The Impact of ChatGPT on the Development of Vocabulary Knowledge of Saudi EFL Students

Ahmad Ibrahim Mugableh
Department of English, College of Arts
Jouf University, Sakaka 72388, Jouf, Kingdom of Saudi Arabia
Email: ahmugableh@ju.edu.sa

Received: 11/20/2023         Accepted: 04/02/2024  Published: 04/24/2024

Abstract
This research was piloted to systematically compare the impact of ChatGPT-generated exercises versus traditional exercises on Saudi EFL students’ vocabulary size and the strength of word families. It aimed to answer the following questions: 1) To what extent does the integration of ChatGPT-generated exercises impact Saudi EFL students’ vocabulary size of word families? 2) To what extent does the integration of ChatGPT-generated exercises impact Saudi EFL students’ vocabulary size and strength of word families? This research holds significance because it provides educators with empirical results of the impact of ChatGPT’s role and function on the second language vocabulary development process. Sixty male Saudi students at Jouf University were selected by stratified random sampling. The design of this research is experimental. Consequently, samples were randomly assigned to different experimental conditions and groups (i.e., control and treatment groups). Computer-Adaptive Test of Size and Strength, as developed by Laufer and Goldstein (2004), was administered to both groups at the beginning and end of their studies in the IELP course with an interval of four months. The collected results from all participants (N=60) from both groups and in both intervals (i.e., pre-test and post-test) were clearly and concisely imported to Statistical Package for the Social Sciences SPSS. Multivariate Analysis of Variance was conducted. Results revealed that Saudi EFL students’ vocabulary development improved significantly in both groups. However, participants in the experimental group who experienced ChatGPT-generated exercises notably surpassed that of the control group in terms of vocabulary size and strength of word families.

Keywords: Computer-Adaptive Test of Size and Strength, ChatGPT, experimental design, Multivariate Analysis of Variance, Saudi EFL students, vocabulary development, vocabulary size and strength of word families

Introduction

Large Language Models (LLMs) such as Chat Generative Pre-Trained Transformers (ChatGPT) are a type of Artificial Intelligence (AI) technology that is nourished by a considerable amount of information, allowing them to mimic human-like conversation (Skrabut, 2023). ChatGPT has gained substantial popularity since its introduction in November 2022 (Teel, 2023; Flanagin et al., 2023). Despite being arguably the most widely embraced model among LLMs, ChatGPT is accompanied by counterparts such as Bing Chat, Google Bard, Meta’s Galactica, and Anthropic’s Claude (Webb, 2023; Hosseini et al., 2023). Microsoft has also launched Co-Pilot, an innovative generative AI applied across Microsoft 365 tools (Webb, 2023).

The versatile applications of ChatGPT quickly became evident. Due to its training in multiple languages, it exhibits a rapid and nuanced translation capability, considering cultural nuances within each language (Tate et al., 2023; Zhang et al., 2023). LLMs like ChatGPT contribute to curriculum development and educational content creation (Kasneci et al., 2023). For scientific writing and research, they assist in preparing manuscripts and analysing results. Notably, ChatGPT's language editing capabilities hold promise for aiding individuals with limited English proficiency (Májovský et al., 2023). Furthermore, it can generate appropriate metadata, indexing, and research findings summaries, thereby enhancing the dissemination and accessibility of information to a broader audience (Lund et al., 2023).

Scholars have delved into using ChatGPT, yielding divergent perspectives within the English language and teaching domain. While some proponents argue for its benefits, others point out its drawbacks and limitations. For instance, Xiao and Zhi (2023) qualitatively investigated students’ experiences with ChatGPT and their perceptions of its role in language learning at one of the international universities in China. Participants’ responses indicate that ChatGPT has the potential to be a valuable learning partner, assisting them in language-related tasks. Moreover, participants demonstrated critical judgment in assessing the quality of ideas generated by ChatGPT and adjusting prompts for enhanced learning. On the other hand, Ali (2023) quantitatively investigated how 58 Saudi EFL instructors in one of the government universities in the Kingdom of Saudi Arabia (KSA) perceived the advantages and difficulties of using ChatGPT. Results revealed that most EFL instructors expressed strong positivity toward ChatGPT's contribution to English language teaching and learning in higher education. However, some expressed concerns about issues such as plagiarism and overreliance on ChatGPT with little effort from the learners.

As mentioned earlier and others to be discussed later in existing literature, the studies focus the most significantly on students' and instructors' perceptions of using ChatGPT in English teaching and learning. Although these studies provide valuable insights into the subjective experiences and opinions surrounding incorporating language models in educational settings, several questions remain unanswered. To mention but a few, what impact does interacting with ChatGPT have on phonological, morphological, syntactic, semantic, and pragmatic development during the second language acquisition process? Teachers, students, academic institutions, and policymakers in a second language context should be aware of the influence of ChatGPT on the second language acquisition process to shape future language learning technology and instructional technique requirements of second language learners.

This research focuses on the semantic structure of second language learners' development, specifically on Second language vocabulary development. This is how individuals expand their vocabulary in a language other than their native or first language. This expansion often involves increasing the size and depth of vocabulary knowledge (Gutiérrez-Colón & Gutiérrez-Clellen,
2022). This research aims to systematically compare the impact of ChatGPT-generated exercises versus traditional exercises on Saudi EFL students’ vocabulary size and word family strength. The significance of this research is derived from the fact that it is inevitable to assess the possibilities, advantages or disadvantages of integrating such technology in the curriculum of teaching vocabulary for second language learners, particularly in this digital age that transforms education every day. This research aims to answer the following questions:

- To what extent does the integration of ChatGPT-generated exercises impact Saudi EFL students’ vocabulary size of word families?
- To what extent does the integration of ChatGPT-generated exercises impact Saudi EFL students’ vocabulary size and strength of word families?

**Literature Review**

**The Role of ChatGPT in Language Learning**

Incorporating technology in language learning has become vital in almost every context. Various technologies nowadays can provide learners with a unique language socialisation hotspot, immersing them in the language acquisition process and thus acquiring it quickly (Krietzberg, 2023). Furthermore, technology was also found to be beneficial from a psychological perspective in that it positively impacts learners’ motivation in language learning (Gikas & Grant, 2013).

Kasneci et al. (2023) indicated that ChatGPT can generate text, answer questions, and complete various language-related tasks with a powerful potential to offer a wide range of benefits and opportunities for language learners with different proficiency levels. Furthermore, ChatGPT can provide learners with a pedagogical tool to develop language skills such as writing and vocabulary acquisition via personalised practice materials and explanations (Kohnke et al., 2023). In addition, ChatGPT can assist language educators in writing, research, reports, and problem-solving tasks while offering discipline-specific language skills (Xiao & Zhi, 2023).

Shaikh et al. (2023) quantitively assessed the usefulness and effectiveness of ChatGPT for performing different English language learning tasks such as conversation, writing, grammar, and vocabulary. Participants were diverse in proficiency levels, educational levels, and nationalities. Analysis revealed that ChatGPT has the potential to expand participants' vocabulary, enhance their grammatical and syntactical structures, and improve their written conversational skills. Furthermore, responses also identified several limitations: While ChatGPT technology is compelling, it cannot replace human instruction entirely. Instead, this technology is designed to work collaboratively with human educators from a complementary perspective.

In another space of research, Athanassopoulos et al. (2023) investigated the effectiveness of ChatGPT in improving foreign language writing in terms of vocabulary and grammar for socially vulnerable populations such as refugees and migrants in one of the schools in south Greece. Participants were asked to write personal emails first and then draft them to ChatGPT to correct their written outputs regarding grammar and vocabulary. Then, all suggestions and revisions generated by ChatGPT were thoroughly examined by participants to prepare them for the following written tasks. Results revealed that ChatGPT feedback improved participants’ L2 writing skills regarding different qualitative characteristics, including total words, average word per sentence, unique words, and most frequent words.

ChatGPT, like other remarkable technologies around, is not exempted from scrutiny, as several researchers highlighted several drawbacks and limitations related to the potential for cheating and its impact on assessment. For instance, Xiao and Zhi (2023) argued that the responses
generated by ChatGPT are paraphrases of its sources without proper citations, which leads to plagiarism issues. Another potential drawback, as argued by Kohnke et al. (2023), is the inconsistency of the texts generated by ChatGPT in that responses generated by ChatGPT vary significantly based on how prompts (i.e., initial text input to ChatGPT to provide a text) are worded.

Expanding on the previous point, Yan (2023) qualitatively examined ChatGPT’s influence on learners participating in an L2 writing practicum. The results unveiled ChatGPT’s advantages and potential in L2 writing instruction. That is, ChatGPT demonstrated an automated workflow capable of enhancing efficiency in writing composition. Nevertheless, participants generally voiced apprehensions regarding its impact on academic integrity and educational equity. Another drawback Baskara (2023) voiced is the ethical implications of substituting human language teachers and using machine learning systems to generate text.

In the Saudi context of education, Mahdi (2023) quantitatively investigated 64 EFL learners’ perceptions of their preference for teacher-mediated versus ChatGPT-created writing mode via factor analysis of generated data from a language learning institution in the eastern region of Saudi Arabia. Responses revealed that most participants believed that ChatGPT could be a supplement educational tool and cannot replace teachers’ role in the educational system regardless of its improvement and modification. Furthermore, responses also revealed that ChatGPT does not help students progress due to a lack of learning satisfaction when using this tool.

Although researchers above and others in the literature have made strides in understanding the impact of ChatGPT in different language learning contexts, there is a continuing need for further studies that empirically provide evidence for its impact on specific linguistic issues such as second language vocabulary knowledge development. As far as the researcher is aware, none of the studies found in the literature investigated the impact of ChatGPT on second language vocabulary development at different levels (i.e., breadth or depth). Therefore, this research fills the gap in the literature by investigating the impact of ChatGPT-generated exercise on Saudi EFL vocabulary development in terms of breadth (i.e., size). Moreover, there is also an urgent need to conduct experimental research on different groups of participants to establish cause-and-effect relationships, provide insight into the interventions of ChatGPT in teaching and learning vocabulary context, and provide evidence that can inform educational policies and practices.

**Vocabulary Knowledge Development**

Measuring L2 vocabulary knowledge serves several important practical and theoretical purposes across disciplines such as education, psychology, linguistics, and cognitive science. Zwier and Boers (2023) indicated two types of instruments for assessing vocabulary knowledge. One attempts to assess learners’ vocabulary knowledge in communication, and the other attempts to evaluate vocabulary knowledge in isolation. The former instrument assesses integrated vocabulary in well-known proficiency tests such as IELTS and TOEFL. In contrast, the latter assesses vocabulary knowledge in isolation by measuring learners’ depth and breadth of vocabulary knowledge, which is the focus of this current research.

Over the past four decades, several discrete-item tests have been proposed to gauge learners’ breadth and depth of vocabulary knowledge. Breadth or size interchangeably refers to how many words we know from a specific segment of a language’s vocabulary. For example, how many words do you know that belong to the 5,000 most frequent words of a particular language or a specific discourse genre (e.g., academic discourse)? (Dóczi & Kormos, 2015). Conversely,
depth refers to how well we know a word (Zwier & Boers, 2023). It can be operationalised as the degree to which you are familiar with all word aspects (i.e., grammatical characteristics, derivatives, collocations, pronunciation and spelling, multiple meanings, and registers) or how fast you recognise the word and retrieve it from your lexical repertoire when needed.

For measuring the breadth or size of learners’ vocabulary knowledge, Nation (1983) proposed the first vocabulary size test, namely the Vocabulary Level Test (VLT), with a multiple-choice format matching test where test takers are required to determine which of the six words in a list correspond to the three definitions. A few years later, Schmitt et al. (2001) adopted the same format of VLT and sampled words from different frequency bands of 1,000 words with 30 words per band. If test takers recognise the meaning of 30 words from a band, they are expected to possess a similar knowledge of the other words in this band. The same test was developed by Web et al. (2017) with the same procedures but with a different layout, and test takers need to check the box under the word that matches the definition. Another test was developed by McLean et al. (2015); however, the presentation of the word is conducted aurally since it is assumed that test takers’ familiarity with the word in written format does not guarantee their familiarity with its spoken form and vice versa (Milton & Hopkins, 2006).

Critics have raised concerns about the validity and reliability of the tests above as well as others not explicitly mentioned in this review of the literature (e.g., Brown et al., 2022; Kremmel, 2016; Laufer, 2021; McLean, 2021; Stoeckel et al., 2021; Webb & 2017). These critics highlighted several issues, such as the small samples of words representing bands as large as 1,000 words and the use of word family as a unit of counting and testing. However, Zwier and Boers (2023), in their review, argued that none of these critics is problematic as long as the purpose of these tests is to compare different student groups’ breadth of vocabulary knowledge or track the vocabulary development growth over time (e.g., at the beginning and the end of a language course).

Moving to the second type of instrument, which assesses the depth of word knowledge, several scholars have suggested various measurable tests. To begin with, the word association tests developed by Read (1998). This test provides test takers with a prompt accompanied by eight other words. Four words are associated with the prompt for test takers to identify these associates. The strength of this test stems from the fact that it can estimate knowledge of different word meanings (i.e., polysemous and homonyms). Therefore, this test can measure and capture aspects of word knowledge beyond the previously discussed tests of breadth of word knowledge. So, it is no surprise that scholars such as Chen and Liu (2020) and Vafaee and Suzuki (2020) employ this test to complement the latter in research on the correlation between vocabulary knowledge and text comprehension.

Another test Wesche and Paribakht (1996) developed was the vocabulary knowledge scale. This test distinguishes between receptive and productive knowledge to assess different degrees of word knowledge. Other researchers focus on the connection between word form and meaning by separately operationalising the distinction between recognition and recall. In this regard, Laufer and Goldstein (2004) developed a Computer-Adaptive Test of Size and Strength (CATSS). This test sampled words from corpus-based frequency bands 1,000 to 14,000. For every word, four test modalities are created, namely (productive recall, receptive recall, productive recognition and receptive recognition). According to Zwier and Boers (2023), recall means the capability to retrieve knowledge from memory generated by the word’s meaning or form. In contrast, recognition refers to the ability to match the form of the word to the appropriate meaning or vice versa.
This research will adopt CATSS, as developed by Laufer and Goldstein (2004), to investigate the vocabulary knowledge development of Saudi EFL learners. The justification beyond employing this test, particularly among several other tests available in the literature, is that this test is adaptive. It is designed to measure the breadth and depth of learners’ vocabulary knowledge, unlike other tests in the literature, which are designed to measure these aspects of word knowledge separately. In this regard, Dóczi and Kormos (2015) stated, “While researching the breadth of vocabulary is very important, only by taking both the breadth and the depth of vocabulary knowledge into account can we gain deeper insights into the actual processes of vocabulary development” (p.1).

**Method**

As stated earlier, the main aim of this research is to compare the impact of ChatGPT-generated exercises versus traditional exercises on Saudi EFL students’ vocabulary development. Therefore, experimental design is best proposed in the context of this research. In this vein, Dörnyei (2015) stated that to answer the cause-and-effect dilemma in the educational and applied linguistics domains, experimental design can be implemented by taking a group of learners and applying the treatment to them while measuring their progress. Then, the results generated will be compared with the data obtained from another group that is similar in every respect to the first group except that it does not receive the special treatment. If the discrepancy is marked in the results of the two groups, these can be attributed to the only difference between them, the treatment variable. Furthermore, Christensen et al. (2011) argued that the main feature of experimental design is its tightly controlled environment since specific consciously manipulated processes take place and only target variables are varied while others are kept constant.

**Participants**

This research was conducted in the first semester of the academic year 2023/2024 (i.e., four months) at Jouf University. One of the public universities in the northern part of the kingdom where students who want to enrol in any academic program must attend an English proficiency test, namely Linguaskill. This test is administered online by the British Council. This test is a quick and easy online examination that measures the target students' English proficiency level. The outcomes are consistent with the Common European Framework of Reference for Languages (CEFR), the international standard for describing language competency. Those students who level C1-Advanced and C2-Proficient are exempted from the Intensive English Language Program (IELP) provided by the English department at Jouf University.

The curriculum and structure of IELP are tailored to meet their student's specific needs and characteristics and align with KSA’s Vision 2030. The program has undergone several essential improvements since its foundation in the academic year 2021 based on ongoing assessments and instructor and student feedback. IELP exposes candidates to 25 hours of teaching per week for one academic semester (i.e., four months). One of the teaching materials employed is English Vocabulary in Use, 4th edition, by Stuart Redman (2017). IELP implements strict assessment procedures to gauge students’ progress and provide constructive feedback on strengths and improvement areas.

Probability sampling procedures and, most specifically, stratified random sampling procedures were employed in this research. Several parameters were identified first for the population of Saudi EFL students who join the IELP context during the first semester of the
academic year 2023/2024. All participants were male students due to gender segregation in the educational system in KSA (Ribetti & Ribetti, 2023), aged between 18 to 20 years old, never had a chance to be schooled overseas, and level A1 to B2 according to CEFR. After that, selected participants (N=60) were randomly distributed equally to an experimental group where treatment was applied (i.e., ChatGPT-generated exercises) and a control group (i.e., traditional exercises) to ensure a balanced distribution of potential confounding factors (Dörnyei, 2015).

Integrate ChatGPT into the curriculum of teaching vocabulary for Saudi EFL students
ChatGPT was sought to assist in building assessment methods such as fill-in-the-blank exercises, generating test questions, and interactive quizzes to fulfil the research questions. The same exercises covered by English Vocabulary in Use, 4th edition by Stuart Redman (2017), were integrated into the newly generated ChatGPT exercises to control research variables and thus avoid discrepancy. As a way of explanation, in Unit One titled “Learning” the same exercises and the same content were integrated by using the ChatGPT platform (Open AI, 2023). Prompts were carefully prepared by following Skrabut’s (2023) models on using ChatGPT for creating assessments. The following figure shows a sample of a prompt prepared by instructors in the IELP context to generate ChatGPT-generated exercises:

Figure 1. A sample of a prompt prepared by IELP instructors to generate ChatGPT-generated exercises

Due to the nature of ChatGPT-generated exercises, which require digital interferences, computer labs at the Department of English were used for the experimental group (i.e., the treatment group in which ChatGPT-generated exercises were employed). Participants engaged with assigned tasks, responded to them and underwent assessment. The following figure shows a sample of interactive exercises between ChatGPT and Saudi EFL students.
The Impact of ChatGPT on the Development of Vocabulary Knowledge

Mugableh

Research Instruments

To systematically compare the impact of ChatGPT-generated exercises versus traditional exercises on Saudi EFL students’ vocabulary knowledge, the CATSS online test, as developed by Laufer and Goldstein (2004), was administered to both groups (i.e., experimental and control groups) at the beginning and end of their studies in the IELP course with an interval of four months (i.e., the first semester of the academic year 2023/2024). CATSS vocabulary test assesses vocabulary size, or knowledge of word meaning, by testing four degrees of strength: productive recall, receptive recall, productive recognition, and receptive recognition (Laufer & Levitzky-Aviad, 2018).

The test aims to overcome the limitations of size tests by testing each word in more detail. The four modalities are a hierarchy of difficulty and are implicationally scaled. If a test-taker answers an item correctly in one strength modality, testing the same word in subsequent strength modalities will not be necessary. Moreover, the test-taker’s answers are matched against a pre-specified marking key and scored as correct or incorrect. A word answered correctly at productive recall receives 1 point, a word answered correctly at receptive recall receives 0.75 points, a word answered correctly at productive recognition receives 0.5 points, and a word answered correctly at receptive recognition receives 0.25 points (Laufer & Levitzky-Aviad, 2017).

At the end of the exams, the results are displayed in a table, providing information on the number of correct answers for each modality at each frequency level, the total vocabulary size score per modality, the total vocabulary size score in the test, and the total vocabulary strength scores per modality (About CATSS, n.d.) The following figures exhibit samples of a test summary.
and score level breakdown generated by CATSS for one of the participants in this study with a vocabulary size of 11500-word families and a vocabulary size and strength of 8200-word families.

Figure 3. A sample of a test summary generated by CATSS for one of the participants

Figure 4. A sample of the score level breakdown generated by CATSS

Data Analysis Methods and Results

The collected results from all participants (N=60) from both groups (i.e., experimental and control groups) and in both intervals (i.e., pre-test and post-test) were imported to SPSS (Statistical Package for the Social Sciences) clearly and concisely. This research aims to systematically compare the impact of ChatGPT-generated exercises versus traditional exercises on Saudi EFL students’ vocabulary size and the strength of word families. The mean difference between two dependent groups (i.e., control and experimental) should be compared in multivariate response variables (i.e., pre-test and post-test). Therefore, Multivariate Analysis of Variance (MANOVA) is
assumed to be the best statistical technique because of the correlated multiple response variables in the context of this research (IBM Documentation, n.d.).

First, before running the MANOVA statistical tests, it is common practice in statistics to run a Test of Normality since MANOVA assumes that dependent variables follow a multivariate normal distribution (IBM Documentation, n.d.). Thus, checking for normality is important because MANOVA is sensitive to violations of the normality assumption. Figure 5 below shows the Test of Normality results in which the assumption of multivariate normality was assessed for each group using the Shapiro-Wilk test. In both cases, the p-values for the Shapiro-Wilk test are less than 0.05 (the standard significance level), indicating evidence to reject the null hypothesis of normality. Additional visual inspection of the data through a Q-Q plot supported these findings, as reflected in Figures 6 and 7 below.

<table>
<thead>
<tr>
<th>Tests of Normality</th>
<th>Kolmogorov-Smirnov</th>
<th>Shapiro-Wilk</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Statistic</td>
<td>df</td>
</tr>
<tr>
<td>Improvement vocabulary size of word families</td>
<td>1.145</td>
<td>60</td>
</tr>
<tr>
<td>Improvement vocabulary size and strength of word families</td>
<td>-2.07</td>
<td>60</td>
</tr>
</tbody>
</table>

* a. Lilliefors Significance Correction

**Figure 5.** Test of normality

**Figure 6.** Normal Q-Q plot of improvement vocabulary size of word families

**Figure 7.** Normal Q-Q plot of improvement vocabulary size and strength of word families
The overall MANOVA in Figure 8 below indicated a Wilks' Lambda value of (.534). The test of Wilks' Lambda was statistically significant, F (2.000, 57.000) = 24.912, p = .000.

**Table: Multivariate Tests**

<table>
<thead>
<tr>
<th>Pillai's trace</th>
<th>Wilks' lambda</th>
<th>Hotelling's trace</th>
<th>Roy's largest root</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>0.466</td>
<td>0.634</td>
<td>0.674</td>
</tr>
<tr>
<td>F</td>
<td>24.912*</td>
<td>24.912*</td>
<td>24.912*</td>
</tr>
<tr>
<td>Hypoth. df</td>
<td>2.000</td>
<td>2.000</td>
<td>2.000</td>
</tr>
<tr>
<td>Error df</td>
<td>57.000</td>
<td>57.000</td>
<td>57.000</td>
</tr>
<tr>
<td>Sig.</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Partial Eta Squared</td>
<td>.466</td>
<td>.466</td>
<td>.466</td>
</tr>
</tbody>
</table>

Each F tests the multivariate effect of Experimental - Control. These tests are based on the linearly independent pairwise comparisons among the estimated marginal means.

**Figure 8. Multivariate tests**

Moreover, MANOVA results in Figures 8 and 9 below demonstrated that participants in the experimental group who experienced ChatGPT-generated exercises notably surpassed that of the control group regarding vocabulary size of word families and vocabulary size and strength of word families. That is, by examining the descriptive statistics generated in Figure 8 below. Results indicated a significant improvement in the vocabulary size of word families in the experimental group, with a mean of (2213.33) compared to (916.67) in the control group. Moreover, a significant improvement in vocabulary size and strength of word families was also tangible in the experimental group, with a mean of (1743.33) compared to (730.00) in the control group.

**Table: Descriptive Statistics**

<table>
<thead>
<tr>
<th></th>
<th>Experimental - Control</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Improvement vocabulary size of word families</td>
<td>Experimental</td>
<td>2213.33</td>
<td>730.160</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>916.67</td>
<td>642.776</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>1555.00</td>
<td>687.405</td>
<td>60</td>
</tr>
<tr>
<td>Improvement vocabulary size and strength of word families</td>
<td>Experimental</td>
<td>1743.33</td>
<td>778.610</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>730.00</td>
<td>647.622</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>1236.67</td>
<td>874.491</td>
<td>60</td>
</tr>
</tbody>
</table>

**Figure 8. Descriptive Statistics**

The Pairwise Comparison test in Figure 9 below reflects the same results as the experimental group in which the treatment was applied. The experimental group improved significantly higher than the control group in the CATSS test at two different intervals, with a mean difference of (1296.67) in vocabulary size and a mean difference of (1013.33) in vocabulary size and strength of word families.

**Table: Pairwise Comparisons**

<table>
<thead>
<tr>
<th>Dependent Variable</th>
<th>i Experimental - Control</th>
<th>j Experimental - Control</th>
<th>Mean Difference</th>
<th>Std Error</th>
<th>Sig</th>
<th>95% Confidence Interval for Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Improvement vocabulary size of word families</td>
<td>Experimental</td>
<td>Control</td>
<td>1296.67</td>
<td>185.664</td>
<td>.000</td>
<td>924.336 - 1668.354</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>Experimental</td>
<td>-1296.67</td>
<td>185.664</td>
<td>.000</td>
<td>-1668.354 - 924.336</td>
</tr>
<tr>
<td>Improvement vocabulary size and strength of word families</td>
<td>Experimental</td>
<td>Control</td>
<td>1013.33</td>
<td>184.816</td>
<td>.000</td>
<td>643.383 - 1383.283</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>Experimental</td>
<td>-1013.33</td>
<td>184.816</td>
<td>.000</td>
<td>-1383.283 - 643.383</td>
</tr>
</tbody>
</table>

Based on estimated marginal means.

* The mean difference is significant at the .05 level.

b. Adjustment for multiple comparisons. Least Significant Difference (equivalent to no adjustments).
Discussion

As mentioned earlier, a thorough examination of the literature revealed a limited number of studies directly addressing the specific nuances of the impact of ChatGPT on the vocabulary development of EFL students. Consequently, in this ensuing discussion, an effort was made to draw insightful comparisons with the closest available literature in varying degrees that touched upon the thematic contours of this research field to distil relevant insights, identify commonalities, understand the research landscape, and provide a foundation for the novel contributions presented.

Results generated in this research indicated that Saudi EFL students' vocabulary knowledge developed over four months in both groups. However, the development where ChatGPT-generated exercises were employed was significantly higher than those who employed traditional exercises. These results are of significant importance due to the different contributions. First, from educational contribution, integrating ChatGPT into the teaching material of IEPL, i.e., English Vocabulary in Use, 4th edition by Stuart Redman (2017), provided Saudi EFL students with an interactive and engaging learning experience that reinforced their vocabulary usage in context. These results align with several researchers (e.g., Athanassopoulos et al., 2023; Guo et al., 2022). For instance, Guo et al. (2022) integrated ChatGPT into argumentative writing classrooms for EFL students at Hong Kong University. They argued that ChatGPT enhances students’ academic and professional vocabularies by providing them with a nontraditional interactive learning environment. Moreover, Athanassopoulos et al. (2023) investigated the effectiveness of ChatGPT in improving foreign language writing in terms of vocabulary and grammar of EFL students in one of the schools in south Greece. Results revealed that ChatGPT provided students real-world context that improved participants’ writing skills regarding different qualitative characteristics, including total words, average word per sentence, unique words, and most frequent words.

Kohnke et al. (2023) argued that one of the most potential drawbacks of the ChatGPT tool is its inconsistency of the texts generated in that responses generated by ChatGPT vary significantly based on how prompts (i.e., initial text input to ChatGPT to provide a text) are worded. However, the text generated by ChatGPT in this research diverges significantly from the perspectives mentioned above. To mitigate this potential inconsistency in this research, all prompts used to generate ChatGPT-generated exercises were prepared and controlled by instructors with no active involvement or influence from the students in the insertion procedures. Prompts were inserted carefully following Skrabut’s (2023) models of using ChatGPT for creating assessments. This, in turn, provides a more controlled and reliable foundation for Saudi EFL students’ experience with the ChatGPT environment.

It should be noted that this current research does not fall under the category of survey research. However, the results generated are of ChatGPT interest rather than traditional teaching methods employed in the context of IELP. Therefore, this research's findings constantly align with others in the literature (e.g., Xiao and Zhi, 2023; Ali, 2023; Yan, 2023; Baskara, 2023). These parallel outcomes enhance the credibility and robustness of the findings within the broader scholarly discourse.

From a methodological contribution, none of the studies found in the literature to date employed the CATSS test as developed by Laufer and Goldstein (2004) in an experimental design context since this test generates two different results and thus addresses two dependent measurable variables, namely vocabulary size of word families and vocabulary size and strength of word families. These dual measurable variables provide researchers with a powerful tool to measure
word form and meaning, thus clearly showing whether test-takers can distinguish between recognition and recall of vocabulary knowledge separately (Zwier & Boers, 2023).

From empirical contribution, this research presents new empirical evidence on the impact of ChatGPT on vocabulary development in teaching English as a foreign language. As stated earlier in the literature review section, the available literature indicates a notable absence of thorough investigations into this matter by previous researchers. Thus, the results generated in this research provide an original research contribution by introducing new findings, data, and experimental results that contribute to the existing body of knowledge.

Conclusion

This research systematically compared the effect of ChatGPT-generated exercises versus traditional exercises on Saudi EFL students' vocabulary size and word family strength. Specifically, it aimed to answer the following questions: 1) To what extent does the integration of ChatGPT-generated exercises impact Saudi EFL students’ vocabulary size of word families? 2) To what extent does the integration of ChatGPT-generated exercises impact Saudi EFL students’ vocabulary size and strength of word families? The CATSS online test, developed by Laufer and Goldstein (2004), was administered to both groups (i.e., experimental and control groups) at the beginning and end of their studies in the IELP course, with an interval of four months. Based on the study's findings, Saudi EFL students’ vocabulary development improved significantly in both groups. However, participants in the experimental group who experienced ChatGPT-generated exercises notably surpassed the control group regarding vocabulary size and word family strength. The results of this research present a novel contribution to literature in that it presents new empirical evidence on the positive impact of ChatGPT on vocabulary development in teaching English as a foreign language. Moreover, it presents new pedagogical evidence for the possibility of integrating ChatGPT-generated exercises into the curriculum of teaching English vocabulary to Saudi EFL students. It provided them with an interactive and engaging learning experience that reinforced their vocabulary usage in context.

Limitations and Future Research Recommendations

While this research contributes valuable insights into the impact of ChatGPT on vocabulary development in teaching English as a foreign language, it is essential to acknowledge several limitations that may impact the interpretation and generalizability of the findings generated. Firstly, the design adopted in this research is experimental, which has been claimed by several scholars to be the only compelling method of establishing a cause-effect relationship and evaluating educational innovations (Dörnyei, 2015). Nevertheless, this design, as argued by Clarke and Braun (2017), possesses specific limitations related to the generalizability aspect. That is, to control all the dependent and independent variables tightly, the researcher might end up with artificial frameworks in laboratory conditions and sacrifice the external validity to enhance the internal validity.

Another limitation of this research is related to the tool employed to assess Saudi EFL students’ vocabulary knowledge development. CATSS online test, as developed by Laufer and Goldstein (2004), has been demonstrated to be one of the most effective tools currently available in the existing literature due to its approachability as it is available online and free of charge for educators and researchers. Moreover, this tool could measure both breadth and depth of word knowledge, unlike other tests developed so far that are merely dedicated to measuring either depth
or breadth of vocabulary knowledge individually. Nevertheless, there is inevitably a practical limitation for this tool, which is related to the risk of fatigue on the part of the test takers since this test assesses knowledge of corpus-based frequency bands 1.000 to 14.000 and for every word, four test modalities are assessed namely (productive recall, receptive recall, productive recognition and receptive recognition).

In light of the findings in this research, several avenues for future research could contribute to a deeper understanding of the effect of ChatGPT on vocabulary development. Future researchers can expand their employment of ChatGPT to prepare a complete curriculum with a focus on vocabulary teaching, which includes several pedagogical phases such as creating course descriptions, creating learning goals and objectives, writing a lesson plan, crafting classroom activities, developing a list of ice breaker questions and activities, and create rubrics (Skrabut, 2023). Moreover, future researchers can explore the impact of ChatGPT on vocabulary knowledge development longitudinally (i.e., different timeframes, repeated measurement, and naturalistic observation), which provides researchers with a design that focuses on observing changes over time in a naturalistic setting. These proposed research paths represent a novel exploration of the impact of ChatGPT on vocabulary development, which, to the best of the researcher's knowledge, no prior studies have delved into such exploration paths.

Funding
This research is not funded.

Acknowledgements
Not applicable.

Conflicts of Interest
The authors declare no conflict of interest.

Authenticity
This manuscript is an original work.

Artificial Intelligence Statement:
AI and AI-assisted technologies were not used.

About Author
Ahmad Ibrahim Mugableh completed his PhD in Applied Linguistics from Universiti Sains Islam Malaysia (USIM) in 2016. He has held various positions at Jouf University in the Kingdom of Saudi Arabia. In 2019, he became the head of the English Department for two years before transitioning to the Deanship of Quality and Accreditation. In 2021, he served as the Head of the Programs Accreditation Unit for one academic year. He has published several papers indexed in SCOPUS and ISI that focus on linguistic issues among Arab students, especially Saudis. His research covers areas like speech recognition, discourse analysis, psycholinguistics, sociolinguistics, and language development. https://orcid.org/0000-0002-5511-553X

References
About CATSS. (n.d.). Available at https://lexisite.com/catss/catss-info
Athanassopoulos, S., Manoli, P., Gouvi, M., Lavidas, K., & Komis, V. (2023). The use of ChatGPT as a learning tool to improve foreign language writing in a multilingual and


McLean, S. (2021). The coverage comprehension model, its importance to pedagogy and research, and threats to the validity with which it is operationalised. *Reading in a Foreign Language, 33* (1), 126–140.


