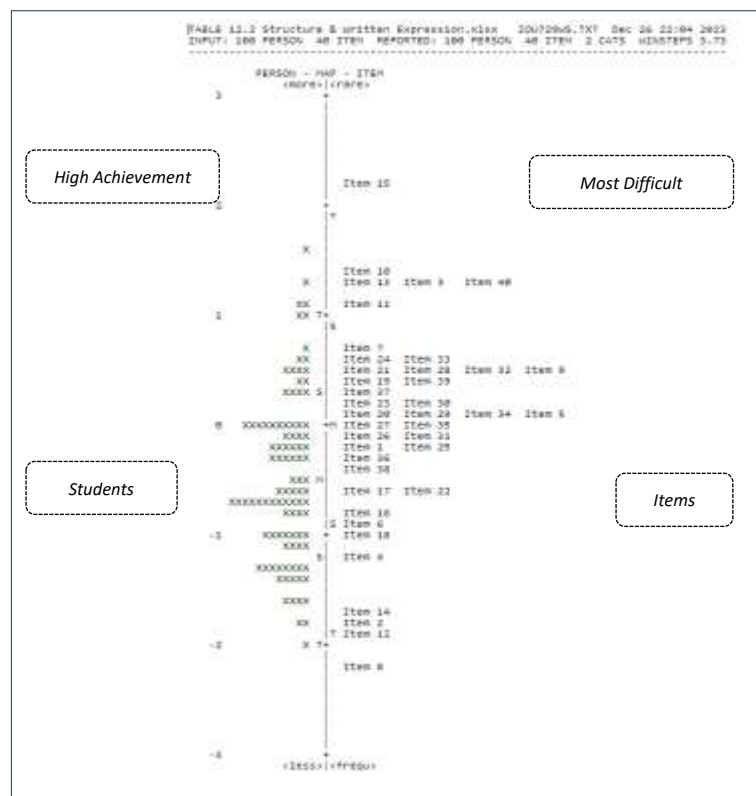


Figure 2. Item map structure and written expression

Low Achievement

Easiest

In addition, the items fit analysis for the structure and written expression questions reveals that both items and individuals have rational probability values. According to Table 2, the reliability column shows that the Cronbach's alpha coefficients for males and females, for both items and individuals, exceed the threshold of 0.70. This observation emphasizes the high reliability of all items and individuals within the examined construct, confirming the assessment instrument's validity and reliability.

Table 2. Reliability and fit statistics of structure & written expression

Construct	Male		Female	
	Items	Persons	Items	Persons
Number	50	28	50	72
Mean	10.9	15.5	29.5	16.4
Standard Deviation	4.4	7.5	14.0	5.0
Reliability (Cronbach's alpha)	0.72	0.85	0.92	0.70
Separation	1.61	2.34	3.43	1.48
MNSQ (infit)	1.00	1.00	1.00	1.00
MNSQ (outfit)	1.03	1.03	1.00	1.00
ZSTD (infit)	0.0	0.0	0.0	0.0
ZSTD (outfit)	0.2	0.0	0.0	0.0
Chi-squared (X^2)	1182.94		3186.44	
df	1053		2769	

The data confirms that the mean logit values for item and person values for males are 10.9 and 15.5, respectively. Females have mean logit values of 29.5 for person and 16.4 for item. Each value has a positive standard deviation. According to Table 2, the infit standardised mean square (MNSQ) value was 1.00 for item and 1.00 for person, while the outfit MNSQ value was 1.03 for item and 1.03 for males. Meanwhile, for women, the MNSQ fit value for item and person was 1.00, as was the MNSQ outfit value for item and person. The MNSQ value is considered



appropriate if it falls between 0.5 and 1.5. For males, the z-standardised test score (ZSTD) showed infit values of 0.0 on item and person, as well as outfit values of 0.2 on item and 0.0 on person. For females, the ZSTD values for item and person were 0.0. These values fall within the range of $-2 < ZSTD < 2.0$, indicating that all items and individuals have rational the probability values. Cronbach's alpha values are reliable (> 0.70) or unreliable (< 0.70). The Table 2 reliability column shows values for item and person in both males and females that are greater than 0.70, indicating that all items and persons are reliable.

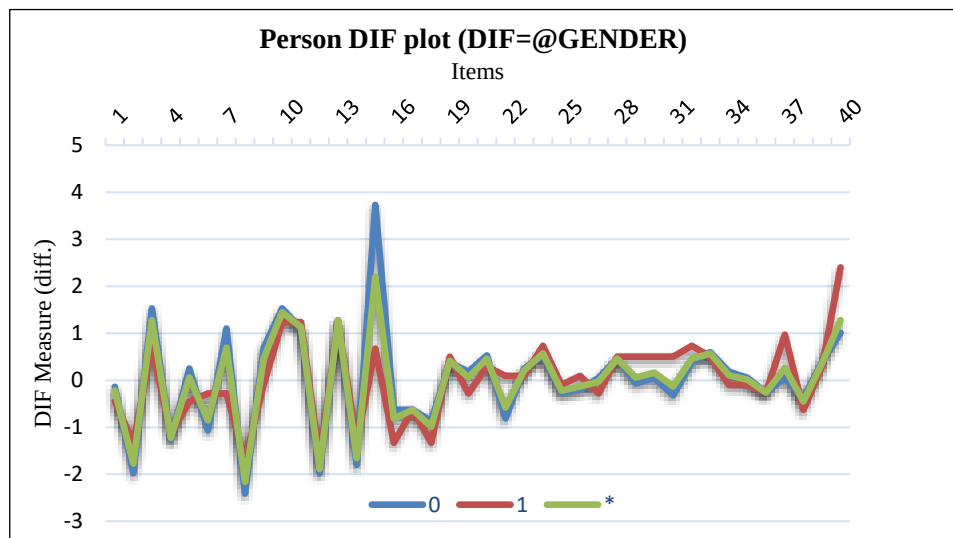


Chart 2. Differential item functioning structure & written expression

Meanwhile, the Differential Item Functioning analysis in chart 2 indicates a significant difference between males and females on items 7, 15, and 40. The number 0 denotes the male group (blue line), the number 1 the female group (orange line), and the number 2 the average (green line). The higher the graph position, the less favourable; the lower the graph position, the more favourable. The graph for items 7 and 15 shows an increase in the male group, indicating that the male group struggled to solve the problems. In contrast, the female group performed better on items 7 and 15. For item 40, the graph shows an increase in the female group, indicating that the male group is more favoured.

Reading comprehension

The investigation of the reading comprehension items established that items 20, 49, 21,



and 18 were the most difficult, while items 17, 9, 15, 12, 25, 33, 38, and 28 were easy. Item numbers 20, 49, 21, and 18 remain within the standard deviation, indicating that students may still answer correctly. Item numbers 17, 9, 15, 12, 25, 33, 38, and 28 fall outside the standard deviation, indicating that students are likely to answer correctly. A question's difficulty level is categorized as low (< -2.0 logits), medium ($-2.0 \text{ logits} \leq b \leq 2.0 \text{ logits}$), and high (> 2.0 logits). The issues number 20 and 49 have logit values of +2.35, 21 +2.24, and 18 +2.14, indicating high difficulty. Problem numbers 17, 9, 15, 12, 25, 33, 38, and 28 had logit values ranging from -1.83 to -1.64, indicating moderate difficulty.

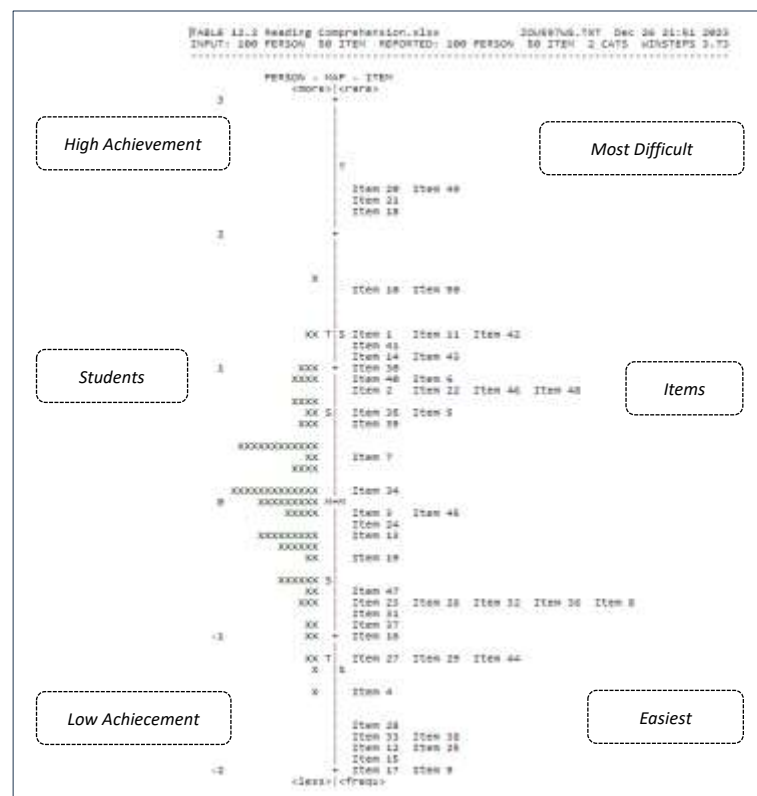


Figure 3. Item map structure and written expression

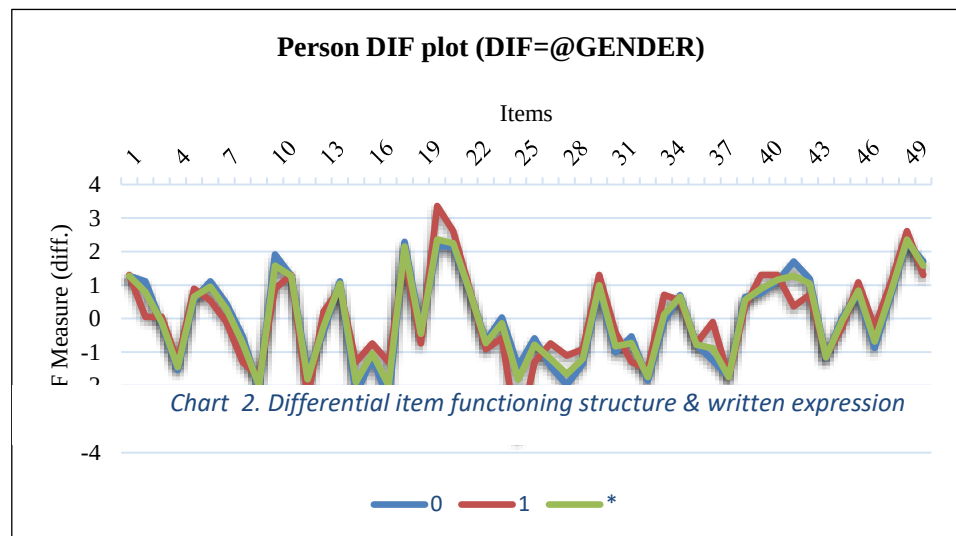
The table below highlights the mean (logit) measures of item and person scores for men and women. These results were obtained from the completed answers to the reading comprehension section, resulting in the following data:

Table 3. Reliability and fit statistics of reading comprehension

Construct	Male		Female	
	Items	Persons	Items	Persons
Number	50	28	50	72
Mean	13.5	24.1	37.2	25.8
Standard Deviation	6.8	5.4	18.4	5.1
Reliability (Cronbach's alpha)	0.86	0.70	0.95	0.70
Separation	2.51	1.50	4.37	1.41
MNSQ (infit)	0.99	1.00	1.00	1.00
MNSQ (outfit)	1.01	1.01	1.00	1.00
ZSTD (infit)	0.0	-0.1	0.0	-0.1
ZSTD (outfit)	0.1	0.0	0.1	-0.1
Chi-squared (X^2)	1488.65		3743.80	
df	1323		3478	

The data above show that the mean (logit) measures for item scores and person scores in males are 13.5 and 24.1, while in females they are 37.2 and 25.8. All of the scores have positive standard deviations. According to Table 3, the infit standardised mean square (MNSQ) value was 0.99 for the item and 1.00 for the person, while the outfit MNSQ value for the item was 1.01 and the person was 1.01 for males. Meanwhile, for women, the MNSQ infit value is 1.00 for both item and person, and the MNSQ outfit value is also 1.00. An MNSQ value is considered fit if it falls between 0.5 and 1.5, indicating that neither the item nor the person is unfit. The z-standardized test score (ZSTD) values revealed an infit on item of 0.0 and -0.1 for males, and an outfit ZSTD of 0.1 on item and 0.0 on person. Meanwhile, for females, the item ZSTD is 0.0 and the person ZSTD is -0.1, while the outfit ZSTD is 0.1 and -0.1, respectively. The values fall within the range of $-2.0 < \text{ZSTD} < 2.0$, indicating rational values for all items and individuals. Reliability (Cronbach's alpha) is considered reliable when the value exceeds 0.70 and unreliable when it is less than 0.70. Table 3's reliability column shows that the male and female values for the item and person are greater than 0.70, indicating that all items and person are reliable.





The analysis Differential Items Functioning (DIF) above reveals a significant difference between males and females on item numbers 2, 17, 20, 25, 37, and 42. The number 0 denotes the male group (blue line), the number 1 the female group (orange line), and the number 2 the average (green line). The higher the graph position, the less beneficial; the lower the graph position, the more beneficial. The graph shows an increase in the male group for items 2, 25, and 42, indicating that the male group had more difficulty with these items and that the female group was more preferred. In contrast, the graph for items 17, 20, and 37 shows an increase in the female group, indicating that the male group has an advantage on these items.

Finally, study examines the item maps, reliability and matching statistics, and differential item functioning (DIF) in the English test's sections on listening comprehension, written structure and expression, and reading comprehension. Results revealed that in the listening comprehension section, items 36 and 18 presented the most significant challenge to students, indicating difficulties in inferential and interpretive listening skills, whereas item 31 was the easiest, indicating better performance in direct recall of facts. In the written structure and expression section, item 15, which involved inversion in conditional sentences, was the most difficult, indicating that students struggled with complex syntactic structures, whereas item 8, a clause question, was the easiest, indicating proficiency with simpler grammatical constructions. Items 20, 49, 21, and 18 were the most difficult in terms of reading comprehension due to they required higher-order thinking skills like inference and detailed context understanding. Items 17, 9, 15, 12,

25, 33, 38, and 28 were classified as easy, involving simple comprehension tasks such as identifying vocabulary in context and unstated details.

The reliability and fit statistics revealed that all sub-tests had positive standard deviation values and reliability values (Cronbach's alpha) higher than 0.70, indicating that the test was reliable. MNSQ values of 0.5 to 1.5, as well as ZSTD values of -2.0 to 2.0, indicated that no items or people were unsuitable. DIF analysis revealed gender differences in listening comprehension; males had more difficulty with items 3 and 35, whereas females found items 31, 37, 38, and 47 more difficult. In written structures and expressions, males struggled with items 7 and 15, while females struggled with item 40. Males found item numbers 2, 25, and 42 more difficult in reading comprehension, whereas females struggled with item numbers 17, 20, and 37. These results are consistent with previous research, which found that students struggled with inferential and complex syntactic questions but performed better on factual and context-based questions. This study offers specific insights into gender-based performance differences, which can help educators and test designers create a balanced and fair assessment of English as a foreign language.

DISCUSSION

This study analyzed the difficulty, reliability, and potential gender bias in a language proficiency test, focusing on listening comprehension, structure and written expression, and reading comprehension. Utilizing the Rasch model, it was found that certain items posed significant challenges to students. For instance, the long dialogue logical inference question (Item 36) and short dialogue restatement question (Item 18) were the most difficult in listening comprehension, while an inversion in a conditional sentence (Item 15) proved challenging in structure and written expression. Conversely, students found list questions in long dialogues (Item 31) and vocabulary in context questions relatively easy. The analysis showed that reliability measures, such as Cronbach's alpha, were above 0.70, indicating satisfactory internal consistency. The study also revealed no misfit items or persons, suggesting that all test components effectively measured student abilities.

Additionally, differential item functioning (DIF) analysis identified gender biases in several test items, with certain questions disadvantaging either male or female students. In listening comprehension, items like 31, 37, 38, and 47 were more difficult for females, while items 3 and 35 disadvantaged males. Similar patterns were observed in structure & written expression and



reading comprehension sections, indicating that gender bias could impact test performance. These findings align with related studies, such as Bordbar (2020), which examined gender-biased items in language proficiency tests, and Dewi et al. (2023), who analyzed reading comprehension questions using the Rasch model. These studies emphasize the importance of addressing biases to ensure fair and valid assessments, highlighting the need for ongoing evaluation and refinement of testing materials to accommodate diverse student backgrounds and abilities.

CONCLUSION

The analysis confirms that all items in the TOEFL preparation questions have satisfactory fit coefficients, demonstrating their suitability for assessing English language skill levels across academic semesters. Particularly, no misfit questions were found, indicating that the TOEFL preparation items were generally effective in measuring students' English language competencies from the first to the seventh semester. Different difficulty levels were noticed among the items, with some questions creating obstacles for students while others were quite easy to navigate. Furthermore, substantial gender discrepancies were seen in students' performance on these questions.

In terms of listening comprehension, students presented excellent ability in interpreting longer dialogues, although experiencing significant obstacles with questions about logical inference in long dialogues and restatement in short dialogues. In the area of structure and written expression, students established weaknesses in addressing inversion in conditional sentences, in contrast to their proficiency in handling clause-based queries. In reading comprehension, students struggled with topic-type questions, inferred questions, and those that required the use of 5W+1H strategies, whereas they were more comfortable with reference, vocabulary in context, unstated detail, and implied detail questions, which are considered less complex. Furthermore, the reliability analysis for both items and individuals emphasizes the instrument's ability to produce dependable measurements, as indicated by the obtained reliability values, so confirming the instrument's overall reliability.

Acknowledgement

The Authors would like to express their heartfelt gratitude to the English Education Program, University of Bandar Lampung, for their invaluable support and assistance throughout the preparation of this research. Special appreciation is extended to our advisor, whose guidance



and expertise have been instrumental in shaping the trajectory of this study. We are profoundly grateful for their unwavering encouragement, insightful feedback, and scholarly mentorship, which have significantly enriched the quality and depth of our research endeavor.

REFERENCES

- Andrich, D., & Marais, I. (2018). *Controlling Bias in Both Constructed Response and Multiple-Choice Items When Analyzed With the Dichotomous Rasch Model*. In *Journal of Educational Measurement Summer* (Vol. 55, Issue 2). [https://doi.org/10.1111/\(ISSN\)1745-3984](https://doi.org/10.1111/(ISSN)1745-3984)
- Anggia, H., & Habók, A. (2023). *Textual complexity adjustments to the English reading comprehension test for undergraduate EFL students*. *Heliyon*, 9(1). <https://doi.org/10.1016/j.heliyon.2023.e12891>
- Aryadoust, V. (2015). *Fitting a Mixture Rasch Model to English as a Foreign Language Listening Tests: The Role of Cognitive and Background Variables in Explaining Latent Differential Item Functioning*. *International Journal of Testing*, 15(3), 216–238. <https://doi.org/10.1080/15305058.2015.1004409>
- Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch Analysis in the Human Sciences*. Springer Netherlands. <https://doi.org/10.1007/978-94-007-6857-4>
- Creswell. (2014). *Research design: Qualitative, quantitative, and mixed method approaches* (4th ed.). SAGE Publications.
- Dewi, H. H., Damio, S. M., & Sukarno, S. (2023). *Item analysis of reading comprehension questions for English proficiency test using Rasch model*. *REID (Research and Evaluation in Education)*, 9(1), 24–36. <https://doi.org/10.21831/reid.v9i1.53514>
- Firmansyah, D., Elvisca, E., & Vegian, V. (2019). *Improving the Students' Speaking Skill Through Debate Technique*. *Professional Journal of English Education*, 2(6). <https://doi.org/10.22460/project.v2i6.p891-895>
- Gonzalez, M. (2013). *The Intricate Relationship Between Measures of Vocabulary Size and Lexical Diversity as Evidenced in Non-native and Native Speaker Academic Compositions [University of Central Florida]*. <http://library.ucf.edu>
- Gu, L., Lockwood, J., & Powers, D. E. (2015). *Evaluating the TOEFL Junior ® Standard Test as a Measure of Progress for Young English Language Learners*. *ETS Research Report Series*, 2015(2), 1–13. <https://doi.org/10.1002/ets2.12064>
- Halim, N., Ypup Makassar, S., Yulianti, S., Stkip, A., & Makassar, Y. (2018). *Difficulties Faced By the Students in Answering TOEFL Test Questions*. *English Teaching Learning and Research*



- Journal*, 4(2). <https://doi.org/10.24252/Eternal.V42.2018.A7>
- Hashemi, M. (2011). Language stress and anxiety among the English language learners. *Procedia - Social and Behavioral Sciences*, 30, 1811–1816. <https://doi.org/10.1016/j.sbspro.2011.10.349>
- Huettig, F., & Pickering, M. J. (2019). Literacy Advantages Beyond Reading: Prediction of Spoken Language. In *Trends in Cognitive Sciences* (Vol. 23, Issue 6, pp. 464–475). Elsevier Ltd. <https://doi.org/10.1016/j.tics.2019.03.008>
- Judd, B. K., Scanlan, J. N., Alison, J. A., Waters, D., & Gordon, C. J. (2016). The validity of a professional competence tool for physiotherapy students in simulation-based clinical education: A Rasch analysis. *BMC Medical Education*, 16(1). <https://doi.org/10.1186/s12909-016-0718-x>
- Kesmodel, U. S. (2018). Cross-sectional studies – what are they good for? In *Acta Obstetrica et Gynecologica Scandinavica* (Vol. 97, Issue 4, pp. 388–393). Wiley-Blackwell. <https://doi.org/10.1111/aogs.13331>
- Köroğlu, Z. Ç. (2021). Using Digital Formative Assessment to Evaluate EFL Learners' English Speaking Skills. *Gist Education and Learning Research Journal*, 103–123. <http://acikerisim.aksaray.edu.tr/xmlui/handle/20.500.12451/9155>
- Özdemir Yilmazer, M. (2022). Direct Access to English-Medium Higher Education in Turkey: Variations in Entry Language Scores. *Dil Eğitimi ve Araştırmaları Dergisi*, 8(2), 325–345. <https://doi.org/10.31464/jlere.1105651>
- Pinatih, I. G. A. D. P. (2021). Improving Students' Speaking Skill through Role-Play Technique in 21st Century. *Journal of Educational Study*, 1(1), 103–108. <https://doi.org/10.36663/joes.v1i1.159>
- Pussegoda, K., Turner, L., Garritty, C., Mayhew, A., Skidmore, B., Stevens, A., Boutron, I., Sarkis-Onofre, R., Bjerre, L. M., Hróbjartsson, A., Altman, D. G., & Moher, D. (2017). Identifying Approaches for Assessing Methodological and Reporting quality of Systematic Reviews: A Descriptive Study. *Systematic Reviews*, 6(1). <https://doi.org/10.1186/s13643-017-0507-6>
- Tenri Ampa, A., & Quraisy, H. (2018). Needs Analysis of the English Writing Skill as the Base to Design the Learning Materials. *SHS Web of Conferences*, 42, 00050. <https://doi.org/10.1051/shsconf/20184200050>

