ISSN: 2278-4012, Volume:15, Issue:2, June 2026

Comparative Lexical Diversity Metrics in High Stakes English Proficiency Examinations including IELTS and TOEFL

Aziza Safarova[1]

[1]Teacher, Gulistan State Pedagogical Institute, Gulistan, Uzbekistan.
E-mail: a.safarova@gspu.uz, Orcid: https://orcid.org/0009-0007-9678-3538

Abstract: Lexical diversity is quite significant in reflecting language proficiency in high-stakes English tests that depict whether a candidate is capable of knowing and using various and appropriate vocabulary. However, there is very little comparative linguistic research on investigating the differences in lexical diversity of the major proficiency tests. This research fills this gap with a comparison and analysis of lexical diversity measurements of IELTS and TOEFL answers. It involves a corpus-based methodology, where with the help of computational tools, Type-Token Ratio (TTR), Measure of Textual Lexical Diversity (MTLD), Hypergeometric Distribution D (HD-D), and VOCD are measured. Results show that IELTS answers are more lexical than TOEFL answers, which are more controlled lexical due to integrated task designs. These results highlight the impacts of the test design on the production of language. Language test evaluation studies also utilize the research paper to offer information on how to improve test scores, scoring rubrics, and how a pedagogical approach to English language learners may be provided in the high-stakes testing environment.

Key Words: lexical diversity, IELTS, TOEFL, corpus linguistics, language assessment, computational linguistics.

(Received: 10 March 2026; Revised: 21 April 2026; Accepted: 14 May 2026; Published: 30 June 2026)

Introduction

Lexical diversity is mostly considered to be a significant indicator of language proficiency, particularly in high-stakes English tests where the language proficiency of the candidates is assessed in different dimensions. It signifies how a learner can effectively use a huge number of vocabulary in the correct manner and in the most efficient manner; thereby, it is a wellspring of fluency, coherence, and general communicative ability. Recent studies emphasize that lexical properties are very crucial in defining the quality and performance of writing, especially when complexity, accuracy, and fluency are their respective lenses of measurement (Qian, 2023). The lexical diversity is, hence, a unit that has always been essential in the modern-day testing model of language.

The English proficiency tests such as the IELTS and TOEFL are high-stakes tests, which are recognized as academic admissions, immigration, and professional certifications internationally. The tests will be used to gauge the ability of the candidate to work under English-speaking conditions. Research has found that these tests have various linguistic activities that not only require one to be able to comprehend but also to produce various lexical objects (Chen & Luo, 2025). In addition, the validity and reliability of such tests, on whether can capture the actual language use and linguistic variations, are highly subject to the quality of such tests (Khan & Javed, 2025). Have become so accustomed that a question is raised as to whether there is a comparability of linguistic demands between such systems of testing.

ISSN: 2278-4012, Volume:15, Issue:2, June 2026

The increase in the application of computational and corpus-based methods has highlighted the need to have objective and measurable measures of linguistic performance in the study of language performance. It has been found that lexical complexity and cohesion have an impact on scoring high-stakes writing activities (Abbaspour & Mathew, 2025). The invention of computerized scoring and speech assessment technologies has enhanced the importance of measurable speech features under mass testing conditions (Gong, 2023). The changes underscore the need to take into account effective lexical diversity indicators in the assessment systems to make it fair and consistent.

There is a considerable gap in comparative studies that examine the lexical diversity in different types of examination in this broad literature on lexical features in language testing. Although some of these studies have been conducted on specific aspects of either IELTS or TOEFL, there is a dearth of research undertaking a systematic comparison of the lexical richness of both of these tests in multiple measures (Le et al., 2025). This study is therefore aimed at making comparisons between lexical diversity in the answers in IELTS and TOEFL and evaluating the effectiveness of the various lexical diversities. To address this gap, the research will address a greater comprehension of language testing and drive reforms in the development of tests and testing to enhance evaluation processes.

The paper has the following structure: section 1 presents the concept of lexical diversity and the research problem, objectives and importance of the comparison between IELTS and TOEFL. In Section 2, the literature review on the lexical diversity measures, language testing, and comparison of previous studies are reviewed, and gaps in research and goals are identified. Section 3 outlines the methodology, which involves research design, data collection, data collection tools and analytical methods. Section 4 includes results and analysis which comprises descriptive statistics, comparative results and statistical significance. Section 5 talks about important lessons and lessons learnt in terms of test design, scoring and pedagogy. Lastly, Section 6 wraps up the research by summarizing the findings, contributions, limitations and future research directions.

Literature Review

The variety and diversity of vocabulary in the spoken or written language is known as "lexical diversity" and is considered to be the significant measure of language proficiency. It is extremely intermingled with such linguistic constructs as fluency, coherence, and lexical sophistication, which are of utmost significance in high-stakes testing. Studies have indicated that most of the time high lexical diversity is related to high quality of writing and high grades in assessments (Holmberg Sjöling, 2025). Also, lexical properties are also part of determining the ability of the candidates to create meaningful and contextually appropriate language, particularly in an academic context.

Lexical diversity is also important in determining the scoring criteria and judgment of the examiners in high-stakes testing conditions. The experiments demonstrate that writing performance is well predicted by lexical complexity and cohesion, especially when one has to write a lengthy response (Abbaspour & Mathew, 2025). Additionally, the lexical analysis as one of the evaluation systems is consistent with the bigger concepts of English as a lingua franca, where the quality of communication is of greater concern than the native-like quality (Saeedi et al., 2025). This leads to the emergence of the importance of lexical diversity as a theoretical and practical concept in the assessment of language.

In a bid to measure lexical diversity, scholars have come up with a number of computational measures, all of which have strengths and weaknesses. The Type-Token Ratio (TTR) is one of the oldest and simplest ones, and the combination of the unique words (types) and the total words (tokens) is provided as the ratio.

ISSN: 2278-4012, Volume:15, Issue:2, June 2026

TTR is, on the other hand, text length sensitive, thus rendering it unreliable when dealing with longer texts. This has led to the recommendation of a number of more advanced measures for overcoming these problems, such as the Measure of Textual Lexical Diversity (MTLD) and Hypergeometric Distribution D (HD-D), more trustworthy estimators of the same at varying text lengths.

The other measure is its VOCD, which is extensively applied and applies the probabilistic modeling to approximate the lexical diversity on the basis of random sampling techniques. Such higher measures have increasingly been utilized in corpus-based studies to make comparisons on the performance of language in high-stakes tests. Using language as an example, distinguishing the linguistic features in listening and reading tasks has been examined using multidimensional corpus analysis that has revealed that differences in the complexity of vocabulary could be drawn in different sections of the test (Tao & Aryadoust, 2024). These methods show how it is possible to integrate multiple measures to come up with a more multifaceted measure of lexical diversity.

Computational and corpus-based approaches to lexical and textual aspects of English proficiency tests are an increasing area of research. Comparative studies have been conducted testing the differences in reading passages and linguistic features and have discovered the differences in the lexical density and complexity in tests such as CET-6, IELTS, and TOEFL (Chen & Luo, 2025). Equally, research conducted on the IELTS reading exams has shown that there are significant variations in the vocabulary use in academic and general training modules, implying that the type of task to be undertaken influences vocabulary use (Le et al., 2025).

In addition to written tests, studies performed to assess the performance in spoken language and the impact of automated scoring systems have been carried out as well. As an illustration, TOEFL iBT studies have been conducted on the impact of speech assessment technologies on lexical choices and orating strategies of students (Gong, 2023). Moreover, test preparation studies have demonstrated that being used to the formats of tests can influence the perception and application of integrated skills in language, such as the use of lexical variation (Wang & Cheng, 2025). These findings point to the dynamic interaction of the test design, learner, and linguistic output behavior.

Other studies have dwelled on the wider perspectives of language testing, such as validity, reliability, and authenticity. Researchers have stressed the need to make test tasks a mirror of the way language is used in the real world and to adequately assess linguistic competence (Khan & Javed, 2025). The experiments conducted on time fluency and authenticity in executing listening activities only indicate that it is a challenge to develop effective assessment tools (Nishizawa, 2024). Besides, issues of standardisation and comparability of the testing systems in the regions have also been brought out through the analysis of the proficiency tests in the regions (Parviz & Azizi, 2025).

Research Gap

Although there is a lot of literature on lexical features in language assessment, a single framework is not available that would allow comparing the lexical diversities of IELTS and TOEFL. Most of the studies have studied individual tests or studied a single measure (TTR), which may not be consistent in a study of varying text lengths. Also, little focus has been directed towards discourse-sensitive and multidimensional analysis that entails the integration of various measures of lexical diversity. This loophole limits an in-depth comprehension of the role of the various test formats in lexical use and could affect the scoring reliability and validity in high-stakes English proficiency tests.

ISSN: 2278-4012, Volume:15, Issue:2, June 2026

Objectives of the Study

- To compare and contrast the lexical diversity of responses in IELTS and TOEFL with different measures of computation.
- To assess the utility and effectiveness of different measures of lexical diversity for high-stakes testing in language.
- To investigate the implications of lexical variation in designing a test, scoring fairness and testing language proficiency.

Methodology

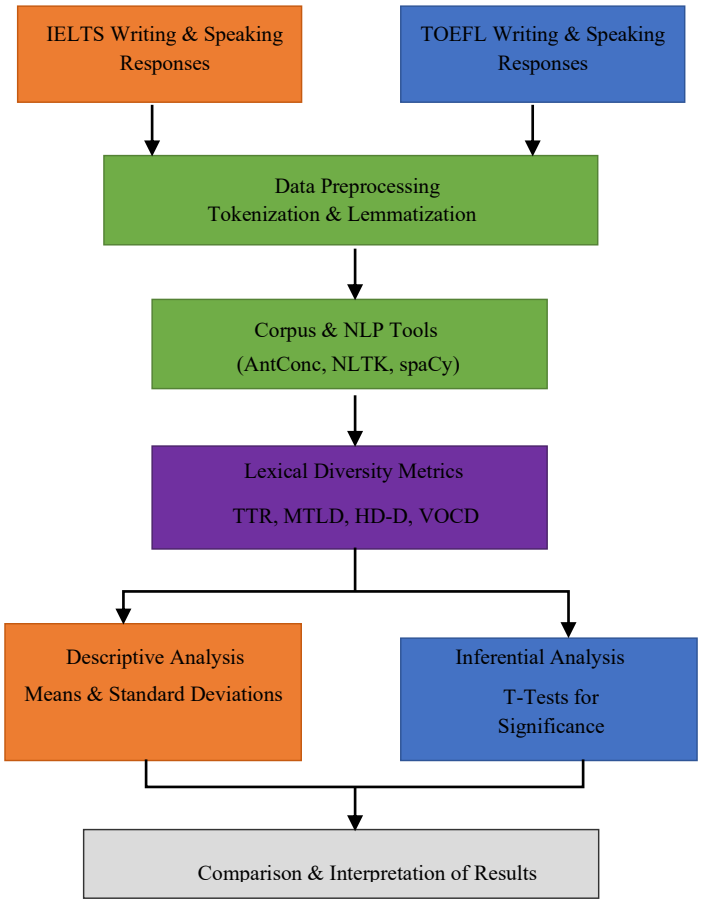

Figure 1: Research Methodology Framework for Comparative Lexical Diversity Analysis

Figure 1 illustrates the overall approach that will be employed in this study, which will have a logical sequence of data collection to statistical analysis. It starts by the compilation of two similar corpora of IELTS and TOEFL writing and speaking responses and then proceeds to the systematic preprocessing with the help of NLP techniques (tokenization and lemmatization). Computations of different lexical diversity metrics using the cleaned datasets in the form of corpus tools and Python-based software are then computed using the clean datasets in terms of TTR, MTLD, HD-D, and VOCD. Descriptive statistics are then used to compare the values and determine the central tendency and variability, and inferential statistics are used to compare the difference between the two types of examination using independent sample t-tests to determine the significance of the difference between the two. Such a concatenated pipeline, as illustrated in Figure 1, provides a promise of methodological rigor, reproducibility, and a comprehensive study of lexical diversity among high-stakes tests of English proficiency.

ISSN: 2278-4012, Volume:15, Issue:2, June 2026

Research Design

The study design is also a quantitative, comparative, corpus research design-based study that follows a corpus research design to methodologically compare the lexical diversity of two high-stakes English proficiency tests, i.e., IELTS and TOEFL. Corpus-based methodology makes possible the objective determination of the linguistic peculiarities by analyzing actual texts produced by learners as compared to the subjective one. The differences and similarities in the lexical patterns of use in the two types of examination, both in writing and speaking, are determined by the use of a comparative approach. The design can be replicated and is statistically rigorous since it involves a combination of computational linguistics and the standardized measures of lexical diversity. The research enables a comparative examination of the lexical richness of two sets of examination in a systematic format, which can be achieved by viewing each of the sets as an independent corpus, and the research adds to the knowledge-based comprehension of the best practices in language assessment.

Data Collection

The data in this study will include the responses of the learners in the IELTS Writing Task 2 and speaking transcripts, as well as the TOEFL Independent Writing and speaking responses. These data are selected as long-term language production activities that include demonstration of lexical range and communicative competence by the candidates. A balanced sampling method will be employed to make the sample statistically reliable and representative, with approximately 100-200 scripts being collected for each examination. The origins of the answers are publicly available test preparation materials, institutional repositories, and anonymized learner corpora that ensure the adherence to ethics and validity of the data. The comparability is ensured by only comparing answers of the same proficiency levels and number of words. The texts have also been pre-processed to remove any extraneous data such as annotations, comments by the examiners, and variations of the transcription, therefore creating uniformity in the datasets.

Tools & Software

This is analyzed using a combination of the corpus linguistics tools in addition to the natural language processing (NLP) models to extract and compute the lexical diversity measures. AntConc is applied to concordance, frequency distribution, and lexical profiling and provides a background of the vocabulary usage patterns. Parallel to these Python-based systems such as NLTK and spaCy, these systems are used to process text, tokenize, lemmatize, and automatically compute lexical indices. Through these programs it is possible to effectively process large volumes of datasets and ensure consistency of linguistic processing. In addition, special lexical diversity calculators and special scripts are embraced in calculating certain more complex measures such as MTLD, HD-D, and VOCD. These tools combined encourage accuracy in analytics and lead to scalable and reproducible research.

Metrics Used

The research employs four popular measures to comprehensively measure the lexical diversity: Type-Token Ratio (TTR), Measure of Textual Lexical Diversity (MTLD), Hypergeometric Distribution D (HD-D), and VOCD. TTR provides a crude measure of the lexicon variance by calculating the number of different words divided by the total words, but it is text length dependent. MTLD manages to overcome this deficiency by measuring the mean length of sequence word strings, which have a constant value of TTR and which are more valid with text of varying sizes. HD-D is a probabilistic model of lexical variety based on random sampling and thus more robust when it comes to shorter texts. VOCD goes an extra mile

ISSN: 2278-4012, Volume:15, Issue:2, June 2026

to simulate vocabulary development using curve-fitting. The various measures employed give a multidimensional measure of lexical richness and minimize the bias of any one measure.

Data Analysis Techniques

The data is analyzed using both descriptive and inferential statistical methods in order to enable its assessment in detail regarding the lexical diversity tendencies. The central tendencies and variability of the data sets in IELTS and TOEFL are summarised through descriptive statistics (mean and standard deviation) for each metric. This provides some preliminary indications of differences in lexical prosperity in the two kinds of examination. The statistical significance of differences is tested using independent sample t-tests in order to perform the inferential analysis. Hypotheses testing the significance of the differences in lexical diversity measures of IELTS and TOEFL responses can be tested using this. The descriptive and inferential methods will be combined to make sure that the results are well-interpreted and can be used to make evidence-based conclusions on the topic of comparative language performance.

Results And Analysis

Descriptive Analysis

The descriptive statistics give a first look at the lexical diversity in the IELTS and TOEFL data sets. The mean of all measures indicates overall lexical richness, and the SDs indicate variability in candidates' responses. Table 1 shows that, across most measures, IELTS responses exhibit slightly higher lexical diversity than TOEFL responses.

Table 1: Lexical Diversity Comparison Between IELTS and TOEFL

Metric	IELTS (Mean ± SD)	TOEFL (Mean ± SD)
TTR	0.62 ± 0.05	0.58 ± 0.04
MTLD	78.4 ± 10.2	70.1 ± 9.5
HD-D	0.84 ± 0.03	0.80 ± 0.04
VOCD	92.6 ± 11.8	85.3 ± 10.7

The findings show that IELTS test takers have a wider vocabulary, as evidenced by higher TTR and MTLD scores. Similarly, scores in HD-D and VOCD show a greater consistency in lexical variation in IELTS answers. However, the standard deviations are quite small in the examination forms, indicating moderate consistency across the datasets.

Comparative Analysis

A comparative lexical diversity analysis indicates that IELTS and TOEFL differ significantly in vocabulary use. IELTS responses are also more fruitful in lexis; this can be attributed to the fact that IELTS writing and speaking and tasks are open-ended. The vocabulary may be more diverse, as applicants are typically required to express their views, justify their beliefs, and elaborate on their ideas.

In contrast, responses on the TOEFL, particularly during integrated tasks, have been found to indicate a more controlled, ordered use of lexical items. The source material often constrains the candidates' lexical opportunities, as will be required to synthesize the information through reading and listening. This leads to relatively fewer lexical diversity measures. In addition, task variation is evident: speaking responses exhibit less lexical diversity than writing tasks due to time constraints and spontaneous language production. Generally, IELTS appears to encourage the more varied lexical production, whereas TOEFL is concerned with the accuracy and the combination of information.

ISSN: 2278-4012, Volume:15, Issue:2, June 2026

Statistical Significance

Independent-samples t-tests were used to determine whether differences in lexical diversity were statistically significant for each metric. The null (H_0) hypothesis will be that there is no considerable difference between IELTS and TOEFL lexical diversity scores, and the alternative hypothesis (H_1) will be that there is a considerable difference.

The t-test results reveal that the differences in scores for MTLD, HD-D, and VOCD are statistically significant ($p < 0.05$), indicating that IELTS responses exhibit greater lexical diversity than TOEFL responses. TTR also shows a difference but is less sensitive to text length, i.e., it is not as statistically sound as other measures. These findings confirm that the difference in lexical abundance between the two examination formats is not due to random variation but rather to differences in test design and task structure.

Interpretation

The findings show that IELTS encourages greater lexical diversity than TOEFL, primarily because of its tasks and assessment framework. Both the IELTS writing and speaking activities offer greater freedom of expression, allowing candidates to use more vocabulary. TOEFL, on the contrary, has incorporated tasks in which the candidates have to rely on the input materials given, and this may suppress the variability in lexical.

Time pressure is another factor that influences lexical production. The speaking tasks in both tests exhibit lower lexical diversity due to limited time for planning. Still, this effect is more pronounced in the TOEFL, as the tasks are designed to elicit specific responses. The scoring rules may also influence candidates' behavior, as during high-stakes tests test-takers are more inclined to focus on clarity and accuracy rather than experimenting with words.

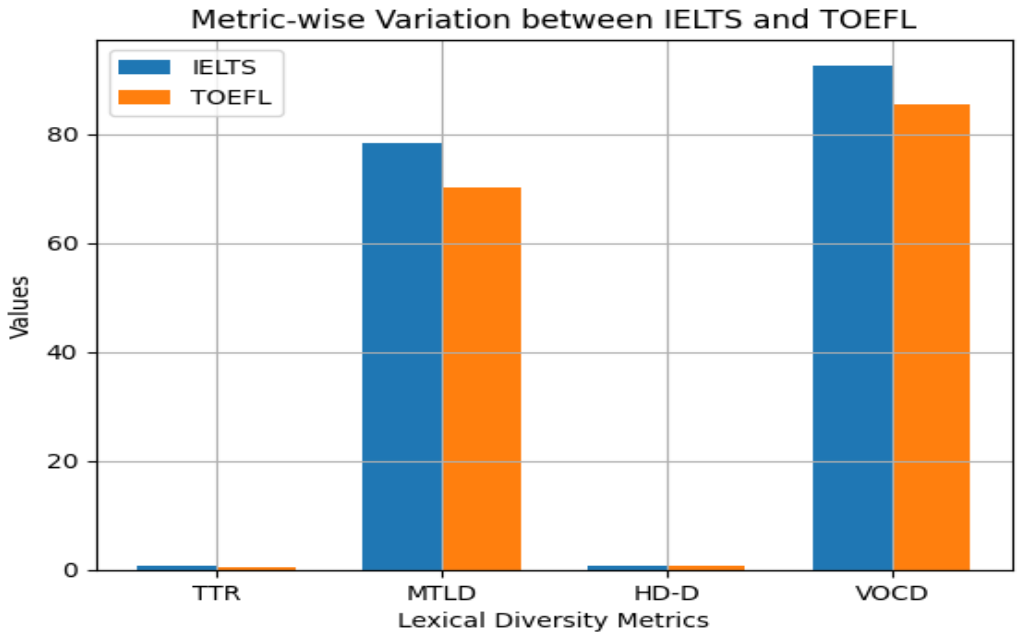

Figure 2: Metric-wise Variation between IELTS and TOEFL

The differences in TTR, MTLD, HD-D, and VOCD shown in Figure 2 between IELTS and TOEFL indicate that IELTS scores much higher on all measures.

ISSN: 2278-4012, Volume:15, Issue:2, June 2026

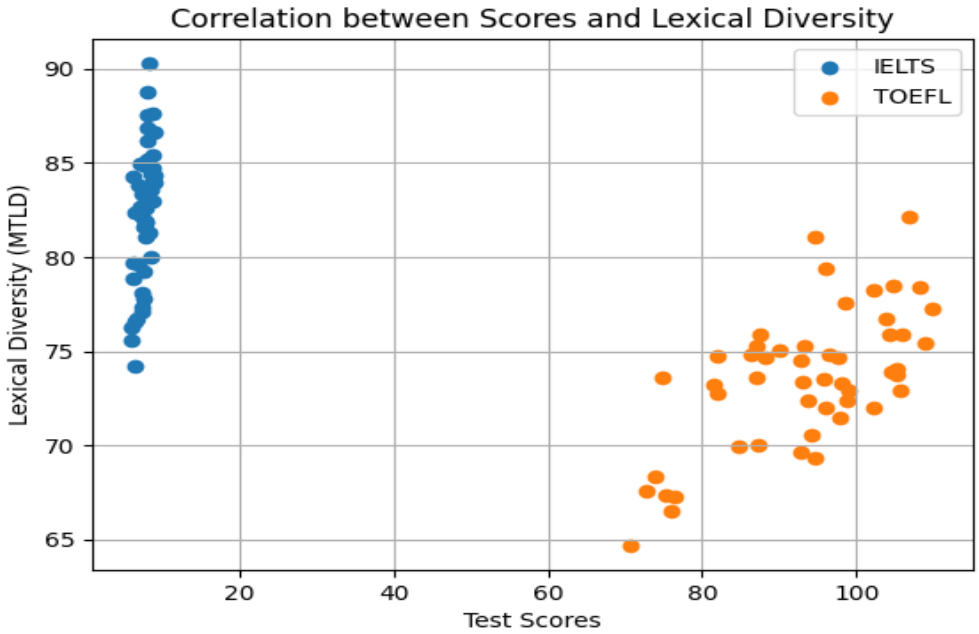

Figure 3: Correlation between Scores and Lexical Diversity

As shown in Figure 3, which depicts the correlation between candidate scores and lexical diversity measures, higher lexical diversity is associated with higher performance, particularly in IELTS responses.

Discussion

The results of the current research suggest that IELTS responses tend to have greater lexical variety than those in TOEFL, differences that differences in task formulation and response flexibility can explain. Activities in the IELTS, especially in the writing and speaking parts, are open-ended, allowing candidates to present their own opinions, make arguments, and elaborate on them without much restriction. The stimulating effect produced by this form provokes the expansion of the lexical repertoire, as the result of which the scores of most measures, such as MTLD, HD-D, and VOCD, increase. In turn, TOEFL activities, and in particular integrated writing and speaking, require that the candidates synthesize the information read and listen to. This inevitably restricts the vocabulary of lexical choice, since the lexical material of the source text is more likely to be employed by the same test takers, and creates greater control over language usage and a repetitive quality. In addition, time constraints and the structured response format in TOEFL also limit opportunities for lexical change, promoting more standardized language production.

Such differences have great implications for language evaluation and pedagogy. Given the test design, the results show that a balanced test must be constructed that measures both the lexical and task-specific use of language and is equitable compared to other tests. The scoring rubrics may require revision to account for the contextual constraints of all exams, particularly by recognizing that, given the narrower lexical range in integrated tasks, this does not necessarily indicate lower competence. The results prove that teachers and ESL students need to consider certain pre-preparation techniques: students who train to take the IELTS test need to be taught to expand their vocabulary, to be expressive, and to mix the material work with and the original information; students who train to take the TOEFL test need to be taught to be accurate, to paraphrase, and to mix the material that use and the original information. Overall, the validity and effectiveness of high-stakes English proficiency assessments may be enhanced by incorporating awareness of lexical diversity into assessment design and English instruction.

Conclusion

The paper has explored IELTS and TOEFL lexical diversity using a corpus-based correlation analysis and various computational measures of lexical diversity. The results establish that IELTS is more likely to promote lexical diversity due to its open-ended nature, in which tasks are set. In contrast, the TOEFL is more controlled and structured, particularly in integrated tasks. The results emphasize the significant role of test format in determining linguistic performance and the need to apply multiple measures to obtain a comprehensive analysis of lexical richness. The study contributes to the practice field of linguistics and language testing by comparing the lexical variety of two globally recognized proficiency tests. It has practical implications for improving test design, ensuring scoring fairness, and offering specific teaching strategies for ESL learners. However, the study has certain limitations, including a relatively small sample size; the results are situational and may not be generalizable to all test groups. Future studies may broaden the context by using larger samples, comparing speaking and writing in detail, and employing AI-based lexical analysis methods to improve the accuracy and scalability of language evaluation studies.

References

Abbaspour, E., & Mathew, P. (2025). Evaluating Cohesion as a Predictor of Writing Quality: An Analysis of Local, Global, and Text-Level Indices in IELTS Writing Task 2. *International Journal of Practical and Pedagogical Issues in English Education, 3*(4), 97-117. https://doi.org/10.22034/ijpie.2025.535152.1116

Chen, L., & Luo, Q. (2025). A comparative study of text characteristics of CET-6, IELTS, and TOEFL reading passages based on computational tools. *Journal of English for Academic Purposes, 77,* 101556. https://doi.org/10.1016/j.jeap.2025.101556

Gong, K. (2023). Challenges and opportunities for spoken English learning and instruction brought by automated speech scoring in large-scale speaking tests: a mixed-method investigation into the washback of SpeechRater in TOEFL iBT. *Asian-Pacific Journal of Second and Foreign Language Education, 8*(1), 25. https://doi.org/10.1186/s40862-023-00197-2

Holmberg Sjöling, C. (2025). The effect of lexical complexity on grading of Swedish EFL learners' texts during high-stakes exams. *International Journal of Learner Corpus Research, 11*(2), 245-275. https://doi.org/10.1075/ijlcr.23038.hol

Khan, A., & Javed, M. (2025). Language Testing and Assessment: Validity and Reliability in English Proficiency Exams. *Contemporary Journal of Social Science Review, 3*(1), 1950-1959. https://doi.org/10.12345/t0449k65

Nishizawa, H. (2024). Authenticity of academic lecture passages in high-stakes tests: A temporal fluency perspective. *Language Testing, 41*(4), 792-816. https://doi.org/10.1177/02655322241262453

Parviz, M., & Azizi, M. (2025). The MSRT: a critical review of english proficiency in Iran. *Discover Education, 4*(1), 226. https://doi.org/10.1007/s44217-025-00662-9

Qian, L. (2023). Use of lexical features in high-stakes tests: Evidence from the perspectives of complexity, accuracy and fluency. *Assessing Writing, 57,* 100758. https://doi.org/10.1016/j.asw.2023.100758

Saeedi, Z., Tajeddin, Z., & Tadayon, F. (2025). Assessment of English as a Lingua Franca and Its Principles: A Research Synthesis. *International Journal of TESOL Studies, 7*(3). https://doi.org/10.58304/ijts.250711

Tao, X., & Aryadoust, V. (2024). A multidimensional analysis of a high-stakes English listening test: A corpus-based approach. *Education Sciences, 14*(2), 137. https://doi.org/10.3390/educsci14020137

Thi Thao Le, L., Ho, N.T.P., Trang, N.H., & Ha, H.T. (2025). Revisiting the Lexical Differences Between Academic and General Training IELTS Reading Tests. *SAGE Open, 15*(3), 21582440251362269. https://doi.org/10.1177/21582440251362269

Wang, P., & Cheng, L. (2025). The impact of TOEFL iBT preparation on Chinese test-takers' perceptions of integrated speaking and writing design. *Critical Inquiry in Language Studies, 22*(1), 64-87. https://doi.org/10.1080/15427587.2024.2448815