# "LEXICAL COMPARISON BETWEEN THE COMMON EUROPEAN FRAMEWORK OF REFERENCE FOR LANGUAGES AND THE FLESCH-KINCAID"

Muhammad Hammad Hussain Shah[1], Dr. Shahid Nawaz*[2], Dr. Shazia Bukhari[3]

**Original Article**

1. Assistant Professor of English, Emerson University Multan.
Email: hammadbukhari335@gmail.com

3. Assistant professor of English, Department of Linguistics, The Islamia University Bahawalpur
Corresponding Email: shahidnawaz@iub.edu.pk

3. Lecturer in English, Department of English, University of Central Punjab, Lahore
Email: shazia.bukhari@ucp.edu.pk

**Abstract**

*This article presents the results of an exploratory study pertaining to the use of the Common European Framework of Reference (CEFR) for Languages and its Text Inspector; an automated tool measuring proficiency level, text complexity, and readability. This tool alsoanalysed text based on Flesch reading Ease and Flesch-Kincaid Grade. The assessment of the text in the current study is based on seven text parameters: CEFR levels, types, tokens, sentence count, Flesch reading ease and Flesch-Kincaid Grades. Ten academic texts at tertiary level (Essays) were analysed to find how specific parameters collaborate with each other. As shown by the findings, conventional readability equations, such as the Flesch-Kincaid Ease and Grades, are not as accurate predictors of text quality as are the CEFR text characteristics. A comprehensive approach for selecting appropriate academic texts for a given audience may be strengthened by using the discovered connection between the text complexity characteristics.*

**Keywords**: CEFR, Text Analysis, Flesch-Kincaid

## Introduction

It is believed that determining the level of complexity of a text can significantly aid in language acquisition. The degree to which a text is both complicated and varied can be characterised by using the term "text complexity." It is founded on aspects of language that can be categorised as morphology, lexicon, syntax, and discourse respectively. Lu 2010; Lu, 2017; McNamara et al., 2014; Kyle & Crossley 2018; Bultéand Housen 2018, conducted studies and found that certain factors are better indicators of language competency than others. It is a well-established fact that the qualities used to judge the level of difficulty of writing done in one's native language as opposed to writing performed in a language for which the writer does not have a native proficiency are distinct from one another.

Although there are several tools to measure a student's proficiency in a foreign language, research shows that these tests often fail to accurately gauge a student's proficiency in English for Speakers of Other Languages (ESL) (Vinogradova et al., 2019). The version of the student essay with feedback produced by the LexInspector system (Figure 1), which was designed for the learner corpus but quickly proved to be of little use to student authors, exemplifies the fact that these programmes do not always provide feedback that is understandable to non-experts.

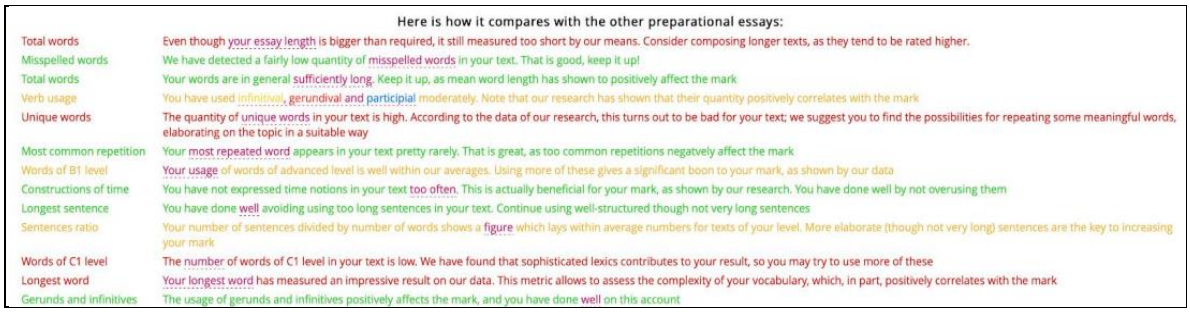



Figure 1. LexInspector Interface Automated feedback for the essay in REALEC presented at https://linghub.ru/inspector

In contrast to it, the Text Inspector designed by the (EVP) English Vocabulary Profileof the (CEFR) Common European Framework of Reference for language is much better and diverse in its mechanics for analysing text complexity and diversity (Figure 2).

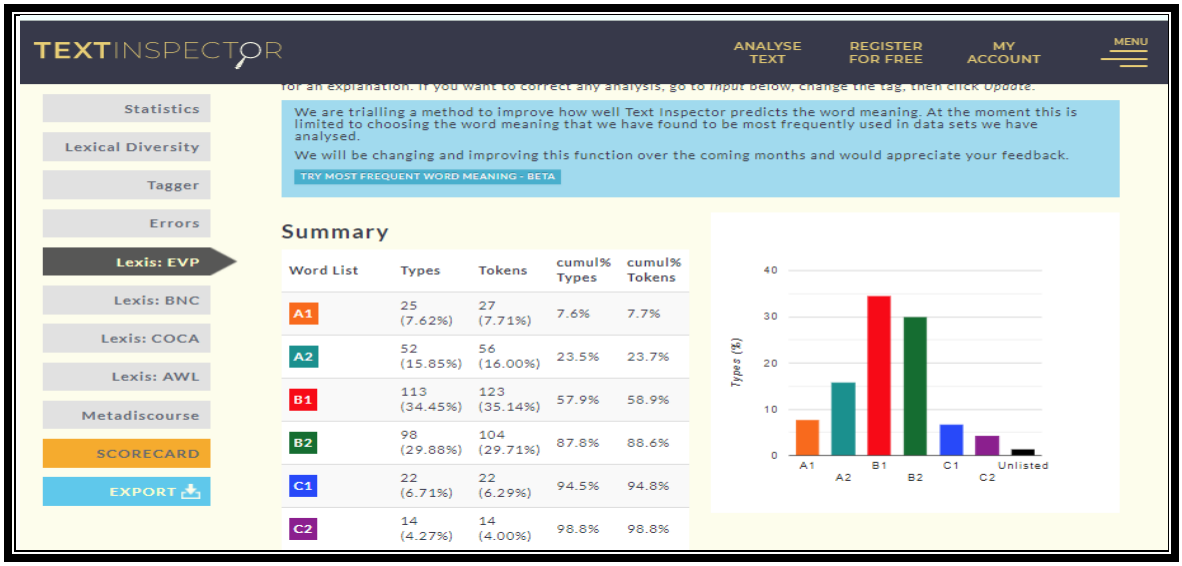



Figure 2. Text Inspector for the written text

Additionally, some automated feedback tools, such as the Coh-Metrix online tool (McNamara et al., 2014), and the L2 Syntactic Complexity Analyzer, just provide the values of features without providing any context or recommendations on how to make a text better (Lu, 2010). In order to show the author of the text how to enhance certain features of the text complexity that have been deemed lacking, this study offers a user-friendly version of the CEFR online tool.

There have been a number of studies that have investigated the use of CEFR descriptors for the authoring of assessments and the creation of rating scales within different national contexts. The use of learner corpora in SLA research seems to have a lot in common with these studies, which appear to have a lot in common with studies on second language acquisition (SLA). Understanding the proficiency level of the learners with some degree of accuracy might help in understanding the results of the study; consequently, one of the challenges in SLA research is determining how to assess both the proficiency level of the learners and the performances that are being examined.Numerous studies

have investigated the usage of CEFR descriptors for writing evaluation and rating scale construction within native context. These studies seem to have a lot in common with second language acquisition (SLA) studies and the use of learner corpora in SLA research. Understanding the learners' proficiency with some degree of accuracy might assist in comprehending the study's results. Hence, one of the challenges in SLA research is assessing both the learners' proficiency level and the performances being investigated. Since the CEFR is so widely used, its proficiency scale has become an extremely attractive tool since it provides a more accurate and dependable approach to determine the stage at which the learners of interest are than utilising general categories of learners like "beginners," "intermediate," or "advanced."As a consequence of this, the CEFR tool for evaluating the performances of individuals learning a second (L2) or foreign language has contributed to an improvement in the quality of student placements. It is more apparent than earlier approaches that were used to characterise learners' competency since it does not look at how the CEFR scale compares in diverse circumstances.

To summarise, there aren't many studies that compare the CEFR descriptors, the Flesch reading score, and the Flesch-Kincaid grade against one another. The present research aims to address the aforementioned concerns by addressing the following objectives and questions.

**Research objectives**

The objectives of the study were:

1. To measure the CEFR proficiency of the students' reflection
2. To analyse the lexical features of the text of the students' reflection
3. To determine the Flesch reading ease of the text

**Research Questions**

1. What is the CEFR proficiency of the students' reflection?
2. What are the lexical features in the written text of students' reflection?
3. What is the Flesch reading ease of the text in students' reflection?

**Background of the Study**

Significant progress has been made with the CEFR in both Europe and the rest of the globe (Byram & Parmenter, 2012). It tries to establish goals for university-level foreign language proficiency (Read, 2019). The CEFR is used by the UK to establish the English language criteria for immigration (Read, 2019). Vietnam adopted the CEFR as its official measure of English language proficiency in 2005. (Wu, 2012). Malaysia adopted the CEFR as a component of its English language curriculum in accordance with the English Language Education Roadmap for Malaysia (2015–2025), ten years after Vietnam (Ahmad, Hamid, & Renshaw, 2019). After being modified to cater to the requirements of Japanese students, the Common European Framework of Reference (CEFR) was given the designation "CEFR-J" in Japan.This framework was adjusted to the Japanese context (Negishi & Tono, 2014). (Read, 2019). China took a different strategy, proposing creating a Common Chinese Framework of Reference for Languages (CCFR) with a focus on teaching English (Jin et al., 2014). Education In response to the CEFR, New Zealand took the initiative to promote educational possibilities across the world (Read & Hirsh, 2005). The CEFR was favoured by Elder and O'Loughlin (2007) as the most suitable one for teaching programmes in Australia. According to Read (2019), its main goal is to increase the effectiveness and coherence of foreign language education, especially English.

According to Ahmad, Hamid, and Renshaw (2019), the CEFR has become a worldwide standard for language teaching. According to Rizvi and Lingard (2010), educational goals have been cited as an issue with human capital expansion. Human capital aspires to match the economic needs of the global marketplace. One of the graduates' potential economic strengths in international forums is their knowledge of and proficiency in many languages (Steiner-Khamsi, 2016). As a result, the Council of Europe (2008) advised that the global forum should outline all exams, examinations, and evaluation processes in accordance with the CEFR.

Several standardized foreign language proficiency tests have had their scores mapped into the CEFR (Figueras & Noijons, 2009; Martyniuk, 2010). The CEFR has also been used more often in teacher training programmes, foreign language curriculum, and instructional resources (Jones & Saville, 2009). In order to apply the CEFR in curriculum design, pedagogy, and assessment, EFL instructors are obliged to modify it (Moonen et al., 2013). Teachers use the CEFR to figure out the grades of their students (Fleckenstein, Leucht, &Köller, 2018). Runnels (2013) discovered that instructors in Japan utilised the CEFR-J to rate students' levels of reading and listening proficiency in a specific EFL course. According to Fleckenstein et al. (2018), assessment practise may use the CEFR.

An online application called Text Inspector measures the lexical variety of each text under study. The authors define lexical diversity as "the variety of various terms employed in a document" (McCarthy & Jarvis, 2010) andText Inspector evaluates MTLD and VOCD. Since all of the texts in the corpus under study are around 400 words, lexical diversity metrics are thought to be trustworthy and insensitive to the length of the texts under study. The MTLD and VOCD measuring Perl modules created by Aris Xanthos serve as the foundation for the Lexical Diversity tool that Text Inspector uses (Text Inspector). A left-to-right text order and a right-to-left text order MTLD are both carried out twice. Each pass produces a weighted average and variance, and the final value is calculated by averaging the two averages (the two variances are also averaged). This option allows the user to choose whether or not the average that is provided should be weighted (values: "within and between," "within only," or not) based on the possibly varying amount of observations in the two passes. The VOCD approach entails selecting 35, 36,...,49, and 50 tokens at random from the data, and then computing the average type-token ratio for each of these lengths of tokens.The matching parameter value is presented as the outcome of the diversity measurement. The process may be averaged after being conducted numerous times (Text Inspector).

**Explanation of Flesch Reading Ease score**

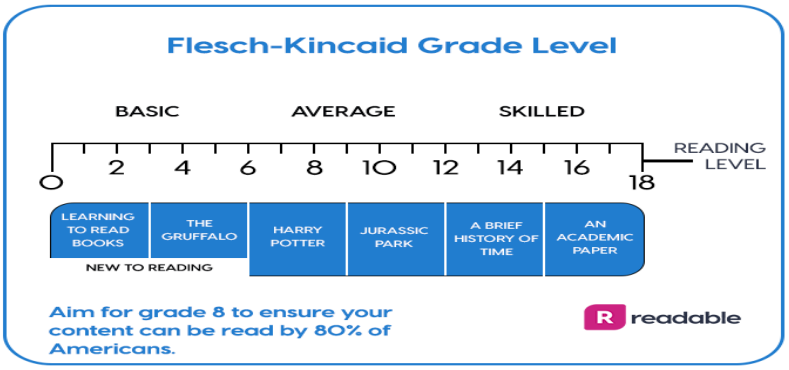
A text is given a readability score using the Flesch Reading Ease scale, with 100 being the best possible score. A score of 70 to 80 corresponds to the eighth grade. This indicates that the material ought to be pretty simple for the typical adult to read. Rudolf Flesch created the equation in the 1940s. He worked as a consultant for the Associated Press, creating strategies to make publications easier to read. Over 70 years later, many people still use the Flesch Reading Ease, including policy writers, marketers, and research communicators. They all use it to determine how easily a text will be understood and engaged with. A piece of content's Flesch score indicates how readable it is. The US educational system is reflected in Flesch reading ease and Kincaid Grade Level. Although the same units are used in both calculations, the weightings for these units vary between the two tests, leading to different readability scores.

A piece of text is easier to read the higher the reading score. Note that this differs from most readability ratings, where a lower score indicates an easier read. A text with a reading score of 60 to 70, for instance, corresponds to a grade level of 8 to 9, thus, 13 to 15-year-olds should be able to

understand it. A conversion table is required in order to understand a Reading Ease score. This translates the grade level of the score.

**The Flesch Kincaid Grade**

A popular readability method that determines the approximate reading grade level of a book is the Flesch-Kincaid Grade Level. It was developed by the US Navy in collaboration with the Flesch Reading Ease. In the past, a table had to be used to convert the Flesch Reading Ease score to the reading grade level. In order to make the revised version simpler to use, it was developed in the 1970s. It was used by the Navy for its training-related technical documents. It now serves a considerably larger range of purposes. A work that has a Flesch Kincaid level of 8 means that the reader must have reading level skills equivalent to or higher than grade 8. Even if they are an accomplished reader, this means that the text will take less time to read. The US grade level of schooling is comparable to the Flesch-Kincaid Grade Level. It demonstrates the level of knowledge needed to comprehend a text.




https://readable.com/readability/flesch-reading-ease-flesch-kincaid-grade-level/

Flesch, 1948) first developed the Flesch-Kincaid Grade Level (FKGL), which has since been widely utilised in the medical field despite never having been proven to affect fundamental understanding (Shardlow, 2014; Kauchak & Leroy, 2016). The average number of words per sentence and the average number of syllables per word are combined by FKGL to get the score.

The first version of the Flesch-Kincaid readability formula was provided in Flesch's doctoral dissertation in 1943. It was co-authored by Kincaid. It considers things like the average number of words used in a phrase, the number of affixes, and the number of times certain people are mentioned. The data from the McCall-Crabbs Standard Test Lessons in Reading (McCall-Crabbs, 1926) were used to develop the formula. This test is a standardised reading exam that is administered to students in grades 3 through 7. An average of eight reading comprehension questions are included in each of the 376 sections that make up the McCall-Crabbs tests. A grade level and an indication of the difficulty level are provided for each lesson.

Flesch created a model using this data to predict the grade of third through seventh graders who correctly answered 75% or more of the passage-related questions. It was originally intended as a tool to help students keep tabs on their progress. Five years later, he came up with a new method he termed the Reading Ease Score (Flesch, 1948). He adjusted the model by recalculating the coefficients and making use of common textual metrics like syllable count and sentence length. Similar to the original research, this new strategy was tested on children and was based on the McCall-Crabbs Standard Test Lessons in Reading.

The formula commonly utilised in text simplification assessment is Flesch-Kincaid Grade Level, a modified Reading Ease formula with revised weights. Three decades later (Kincaid et al., 1975), the formula was explicitly derived to evaluate the readability of technical texts for military personnel. The reading comprehension of 531 Navy personnel in four technical training schools at Navy bases was evaluated using the Gates-McGinitie reading test's comprehension portion and 18 texts from rate training manuals. This formula was developed specifically for use by the Navy by professionals from the Navy using facts pertaining to the military. The readability of text in a variety of settings has, on the other hand, consistently been evaluated using this method. For instance, it is often used by medical writers as a guide for the development of writing in the medical profession, and even Microsoft Word integrates both the Flesch Reading Ease and the FKGL ratings (Shedlosky-Shoemaker et al., 2009).

**Data Collection and Research Procedure**

Text complexity testing is a crucial component in assessing learners' proficiency in foreign language output. It is based on lexical characteristics, which may be further broken down into tokens, kinds, Flesch Reading Ease, Flesch-Kincaid Grade, and CEFR Proficiency Level. The study was conducted using a corpus of essays authored by students studying English at a Pakistani public institution. The method of convenience sampling was utilised to get the data. Only 10 of the original group students consented to write a reflection paper on the topic of technology. The gathered data (in the form of essays) was first typed into the computer using the transcribed text from the papers in order to answer the study's goals. By comparing every text in written and typed form, the data's veracity was ascertained. It is helpful to instructors of English as a second or foreign language in addition to learners. The analysis generated sufficient data for analysing L2 interference in learners' texts in addition to its primary objective of providing automated lexical analysis to English text learners' texts.

**Results and Discussion**

Table 1: CEFR EVP Proficiency Level and Lexical types used in the essays

| | CEFR Level | A1 type count | A2 type count | B1 type count | B2 type count | C1 type count | C2 type count |
|---|---|---|---|---|---|---|---|
| N1 | C1+ | 70 | 23 | 26 | 12 | 4 | 1 |
| N2 | C1 | 79 | 27 | 28 | 20 | 2 | 1 |
| N3 | C2 | 93 | 39 | 33 | 18 | 20 | 0 |
| N4 | C1+ | 74 | 22 | 33 | 9 | 4 | 0 |
| N5 | C1+ | 83 | 32 | 28 | 17 | 4 | 3 |
| N6 | C2 | 99 | 40 | 30 | 24 | 6 | 1 |
| N7 | C1 | 101 | 34 | 31 | 20 | 3 | 0 |
| N8 | C1+ | 89 | 27 | 22 | 15 | 2 | 3 |
| N9 | C2 | 81 | 37 | 36 | 38 | 6 | 0 |
| N10 | C1+ | 104 | 36 | 37 | 18 | 6 | 1 |

The data in the Table 1 demonstrates that only three students coded as N3, N6, and N9 scored C2 proficiency Level on the CEFR scorecard, and five students N1, N4, N5, N8, and N10 scored C1+ level on the scorecard. In comparison, only two students scored C1 levels coded as N2 and N7. The

data shows the students who got C2 level had diversity in their language use compared with the others who scored C1+ and C1 level. The table demonstrates that students who achieved C2 level contrastively did not use C2 word types in their essays in comparison N1, N5, N8, and N10 who used word types of C2 level in their essays, while most of the words (3-word types C2 type count) were used by N5 and N8.

It is interesting to note that the students who achieved the highest scores in C2 did not use C2 type counts or rarely used one-word C2 type count (N6). Other columns of the data demonstrate that most of the words were used at level A1 type count, and the use of word type count decreased as the difficulty level, or proficiency level increased from A1 to C2. As N1 used 70 A1 type count and 23 A2 type count but contrastively use of B1 type count increased to 26, but again word type count deceased respectively in B2, C1 and C2 type count from 12, 4 1 respectively. In the lexical analysis of N2, it can be observed that the use of words according to the proficiency level decreased from 79, 27, 28, and 20, 2 to 1, respectively from A1 to C2 type count. Lexical analysis of cases of N3, N4, N5, N6, N7, N8, and N 9 are quite similar, where word type count level decreased as the proficiency level increased.

Yusoff et al. (2022) mentioned in their study that the word difficulty level also increases when proficiency level increases. Yusoff et al. (2022) measured the proficiency of students' vocabulary based on Nation and Beglar's (2007) vocabulary size test, and they measured VST (2007) vocabulary proficiency level with the CEFR testing tool of the text-inspector. It can be incurred from the data that proficiency on the scorecard is not directly proportional to the individual lexical level of the word type count used in the CEFR descriptors. Hence, CEFR uses several metrics to analyse the text. A few metrics have been discussed and described in table 2 in the current study.

**Table 2: CEFR EVP Proficiency Level and Lexical Tokens and Types used in the essays**

| | CEFR Level | Sentence count | Token count | Type count |
|---|---|---|---|---|
| N1 | C1+ | 19 | 305 | 152 |
| N2 | C1 | 26 | 409 | 179 |
| N3 | C2 | 14 | 473 | 213 |
| N4 | C1+ | 22 | 300 | 161 |
| N5 | C1+ | 28 | 454 | 179 |
| N6 | C2 | 31 | 462 | 212 |
| N7 | C1 | 43 | 484 | 193 |
| N8 | C1+ | 27 | 411 | 168 |
| N9 | C2 | 28 | 456 | 228 |
| N10 | C1+ | 31 | 448 | 216 |

Table 2 demonstrates lexical features and diversification of the text. All students in the study used a different number of sentences in the essays. The total number of sentences used by the students is mentioned in this table that demonstrates N1 used 19, N2 used 26, N3 14, N4 22, N5 28, N6 31, N7 43, N8 27, N9 28, and N10 wrote 31 sentences in their essays. However, the CEFR level comparison shows that N3 wrote an essay using 14 sentences and scored C2 level. In contrast to N3, N7 utilized 43 sentences and scored C1 level, while N7 sentences comprised three times higher than the N3.

Most of the token counts were used by N7, who used 484 tokens and 193 type counts and scored C1 level, contrastively N3, N6, and N9 used 473, 462, 456 tokens and higher proficiency levels than those who used the highest token counts in the essays.

The lexical features of the data conclude that higher proficiency is not dependent on more token and type count, not it is based on the number of sentences. However, it is based on the choice of words the students use in writing their essay or text material.

**Table 3: CEFR EVP Proficiency Level and Flesch Reading Ease and Grade**

| | CEFR Level | Flesch Reading Ease | Flesch-Kincaid Grade |
|---|---|---|---|
| N1 | C1+ | 49.63 | 10.32 |
| N2 | C1 | 51.87 | 9.93 |
| N3 | C2 | 37.15 | 16.47 |
| N4 | C1+ | 45.79 | 10.26 |
| N5 | C1+ | 51.74 | 10.07 |
| N6 | C2 | 48.33 | 10.22 |
| N7 | C1 | 65.54 | 6.91 |
| N8 | C1+ | 54.3 | 9.47 |
| N9 | C2 | 57.28 | 9.32 |
| N10 | C1+ | 57.52 | 8.83 |

Table 3 demonstrates the data based on Flesch reading ease and Flesch – Kincaid grade and its comparison with the CEFR level. As it has been discussed earlier in the literature, the higher the Flesch reading score, the easier the text to read, and the lower the Flesch – Kincaid reading score, the complex and difficult it is to read the piece of the text. However, the case is not the same here; N2 and N7 scored 51.87 and 65.54 on the Flesch reading ease scale respectively, while both scored C1 on the CEFR proficiency scale. There is a difference of 13.76 scores between both students. On the other hand, N3, N6, and N9 scored C2 on the CEFR scorecard while they scored 37.15, 48.33, and 57.28 Fleschreading scores, which shows there is the disparity in the results of Flesch scores as compared to the CEFR scores.

The same is the case with the Flesch – Kincaid grades, as N3, N6 and N9 are at the same level on the CEFR scorecard and scored C2 proficiency level, but the score on the Flesch – Kincaid reading grades is diversified from 16.47 to 10.22 and 9.32.So it can be concluded that the results of research question 3 are in line with the results of Tanprasert and Kauchak (2021) who examined the use of automated Flesch-Kincaid Grade Level (FKGL) metric combination for testing system output readability and concluded that FKGL should not be used to assess text simplification systems. They carried out a series of experiments to demonstrate that the FKGL score is susceptible to easy manipulation, which would result in a significant rise in the score while having no impact on the results of other automated tests.

Providing a standard foundation for creating language tests across Europe is one of the CEFR's key goals (Council of Europe, 2001). Consequently, using the CEFR descriptors for rating student performances with other rating scales, such as Flesch reading and Flesch - Kincaid grades, should ideally provide similar outcomes. Even with reference to the lexical level CEFR writing descriptors in non-native situations, there are few studies on this issue. It made sense to investigate if employing CEFR descriptors to evaluate students' writing yielded similar performances to other techniques. Scoring written essays is primarily interpretive and evaluative, depending on current educational practises and the individual's past experiences. Cumming, Kantor, and Powers (2001) Rating scales comprised of descriptors similar to those found in the various CEFR tables have been criticised for their use of "impressionistic terminology that is open to subjective interpretations" (Knoch, 2009, p. 277, citing Brindley, 1998; Watson Todd, Thienpermpool & Keyuravong, 2004).

The CEFR scales are not meant to be used in direct rating systems and should not be interpreted as such. These are known as proficiency scales instead of rating scales, because they serve an entirely different purpose. However, research has shown that raters who have a comprehensive knowledge of the CEFR scales and the meaning of the CEFR levels can confidently apply the CEFR descriptors to student ratings of their performances (Alanen et al., 2012; Huhta et al., 2014; Holzknecht, Huhta, & Lamprianou, 2018). The purpose of this study was to determine the extent to which two distinct and highly acclaimed rating scales, each of which used its own scales of assessment, agreed on the CEFR levels of writing performances written by Pakistani students who were engaged in higher education. The data indicate that rating scales differed depending on the individual.

**Conclusion**

In our study, we included an online application tool called CEFR, based on the outcomes of our previous research. This application is intended to be of use to English instructors and college students currently preparing for an English exam. ESL instructors, students, and linguists will hopefully find student text lexical analysis measures even more valuable. This research is the first one to our knowledge that explicitly compares CEFR-based evaluations of writing performances with the Flesch reading ease and Flesch - Kincaid grades rating at the tertiary level in Pakistan. This was done by Flesch and Kincaid. The findings offer us reason to believe that using CEFR descriptors associated with writing for the purpose of rating may generate results that are incomparable to those produced by other rating systems. In order to make direct use of CEFR descriptors for rating purposes, which require additional parameters and cannot be predicted from other online tools, this needs to be provided. The current research is just the first stage in assessing whether or not the CEFR descriptors produce equivalent assessment findings in a variety of educational settings. Despite having optimistic results, the study is still only the first step. Research along these lines seems particularly essential in light of the growing use of the CEFR in a wide range of language evaluations around the globe. As a result, subsequent research might apply the study's methodology to various contexts of assessment, focussing instead on linguistic talents other than writing.

**References**

Ahmad Afip, L., Hamid, M. O., & Renshaw, P. (2019). Common European framework of reference for languages (CEFR): insights into global policy borrowing in Malaysian higher education. *Globalisation, Societies and Education*, *17*(3), 378-393.

Alanen, R., Huhta, A., Jarvis, S., Martin, M., &Tarnanen, M. (2012). Issues and challenges in combining SLA research and language testing. *Collaboration in language testing and assessment*, 15-30.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

Bulté, B., &Housen, A. (2018). Syntactic complexity in L2 writing: Individual pathways and group trends. International Journal of Applied Linguistics, 28, 147–164.

Byram, M., & Parmenter, L. (Eds.). (2012). *The Common European Framework of Reference: The globalisation of language education policy* (Vol. 23). Multilingual matters.

Council of Europe. (2008). *Recommendation CM/Rec(2008)7 of the Committee of Ministers to member states on the use of the Council of Europe's Common European Framework of Reference for Languages (CEFR) and the promotion of plurilingualism*. Strasbourg, France: Council of Europe.

Council of Europe. Council for Cultural Co-operation. Education Committee. Modern Languages Division. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.

Cumming, A. H., Kantor, R., & Powers, D. E. (2001). *Scoring TOEFL essays and TOEFL 2000 prototype writing tasks: An investigation into raters' decision making and development of a preliminary analytic framework*. Educational Testing Service.

Elder, C., & O'Loughlin, K. (2007). ELICOS language levels feasibility study: Final report. Canberra: Department of Education. *Science and Training*.

Figueras, N., &Noijons, J. (2009). Linking to the CEFR levels: Research perspectives.

Fleckenstein, J., Leucht, M., &Köller, O. (2018). Teachers' judgement accuracy concerning CEFR levels of prospective university students. *Language Assessment Quarterly*, *15*(1), 90-101.

Flesch, R. (1943). Marks of readable style; a study in adult education. *Teachers College Contributions to Education*.

Flesch, udolph. (1948). A new readability yardstick. Journal of applied psychology, 32(3):221.

Holzknecht, F., Huhta, A., &Lamprianou, I. (2018). Comparing the outcomes of two different approaches to CEFR-based rating of students' writing performances across two European countries. *Assessing Writing*, *37*, 57-67.

Huhta, A., Alanen, R., Tarnanen, M., Martin, M., &Hirvelä, T. (2014). Assessing learners' writing skills in a SLA study: Validating the rating process across tasks, scales and languages. *Language Testing*, *31*(3), 307-328.

Jin, Y., Wu, Z., Anderson, C., & Song, W. (2014). Developing the Common Chinese Framework of Reference for Languages: Challenges at macro and micropolitical levels. In *Language Testing Research Colloquium, Amsterdam, the Netherlands, June 4* (Vol. 6).

Jones, N., & Saville, N. (2009). European language policy: Assessment, learning, and the CEFR. *Annual Review of Applied Linguistics*, *29*, 51.

Kauchak, D., & Leroy, G. (2016). Moving beyond readability metrics for health-related text simplification. *IT professional*, *18*(3), 45-51.

Kincaid, J. P., & Robert Jr, P. (1975). Fishburne, Richard L. Rogers, and Brad S. Chissom,"Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel". *Naval Technical Training Command Millington TN Research Branch*.

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing*, *26*(2), 275-304.

Kyle, K., & Crossley, A. S. (2018). Measuring Syntactic Complexity in L2 Writing Using FineGrained Clausal and Phrasal Indices. The Modern Language Journal, 102, 333–349

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. International Journal of Corpus Linguistics, 15, 474–496.

Lu, X. (2017). Automated measurement of syntactic complexity in corpus-based L2 writing research and implications for writing assessment. Language Testing, 34 (4), 493–511.

Martyniuk, W. (2010). Relating language examinations to the Common European Framework of Reference for Languages: Case studies and reflections on the use of the Council of Europe's draft manual. *Studies in language testing*.

McCarthy, P.M. & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. Behavior Research Methods, 42: 381. https://doi.org/10.3758/BRM.42.2.381

McNamara, D. S. ,Graesser, A. C., McCarthy, P., & Cai, Z. (2014). Automated evaluation of text and discourse with Coh-Metrix. Cambridge: Cambridge University Press.

Moonen, M., Stoutjesdijk, E., de Graaff, R., & Corda, A. (2013). Implementing CEFR in secondary education: impact on FL teachers' educational and assessment practice. *International Journal of Applied Linguistics*, *23*(2), 226-246.

Negishi, M., & Tono, Y. (2014, April). An update on the CEFR-J project and its impact on English language education in Japan. In *5th International Conference of the Association of Language Testers in Europe (ALTE), Paris, France, April* (pp. 10-11).

Read, J. (2019). The influence of the Common European Framework of Reference (CEFR) in the Asia-Pacific region. *LEARN Journal: Language Education and Acquisition Research Network*, *12*(1), 12-18.

Read, J., & Hirsh, D. (2005). English language levels for international students in tertiary institutions.

Rizvi, F., & Lingard, B. (2009). *Globalizing education policy*. Routledge.

Runnels, J. (2013). Student ability, self–assessment and teacher assessment on the CEFR–J's can do statements. *The Language Teacher Journal. JALT SIG Special Issue. Vol. 3 (5)*, 3-5.

Shardlow, M. (2014). A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications*, *4*(1), 58-70.

Shedlosky-Shoemaker, R., Sturm, A. C., Saleem, M., & Kelly, K. M. (2009). Tools for assessing readability and quality of health-related web sites. *Journal of genetic counseling*, *18*(1), 49-59.

Steiner-Khamsi, G. (2016). New directions in policy borrowing research. *Asia Pacific Education Review*, *17*(3), 381-390.

Tanprasert, T., & Kauchak, D. (2021, August). Flesch-Kincaid is Not a Text Simplification Evaluation Metric. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)* (pp. 1-14).

Text Inspector. Retrieved from http://textinspector.com/workflow/B3021C1A-706A-11E7-B233-AB44AFCE53D3.

Todd, R. W., Thienpermpool, P., &Keyuravong, S. (2004). Measuring the coherence of writing using topic-based analysis. *Assessing writing*, *9*(2), 85-104.

Vinogradova, O., Ershova, E., Sergienko, A., Generalova, S. (2019). AWARL (Automated Writing Assistant for Russian Learners) As a Computer-Assisted Language Learning Tool. Paper to be presented at Eurocall 2019, Louvain-la-Neuv, August, 2019.

William, A., & McCall, C. S. L. (1926). *McCall-Crabbs: Standard Test Lessons in Reading: Teacher's Guide and Answer Key*. Teachers College.

Wu, J. (2012). 18 Policy Perspectives from Taiwan. *The Common European Framework of Reference: The globalisation of language education policy*, *23*, 213.

Yusoff, Z. S., Gurmani, M. T. A., Sanif, S., & Noor, S. N. F. M. (2022). The Effect of Mobile-Assisted CEFR English Vocabulary Profile Word Lists on L2 Students' Vocabulary Knowledge. *Asian Journal of University Education*, *18*(2), 526-543.