Education and Information Technologies
https://doi.org/10.1007/s10639-025-13342-w

AI-powered language learning: Developing the chatGPT usage scale for foreign language learners

Ferdiye ÇOBANOĞULLARI[1] · **Özge ÖZBEK**[2]

Received: 4 August 2024 / Accepted: 3 January 2025

Abstract

This study introduces the "ChatGPT Usage Scale for Foreign Language Learners," designed to evaluate the usage of AI chatbots in language learning. The scale was developed based on a comprehensive literature review and expert evaluations, ensuring a strong theoretical foundation and content validity. It comprises three subdimensions Usability, Learning, and Development that collectively explain 62.896% of the total variance, exceeding standard expectations. Data were collected from 1006 university students (657 female, 349 male) from foreign language departments at three Turkish state universities during the 2023–2024 fall semester. The scale demonstrated excellent internal consistency (Cronbach's Alpha = 0.93) and validity, supported by robust Confirmatory Factor Analysis results. A test–retest analysis with a subset of participants confirmed the scale's reliability over time. Future research should apply this measurement tool in diverse educational contexts to enhance its generalizability and further refine its psychometric properties, providing deeper insights into the role of AI tools in language education.

Keywords ChatGPT · Foreign language learning · Artificial intelligence · Scale development

✉ Ferdiye ÇOBANOĞULLARI
ferdiyec@gazi.edu.tr

Özge ÖZBEK
ozgekaracadal@gazi.edu.tr

[1] Gazi Faculty of Education, Division of German Language Education, Gazi University, Ankara, Turkey

[2] Gazi Faculty of Education, Division of French Language Education, Gazi University, Ankara, Turkey

Published online: 13 January 2025

1 Introduction

In the twenty-first century, we have witnessed numerous technological advancements and their integration into foreign language education. Concepts such as 'Technology-Enhanced Language Teaching' and 'Computer-Assisted Language Learning' have emerged, placing technology at the core of education. Subsequently, the use of machine learning and digital content in education has gained attention, leading to the integration of new digital systems like augmented reality and virtual reality into educational settings. In this context, generative artificial intelligence (AI) and AI-assisted language Learning have been developed (Baidoo-Anu & Owusu Ansah, 2023).

In this context, various applications or chatbots supported by artificial intelligence, such as ChatGPT, Google Bard (Gemini), Microsoft Bing AI, Chinese Ernie Bot, Korean SearchGPT, Russian YaLM 2.0, Chatsonic, Jasper Chat, Character AI, Perplexity AI, and YouChat, have been developed by different countries (Singh, 2023). Thanks to these tools, researchers, teachers, and students worldwide working in various scientific fields can exchange ideas, request examples, generate texts, experience a new foreign language, perform interlingual translations, and converse with artificial intelligence in multiple languages. This situation can be considered a significant opportunity in foreign language education. These tools have been tested many times through large datasets to produce human-like and creative texts; hence, they are referred to as large language models (LLMs) (Kasneci et al., 2023; Topsakal & Topsakal, 2022; Yu, 2023).

Just a few hours after Google launched Gemini 1.5 Pro, an enhanced version of its artificial intelligence language model with 1 million new tokens (15 February 2024), OpenAI announced the development of its new AI language model named Sora (Pichai & Hassabis, 2024). This situation demonstrates the technological and scientific rivalry between two tech giants, Google and OpenAI. The rapidly advancing field of generative AI, which already impacts text, audio, image, and video, will be used and developed by many companies in various other areas in the coming years.

AI-based tools are utilized in various fields, including computer science, health sciences, social sciences, natural sciences, commerce, gaming, and many others. In the field of education, and particularly in the process of foreign language learning and teaching, there are numerous studies examining the impact and usability of Chat GPT, an AI language model, from various perspectives (Farhi et al., 2023; Fitria, 2023; Hong, 2023a, 2023b; Huang & Li, 2023; Kim et al., 2023).

Researchers have differing opinions and debates regarding using this new technology in foreign language education. It can be noted that most of the research on ChatGPT consists of general informative qualitative studies, and some are specific to some areas of various countries. In the research conducted by the authors, only one scale study was identified, which was carried out in the United Arab Emirates; it explored students' use of ChatGPT, their opinions, concerns, and ethical perceptions (Farhi et al., 2023). However, the abovementioned study was not applied to students studying in foreign language departments.

It has been determined that there is no scale study regarding the use of ChatGPT among students in foreign language departments. Addressing this research gap, a scale named "ChatGPT Usage Scale for Foreign Language Learners" was developed and applied to students studying Arabic, German, French, and English at three universities in Turkey. The aim of the study is primarily to (i) uncover the thoughts of students in foreign language departments on the general usability of ChatGPT, (ii) the direct contribution of using ChatGPT to the language learning process, and (iii) ChatGPT's ability to support and enhance the language learning process.

We believe developing a scale to measure ChatGPT's usage among foreign language learners is crucial for several reasons. It allows us to systematically assess students' experiences and perceptions, providing insights into the tool's effectiveness and areas for improvement. A standardized scale will also enable longitudinal studies to track changes in usage and educational outcomes over time. Additionally, it helps address ethical considerations and challenges, supporting informed decisions by educators and policymakers. Our research fills a gap by creating the "ChatGPT Usage Scale for Foreign Language Learners," ensuring AI's ethical and effective use in language education.

This study focuses on ChatGPT, an artificial intelligence product capable of automatically generating human-like conversations and possessing extraordinary capabilities as a chatbot. This paper highlights ChatGPT due to its capacity to generate responses that are more human-like compared to other natural language processing models, such as rule-based systems, which enable more meaningful and captivating interactions with users, thereby enhancing user experience and contentment (Kalla et al., 2023). The following section covers the emergence of ChatGPT and its use in education and foreign language teaching, and it contains various opinions about its use in education. The third section details the methodology, research design, participants, scale development process, data collection, and data analysis. The fourth section emphasizes the results, showcasing content validity, pilot study outcomes, scale reliability, and analysis results. The final section includes the discussion and limitations of the study.

2 Review of literature

This section discusses the literature on ChatGPT, one of the widely used models of digital technologies and artificial intelligence, and its use in foreign language education.

2.1 The emergence of chatGPT

ChatGPT (Chat Generative Pre-training Transformer), a generative artificial intelligence product created by OpenAI on November 30, 2022, offers applications that can be used in various areas such as text generation, interlingual translation, summarisation, chatbots, virtual assistants, etc. (https://chat.openai.com/) (Singh, 2023). As the name implies, ChatGPT has been pre-trained; it has read and assimilated data

up to a particular year and can be updated with new data in the future. This model operates based on Natural Language Processing (NLP), meaning it can engage in human-like communication and generate dialogues, and thus is considered one of the most significant technological advancements today (Bozkurt, 2023; Fuchs, 2023; Meyer et al., 2023). Although launched in late 2022, this AI language model has surpassed its predecessors, reaching millions of users today and achieving significant success (Singh, 2023; Zhang, 2023).

Initially emerging as GPT-1, the model was seen as foundational but lacking in several areas, such as providing excessive information, struggling to maintain long conversations within dialogues, and lacking fluency and coherence in extended texts (Hadi Mogavi et al., 2024). When the second version, GPT-2, was released in February 2019, some improvements over the original model were noted, though it still had significant limitations. A year later, in 2020, the new model GPT-3 emerged with 175 billion new parameters and exhibited an incredible ability to generate human-like text. This development was followed by the still-free ChatGPT 3.5 on November 30, 2022, which included better enhancements and developments. The latest version, ChatGPT 4 (fourth-generation generative pre-trained transformer), which also accommodates and interprets visual inputs, was launched on March 14, 2023, as a paid service. The ongoing potential for development and improvement in ChatGPT and the imminent release of the GPT-5 version indicates that it will not remain static (Hadi Mogavi et al., 2024; Ray, 2023).

2.2 The use of chatGPT in education and foreign language teaching

Although there are not many scale studies or experimental research on the new model ChatGPT in the field of foreign language education, there is a qualitative study that conducted a sensitivity analysis to reveal users' perceptions of ChatGPT in education (Tlili et al., 2023). According to the results of this study, ChatGPT has the potential to revolutionize education by being used in various ways. Additionally, it is emphasized that a foreign language student using ChatGPT can receive personalized support, guide themselves, access information from many open sources, and perform self-assessments related to their language learning process (Firat, 2023; Huang & Li, 2023).

Some experts have demonstrated how both foreign language students and teachers can use ChatGPT by giving instructions in various contexts (Athanassopoulos et al., 2023; Kohnke et al., 2023). Instructions can be given to ChatGPT in the form of questions or sentences to be completed, similar to everyday communication. ChatGPT also has a feature to remember what is asked, allowing users to revisit and read previous queries (Bonner et al., 2023). In the field of foreign language education, ChatGPT can be used in many different ways, such as writing emails, creating dialogues, adjusting existing text or dialogue to a desired language level (e.g., beginner or advanced), translating a given dialogue into another language (e.g., from English to Chinese), summarising texts, and taking vocabulary notes on translated texts (Bonner et al., 2023; Jiao et al., 2023; Kohnke et al., 2023). Additionally, many studies have shown that foreign language teachers can use ChatGPT to prepare questions

or exercises (open-ended or multiple-choice) for reading comprehension and to work on writing skills and grammar for non-native speakers (Athanassopoulos et al., 2023; Kohnke et al., 2023).

Kostka and Toncelli (2023) used ChatGPT in two English Language Teaching (ELT) classes to teach topics such as persuasive argumentation and formal presentation skills. They found that the students had many positive approaches to using ChatGPT. As mentioned below, ChatGPT in foreign language education provides pedagogical flexibility and encourages students to think critically.

2.3 Different opinions on the use of chatGPT in education

Using ChatGPT, a chatbot that produces consistent texts and dialogues, provides rich input and output in multiple languages and fields, and interacts human-like, is highly beneficial in foreign language education. It is accessible to everyone worldwide as an open resource, and versions up to ChatGPT 3.5 are free, promoting educational equity for students. The texts it generates are human-like and consistent, seamlessly connecting sentences. ChatGPT can be helpful as a chatbot in attracting foreign language learners' interest and motivating them. This chatbot can also create various types of original texts, correct grammar, vocabulary, syntax, and conjunction errors in texts written by students, provide examples on topics requested by teachers, and prepare exams (Fryer et al., 2017; Kohnke et al., 2023).

When a student asks ChatGPT about a topic they do not understand, they can receive assistance in a foreign language or their native language, including explanations, examples, dictionary definitions of words, or how to use them in different contexts and speech types (Kohnke et al., 2023). Additionally, by generating different content, providing new ideas, presenting new words and examples, and correcting errors in students' texts, ChatGPT helps enhance their creative writing skills in a foreign language (Hong, 2023a, 2023b). Students involved in studies generally find using ChatGPT in the educational process fun and motivating (Kostka & Toncelli, 2023).

While ChatGPT is considered an opportunity in foreign language education, it also brings its share of concerns. There are worries about the accuracy of texts generated by ChatGPT in different languages, and some authors even suggest that ChatGPT is prone to 'hallucinations' (Meyer et al., 2023). It is noted that English is the language with the most data and the most frequent training for ChatGPT, and responses in languages other than English may contain more errors or inconsistencies, as indicated by some authors (Houston & Corrado, 2023). Among the criticisms of ChatGPT is that the words used in its responses are more common in written language and do not cater much to spoken language. Additionally, this chatbot may provide biased responses when dealing with different cultures and societies. ChatGPT may sometimes make incorrect corrections or create repetitive sentences when instructed to correct linguistic errors, leading to confusion among students (Kohnke et al., 2023). Therefore, some students approach ChatGPT skeptically and do not fully trust it in foreign language learning (Kostka & Toncelli, 2023). ChatGPT also poses a significant threat to examination processes in learning

environments, even causing some universities in Australia to revert to paper-and-pencil exams (Cassidy, 2023). Additionally, there are concerns that chatbots provide data privacy problems (Berşe et al., 2023).

In conclusion, like other chatbots that contribute to foreign language education, ChatGPT has many advantages. However, the learning environments and tasks assigned to students during its use are critically important (Fryer et al., 2017). It is recommended that educators update all teaching, learning, and assessment processes and modify course materials with this tool, which is accessible 24/7 (Houston & Corrado, 2023).

3 Method

3.1 Research design

This research undertakes the development of a scale utilizing a survey methodology. Such an approach is frequently employed to delineate phenomena and occurrences by collecting data from sizable cohorts (Karakaya, 2012).

3.2 Participants

After obtaining the necessary ethical approvals, the study was conducted in Turkey during the fall semester of 2023–2024 at three different state universities, involving 1006 university students (657 female, 349 male) enrolled in foreign language departments. These students receive education in German (n: 171), French (n: 292), Arabic (n:179), and English (n:364) languages. Detailed information about the participants is provided in Table 1.

Şencan (2005) and Tabachnick et al. (2019) suggest that, as a general rule, the sample size for scale development studies should be at least five times the number of items included. Thus, for a scale with 18 items, a minimum of 90 participants is recommended. However, this study exceeded this recommendation, reaching a sample size of 1006 participants. In sampling, the selected students were not chosen based on any criteria but instead randomly selected. The inclusion criteria for this study are as follows: Participants must be enrolled in programs offering foreign language education. Additionally, they should have used ChatGPT more than once in the language learning process and must sign a consent form indicating their voluntary participation in the study. The exclusion criteria are as follows: Students who have never used ChatGPT before and participants who provide incomplete information during the data collection process will be excluded from the study.

3.3 Scale development process

The scale items were developed systematically and grounded in a comprehensive literature review. This review focused on previous studies and theoretical discussions related to attitudes toward artificial intelligence in education, particularly

Table 1 Participant characteristics

Participant Characteristics of Exploratory Factor Analysis		n	%
Gender	Female	256	64,65
	Male	140	35,35
Age	Max	66	
	Min	17	
	Mean	21.67 ± 4,25	
Grade	Preparatory	57	14.36
	1st Class	108	27,27
	2nd Class	94	23,74
	3rd Class	89	22,47
	4th Class	49	12.36
Keeping up with technological advancements	Yes	322	81,31
	No	74	18,69
The belief in the necessity of using artificial intelligence in foreign language education	Yes	341	86,11
	No	55	13,89
Participant Characteristics of Confirmatory Factor Analysis		**n**	%
Gender	Female	401	65,7
	Male	209	34,3
Age	Max	49	
	Min	17	
	Mean	20.44 ± 3.73	
Grade	Preparatory	366	60.0
	1st Class	92	15.1
	2nd Class	55	9.0
	3rd Class	84	13.8
	4th Class	13	2.1
Keeping up with technological advancements	Yes	459	75.2
	No	151	24.8
The belief in the necessity of using artificial intelligence in foreign language education	Yes	486	79.7
	No	124	20.3

within the context of language learning. The review facilitated the identification of key dimensions and constructs relevant to the study's objectives.

Based on these dimensions, an initial pool of 30 items was created, ensuring that each item corresponded to a specific aspect of attitudes toward AI and comprehensively reflected the construct the scale aimed to measure. Expert evaluations were then employed to refine these items, ensuring both theoretical coherence and practical applicability. Ten different experts from the fields of foreign language education (5 experts), measurement and evaluation (2 experts), and computer and instructional technology education (3 experts) then evaluated these items. The content validity was assessed using the Lawshe method, and an expert opinion form was created. This form described the objectives of the scale, and experts were asked to rate each item as essential, helpful but not essential, or unnecessary. Experts were also encouraged to provide suggestions for any improvements.

To evaluate the clarity of the items and the internal validity of the scale, a pilot study was carried out with 32 students. Şeker and Gençdoğan (2014) recommend selecting 30–50 participants from the target group for a pilot study. Following the pilot study, the scale was administered to a larger sample of 1006 participants, and the collected data was analyzed. Of the students who participated in the main application, 102 were selected for a test–retest conducted four weeks later. According to Tavşancıl (2010), the test–retest should be administered to at least 30 individuals. Additionally, Seçer (2015) suggests that the ideal time interval for test–retest reliability should be between 15 and 30 days.

3.4 Measurement instrument and data collection

This research used a 5-point Likert-type preliminary scale consisting of 24 items to assess students' usage of ChatGPT. The scale's response options ranged from strongly disagree (1) to strongly agree (5), with item 4 as a reverse-coded statement. Additionally, demographic and background information, including gender, age, department, and university affiliation, were collected on the scale form. The survey was administered in person to all accessible students over five months. Students were briefed on the study's objectives before participation, and voluntary consent forms were provided. Participation in the survey was strictly voluntary. Students were instructed to select the response option that best represented their views on the scale form.

3.5 Data analysis

For data analysis, SPSS 25 and AMOS statistical software were used. Two separate samples were utilized for the exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). The EFA sample comprised 396 participants, and the CFA sample comprised 610 participants, totaling 1006 individuals. EFA was performed to determine the scale's construct validity, employing principal components analysis with the direct noblemen method to investigate its factor structure. Due to the correlated nature of the factors, the direct noblemen rotation method, known for its oblique nature, was selected. The suitability of the data for factor analysis was assessed using the Kaiser–Meyer–Olkin (KMO) measure and Bartlett's test of sphericity. Additionally, item-total correlation coefficients were recalculated to assess the collected data, and Cronbach's alpha was computed to evaluate reliability. Test–retest reliability analysis assessed the measurement tool's consistency over time. The Pearson correlation coefficient was calculated to examine the relationship between measurements taken at two different time points on the same sample, and ICC values were also computed. CFA was then employed to validate the sub-dimensions identified by EFA. Alongside CFA, convergent and discriminant validity analyses were conducted to evaluate construct validity comprehensively.

4 Results

4.1 Content validity and pilot application results

Utilizing the Lawshe method, the content validity ratio was determined based on the evaluations of 10 experts. With an acceptable content validity ratio of 0.62 for each item, it was agreed to omit six items (Items 6, 15, 17, 18, 27, 29) falling below this threshold from the scale. Subsequently, the content validity ratios of the remaining items were computed and yielded an average of 0.852. Thus, a draft scale comprising 24 items was formulated. In the pilot study conducted with the participation of 32 students, it was reported that they did not encounter any problems understanding the meaning of all the scale items. Additionally, the Cronbach Alpha value of the scale was calculated as 0.90.

4.2 Reliability and EFA results of the scale

As part of the test–retest analysis conducted to assess the consistency and reliability of the scale over time, the Pearson correlation coefficient was found to be 0.82 ($p < 0.001$). The ICC value for the Single Measures model was determined to be 0.785 (95% CI [0.697, 0.849]), and for the Average Measures model, the ICC value was 0.897 (95% CI [0.822, 0.919]). In both cases, the p-value was found to be < 0.001, indicating that the obtained ICC values are statistically significant.

The EFA determined that the KMO value was 0.927, and Bartlett's Test result was significant ($p < 0.05$). These results indicate that the sample size and the test are highly consistent. As a result of the analysis aimed at revealing the factor structure of the ChatGPT Usage Scale for foreign language learners, it was found that the scale consists of 18 items and three dimensions. The emerging structure explains 62.896% of the total variance. Factor loadings for scale items range from 0.57 to 0.86, indicating a high level of scale reliability (Cronbach's Alpha of the Scale: 0.930, Development Sub-dimension: 0.84, Learning Sub-dimension: 0.89, Overall Usage Sub-dimension: 0.829). Detailed information on the EFA findings is presented in Table 2.

As shown in Table 3, for the ChatGPT Usage Scale for foreign language learners, item-total correlation coefficients range from 0.420 to 0.723. These coefficients are more significant than 0.30 (Büyüköztürk, 2007), which indicates that the items are consistent with the entire scale and contribute to its internal consistency.

4.3 The CFA results of the scale

CFA was conducted to assess the alignment of the ChatGPT Usage Scale for foreign language learners with its factors. As a result, it was determined that certain items (4, 5, 8, 16, 20, and 21) exhibited cross-loadings, prompting their exclusion from the analysis for refinement. Covariance values were examined to facilitate the refinement process, and variables within the same factor with the highest

Table 2 EFA results for the chatGPT usage Scale for foreign language learners

Scale Items	Development	Learning	Overall Usability
1. Using ChatGPT is easy for me			0,863
2. I use ChatGPT regularly			0,709
3. I use ChatGPT effectively			0,828
7. I use ChatGPT to learn new information		0,632	
9. I use ChatGPT to study for my exams		0,753	
10. I use ChatGPT to learn subjects I am lacking in		0,815	
11. I use ChatGPT to access learning materials		0,737	
12. I use ChatGPT because it provides versatile learning		0,647	
13. I use ChatGPT to enhance my linguistic skills		0,651	
14. I use ChatGPT to learn communication skills		0,584	
19. I use ChatGPT to enhance my analytical and critical thinking skills	0,609		
22. I use ChatGPT to improve my creative thinking skills	0,663		
23. I use ChatGPT to obtain different views and ideas	0,628		
24. I use ChatGPT to increase my learning motivation	0,787		
25. I use ChatGPT to make my learning process more enjoyable	0,721		
26. I use ChatGPT to boost my self-confidence in language learning	0,687		
28. I use ChatGPT to enhance my learning self-efficacy	0,668		
30. I use ChatGPT for inspiration	0,575		
Eigenvalue	8,284	1,745	1,292
Explained Variance	46,022	9,696	7,178
Total Variance %	46,022	55,718	62,896
Kaiser–Meyer–Olkin Measure	0,927		
Bartlett's Test	X^2	4104,285	
	df	153	
	Sig	0,001	
Cronbach's Alpha Total: 0,930	0,84	0,89	0,89

values were associated. As seen in Fig. 1, modification indices were also examined, and associations were established between variables with the highest Modification Index (M.I.) scores (e9 < – > e10 and e13 < – > e14).

In CFA, the factor loading of items is examined considering the standard regression coefficients. Additionally, the adequacy of factor loadings and the significance of standard regression coefficients are assessed to make decisions (Kartal & Bardakçı, 2018). Table 4 below shows that the standard regression coefficients of the items are statistically significant.

When examining the fit indices obtained due to the modification, it is observed that the 3-factor structure with 18 items created in Fig. 1 is compatible with the data in Table 5. According to Table 5, the model's goodness of fit obtained from the EFA has been confirmed through structural equation modelling. The model's

Table 3 Item-total score correlations for chatGPT usage scale

Items	r
Item 1	0,420
Item 2	0,550
Item 3	0,546
Item 7	0,622
Item 9	0,619
Item 10	0,678
Item 11	0,659
Item 12	0,723
Item 13	0,624
Item 14	0,606
Item 19	0,686
Item 22	0,701
Item 23	0,621
Item 24	0,679
Item 25	0,654
Item 26	0,649
Item 28	0,698
Item 30	0,554

significance and fit indices are at acceptable levels, and CMIN/df is close to acceptable levels.

For assessing the reliability of the factors in the ChatGPT Usage Scale for foreign language learners, CR (Composite Reliability) was preferred. For convergent and discriminant validity, AVE (Average Variance Extracted), MSV (Maximum Shared Variance), and ASV (Average Shared Variance) values were calculated. Convergent validity suggests AVE > 0.5, CR > 0.7, and CR > AVE, while discriminant validity requires MSV < AVE, ASV < AVE, and $\sqrt{\text{AVE}}$ > inter-factor correlations (Gürbüz, 2019). Table 6 presents the CR, AVE, MSV, and ASV values and correlation coefficients for the factors in the ChatGPT Usage Scale for foreign language learners.

According to the results in Table 6, all factors exhibit high reliability, with CR values exceeding 0.70. The AVE values for the factors are lower than the CR values, above 0.5, indicating convergent validity. Moreover, the AVE values for the factors are higher than the MSV and ASV values, indicating the presence of discriminant validity for the factors. Additionally, the $\sqrt{\text{AVE}}$ scores for the factors are higher than the inter-factor correlations, indicating the presence of discriminant validity.

5 Discussion

Recently, AI-powered chatbots have garnered increasing interest, particularly in foreign language education. In this context, popular AI tools such as ChatGPT have been the subject of various studies in foreign language teaching and have been evaluated as significant tools (Hong, 2023a, 2023b; Kim et al., 2023). However,

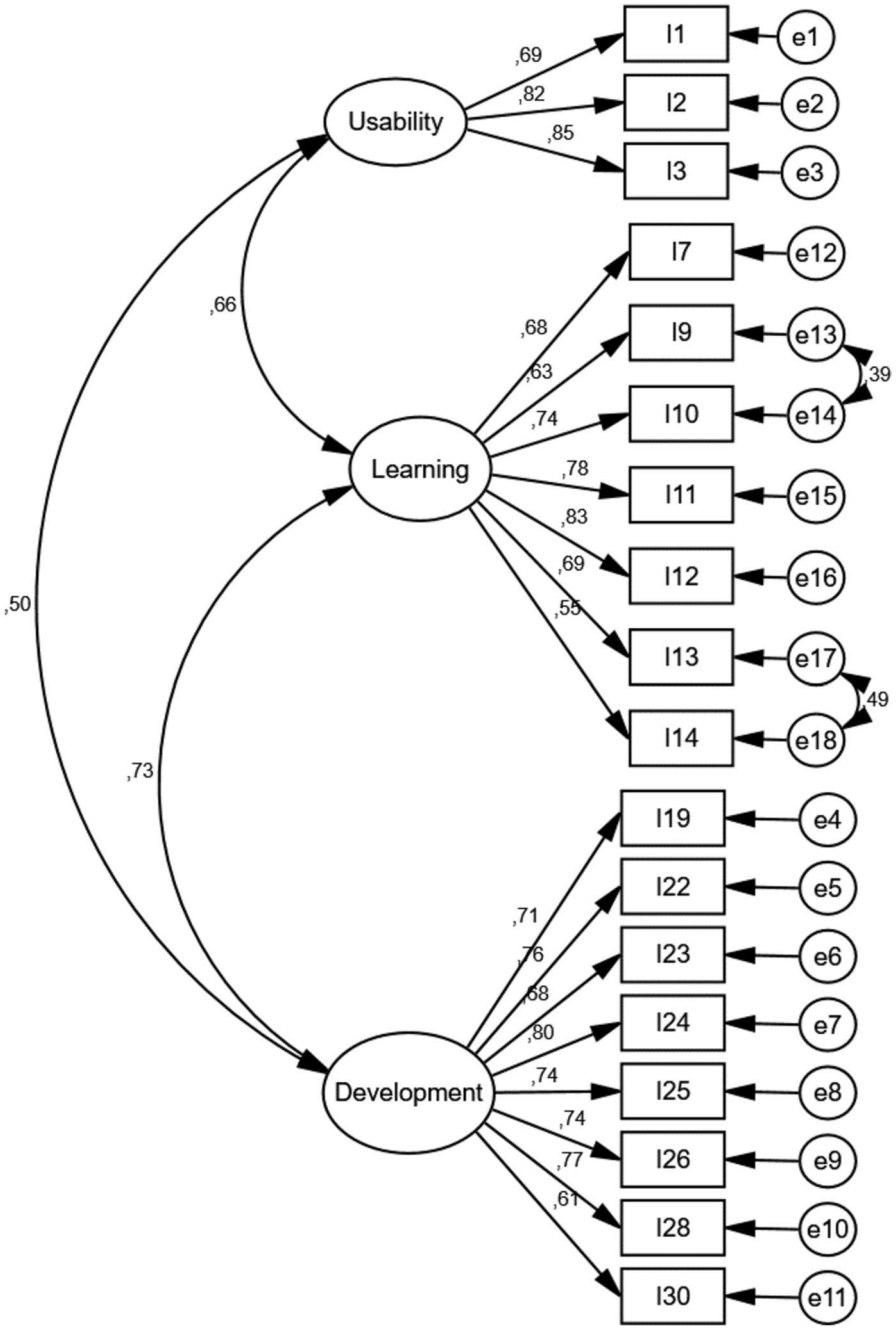

Fig. 1 Multifactorial model of the ChatGPT usage scale for foreign language learners

Table 4 Item/factor loading values

Relationships			Standard, Regression, Weight	Regression Weight	S.E	C.R	P
I1	<—	Usability	0,687	1			
I2	<—	Usability	0,818	1,121	0,066	17,031	***
I3	<—	Usability	0,847	1,284	0,074	17,271	***
I7	<—	Learning	0,683	1			
I9	<—	Learning	0,631	0,975	0,069	14,115	***
I10	<—	Learning	0,741	1,148	0,07	16,36	***
I11	<—	Learning	0,782	1,176	0,068	17,178	***
I12	<—	Learning	0,832	1,248	0,069	18,083	***
I13	<—	Learning	0,695	1,092	0,071	15,464	***
I14	<—	Learning	0,549	0,807	0,065	12,411	***
I19	<—	Development	0,707	1			
I22	<—	Development	0,757	1,152	0,065	17,65	***
I23	<—	Development	0,68	1,035	0,065	15,894	***
I24	<—	Development	0,804	1,159	0,062	18,692	***
I25	<—	Development	0,74	1,096	0,063	17,264	***
I26	<—	Development	0,741	1,074	0,062	17,27	***
I28	<—	Development	0,772	1,12	0,062	17,989	***
I30	<—	Development	0,612	0,972	0,068	14,331	***

Table 5 Fit Indices of model

Fit Indices	After Modification	Acceptable Fit	Good Fit
CMIN/df	5,542	≤5	≤3
GFI	0,87	≥0.85	≥0.90
IFI	0,906	≥0.90	≥0.95
CFI	0,906	≥0.95	≥0.97
RMSEA	0,086	≤0.08	≤0.05
NFI	0,90	≥0.90	≥0.95

Hooper et al. (2008). Structural Equation Modelling: Guidelines for Determining Model Fit, Dublin Institute of Technology Articles, s.55

Table 6 Reliability and validity values for the factors

Factors	CR	AVE	MSV	ASV	$\sqrt{AVE}$		
					1	2	3
1. Usability	0,62	0,72	0,44	0,34	**[0,85]**		
2. Learning	0,50	0,87	0,54	0,49	(0,66)	**[0,93]**	
3. Development	0,53	0,90	0,54	0,40	(0,50)	(0,74)	**[0,94]**

The values in square brackets [] represent the $\sqrt{AVE}$ scores, while those in parentheses () indicate the correlation coefficients

reviewing the current literature reveals a lack of measurement tools explicitly designed to assess the use of AI-based chatbots like ChatGPT in foreign language learning among students. This gap highlights the need for a robust measurement tool to evaluate and assess the impact of AI-supported tools on foreign language learning. This study aims to establish a foundation for further groundbreaking research in AI-supported foreign language learning by developing such a tool.

EFA of the ChatGPT Usage Scale revealed a three-factor structure: "Overall Usability," "Learning," and "Development." The first factor reflects ChatGPT's general usability and widespread usage level. The second factor evaluates the direct contribution of using ChatGPT to the language learning process, including acquiring new knowledge, studying, and learning unfamiliar subjects. The third factor assesses ChatGPT's ability to support and enhance the language learning process outside traditional learning activities. According to EFA, the ChatGPT Usage Scale explains 62.896% of the total variance, which is significantly higher than the recommended 30% for multi-factor scales (Büyüköztürk, 2007). Therefore, the explained variance ratio for the three-dimensional structure is considered sufficient.

The Development factor contributes the most to the total variance (46.022%), followed by the Learning factor (9.696%) and the Overall Usability factor (7.178%). This substantial contribution of the Development factor suggests that ChatGPT significantly influences various aspects of users' language learning processes, such as enhancing language skills, boosting confidence, and improving communication. The high and significant item-total correlation coefficients (ranging from 0.420 to 0.723) indicate that the items are correctly assigned to their respective factors and demonstrate structural consistency. Additionally, the high KMO measure of 0.927 and the significant Bartlett's Test ($p < 0.001$) confirm that the data is suitable for factor analysis. The high Cronbach's Alpha value of 0.93 indicates the scale's strong internal consistency and reliability (George & Mallery, 2003). In the literature, other ChatGPT scales show similar results. In the study by Sallam et al. (2023), the ChatGPT usage scale for health students explained 72% of the total variance with four factors. The subscales demonstrated good reliability with Cronbach's α values > 0.78. In the study by Lee and Park (2024), the five-factor ChatGPT Literacy Scale explained 61% of the total variance, and Cronbach's alpha values were above 0.70. These results are similar to the ChatGPT usage scale we developed for foreign language learners. The high-reliability values obtained in all three studies support the reliability of the scales.

To assess the measurement tool's reliability over time, a test–retest was conducted, yielding a Pearson correlation coefficient of 0.82 ($p < 0.001$) and ICC values of 0.785 (95% CI [0.697, 0.849]) for Single Measures and 0.897 (95% CI [0.822, 0.919]) for Average Measures. According to Cohen (1988), Pearson correlation coefficients above 0.80 indicate strong reliability. Additionally, Koo and Li (2016) suggest that ICC values above 0.75 denote good reliability, supporting this study's findings.

CFA was conducted to test the three-factor structure obtained from the EFA. Two modifications improved the model fit, resulting in the following indices: CMIN/df = 5.542, GFI = 0.87, IFI = 0.906, CFI = 0.906, RMSEA = 0.086, and NFI = 0.90. These values suggest that the model fits the data well (Hooper et al., 2008). CMIN/df is

one of the fundamental fit indices testing the overall adequacy of the obtained model. The resulting outcome is marginally above the acceptable limit (≤ 5). According to McIntosh (2007), even if the model is correct, there is a possibility that the CMIN/df index may indicate that the model is inadequate. This is because CMIN/df is highly sensitive to sample size. Therefore, when large samples are used, it is likely that the model will be rejected almost every time (Bentler & Bonnet, 1980). Accordingly, we can attribute the non-attainment of the desired outcome from our CMIN/df value not to the model's inadequacy but rather to using a large sample size.

The strong positive relationships between measurement items and factors, as evidenced by significant standard regression coefficients, indicate that the scale accurately measures the intended variables. To ensure scale validation, convergent and discriminant validity analyses were performed. All factors exhibited high reliability, with CR values exceeding 0.70. The factors' AVE values were higher than the CR values and above 0.5, indicating convergent validity. Additionally, the AVE values exceeded the MSV and ASV values, demonstrating discriminant validity among the factors. The $\sqrt{\text{AVE}}$ scores for the factors also surpassed inter-factor correlations, further reinforcing discriminant validity (Gürbüz, 2019).

Increasing scores from each subscale indicate that the individual possesses more of the behavior assessed by that subscale. A total score can also be obtained from the scale, with higher total scores indicating a greater tendency to use ChatGPT. The scale's final version with 18 items includes no reverse-scored items, with a possible total score range from 18 to 90.

6 Limitations

This study was conducted with university students from three state universities in Turkey, and the scale's validity and applicability in different cultural contexts have not been explicitly examined. Therefore, the generalizability of the findings to different populations, such as individuals from various age groups, educational backgrounds, or cultural environments, may be limited. Future research should explore the scale's validity and applicability across diverse cultural settings and conduct additional psychometric analyses to enhance its reliability and validity.

The scale items do not focus on any specific version of ChatGPT. Currently, we believe that the scale is applicable to all versions of ChatGPT. Therefore, we do not anticipate that technological advancements will negatively impact the scale significantly for a while. However, considering that ChatGPT and similar tools are constantly evolving, we still believe it would be beneficial to focus on how the scale can be updated and improved in line with future technological advancements.

7 Conclusion

The study aimed to develop a scale to measure and evaluate the usage of ChatGPT for foreign language learners. Following the analyses, the "ChatGPT Usage Scale for Foreign Language Learners" was developed as a valid and reliable measurement

tool. We believe this scale will advance the field, provide valid and reliable data collection opportunities, enable cross-cultural usage, offer a new perspective to the literature, and shed light on future research. Educators can use this scale to assess how and to what extent students effectively use ChatGPT and similar AI tools. These assessments can provide valuable insights for determining educational strategies and developing personalized learning plans. Additionally, this scale can help educational institutions measure the effectiveness of AI-supported educational tools and optimize their integration into educational programs. Lastly, it is recommended that the validity and applicability of the scale in different cultural contexts be examined in future studies. Additionally, future research could conduct additional psychometric analyses to enhance the reliability and validity of the scale.

Authors' contributions FÇ analyzed and interpreted the research data and was a major contributor in writing the manuscript**.** ÖÖ was a contributor in writing the manuscript and a major contributor in collecting data. All authors read and approved the final manuscript.

Funding Open access funding provided by the Scientific and Technological Research Council of Türkiye (TÜBİTAK). No funding was received to assist with the preparation of this manuscript.

Data availability The datasets generated and/or analyzed during the current study are available in the OSF repository, [https://osf.io/w9pk4/?view_only=c08dca76ee8e4cf5bc0bfba73046d568].

Declarations

Ethical approval After receiving the required permissions from the Gazi University Research Ethics Committee (Date: 23.05.2023, Decision No: 2023—723), students participated voluntarily and were informed about the research. Their data was kept confidential and only used for the study. They had the right to withdraw from the study at any time.

Competing interests The authors declare that they have no competing interests.

References

Athanassopoulos, S., Manoli, P., Gouvi, M., Lavidas, K., & Kovis, V. (2023). The use of ChatGPT as a learning tool to improve foreign language writing in a multilingual and multicultural classroom. *Advances in Mobile Learning Educational Research, 3*(2), 818–824. https://doi.org/10.25082/AMLER.2023.02.009

Baidoo-Anu, D., & OwusuAnsah, L. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of AI, 7*(1), 52–62. https://doi.org/10.61969/jai.1337500

Bentler, P. M., & Bonnet, D. C. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*(3), 588–606. https://doi.org/10.1037/0033-2909.88.3.588

Berşe, S., Akça, K., Dirgar, E., & Kaplan, S. E. (2023). The role and potential contributions of the artificial intelligence language model ChatGPT. *Annals of Biomedical Engineering, 52*(2), 130–133. https://doi.org/10.1007/s10439-023-03296-w

Bonner, E., Lege, R., & Frazier, E. (2023). Large language model-based artificial intelligence in the language classroom: practical ideas for teaching. *Teaching English with Technology, 23*(1), 23–41. https://doi.org/10.56297/BKAM1691/WIEO1749

Bozkurt, A. (2023). Generative artificial intelligence (AI) powered conversational educational agents: The inevitable paradigm shift. *Asian Journal of Distance Education, 18*(1), 198–204. https://doi.org/10.5281/zenodo.7716416

Büyüköztürk, Ş. (2007). *Sosyal bilimler için veri analizi el kitabı* (7th ed.). Pegem Akademi.

Cassidy, C. (2023). Australian universities to return to 'pen and paper' exams after students caught using AI to write essays. The Guardian Online. Available at https://www.theguardian.com/australia-news/2023/jan/10/universities-to-return-to-pen-and-paper-exams-after-students-caughtusing-ai-to-write-essays (Accessed 20/03/2024).

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Farhi, F., Jeljeli, R., Aburezeq, I., Dweikat, F. F., Al-shami, S. A., & Slamene, R. (2023). Analyzing the students' views, concerns, and perceived ethics about ChatGPT usage. *Computers and Education: Artificial Intelligence, 5*, 1–8. https://doi.org/10.1016/j.caeai.2023.100180

Firat, M. (2023). How Chat GPT can transform autodidactic experiences and open education?. https://doi.org/10.31219/osf.io/9ge8m

Fitria, T. N. (2023). Artificial intelligence (AI) technology in OpenAI ChatGPT application: A review of ChatGPT in writing English essay. *Journal of English Language Teaching, 12*(1), 44–58. https://doi.org/10.15294/elt.v12i1.64069

Fryer, L. K., Ainley, M., Thompson, A., Gibson, A., & Sherlock, Z. (2017). Stimulating and sustaining interest in a language course: An experimental comparison of Chatbot and human task partners. *Computers in Human Behavior, 75*, 461–468. https://doi.org/10.1016/j.chb.2017.05.045

Fuchs, K. (2023). Exploring the opportunities and challenges of NLP models in higher education: is Chat GPT a blessing or a curse? *Frontiers in Education. 8*:1166682. https://.3389/feduc.2023.1166682

George, D., & Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference. 11.0 update* (4th ed.). Allyn & Bacon.

Gürbüz, S. (2019). *Structural equation modeling with AMOS* (1st ed.). Seçkin Yayıncılık.

HadiMogavi, R., Deng, C., Kim, J. J., Zhou, P., Kwone, Y. D., Metwallya, A. H. S., Tlili, A., Bassanelli, S., Bucchiarone, A., Gujarh, S., Nackea, L. E., & Hui, P. (2024). ChatGPT in education: A blessing or a curse? A qualitative study exploring early adopters' utilization and perceptions. *Computers in Human Behavior: Artificial Humans, 2*(1), 100027. https://doi.org/10.1016/j.chbah.2023.100027

Hong, W. C. H. (2023). The impact of ChatGPT on foreign language teaching and learning: Opportunities in education and research, *Journal of Educational Technology and Innovation, 5*(1), 37–45. https://jeti.thewsu.org/index.php/cieti/article/view/103/64

Hong, W. C. H. (2023). The impact of ChatGPT on foreign language teaching and learning: Opportunities in education and research, *Journal of educational technology and innovation, 5*(1), 37–45. https://jeti.thewsu.org/index.php/cieti/article/view/103/64 (Accessed on 3 April 2024).

Hooper, D., Coughlan, J., & Mullen, M. R. (2008). Structural equation modeling: Guidelines for determining model fit structural equation modeling. *Dublin Institute of Technology ARROW@ DIT, 6*(1), 53–60. https://doi.org/10.21427/D7CF7R

Houston, A. B., & Corrado, E. M. (2023). Embracing ChatGPT: Implications of emergent language models for academia and libraries. *Technical Services Quarterly, 40*(2), 76–91. https://doi.org/10.1080/07317131.2023.2187110

Huang, J., & Li, S. (2023). Opportunities and challenges in the application of ChatGPT in foreign language teaching. *International Journal of Education and Social Science Research, 6*(4), 75–89. https://doi.org/10.37500/IJESSR.2023.6406

Jiao, W., Wang, W., Huang, J., Wang, X., Shi, S. & Tu, Z. (2023). *Is ChatGPT a good translator? Yes, with GPT-4 as the engine*. arXiv:2301.08745. https://doi.org/10.48550/arXiv.2301.08745

Kalla, D., Smith, N., Samaah, F., & Kuraku, S. (2023). Study and Analysis of Chat GPT and its Impact on Different Fields of Study. *International Journal of Innovative Science and Research Technology, 8*(3), 827–833. https://doi.org/10.5281/zenodo.7767675

Karakaya, İ. (2012). Bilimsel araştırma yöntemleri [Scientific research methods]. In A. Tanrıöğen (Ed.), *Bilimsel araştırma yöntemleri [Scientific research methods]* (pp. 55–84). Anı.

Kartal, M., & Bardakçı, S. (2018). *SPSS ve AMOS uygulamalı örneklerle güvenirlik ve geçerlik analizleri* (1st ed.). Akademisyen Kitabevi

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*(102274), 1–9. https://doi.org/10.1016/j.lindif.2023.102274

Kim, S., Shim, J., & Shim, J. (2023). A study on the utilization of OpenAI ChatGPT as a second language learning tool. *Journal of Multimedia Information System, 10*(1), 79–88. https://doi.org/10.33851/JMIS.2023.10.1.79

Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). ChatGPT for language teaching and learning. *RELC Journal, 54*(2), 1–14. https://doi.org/10.1177/0033688223116286

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Kostka, I. & Toncelli, R. (2023). Exploring applications of chatgpt to english language teaching: Opportunities, challenges, and recommendations. *The Electronic Journal for English as a Second Language (TESL-EJ)*, *27*(3), 1-19. https://doi.org/10.55593/ej.27107int

Lee, S., & Park, G. (2024). Development and validation of ChatGPT literacy scale. *Current Psychology, 43*, 18992–190041. https://doi.org/10.1007/s12144-024-05723-0

McIntosh, C. N. (2007). Rethinking fit assessment in structural equation modelling: A commentary and elaboration on Barrett (2007). *Personality and Individual Differences, 42*(5), 859–867. https://doi.org/10.1016/j.paid.2006.09.020

Meyer, J. G., Urbanowicz, R. J., Martin, P. C. N., O'Connor, K., Li, R., Peng, P. C., Bright, T. J., Tatonetti, N., Won, K. J., Gonzalez-Hernandez, G., & Moore, J. H. (2023). ChatGPT and large language models in academia: opportunities and challenges. *BioData Mining, 16*(20), 1–11. https://doi.org/10.1186/s13040-023-00339-9

Pichai, S. & Hassabis, D. (2024). Our next-generation model: Gemini 1.5. https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/#sundar-note

Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems, 3*, 121–154. https://doi.org/10.1016/j.iotcps.2023.04.003

Sallam, M., Salim, N. A., Barakat, M., Al-Mahzoum, K., Ala'a, B., Malaeb, D., Hallit, R., & Hallit, S. (2023). Assessing health students' attitudes and usage of ChatGPT in Jordan: Validation study. *J MIR Medical Education, 9*(1), e48254. https://doi.org/10.2196/48254

Seçer, İ. (2015). SPSS ve LISREL ile pratik veri analizi: Analiz ve raporlaştırma (2nd ed.). Anı Yayıncılık.

Şencan, H. (2005). *Validity and reliability in social and behavioural measurements* (1st ed.). Seçkin.

Singh, O. (2023). Artificial intelligence in the era of ChatGPT - Opportunities and challenges in mental health care. *Indian Journal of Psychiatry, 65*(3), 297–298. https://doi.org/10.4103/Indianpsychiatry.indianjpsychiatry_112_23

Tabachnick, B. G., Fidell, L. S., & Ullman, J. B. (2019). *Using multivariate statistics* (7th ed.). Pearson.

Tavşancıl E. (2010). *Tutumların ölçülmesi ve SPSS ile veri analizi* (6th ed.). Nobel Yayın Dağıtım.

Tlili, A., Shehata, B., Adarkwah, M. A., Bozkurt, A., Hickey, D. T., Huang, R., & Agyemang, B. (2023). What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. *Smart Learn. Environ., 10*, 15. https://doi.org/10.1186/s40561-023-00237-x

Topsakal, O., & Topsakal, E. (2022). Framework for a foreign language teaching software for children utilizing AR, voicebots and ChatGPT (large language models). *The Journal of Cognitive Systems, 7*(2), 33–38. https://doi.org/10.52876/jcs.1227392

Yu, H. (2023). Reflection on whether Chat GPT should be banned by academia from the perspective of education and teaching. *Frontiers in Psychology, 14*, 1181712. https://doi.org/10.3389/fpsyg.2023.1181712

Zhang, B. (2023). ChatGPT, an opportunity to understand more about language models. *Medical Reference Services Quarterly, 42*(2), 194–201. https://doi.org/10.1080/02763869.2023.2194149