



**THE REPUBLIC OF TÜRKİYE
SOCIAL SCIENCES UNIVERSITY OF ANKARA
THE GRADUATE SCHOOL OF SOCIAL SCIENCES**

**The Use of Artificial Intelligence in L2 Reading Test Development: A Study of
Language Instructors' Perceptions and Experiences in an English Preparatory School**

Master's Thesis

Gülçin AKYOL

Foreign Languages Education (Interdisciplinary)

Master of Science in English Language Education

August 2025



**THE REPUBLIC OF TÜRKİYE
SOCIAL SCIENCES UNIVERSITY OF ANKARA
THE GRADUATE SCHOOL OF SOCIAL SCIENCES**

**The Use of Artificial Intelligence in L2 Reading Test Development: A Study of
Language Instructors' Perceptions and Experiences in an English Preparatory School**

Master's Thesis

Gülçin AKYOL

Advisor

Assoc. Prof. Dr. Aysel SARICA OGLU AYGAN

Foreign Languages Education (Interdisciplinary)

Master of Science in English Language Education

August 2025

ACKNOWLEDGEMENTS

I would first like to express my deepest gratitude to my thesis advisor, **Assoc. Prof. Dr. Aysel Sarıcaoğlu Aygan**, whose support went far beyond the role of an academic advisor. She was a mentor, a guide, and a source of immense understanding throughout this journey, especially after the birth of my daughter. Her encouragement and unwavering belief in me have been invaluable.

I am also grateful to my former advisor, **Prof. Dr. Julie Ann Aydınlı**, who initially appreciated and supported my idea of working on this topic. Her encouragement at the early stages of this study and her kindness will always be remembered with gratitude.

My sincere thanks also go to the esteemed jury members, **Assoc. Prof. Dr. Hakan Demiröz** and **Assoc. Prof. Dr. Zeynep Bilki**, for their time, their invaluable feedback, and their thoughtful engagement during my defense.

I would also like to thank my classmates **Özlem** and **Merve**, with whom I started this journey. We shared not only our studies but also the struggles along the way. I am also thankful to my lifelong friend **Mediha** for her constant presence and support.

Special thanks go to my sister **Elçin** and my sister-in-law **Ülkü**, to whom I could not allocate as much time as I would have wished during this demanding process. I deeply appreciate their consistent presence even when I could not be fully present for them.

Most importantly, I owe my deepest gratitude to my husband, **Sinan** - my partner, my friend, and my everything. Throughout this journey, he not only gave up his personal and professional time to stand by me but also cared for both me and our daughter with endless devotion. His quiet strength and constant presence made it possible for me to continue when it felt impossible. Without his support, this thesis could not have been completed.

Finally, I dedicate this work to my precious twins, **İmge** and **Tibet**, who arrived too early and left this world far too soon, yet whose presence will always be cherished in my heart, and to my beloved daughter **Devin**, whose arrival has brought me renewed strength, hope and purpose. In her first months of life, I often felt as though I had stolen moments that should have been hers. Yet Devin is my greatest gift. Her presence makes each day brighter, and I eagerly await every new phase and journey we will experience together.

To my beloved twins - Ímge and Tibet,
whose brief time in this world will forever remain with me,

and to my little miracle Devin,
who has given me a new beginning filled with peace, joy, and light.

This work is dedicated to you, with all my love.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS	i
TABLE OF CONTENTS	iii
ABSTRACT	vi
ÖZET	vii
LIST OF FIGURES	viii
LIST OF TABLES	x
LIST OF ABBREVIATIONS	xi
1. INTRODUCTION	1
1.1. Introduction	1
1.2. Background of the Study	2
1.3. Statement of the Problem	4
1.4. Research Questions	5
1.5. Significance of the Study	5
1.6. Definition of Key Terms	7
2. LITERATURE REVIEW	8
2.1. Introduction	8
2.1.1. Theoretical Framework	9
2.1.2. TAM in CALL Studies	11
2.2. Artificial Intelligence (AI)	14
2.2.1. ChatGPT	15
2.2.2. Artificial Intelligence in Education (AIEd)	17
2.2.3. Artificial Intelligence in English Language Teaching (ELT)	17
2.3. English Language Testing	19
2.3.1. Testing Reading	19
2.3.2. Multiple-Choice Test Items	20
2.3.3. Artificial Intelligence in English Language Testing	20
2.4. Teachers' Perceptions of Artificial Intelligence in English Language Testing	21
2.5. Conclusion	22

3.	METHODS	23
3.1.	Introduction	23
3.2.	Research Design	23
3.3.	Participants	24
3.4.	Data Collection	26
3.5.	Data Analysis	29
3.6.	Ethical Considerations	35
3.7.	Conclusion	36
4.	FINDINGS	37
4.1.	Introduction	37
4.2.	EFL Instructors' ChatGPT Use for L2 Reading Test Development	37
4.3.	Differences in EFL Instructors' Use of ChatGPT for Elementary-level and Intermediate-level L2 Reading Task Development	103
4.4.	EFL Instructors' Perceptions of Using ChatGPT for L2 Reading Test Development	105
4.4.1.	Case 1: Perceptions of Using ChatGPT for Reading Test Development	106
4.4.2.	Case 2: Perceptions of Using ChatGPT for Reading Test Development	107
4.4.3.	Case 3: Perceptions of Using ChatGPT for Reading Test Development	109
4.4.4.	Case 4: Perceptions of Using ChatGPT for Reading Test Development	110
4.5.	Differences in the Perceptions of the Instructors When Developing Elementary-level and Intermediate-level Reading Tasks	111
5.	DISCUSSION	113
5.1.	Introduction	113
5.2.	EFL Instructors' Use of ChatGPT for L2 Reading Test Development	113
5.3.	EFL Instructors' Perceptions of Using ChatGPT for L2 Reading Test Development	116
5.4.	Implications	117
6.	CONCLUSIONS	119

REFERENCES	121
APPENDICES	145
Appendix I: Informed Consent Form	145
Appendix II: General Overview of ChatGPT for Test Development	146
Appendix III: General Overview of Screen Recording	148
Appendix IV: General Overview of Think-aloud Protocols (TAP)	150
Appendix V: ChatGPT Conversation Transcripts	152



**THE USE OF ARTIFICIAL INTELLIGENCE IN L2 READING TEST
DEVELOPMENT: A STUDY OF LANGUAGE INSTRUCTORS' PERCEPTIONS
AND EXPERIENCES IN AN ENGLISH PREPARATORY SCHOOL**

ABSTRACT

This exploratory case study aimed to investigate English language instructors' experiences and perceptions of using ChatGPT-4 for L2 reading test development. To this end, four language instructors working in an English preparatory school of a state university in Türkiye, with testing experience, participated in the study. They were asked to use ChatGPT-4 to prepare an L2 reading test, including two reading tasks with five multiple-choice questions: one at the elementary level and the other at the intermediate level. Data were collected through screen recording of instructors' test development process, think-aloud protocols, and ChatGPT conversation transcripts. Data were analyzed using a content analysis approach.

The findings demonstrated some variation among the instructors in how they engaged with ChatGPT-4 for L2 reading test development, despite the small number of participants. A large proportion of all instructors' prompts dealt with item quality and distractor quality, rather than text quality, where they tried to fix more item and distractor issues in the intermediate-level reading task than in the elementary-level task. The way the instructors carefully checked the ChatGPT output for several aspects made it obvious that human agency and expertise have a crucial role in ensuring that the AI-generated test materials meet the required standards. While their perceptions were mostly positive about the reading texts that ChatGPT generated, except for a few length issues, their comments during test development with ChatGPT-4 revealed negative perceptions regarding several aspects. The instructors perceived their intermediate-level reading task development experience more negatively than their elementary-level reading task development experience.

Keywords: Artificial intelligence, ChatGPT-4, English language testing, Reading, Perceptions, Case study

İKİNCİ DİLDE OKUMA TESTİ GELİŞTİRMEDE YAPAY ZEKA KULLANIMI: BİR İNGİLİZCE HAZIRLIK OKULUNDAKİ ÖĞRETİM ELEMANLARININ ALGILARI VE DENEYİMLERİ ÜZERİNE BİR ÇALIŞMA

ÖZET

Bu keşif amaçlı durum çalışması, İngilizce öğretim elemanlarının ChatGPT-4 kullanarak ikinci dilde (L2) okuma testi geliştirme konusundaki deneyimlerini ve algılarını incelemeyi amaçlamaktadır. Bu doğrultuda, Türkiye’de bir devlet üniversitesinin İngilizce hazırlık okulunda görev yapan ve ölçme-değerlendirme alanında deneyime sahip dört dil öğretim elemanı çalışmaya katılmıştır. Katılımcılardan, ChatGPT-4 kullanarak biri başlangıç düzeyinde, diğeri ise orta düzeyde olmak üzere iki okuma görevini ve her biri beş çoktan seçmeli sorudan oluşan bir İngilizce okuma testi hazırlamaları istenmiştir. Veriler, öğretim elemanlarının test geliştirme sürecine ilişkin ekran kayıtları, sesli düşünme protokolleri ve ChatGPT ile gerçekleştirilen yazışma dökümleri aracılığıyla toplanmıştır. Veriler, içerik analizi yöntemiyle incelenmiştir.

Bulgular, katılımcı sayısının az olmasına rağmen, öğretim elemanlarının ChatGPT-4 ile ikinci dilde okuma testi geliştirme sürecine katılım biçimlerinde belirli farklılıklar olduğunu göstermiştir. Tüm öğretim elemanlarının istemlerinin büyük bir bölümü, metin kalitesinden ziyade soru kalitesi ve çeldirici kalitesine odaklanmış; bu bağlamda, başlangıç düzeyindeki görevden daha fazla olarak, orta düzey okuma görevinde soru ve çeldirici ile ilgili sorunları düzeltmeye çalışmışlardır. Öğretim elemanlarının, ChatGPT çıktısını çeşitli açılardan titizlikle kontrol etme biçimleri, yapay zekâ tarafından üretilen test materyallerinin gerekli standartları karşılamasını sağlamak için insan inisiyatifinin ve uzmanlığının kritik bir rol oynadığını açıkça ortaya koymuştur. ChatGPT tarafından üretilen okuma metinleri hakkındaki algıları, birkaç metin uzunluğu sorunu dışında genel olarak olumlu olmakla birlikte, ChatGPT-4 ile test geliştirme süreci sırasında yaptıkları yorumlar, çeşitli açılara ilişkin olumsuz algılar ortaya koymuştur. Öğretim elemanları, orta düzey okuma görevi geliştirme deneyimlerini, başlangıç düzeyindeki okuma görevi geliştirme deneyimlerine kıyasla daha olumsuz değerlendirmişlerdir.

Anahtar Kelimeler: Yapay zeka, ChatGPT-4, İngilizce dil testi, Okuma, Algılar, Vaka çalışması

LIST OF FIGURES

Figure	Page
1. The Original TAM as Proposed by Davis (1985)	9
2. The Extended Version of TAM as Proposed by Davis, Bagozzi, and Warshaw (1989)	10
3. The Coding of the Screen Recording Data and the Think-aloud Data in Excel	34
4. Instructor 1's 1st Prompt	38
5. The Initial Text ChatGPT Created	39
6. The Initial Questions ChatGPT Created	40
7. The Revised Text ChatGPT Created	41
8. The Text the Instructor Edited	42
9. The Final Draft of the Reading Task Instructor 1 Prepared	43
10. Instructor 1's 1st Prompt	44
11. The Initial Text ChatGPT Created	45
12. The Questions of the Initial Text ChatGPT Created	46
13. The Final Draft of the Text Instructor 1 Prepared	48
14. The Final Draft of the Questions Instructor 1 Prepared	49
15. ChatGPT's Initial Response	51
16. The Initial Text ChatGPT Created	52
17. The Questions of the Initial Text ChatGPT Created	53
18. The Revised Text ChatGPT Created	54
19. ChatGPT's Response	55
20. Instructor 2's 1st Prompt	57
21. Instructor 2's 2nd Prompt	58
22. ChatGPT's Response	58
23. ChatGPT's Response	60
24. ChatGPT's Response	61
25. Instructor 3's 1st Prompt	62
26. Turkish Notification on ChatGPT Screen	63
27. The First Text ChatGPT Created	64
28. The Revised Text ChatGPT Created	65
29. ChatGPT's Response for Alignment of the Text Level with CEFR	66

30. Instructor 3's 1st Prompt	73
31. The Initial Text ChatGPT Created	74
32. The Two of the Initial Questions ChatGPT Created	75
33. One of The Revised Questions ChatGPT Created	76
34. The Revised Text ChatGPT Created	77
35. The Two of the Revised Questions ChatGPT Created	78
36. Turkish Notification on ChatGPT Screen	79
37. The Revised Questions ChatGPT Created	80
38. Instructor 4's 1st Prompt	84
39. Possible Topics ChatGPT Gave for the Reading Text	84
40. ChatGPT's Response	85
41. Instructor 4's 3rd Prompt	85
42. The Initial Text ChatGPT Created	86
43. Instructor 4's 4th Prompt	87
44. The Questions ChatGPT Created	88
45. The Revised Questions ChatGPT Created	89
46. ChatGPT's Response	89
47. ChatGPT's Response	90
48. ChatGPT's Response	91
49. ChatGPT's Response	92
50. The Question the Instructor Adapted	92
51. Turkish Notification on ChatGPT Screen	93
52. Turkish Notification on ChatGPT Screen	93
53. The Final Draft of the Reading Task Instructor 4 Prepared	94
54. Instructor 4's 1st Prompt	96
55. The Initial Text ChatGPT Created	97
56. The Initial Questions ChatGPT Created	98
57. The Revised Text ChatGPT Created	99
58. The Revised Questions ChatGPT Created	101
59. ChatGPT's Response	102
60. ChatGPT's Response	102
61. Instructor 4's Prompt	103

LIST OF TABLES

Table	Page
1. Demographic Information of the Participants	25
2. Summary of the Video Recordings of Instructors’ Use of ChatGPT	27
3. The Coding Categories, Subcategories, and Specific Codes Used for the Screen Recording Data	30
4. The Coding Categories, Subcategories, and Specific Codes Used for the Think-aloud Data	33



LIST OF ABBREVIATIONS

AI: Artificial Intelligence

L2: Second Language

LLM: Large Language Model

TAM: Technology Acceptance Model

CALL: Computer-Assisted Language Learning

MCQ: Multiple Choice Question

TRA: Theory of Reasoned Action

TPB: Theory of Planned Behavior

PU: Perceived Usefulness

PEU: Perceived Ease of Use

ATU: Attitude Toward Using

ASU: Actual System Use

BI: Behavioral Intention

AIGC: Artificial Intelligence-Generated Content

GPT: Generative Pre-trained Transformer

AIEd: Artificial Intelligence in Education

ELT: English Language Teaching

EFL: English as a Foreign Language

CEFR: Common European Framework of Reference for Languages

EPT: English Placement Test

EPE: English Proficiency Exam

ECT: End-of-Course Test

TAP: Think-Aloud Protocol

CHAPTER 1: INTRODUCTION

1.1 Introduction

As Henning (1987) puts it, we would not be able to talk about science without measurement. Language testing is an indispensable aspect of language education as it enables learners to show their proficiency in a given language and helps teachers to learn about their students' proficiency. English language testing is an essential part of language education, especially in English preparatory schools, where students must prove their proficiency to carry on their academic studies in their departments. As reading is one of the fundamental skills required in real-life situations involving academic and professional settings, it is essential to include reading tests in language assessment. It is also crucial to focus on reading tests with multiple-choice items as standardized English language tests frequently include reading sections in this format. Test developers play a critical role in designing language tests, having to ensure validity, reliability, and fairness, as it is crucial to verify that the generated tests accurately reflect learners' language proficiency.

The technology utilized to design and deliver language tests has evolved alongside language testing over time. One of such advancements that has the potential to transform language testing practices is artificial intelligence (AI). Natural language processing technology has advanced recently, opening up new areas and possibilities for language education and testing. One such technology is ChatGPT, a new large language model that has been trained on a vast database and can generate human-like responses .

However, the use of AI, specifically ChatGPT in the scope of the current study, which is language testing and particularly L2 reading test development, is still a relatively new area of research, and there is no empirical evidence about the perceptions and experiences of language instructors using ChatGPT for L2 reading test development. Thus, this study aims to explore the perceptions and experiences of language instructors with testing experience in an English preparatory school regarding the use of ChatGPT for test development while preparing reading parts of language achievement tests with multiple-choice items.

1.2 Background of the Study

Language tests are as old as language teaching and are commonly used in language programs for different purposes, such as diagnosis, feedback, selection, placement, and program evaluation. Hence, there are many different types of language tests that can be classified distinctly. These can be objective vs. subjective tests, direct vs. indirect tests, discrete-point vs. integrative tests, criterion-referenced vs. norm-referenced tests, speed vs. power tests, as well as diagnostic, placement, achievement, and proficiency tests (Hughes, 2002). The focus of this study is on two different levels of the reading test task of the ECT (End-of-Course Test).

Achievement tests are used to measure the degree to which learners have mastered content within a specified domain, typically aligned with clearly defined objectives of an educational program (Henning, 1987). Designed in line with specific learning objectives of a course or program, achievement tests are generally administered at the end of a course to evaluate how effectively students have mastered the instructional objectives and to help teachers effectively act on what they learn from the test about the students' abilities, needs, etc. (Brown, 1996).

In many educational institutions, including the state university where the current study was conducted, test developers (or test writers or testers) usually bear the responsibility for the preparation of language tests to evaluate the attainment of the goals and objectives of the language program. Test developers, who are to prepare the tests that aim to serve the purpose of the language program in the educational institution, play a pivotal role in the test development process (Aydın, 2004). Test development is the process of creating a measure of some aspect of an individual's knowledge, skills, interests, attitudes, or other traits by developing items to form a test according to a predetermined plan, requiring a methodological approach to validating test score interpretations (Lane et al., 2015).

The test development process involves careful consideration of content, format, the intended use of context, and the potential consequences of using the test. Carefully planning the test development process is very important to ensure that the test achieves its intended purpose (Bachman & Palmer, 1996). Therefore, the construction of test specifications is starting point in test development (Alderson et al., 1995; Bachman & Palmer, 1996; Brown & Hudson, 2002; Davidson & Lynch, 2002; McNamara, 2000). Test specifications, also called

specs or blueprints, are documents that describe the features of the tasks to be created, including goals and objectives, prompts and instructions, length, sample items, and features of the expected responses, or the scoring criteria (Brown & Hudson, 2002; Davidson & Lynch, 2002; Fulcher & Davidson, 2007).

The test tasks and items are written by item writers. Thus, item writing is the next important stage in test development. An item is the basic unit of language testing (Brown, 1996). Test writers can write successful items only with clearly specified learning objectives (Genesee & Upshur, 1996). In the present study, as the test developers were already provided with all the necessary information and documents including the test specifications, learning objectives to be covered, potential themes and topics including the ones covered in the relevant coursebooks, wordlists that students are responsible for, etc., they are only required to create the test task, which is a reading passage within a given word limit for the specified level, and to write the questions, which are multiple-choice items consisting of three options.

Multiple-choice items are the backbone of achievement testing (Rodriguez, 2005) and one major type of item that can be scored quite objectively (Farhady & Shakery, 2000). A multiple-choice item includes a stem that is aimed to elicit information and a number of options. The test taker is required to read the stem and choose one of the options as the correct or best answer. The number of options varies from two (the minimum number) to five (the common maximum number), depending on the purpose of the test and the intention of the test (Farhady & Shakery, 2000). In some situations, three alternatives instead of four or five may be sufficient because it takes students less time to read and answer, and it is also easier to create two good distractors rather than three or four (Bachman & Palmer, 1996). Many studies are also in favor of three-option multiple-choice tests (e.g., Farhady & Shakery, 2000; Landrum et al., 1993; Lee & Winke, 2013; Rodriguez, 2005; Sidick et al., 1994; Vyas & Supe, 2008). In the present study, the language instructors who have experience in test development were asked to prepare three-option multiple-choice questions as they are already familiar with their own test in the institution.

Tests developed and used in educational programs need to be assessed for quality by the stakeholders to ensure they fulfill the program's goals and objectives (Aydın, 2004). The test qualities are "reliability, validity, authenticity, interactiveness, practicality, and impact" (Bachman & Palmer, 1996, p.17). They argue that these qualities contribute to the overall usefulness of a test in a combined way, and that test quality is context-dependent. This is why

these test qualities present a foundation for evaluating language tests within their applicable contexts.

With the recent technological growth, AI-based tools have been created that can assist in diverse aspects of language testing, including item generation, scoring, and feedback provision. One of these AI-based tools is ChatGPT. However, the use of ChatGPT in language testing may also come with its unique concerns and challenges, as well as the benefits and opportunities. Thus, it is vital to explore the perceptions and experiences of language instructors regarding their use of AI in L2 reading test development.

1.3 Statement of the Problem

There is an enormous body of scholarly work on the use of technology in language education (e.g., Bax, 2011; Karabulut-İlgü, 2013; Kessler, 2018) and in language testing (e.g., Álvarez, 2016; Dooley, 2008; Jimenez & Boser, 2021; Suvorov & Hegelheimer, 2013). A number of other studies have explored teachers' perceptions towards the use of technology when teaching languages (e.g., Aljohani, 2021; Alzubi, 2019; Arnold & Ducate, 2015; Ding et al., 2019; Djiwandono, 2019; Huang et al., 2019; Sumakul et al., 2022; Yaşar, 2020; Wilson et al., 2021).

Studies into the use of AI in language teaching (e.g., Ali, 2020; Annamalai et al., 2023; Haristiani, 2019; Hockly, 2023; Huang et al., 2023; Mijwil et al., 2022; Sharadgah & Sa'di, 2022; Yanhua, 2020) began to emerge. There are also some recent studies on the use of AI in language testing, although fewer. One of these studies (Hz et al., 2023) has investigated the potential impacts of AI writing tools on students' writing skills within English proficiency tests, whereas another one (Yunjiu et al., 2022) has examined the use of AI in constructing Chinese vocabulary test items through comparative analysis of human-designed items and AI-generated items. Another recent study (Zhang et al., 2023) has explored the impact of AI integration into language testing through automated scoring systems from the perspective of test-takers.

However, research that specifically explores EFL instructors' perceptions and experiences of AI use in language test development is still lacking. This qualitative case study aims to fill this gap by seeking to investigate the perceptions and experiences of EFL instructors with testing experience regarding the use of ChatGPT in language test development,

specifically reading language tests with multiple-choice items, and to identify potential benefits and challenges associated with using an AI tool for reading test development. After all, the first-hand experiences and perspectives of language instructors are essential for a comprehensive understanding of the impact, benefits, and challenges associated with AI use in the test development process.

1.4 Research Questions

The aim of this study is to explore the perceptions and experiences of EFL instructors regarding the use of ChatGPT in L2 reading test development. In this respect, the study addresses the following research questions:

- 1) How do EFL instructors use ChatGPT-4 for developing L2 reading tasks?
 - 1a) If any, what are the differences in the EFL instructors' use of ChatGPT-4 when developing elementary-level and intermediate-level L2 reading tasks?
- 2) What are the EFL instructors' perceptions while they are using ChatGPT-4 for developing L2 reading tasks?
 - 2a) If any, what are the differences in the perceptions of the EFL instructors when developing elementary-level and intermediate-level reading tasks?

1.5 Significance of the Study

Notwithstanding the rapid progress in artificial intelligence (AI), its practical potential in language testing remains predominantly unexplored. While AI has been increasingly incorporated into various aspects of language education, there is still a limited comprehension of its practical application in language test development, especially in crafting reading tests. While previous studies have explored the potential benefits and challenges of AI tools in language education in general (e.g., Ali, 2023; De la Vall, & Araya, 2023; Godwin-Jones et al., 2024), one of the gaps in the existing research is the lack of qualitative process-oriented studies focusing on the development of reading language tests using AI tools. Additionally, there is

limited focus in the literature on the perceptions and experiences of language instructors on the use of AI tools, such as ChatGPT, in the development of L2 reading tests.

Addressing important gaps in the literature, this study explores how EFL instructors perceive AI use in language test development, how they use ChatGPT-4 to develop L2 reading tasks, and how they perceive this experience. The findings of this study will provide insights into the successful integration of AI in L2 reading test development. It will help to document potential issues of ChatGPT use in creating reading tasks at different proficiency levels. The findings will also offer implications regarding successful pedagogical strategies for incorporating AI tools in L2 reading test development.

Overall, this study will contribute to the growing body of literature on the use of AI tools in English language testing. The findings of this study will have implications for language instructors, test developers, and assessment specialists who are interested in using AI technologies in language testing. By understanding the potential of ChatGPT in L2 reading test development, educators and researchers can better develop effective language tests that meet the needs of language learners and assess their language proficiency accurately and efficiently.

1.6 Definition of Key Terms

Artificial Intelligence (AI) is a branch of computer science that uses algorithms and machine learning techniques to replicate or simulate human intelligence (Helm et al. 2020).

ChatGPT is a public-facing chatbot using generative artificial intelligence (Lund & Wang, 2023).

Large Language Model (LLM) is a type of advanced artificial intelligence system with billions of parameters pre-trained on a large amount of language corpora, providing new opportunities for conversational agents (Wei et al., 2024). Some popular examples of LLMs are ChatGPT (by OpenAI), Copilot (by Microsoft), and Gemini (by Google).

Technology Acceptance Model (TAM) is a powerful tool that explains the influential factors when users adopt new devices or technologies in the field (Davis, 1985, 1989).

Computer-Assisted Language Learning (CALL) is the search for and study of applications of computer technology in language learning and teaching (Levy, 1997).

Language Testing is a subfield within applied linguistics (Bachman, 2000).

Test is a method of measuring a person's ability, knowledge, or performance in a given domain (Hendriane & Suzanne, 2013).

Multiple-choice question (MCQ) consists of a stem followed by several alternatives, and test takers need to select the correct response among other alternatives known as distractors (Ganji & Esfandiari, 2020).

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

Recent advances in science and technology have had profound impacts on societies, with one major development being the successful integration of artificial intelligence into various industries. Over many years, the field of artificial intelligence has made tremendous progress, becoming increasingly significant and sparking a variety of interdisciplinary studies. Many have recognized the revolutionary function of artificial intelligence (AI) in various fields such as advertising (e.g., Kietzmann et al., 2018), agriculture (e.g., Sparrow et al., 2021), art (e.g., Chatterjee, 2022), automotive (e.g., Soegoto et al., 2019), business and industry (e.g., Dirican, 2015; Saxena et al., 2021), defense (e.g., Mori, 2018; Taylor, 2019), law (e.g., Doomen, 2023; Alarie et al., 2018), manufacturing (e.g., Zhou & Bacanin, 2024), media and entertainment (e.g., Lachman & Joffe, 2021), medicine and health (e.g., Hamet & Trambley, 2017; Mintz & Brodie, 2019; Singh, 2024), technology (e.g., Liu et al., 2020) and tourism (e.g., Nannelli et al., 2023). Thus, the field of education has also begun experiencing the ripple effects of AI, and its potential has received academic attention (Self, 2016).

Even though scholars have highlighted the necessity of incorporating AI in education (AIED) and even discussed the potential trends of AIED in the future, the perspectives of teachers have remained less explored (Zawacki-Richter et al., 2019). The exploration of language teachers' perspectives on the use of AI is a relatively emerging area of inquiry, with only a limited body of research available, despite the recent surge in interest and growing studies within the field (e.g., Almanea, 2024; Arslan, 2020; Dilzhan, 2024; Gücük, 2022; Kalra, 2024; Mohamed, 2023; Üretmen, 2024; Varsamidou, 2024).

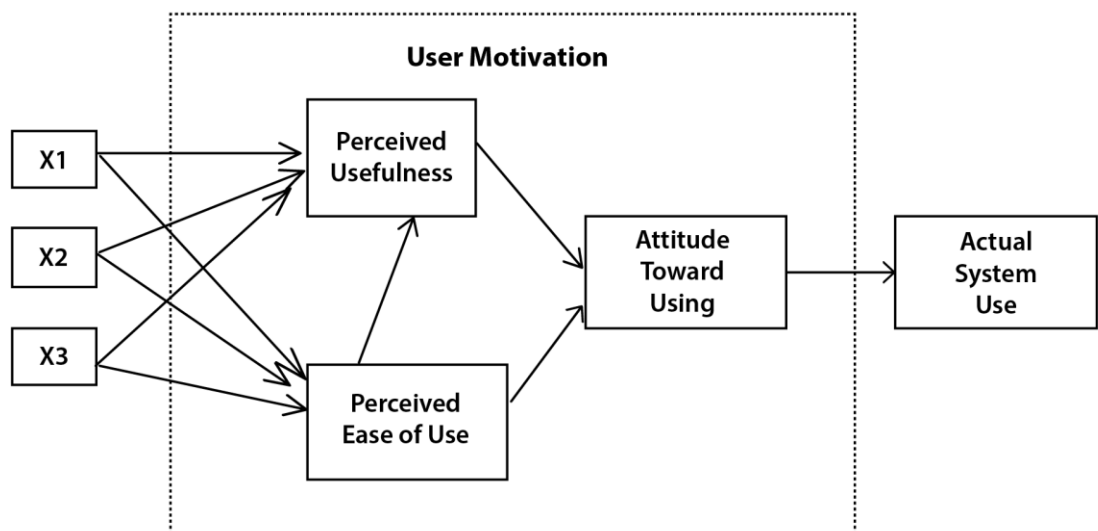
Although a few more studies have begun to examine EFL educators' application and perceptions of AI use in test development and assessment (e.g., Nguyen, 2023, 2024; Nguyen et al., 2024), the literature remains sparse, and none has yet directly addressed EFL instructors' first-hand experiences and perceptions of using ChatGPT in L2 reading test development. This is the main contribution of the current study to the literature.

2.1.1 Theoretical Framework

The theoretical framework of this study is the Technology Acceptance Model (TAM) by Davis (1985), a well-established and widely used model for understanding teachers' technology adoption and user acceptance. As Marangunić and Granić (2015) state, a substantial body of work concerning TAM, from its first appearance many years ago to their extensive review in 2015, demonstrated the model's widespread popularity in the field of technology acceptance.

Davis (1985) developed TAM by relying on the foundational studies by Fishbein and Ajzen (1975), who formulated the Theory of Reasoned Action (TRA), and by Ajzen (1985), who formulated the Theory of Planned Behavior (TPB), an extension of the TRA. The original TAM is illustrated in Figure 1.

Figure 1.
The Original TAM as Proposed by Davis (1985)



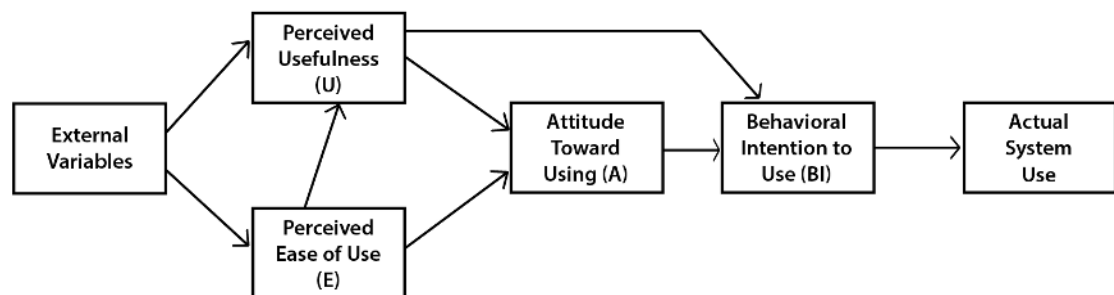
In this original version of TAM, the central constructs are Perceived Usefulness (PU) and Perceived Ease of Use (PEU), which shape the other two tenets of the model: Attitude Toward Using (ATU) and Actual System Use (ASU) (Davis, 1989). While PEU is defined as “the degree to which a person believes that using a particular information system would enhance his/her job performance, PU is defined as “the degree to which an individual believes

that using a particular system is free of effort” (Davis, 1989, p.320). TAM is based on the premise that an individual’s attitude towards technology is shaped by their views of its Perceived Usefulness and Perceived Ease of Use. Hence, these two central constructs of TAM, which are also believed to be directly affected by the system design characteristics represented X1, X2, and X3 in Figure 1, have been extensively researched to determine their influence on users’ ATU and ASU towards new technological adoption (e.g., Claudio et al., 2015; Eze et al., 2021; Habes et al., 2022; Hendrickson & Latta, 1996; Tubaishat, 2017; Vigoroso et al., 2024).

Davis, Bagozzi, and Warshaw (1989) further developed the original TAM by introducing Behavioral Intention (BI) as a mediating factor between attitude and actual system use, which can provide a more structured approach to examining the technology adoption decisions through future tendencies (see Figure 2).

Figure 2.

The Extended Version of TAM as Proposed by Davis, Bagozzi, and Warshaw (1989)



TAM has been employed by researchers to examine the acceptance of many technologies in several different fields such as e-commerce (e.g., McCloskey, 2016; Pavlou, 2003), online shopping (e.g., Ofori & Appiah-Nimo, 2019; Savitskie et al., 2007), e-service (e.g., Taherdoost, 2018), e-comments (e.g., Memarzadeh et al, 2016), cryptocurrency (e.g., Folkinshteyn & Lennon, 2016; Kocabas et al., 2024), healthcare technologies (e.g., Argan et al., 2024; Fleming et al., 2014; Hsiao & Tang, 2015), information technologies (e.g., Niehaves & Plattfaut 2014), navigation systems (e.g., Park et al., 2014), metaverse (e.g., Wu & Yu, 2023), virtual reality (e.g., Fussell & Truong, 2021; Iftikhar et al., 2022; Sagnier et al., 2020; Vallade et al., 2020), and online education (e.g., Esteban-Millat et al., 2018; Gibson et al., 2008; Habes et al., 2022; Kamble et al., 2022; Khasawneh, 2022; Naidoo, 2022; Rienties et al., 2014). Building on its broad application across various fields and technological domains, TAM

has also garnered increasing interest in the field of artificial intelligence, where researchers have begun utilizing the model to investigate user acceptance and adoption of AI-driven technologies (e.g., Belda-Medina & Kokošková, 2024; Esiyok et al., 2024; Kopplin, 2022; Lin & Yu, 2024; Ma et al., 2025; Yang & Appleget, 2024). However, the integration of TAM in studies on AI-driven language testing remains limited, with no existing research particularly examining the use of ChatGPT in L2 reading test development.. In this study, the chosen model offers a thorough lens to find out about the perceptions and experiences of language instructors regarding the use of ChatGPT for reading test development purposes.

The original TAM, as proposed by Davis (1985), was chosen as the fundamental framework for this study since it includes all the essential elements to examine the perceptions and experiences of language instructors who develop L2 reading tasks using ChatGPT-4. The chosen theoretical framework ensures that the study not only explores teachers' perceptions of ChatGPT for testing purposes but also examines the practical experiences they have during its implementation, contributing practical, useful, and deeper insights for language test development with AI. This is especially crucial as, to date, no research has explicitly looked into how language instructors use ChatGPT for developing L2 reading tasks and how they perceive this experience.

2.1.2 TAM in CALL Studies

Computer-Assisted Language Learning (CALL), defined by Levy (1997) as the exploration and examination of computer applications in language teaching and learning, has been widely considered as the principal acronym to refer to the use of technology in second language learning. Levy and Hubbard (2005) also highlight that CALL is the preferred label as it is inclusive, not exclusive, and it is not helpful to create new labels every time technology takes a step forward. In general, CALL is facilitated by a reciprocal relationship among computers, learners, and teachers (Gruba, 2004).

Many researchers have explored various aspects of CALL research through reviews or meta-studies of collections of papers (e.g., Chapelle, 2001; Debski, 2003; Felix, 2005; Levy, 2002; Levy & Stockwell, 2013; Pérez-Paredes, 2022; Zhao, 2003). Hubbard (2003) also states that CALL research can involve almost every aspect of instructed second language learning

due to the computer's potential interaction with all the essential variables in language learning, such as teachers, learners, methods, resources, and contexts. CALL has pursued various paths, encompassing the interactions between learners and teachers with digital technologies, as well as analyzing the impact of these digital tools on the language, literacy, and culture both within and beyond classroom contexts. (e.g., Aljohani, 2024; Chen et al., 2024; Lo, 2023; Shen & Chen, 2025)

The Technology Acceptance Model (TAM), developed by Davis et al. (1989) to predict user adoption of technology, has been extensively used in CALL studies to investigate the adoption and integration of technology by language learners and teachers in learning and teaching practices. Since its introduction, TAM has offered teachers and researchers valuable insights into 'perceived usefulness', 'perceived ease of use', 'attitudes', 'intention to use', and 'actual behaviors of technology users' (Davis, 1989). In CALL studies, TAM has been utilized to examine various technologies such as mobile-assisted language learning (e.g., Barrett et al., 2022; Zhang & Pérez-Paredes, 2021), flipped learning (e.g., Khlaisang et al., 2021), learning management systems (e.g., Al-Mamary et al., 2024; Ashrafi et al., 2022; Saroia & Gao, 2019), intelligent tutoring systems (e.g., Han, 2024; Huang et al., 2022; Karaci et al., 2018), speech recognition (e.g., Hsu, 2016; Hsu, 2024; Johnson, 2021), automated feedback tools (e.g., Li et al., 2019; Zhai & Ma, 2022; Zou et al., 2023), classroom response systems (e.g., Luo & Cao, 2024), mobile dictionaries (e.g., Zhang, Hennessy, & Pérez-Paredes, 2023), social media and communication applications (e.g., Habes et al., 2022; Teo, 2016), and AI-driven chatbots (e.g., Abdi et al., 2025; Belda-Medina & Kokošková, 2024; Chocarro et al., 2023; Esiyok et al., 2024; Liu et al., 2024; Liu & Ma, 2024; Ma et al., 2025; Peng & Liang, 2025; Ziqi et al., 2024).

Among the studies published on AI-driven chatbots in language learning, Belda-Medina and Kokošková (2024) examined teacher candidates' critical skills and perceptions of ChatGPT integration into language learning using TAM. Their mixed-methods study, utilizing convenience sampling (n=58), employed ChatGPT-4 for producing the machine-generated samples. The findings indicated no correlation between language nativeness (L1, L2) and participants' critical skills in differentiating human-produced content from machine-written content. The results demonstrated positive perceptions regarding Perceived Usefulness (PU) and Behavioral Intention (BI), although concerns about accuracy and reliability were expressed. According to the authors, this research is original as it comprehensively examined EFL teacher candidates' critical skills and perceptions of ChatGPT, highlighting the implications and promise of AI technology in language learning and teacher education

programmes. Similarly, Chocarro, Cortiñas, and Marcos-Matás (2023) examined teachers' attitudes towards chatbots in education through a TAM approach. The analysis of responses from 225 primary and secondary education teachers revealed that Perceived Ease of Use and Perceived Usefulness considerably enhanced teachers' intentions to adopt chatbots in their teaching, and chatbots utilizing formal language were received more positively than those employing informal language. Additionally, individual differences, including age and digital proficiency, were analyzed, yielding insights into the influence of user characteristics on technology acceptance.

Some TAM-based CALL studies examined the factors affecting learners' and/or teachers' adoption of and preferences for various digital language learning technologies such as flipped smart applications, classroom response systems, online teaching, and automated writing evaluation (e.g., Chocarro et al., 2023; Khlaisang et al., 2021; Luo & Cao, 2024; Sun & Zou, 2022; Zhai & Ma, 2022). The common factors affecting learners' and/or teachers' acceptance have been found to be the participants' enjoyable experiences in using online technologies, social influences, technological pedagogical content knowledge, complexity, facilitating complexity, etc. Besides, many CALL studies employing TAM explored learners' attitudes and acceptance of technology in their learning process (e.g., Barrett et al., 2022; Liu et al., 2024; Creely et al., 2025; Saroia & Gao, 2019) as well as teachers' perceptions and acceptance of the technology use in their teaching practices with a focus on perceived usefulness, perceived ease of use, and their influence on intention to use and actual use (e.g., Chocarro et al., 2023; Liu et al., 2017; Sharma & Saini, 2022; Yunus, 2007). Findings from these studies show that while learners generally acknowledge the potential benefits of mobile technologies for language learning, system design limitations and usability issues may impede their effectiveness. Similarly, many educators recognize the promise of these technologies to enhance language instruction; however, they express their concerns related to the practical challenges, including insufficient training, infrastructure constraints, and pedagogical beliefs that might hinder full adoption and implementation (e.g., Ansari et al., 2024).

Methodologically, the majority of existing TAM-based CALL studies have employed quantitative approaches, relying on questionnaires and/or surveys (e.g., Cárdenas-Claros et al., 2024; Hsu, 2016; Hsu, 2024; Khlaisang et al., 2021; Li et al., 2019; Liu et al., 2017; Liu & Ma, 2024; Ma et al., 2025; Saroia & Gao, 2019). Many studies have also utilized mixed-methods approaches through questionnaires and/or surveys (pre-, post-, delayed) with semi-structured interviews, and/or focus group interviews (e.g., Barrett et al., 2022; Belda-Medina &

Kokošková, 2024; Creely et al., 2025; Sun & Zou, 2022; Rafiee & Abbasian-Naghneh, 2021; Zhang, Hennessy & Pérez-Paredes, 2023). Purely qualitative methodologies in TAM-based CALL research are still very uncommon, despite their capacity to provide deeper, richer, and more nuanced insights into the experiences of learners and teachers using CALL technologies. This methodological gap is especially apparent in research investigating technology-mediated test development, where comprehensive and in-depth qualitative analysis could yield significant insights into decision-making processes and pedagogical concerns. By utilizing a qualitative case study methodology, the present study aims to address this gap and offer a detailed examination of teachers' perceptions of and experiences with an AI-driven tool for L2 reading test development.

2.2 Artificial Intelligence (AI)

The concept of artificial intelligence (AI) is now familiar to many in various fields. However, over the course of its development, definitions of AI have been inconsistent because of its constantly evolving and complex interdisciplinary nature (Kaplan & Haenlein, 2019; Renz et al., 2020). In other words, the difficulty in defining AI lies in its extensive reach across multiple disciplines including anthropology, biology, computer science, education, finance, law, linguistics, medicine, neuroscience, philosophy, and psychology, where each group offers different terminology and perspective (Luckin et al., 2016; Popenici & Kerr, 2017; Russell & Norvig, 2010).

Although the birth of AI is usually placed at approximately 1956, the year of John McCarthy's Dartmouth Summer Research Project on Artificial Intelligence, the conference which gave the emerging field its name, Alan Turing was thinking about machine intelligence at least as early as the 1940s (Moor, 2003). Turing (1950) asked "Can machines think?" in what was arguably the most influential and widely read article in the philosophy of artificial intelligence, and suggested a definition for deciding whether software is intelligent or not. In his theory, the software's intelligent behavior can be measured like a human intellectual efficiency, and the software is intelligent when a human being does not know whether he/she is chatting with the software or with another human. That test was called the Turing Test. If the software proved able to pass the test, it would be called intelligent software, also called an intelligent agent, which was capable of using sensors to perceive its surroundings and acting

on them (Benko & Lányi, 2009). Turing's 1950 paper and his arguments can be regarded as inspirational to the inception and development of artificial intelligence. Another important test to discuss and determine the extent of machines' human-like acts is the Chinese Room thought experiment by Searle (1980). In this argument, a person was assigned a task while being locked in a room with no access to external information, and by following detailed instructions, produced sophisticated outputs despite not having any specific skills or knowledge.

The field of AI research officially started in 1956 at a workshop at Dartmouth College where McCarthy first coined the term “artificial intelligence” and declared that “the study of AI is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it” (Russell & Norvig, 2010, p. 17). The Dartmouth Summer Research Project defined AI as “making a machine behave in ways that would be called intelligent if a human were so behaving” (McCarthy et al., 1955). Likewise, Minsky (1968) described AI as the discipline focusing on making machines capable of performing tasks that would necessitate intelligence if carried out by humans. Genesereth and Nilsson (1987) also define the notion of artificial intelligence as a field that aims to artificially produce intelligent behaviors of entities found in nature. In recent studies, researchers Kaplan and Haenlein (2019) have come up with a redefinition of AI as “a system's ability to correctly interpret external data, to learn from such data, and to use those learnings to achieve specific goals and tasks through flexible adaptation” (p. 17). Whichever definition is taken or whatever perspective is accepted, AI has and will continue to have its various impacts on education, especially language teaching and learning, as in many other fields.

2.2.1 ChatGPT

Artificial intelligence-generated content (AIGC), which is one of the most fascinating frontiers of technology, refers to the fact that users can use AI to create content (e.g., images, text, and videos) automatically according to their personalized requirements (Wu et al., 2023). With the development of AI algorithms and network structures, noteworthy progress has been made in AIGC. Developed by OpenAI, the Generative Pre-trained Transformer (GPT) model has the ability to carry out a wide range of tasks, including language translation, question answering, and text completion.

One prominent innovation in the AI landscape is ChatGPT, a publicly accessible tool developed by OpenAI (<https://chat.openai.com>). ChatGPT is a chatbot powered by a large language model (LLM) and designed to engage in conversations with users and generate responses that can be considered human-like and meaningful. As a member of AIGC, ChatGPT has shown powerful functions on a variety of language understanding and generation tasks, such as multilingual machine translation, code debugging, story writing, admitting mistakes, and even rejecting inappropriate requests according to the official statement (Wu et al., 2023). Unlike previous chatting robots, ChatGPT is able to recall what the user has written earlier in the conversation, which aids continuous dialogue. After its release in November 2022, it has attracted widespread attention due to its capability to generate coherent output across a wide range of topics and fields. The texts generated by ChatGPT can be virtually indistinguishable from those written by people, as it interacts with users in an apparently natural and intuitive conversational manner (Rudolph et al., 2023).

ChatGPT, with wide-ranging capabilities, is an integration of diverse technologies including “deep learning, unsupervised learning, instruction fine-tuning, multi-task learning, in-context learning, and reinforcement learning” (Wu et al., 2023, p.1123). It has been developed upon the initial GPT model, which has been iteratively updated from GPT-1 to GPT-4o and then GPT-4o mini from 2018 to today. GPT-4, which aimed to enhance the capabilities of GPT-3, was released in March 2023, while GPT-4 was introduced in May 2024, and GPT-4 mini was introduced in June 2024, in addition to many updates and added features to these versions.

The development of ChatGPT can be regarded as a significant milestone in the advancement of AI-powered chatbots. The model is trained on extensive text datasets, allowing it to grasp human language patterns, nuances, and complexities (Roumeliotis & Tselikas, 2023). ChatGPT has many benefits, including but not limited to encouraging personalized and interactive learning and creating prompts for formative assessment activities that offer ongoing feedback to inform teaching and learning.

2.2.2. Artificial Intelligence in Education (AIEd)

New technologies powered by AI have been applied in different sectors and fields, and the educational context is no exception. Artificial Intelligence is dramatically impacting a growing number of fields, including education (Gabriel et al., 2022; Song & Wang, 2020) and the research on the integration of AI technology into the field of education has been the subject of interest (Zawacki-Richter et al., 2019). It is possible to argue that AI is one of the biggest contributions of digital transformation to education. Artificial intelligence, as it relates to education, can be defined as information processing systems that are capable of participating in processes carried out by humans such as learning, adaptation, synthesis, self-correction, and the use of data for complicated processing tasks (Popenici & Kerr, 2017; Tasci & Celebi, 2020). Artificial intelligence in education is extensively used by learners and educators today, and involves various tools and applications such as smart lesson systems, teaching robots and adaptive learning systems (Chen et al., 2020).

A growing body of recent research is emerging regarding AIEd, particularly focusing on the use of chatbots in education. Rudolph, Tan and Tan (2023) conducted a study exploring the influence of AI chatbots, including ChatGPT, on higher education, highlighting their potential advancements and implications. Similarly, Tlili et al. (2023) investigated the issues and concerns associated with the integration of ChatGPT into educational settings. Shoufan (2023) also carried out a study to understand how students perceive the use of ChatGPT and assess its educational potential and challenges. This increasing attention highlights the expanding role of artificial intelligence, specifically chatbots, as transformative tools in various educational contexts, warranting further exploration and examination.

2.2.3 Artificial Intelligence in English Language Teaching (ELT)

The integration of AI into English language education has been driven by the appeal to improve learning outcomes, offer personalized instruction, and benefit from technological developments. AI has become a groundbreaking development in the field of English language teaching, offering innovative tools and approaches to improve both teaching and learning experiences (Law, 2024; Tian, 2024). Jiang (2022) gave a profound overview of AI in the EFL

context by including dominant forms of AI application such as automatic evaluation systems, neural machine translation, intelligent tutoring systems, chatbots, intelligent virtual environments, and affective computing. In their bibliometric analysis study, Huang et al. (2023) found that the frequency of studies on AI-enhanced language education increased over the period, and some of the most popular topics were automated writing evaluation, intelligent tutoring systems for reading and writing, automated error detection, computer-mediated communication, and personalized systems for language learning. Chen et al. (2021) also highlighted the increasing attention artificial intelligence for personalized learning has attracted in many educational contexts, including language learning through intelligent tutoring systems, natural language processing and artificial neural networks. Huang et al. (2021) also conducted a study to discover the possible technological, pedagogical, and social affordances enabled by artificial intelligence, specifically chatbots, in language learning. There are other similar studies reviewing the use of chatbots in language learning by focusing on the different types (Fryer et al., 2020; Haristiani, 2019).

Ongoing studies and advancements in AI, along with pedagogical considerations, have begun and will continue to shape the future of AI integration in the field, opening new opportunities and possibilities for language learners and educators. AI technologies, which are developing and will probably continue to develop at an exponential speed, offer the potential of “learning that is more personalized, flexible, inclusive, and engaging” (Luckin et al., 2016, p. 11). Godwin-Jones et al. (2024) claim that, based on available research studies and reports on teaching practices involving AI tools, L2 instructors are encouraged to adopt a critical, balanced approach to AI integration in L2 instruction, utilizing the strengths of generative AI while taking into account its downsides. Hence, one way to comprehend the effects of AI on language education is to look at them from the perspectives of language instructors. By examining the perceptions of language instructors and test developers on the integration of AI in language test development, professionals can be informed about how practitioners, in this case, language instructors, actually use and feel about such technologies. This, in turn, can help the educational adaptation of artificial intelligence into teaching methods and curriculum, as well as language testing.

2.3 English Language Testing

Language testing is an essential field in the literature on second language education as it deals with fundamental issues concerning language evaluation and assessment, such as test design, test administration, and score interpretation (Bachman & Palmer, 1996; Brown, 2010; Hughes & Scott-Clayton, 2011; McNamara & Roever, 2006). The purpose of language testing is always to “render information to aid in making intelligent decisions about possible courses of action” (Carroll, 1961, p. 314). This could imply that it is unlikely to justify the extent to which L2 student learning outcomes are attained without any form of testing.

Language testing is a “notoriously difficult domain of knowledge for students in second language teacher education programs” because of the complex balance between its highly abstract theoretical notions and their actual applications (O’Loughlin, 2006, p.71). Still, one of the primary concerns of test development is to warrant test usefulness. For tests to be useful, Bachman and Palmer (1996) suggest that test developers should consider six test qualities to ensure test usefulness: reliability, construct validity, authenticity, interactiveness, impact, and practicality. Taking these qualities into consideration for the overall usefulness of a particular test, language test developers should be able to define the test purpose to align with the educational goals and learning outcomes. English language testing also requires test developers to be well-versed in the proficiency standards and frameworks such as the Common European Framework of Reference for Languages (CEFR).

2.3.1 Testing Reading

Reading is regarded as an essential skill for many English learners, so reading proficiency will continue to be tested as long as reading courses are offered. Measuring reading proficiency has been one of the main activities among language teachers and researchers, although it is more challenging to assess reading along with listening from the perspective of obtaining observable protocols than it is to assess speaking and writing. It is not easy to obtain and test samples of test takers’ reading ability because the process of reading cannot be observed (Wolf, 1993), whereas measuring speaking and writing is relatively easier as these types of assessment are subject to observable protocols. There is no doubt that reading must be

tested in a valid and reliable manner. For this aim, numerous reading assessment tools are used, such as multiple-choice, written recall, sentence completion, open-ended, true/false, matching, and fill-in-the-blanks (Bernhardt, 1991). Among these, in language testing, response types for receptive skills often take the form of multiple-choice or constrained response (Liu et al., 2005), whether the means of assessment is paper-based or on screen. It is not surprising that multiple-choice questions (MCQs) appear to be one of the most popular means of testing because of particular features that make them highly objective and easy instruments for measurement. Additionally, they are mostly used in many standardized English exams.

2.3.2 Multiple-Choice Test Items

Various approaches to assessing reading comprehension have been developed and used in education, such as cloze tasks, true/false tasks, comprehension questions, essay tasks, and some other online procedures, including eye-tracking and think-aloud protocols (Cain & Oakhill, 2006; Magliano et al., 2007). Comprehension questions, ranging in format from open-ended to multiple-choice, are probably the most widely used to assess reading (Ozuru et al., 2007). Multiple-choice tests are arguably the most common of all approaches to reading comprehension assessment (Magliano et al., 2007), in which readers need to choose which of three or more response options is the correct answer to the question. A multiple-choice question consists of a stem, options, and additional information in the stem or options, such as text, images, or audio (Shin et al., 2019). A stem is the context or the question being asked, whereas the options, which typically consist of two to five possible answers, include the correct answer amongst several plausible and partially correct but misleading options (Shin et al., 2019).

2.3.3 Artificial Intelligence in English Language Testing

Automated educational assessment is a growing area of research with the advent of the integration of artificial intelligence, as it has the potential to increase the efficiency and accuracy of grading student work. The use of AI is proving to be a game changer internationally in the field of language assessment, particularly through digitally administered tests, automated

scoring technology, and automated written evaluation (e.g., Saricaoglu & Bilki, 2021), but what about the language test development? AI technologies are being utilized to build specialized English language tests and to analyze their effectiveness in terms of validity and reliability of both test delivery and scoring (Richardson & Clesham, 2021). These technologies also have the potential to lessen the workload for the second language (L2) teachers and test developers (Yunjiu et al., 2022). In their study, Chomphooyod et al. (2023) introduced a multiple-choice question generation system that can effectively produce English grammar questions while ensuring the control over both the content and targeted grammar topics. In language testing, AI technologies have taken many forms and influenced all phases of the test development cycle, including task design, delivery, scoring, and reporting.

2.4 Teachers' Perceptions of Artificial Intelligence in English Language Testing

AI developers often fail to consider the expectations of AI end-users in education, that is, of teachers (Luckin & Cukurova, 2019), although they are considered among the most important stakeholders in AI-based teaching (Seufert et al., 2021). To ensure the successful implementation of AI into education, various stakeholders, particularly teachers, should participate in AI creation, development, and integration (Qin et al., 2020). Therefore, there is an increasing attention to the perceptions of teachers on the integration of artificial intelligence into their educational practices. One such study was conducted by Abdelatif and Siddiqui (2021) with EFL faculty members, emphasizing the need for universities to provide targeted support and enough resources to facilitate the integration of AI into educational practices. Belda-Medina and Calvo-Ferrer (2022) also examined the knowledge, level of satisfaction, and perceptions concerning the use of AI in language learning among future educators, highlighting the need for better training integrated into their curricula. Bekou et al. (2024) carried out a study to explore English language teachers' views on the use of ChatGPT in ELT, highlighting the potential benefits of ChatGPT in ELT such as individualized learning, prompt feedback, enhanced professional support for educators, and improved access to authentic resources, whereas pointing out issues regarding accuracy, cultural appropriateness, and overdependence on technology.

As evident from the studies mentioned above and the growing interest in this area, understanding language instructors' perceptions plays a crucial role in effectively integrating

AI tools like ChatGPT into language education. Likewise, taking on the role of test developers, language instructors' views, experiences, and expectations should be of great importance to the literature on language testing, as relevant to the scope of this current study. However, no prior work has explored the perceptions and experiences of language instructors on the integration of ChatGPT into L2 reading test development.

2.5 Conclusion

This literature review has provided a comprehensive overview of artificial intelligence, its use in education, including the field of ELT, specifically English language test development. The implementation of AI technologies into language testing has offered significant potential in improving test creation processes. The studies examined have highlighted the benefits of technology integration into language education. However, it is quite clear that there is a lack of research into the integration of artificial intelligence into English language testing. Therefore, this study is significant as it will be one of the first of its kind to be done on the use of an AI-based tool in the test development process in ELT, along with the teachers' perceptions.

CHAPTER 3: METHODOLOGY

3.1 Introduction

This qualitative case study seeks to explore English as a foreign language (EFL) instructors' perceptions and experiences of using ChatGPT-4 in L2 reading test development. To this end, the study seeks to answer the following research questions:

- 1) How do EFL instructors use ChatGPT-4 for developing L2 reading tasks?
 - 1a) If any, what are the differences in the EFL instructors' use of ChatGPT-4 when developing elementary-level and intermediate-level L2 reading tasks?
- 2) What are the EFL instructors' perceptions while they are using ChatGPT-4 for developing L2 reading tasks?
 - 2a) If any, what are the differences in the perceptions of the EFL instructors when developing elementary-level and intermediate-level reading tasks?

This chapter presents the research methodology employed in this study to answer the above research questions. It introduces the research design, setting, and participants, data collection, data analysis, and ethical considerations.

3.2 Research Design

The approach selected for this study is a qualitative case study approach, in which the researcher develops an in-depth analysis of a case, often a program, event, activity, process, or one or more individuals (Creswell & Creswell, 2017). In this study, "case" refers to four EFL instructors who work in the English preparatory school of a state university in Türkiye and also have secondary roles as test developers. Through this specific approach, the research particularly aims to delve into the insights of the participants related to the use of AI in test generation in the specific context of an English preparatory school. The use of a case study design contributes to uncovering rich, comprehensive, and context-specific data, allowing for

a detailed exploration of the complexities and dynamics regarding the integration of AI in language test development.

3.3 Participants

The participants of this study consist of two female and two male EFL instructors who work in the English preparatory school of a state university in Türkiye. The medium of instruction is English, and the school offers a year-round English program to prepare students for their undergraduate studies in their departments. All newly admitted students are required to first take the English Placement Test (EPT) and get 60 points out of 100 in order to sit the English Proficiency Exam (EPE). The students who pass the EPE by getting at least 60 points out of 100 are eligible to continue to their departmental studies, but the others, according to the results of these two language exams, are placed into one of four levels of language instruction: elementary, pre-intermediate, intermediate, and upper-intermediate in the preparatory school. Whereas each of the elementary, pre-intermediate and intermediate levels has two quizzes (Quiz 1 and Quiz 2) and one end-of-course test (ECT) during the 8-week period, the upper-intermediate level has only two quizzes, but not an ECT as students in this level are eligible to take the EPE at the end of the semester if they meet the criteria of getting 60 points in total out of 100 from the two quizzes and other components of continuous assessment such as assignments, tasks and instructor's evaluation.

Participants were selected through convenience sampling due to reasons such as easy accessibility, geographical proximity, availability at a set time, and the willingness of participants to participate in the study (Dörnyei, 2007). The participants were invited to the study as they worked in the testing office when this study was being conducted or had previously worked in the testing office and had experience in designing language exams (e.g., Quiz, ECT, EPT, and EPE) at different levels. Their professional experience ranged from 8 to 18 years in teaching and from 2 to 6 years in language testing. Table 1 provides the demographic information of the participants.

Table 1.
Demographic Information of the Participants

Participant	Gender	Teaching Experience	Testing Experience
Instructor 1	Male	8 years	2 years
Instructor 2	Male	14 years	3 years
Instructor 3	Female	17 years	5 years
Instructor 4	Female	18 years	6 years

In order to gain some information about the instructors' familiarity with AI tools and background in using technology in language testing, they were asked a few questions in a short, informal online meeting before the data collection started. Instructor 1 reported prior experience with several technological and AI-based tools for test generation such as 'Twee' and 'Claude'. Regarding ChatGPT, he noted using it for approximately two years, initially for entertainment or personal purposes, but later for professional use after joining the testing unit. His professional use focused on ChatGPT's ability to generate reading texts across different exam sections (reading, listening, vocabulary, grammar), which he described as a major advantage compared to his earlier practice of searching through websites and books to locate suitable texts. Instructor 2 reported extensive experience in using AI and digital tools for test development. He used 'Claude' for text generation and 'Fraise.it' to check vocabulary in context for item writing. In addition, he used 'Grammarly' for language refinement, and 'Voiser' to convert texts into listening materials, reflecting his broad adoption of AI in different test-related tasks. Regarding ChatGPT, he had experience with the free version, describing it as manageable, practical, and accurate in understanding prompts. Instructor 3 reported using a variety of AI and digital tools for different purposes in both academic and test-related tasks. These included 'Claude' for item development, 'Quillbot' for summarizing, paraphrasing, and grammar checking, 'Grammarly' for language accuracy, and 'Copilot' for uploading and analyzing documents or images. Regarding ChatGPT, she expressed that she occasionally used ChatGPT for daily purposes, describing it as a general assistant similar to Google, and she mainly relied on the free version, which she found limited for test development purposes. While she occasionally used ChatGPT to double-check reading texts for level appropriateness (e.g., CEFR alignment), she did not regularly use it to create full tests. Instructor 4 reported that she

regularly used AI tools, particularly for text generation, as she considered it a safer and more practical alternative to sourcing texts from books, websites, or magazines. She emphasized that AI-generated texts provided her with suitable material to adapt for tests. For other skills, she mentioned using a voice recording program to prepare listening test audios. Her experience with ChatGPT was limited to the free version, which she primarily used for text generation and occasionally for producing discussion topics for speaking activities, but not for item generation.

3.4 Data Collection

Three data collection tools were used for this case study: 1) screen recordings of the instructors' use of ChatGPT-4, 2) think-aloud protocols while instructors used ChatGPT-4 and 3) the ChatGPT-4 conversation transcripts. Data were collected online at the instructors' own convenient time and in their own environment at the beginning of the spring semester during the 3rd period of the 2024-2025 academic year. The participants were informed both in written form and orally before they actually used ChatGPT and developed L2 reading tasks. Prior to data collection, each participant was informed about the objectives of the study through a consent form (Appendix I) in a Google Doc that they needed to sign to indicate their willingness to participate in the study.

Along with the consent form, in the same Google Doc, the participants were also provided with a written general overview of ChatGPT (Appendix II) regardless of whether they were familiar with it or not; a general overview of screen recording extension (Appendix III) that they were expected to use to record their use of ChatGPT for L2 reading test development; and a general overview of think-aloud protocols (Appendix IV) in order to ensure the standards among the participants at the beginning of the study as much as possible. However, due to the exploratory nature of this study, there was no further training on how to use ChatGPT to create tests, as teachers were encouraged to discover the integration of ChatGPT into L2 reading test development by themselves. There was no further training on how to conduct think-aloud protocols, either.

These four language instructors were requested to use ChatGPT for creating reading tasks with multiple-choice items. That is, the participants were to create their reading tasks following the school's requirements, test specifications, language program curriculum, and the

relevant wordlists for each reading test of two different levels as they usually do in the testing unit. The participants were provided with all the necessary documents for test development, including test specifications, coursebook units, and relevant wordlists; however, these materials were excluded from the appendix due to institutional privacy restrictions. Each participant was asked to prepare each reading task online, individually at their convenience. There were no specific instructions on how to use ChatGPT or whether they should sign into their account or not. There was no time limit, either. Given the exploratory nature of the study, the instructors were encouraged to independently use ChatGPT in their own way for L2 reading task development. They were asked to use a screen recording extension, which allowed them to record both their screens and think-alouds at the same time during the test development process.

To answer research question 1, which focused on how EFL instructors engaged with ChatGPT-4 for L2 reading task development, screen recordings were collected to capture the instructors' real-time interaction with ChatGPT-4. There were eight videos in total: two videos from each instructor, with one of them capturing the elementary-level reading task development process and the other capturing the intermediate-level reading task development process. The instructors recorded their own screens using the screen recorder extension on Google Chrome (Screen Recorder) and shared their recordings with the researcher via a Drive link. Table 2 summarizes the screen recordings of instructors' test development process using ChatGPT.

Table 2.
Summary of the Video Recordings of Instructors' Use of ChatGPT

Participant	Duration (Elementary)	Duration (Intermediate)
Instructor 1	52 minutes	47 minutes
Instructor 2	20 minutes	25 minutes
Instructor 3	73 minutes	51 minutes
Instructor 4	46 minutes	36 minutes

To answer the research question 2, focusing on EFL instructors' perceptions of their experiences using ChatGPT-4 for L2 reading task development, the instructors were also requested to think aloud, verbalizing what they were doing and why during the task development process. The think-aloud protocol (TAP), from the principle of cognitive psychology, required the instructors to express their mental processes verbally while performing the task, inventing reading tests in the scope of this study (Ericsson & Simon, 1993). TAP, as an empirical data collection method, has been employed in numerous studies, in many disciplines, including education (Cumming et al., 1989). Think-alouds in this study were collected together with the screen recordings to capture the instructors' simultaneous verbalizations of their thoughts while engaging with ChatGPT-4 to develop L2 reading tasks.

The instructors were informed about the details of TAP in the general overview document (Appendix IV) (i.e., talking continuously throughout the entire task without stopping and talking naturally and freely). However, there was no further training on this. The instructors were asked to articulate their reasoning, reflections, and decision-making processes aloud as they were creating L2 reading tasks. In total, eight think-aloud protocols were collected, two from each instructor, one capturing the elementary-level reading task development process and perceptions, and the other capturing the intermediate-level reading task development process and perceptions.

In addition to screen recordings and think-aloud protocols, the ChatGPT-4 conversation transcripts of the instructors were also collected as supplementary data, offering a direct record of the instructors' interactions with ChatGPT. These chats included the precise prompts the instructors wrote, the revisions they requested, the editions they made, the strategies they used when generating L2 reading texts and tasks, etc. The transcripts were not analyzed separately as a data source, but were used to provide illustrative examples of the instructors' ChatGPT-4 use for reading task creation. However, only two instructors' conversations with ChatGPT were obtained since one instructor used ChatGPT in temporary mode, and another instructor worked on ChatGPT without signing into his/her account, which made it impossible to save the chats. The two instructors whose chats were available and accessible shared the transcripts with the researcher via the 'share' feature embedded in ChatGPT, which generated a secure link to the available conversation. For the remaining instructors, large portions of their conversations were extracted from the video recordings and were incorporated into the findings section to demonstrate their interaction patterns and strategies.

3.5 Data Analysis

The data analysis was conducted through a content analysis approach, which aided in the development and interpretation of codes that emerged from both the screen recordings and think-aloud protocols. Content analysis is described as a research technique for drawing replicable and valid conclusions from texts or other meaningful means by analyzing them within the contexts of their use (Krippendorff, 2004), and also a research method for the subjective interpretation of the content of text data through the systematic classification process of coding (Hsieh & Shannon, 2005). Similarly, Elo et al. (2014) described it as a flexible method for making valid inferences from data in order to provide new insight, describe a phenomenon through concepts or categories, and develop a comprehension of the meaning of communications with a concern for intentions, outcomes, and context. Qualitative content analysis is one of the several qualitative methods available for analyzing data and interpreting its meaning. As a research method, it serves as a systematic and objective means of describing and quantifying phenomena (Schreier, 2012). For content analysis to be successful, it is essential that data can be reduced to concepts that define the research phenomenon (Cavanagh, 1997; Elo & Kyngäs, 2008; Hsieh & Shannon, 2005) by generating categories, a model, conceptual system, or conceptual map (Elo & Kyngäs, 2008; Morgan, 1993; Weber, 1990).

The data analysis started with the video content of the screen recordings to come up with the common codes, subcategories, and categories. The researcher carefully watched the videos multiple times to write down each and every step the teachers took during the test development process as well as the comments they made as they conducted think-aloud protocols. Then, the researcher analyzed the data numerous times to acquire an in-depth understanding of the content. Hence, she could generate the initial codes by recognizing significant concepts and comments. Then, such initial codes were turned into broader subcategories and categories that captured the central patterns and ideas available in the data. Next, the researcher reviewed the codes along with the subcategories and categories to ensure coherence and relevance to the research questions and/or objectives. Lastly, the refined codes were interpreted and supported with relevant quotes of the participants during their think-aloud protocols to provide broader and richer understanding of their perceptions and practices considering the use of AI in English language test development.

For the analysis of the data from the screen recordings, three coding categories (i.e., Pre-ChatGPT use process, ChatGPT use process, and Post-ChatGPT use process), seven subcategories (i.e., External resource use, Prompt design strategies, Quality check, Editing ChatGPT-generated output, Providing context, Opening chatgpt.com, and Rationalizing - some recurring across different categories), and 69 codes emerged revealing how the instructors engaged with ChatGPT-4 when developing L2 reading tasks. Table 3 lists the specific codes used for these categories and sub-categories:

Table 3.
The Coding Categories, Subcategories, and Specific Codes Used for the Screen Recording Data

Coding Categories	Subcategories	Codes
Pre-ChatGPT use process	External resource use	Test specifications
		Coursebook: theme selection
		Coursebook: text type selection
		Coursebook: word selection
		Wordlist: word selection
ChatGPT use process	External resource use	Word document: word count check
		Word document: editing ChatGPT-generated output
		Word document: adjusting format and layout
		Coursebook: word selection
		Wordlist: word selection
		Coursebook: text type selection
		Coursebook: theme selection
		Readability website: level check
		Coursebook: structure selection
		Test specifications

Prompt design strategies	Providing task details
	Adjusting word count
	Adjusting the level
	Asking for level check
	Rationalizing
	Asking for an answer key
	Asking for revision: items
	Asking for revision: options
	Asking for revision: text
	Asking for revision: title
	Repeating
	Giving positive feedback
	Giving negative feedback
	Asking for self-evaluation
	Asking for alternatives
	Correcting previous prompts
	Asking for a new text
	Uploading files
Quality check	Text
	Word count
	Items
	Options
	Answer key
	Self-evaluation

Editing ChatGPT-generated output	Reorganizing
	Paraphrasing
	Deleting
	Adding
	Rewriting items
	Rewriting the title
	Rewriting options
	Reordering options
Providing context	Giving ChatGPT no role
	Explaining self-role
	Giving ChatGPT a role
Opening chatgpt.com	Signing in to ChatGPT account
	Signing in to a different ChatGPT account
	Using ChatGPT without an account
Rationalizing	Checking the text
	Adjusting the level
	Providing task details
	Checking self-evaluation
	Choosing familiar text/task
	Writing items
	Using word document
	Giving positive feedback
	Word document: adjusting format and layout

Post-ChatGPT use process	External resource use	Word document: editing ChatGPT-generated output
		Rewriting items
	Editing ChatGPT-generated output	Rewriting options
		Reordering options
		Writing items
		Deleting
		Adding

For the analysis of the think-aloud data, one coding category (i.e., Perceptions of ChatGPT use), one subcategory (i.e., Negative/Positive), and 16 codes emerged revealing the instructors' perceptions of their experiences of creating L2 reading tasks with ChatGPT-4. Table 4 lists the specific codes used for these categories and sub-categories:

Table 4.
The Coding Categories, Subcategories, and Specific Codes Used for the Think-aloud Data

Coding Categories	Subcategories	Codes
Perceptions of ChatGPT use	Negative/Positive	Word count
		Grammatical accuracy
		Level appropriateness: items
		Level appropriateness: text
		Level appropriateness: vocabulary
		Text organization
		Text generation
		Item generation
		Response speed

Title generation
Option generation
Following instructions
Theme selection
Self-evaluation
File upload
Usage limitation

Figure 3 below demonstrates how the screen recording data and the think-aloud data were transcribed and coded simultaneously in Excel:

Figure 3.
The Coding of the Screen Recording Data and the Think-aloud Data in Excel

Description of ChatGPT Use	Instructor's Statements	Codes	Sub-categories	Categories
He gives another prompt: 'I want you to add the details about his parents' opinions/ feelings about their jobs in the first paragraph just after giving information about their jobs.'		Asking for revision: text	Prompt design strategies	ChatGPT use process
	"It placed the details at the end of the text. I want to take them into the middle of the text, for example. Let's tell this to ChatGPT."	Text organization	Negative	Perception of chatgpt
ChatGPT generates the new version.				
He goes over the new version.		Text	Quality check	ChatGPT use process
	"In general' the text seems OK."	Text generation	Positive	Perception of ChatGPT use
	"There are some high level words such as 'rewarding' that I can change manually."	Level appropriateness: vocabulary	Negative	Perception of ChatGPT use
He copies the new version onto the Word and starts making some changes manually. He changes the place of a paragraph.		Reorganizing	Editing ChatGPT-generated output	ChatGPT use process
He paraphrases some sentences so that they look different from the ones in the options.		Paraphrasing	Editing ChatGPT-generated output	ChatGPT use process
	"When preparing the questions, I make sure that they are not the same verb in the text, or not in the same sentence structure, so that it is not easy."	Rationalizing	Editing ChatGPT-generated output	ChatGPT use process

To ensure the intercoder reliability for coding, a colleague of the researcher, who had been working as an English language teacher in the Ministry of National Education for 15 years and held an MA in English Language Education, served as an independent coder. After the

second coder was trained based on the coding scheme and the coded sample of one instructor's elementary-level task development data with 48 codes (which was excluded from reliability calculations), she independently coded two screen recordings: Instructor 1's intermediate-level reading task development video (with 54 codes) and Instructor 4's elementary-level reading task development video (with 60 codes). The second coder made a total of 342 coding decisions ($54 \times 3 + 60 \times 3$ for code, subcategory, and category).

The coding of the researcher was compared with that of the second researcher. For each code, a score of 1 was assigned for agreement and 0 for disagreement. The level of reliability was then calculated using percentage agreement, which is a straightforward and commonly used measure in studies utilizing qualitative content analysis (Lombard et al., 2002; Miles & Huberman, 1994). There were 31 mismatches (Intermediate: 8 codes, 5 subcategories, 2 categories; Elementary: 8 codes, 5 subcategories, 3 categories), yielding 311 agreements in total. The overall percent agreement was 90.9%. (90.7% for the intermediate level task development codes and 91.1% for the elementary level). Agreement between the two coders was 86.0% for codes, 91.2% for subcategories, and 95.6% for categories.

3.6 Ethical Considerations

In order to ensure the study is conducted ethically, first of all, permission to conduct research was sought from the university. Then, written informed consent forms (Appendix I) were obtained from all the participants – language instructors. Participants were provided with detailed information about the nature and purpose of the study as well as about their rights to refuse to participate or to withdraw from participating at any time. They were also informed that confidentiality and anonymity were assured.

Reliability is an important issue in content analysis as substantial interpretation goes into defining the data items (i.e., codes) as well as applying the codes to chunks of text. Thus, the researcher is responsible for the coding and the data analysis and another colleague carefully followed up on the analysis process and categorization to subsequently get together and discuss any divergent opinions concerning the categorization (Elo et al., 2014). Intercoder reliability was recognized in this study by having two coders code the data separately, and then match the codes through deep reflexivity and further discussion. That is because deep

reflection, discussion of biases, and a meticulous coding process for interview transcripts are among the steps taken to ensure that the findings are accurate from the standpoints of both the researchers and the participants (Creswell & Creswell, 2017).

3.7 Conclusion

Overall, this chosen methodology, which aligns with the research questions and objectives, contributed to gaining a holistic comprehension of the research topic, which will also shed light on the potential implications of AI in language testing and offer useful insights for test developers, educators, and policymakers.

CHAPTER 4: FINDINGS

4.1 Introduction

This chapter presents the findings of the study, which aimed to investigate the perceptions and practices of English language instructors in the use of ChatGPT for L2 reading test development in an English preparatory school context. The results are derived from data collected through screen-recorded video observations of the participants' test development processes along with their think-aloud protocols.

4.2 EFL Instructors' ChatGPT Use for L2 Reading Test Development

In total, the four instructors spent 5 hours and 50 minutes on reading test development in this study. Instructor 1's reading test development process lasted 1 hour and 39 minutes. He spent 52 minutes creating the elementary level reading task using ChatGPT and 47 minutes for the intermediate level reading task. Instructor 2 spent a total of 45 minutes: 20 minutes for the elementary level reading task and 25 minutes for the intermediate level reading task. Instructor 3's reading test development process lasted 2 hours and 4 minutes: 73 minutes for the elementary level reading task and 51 minutes for the intermediate level reading task. Instructor 4 spent a total of 1 hour 22 minutes: 46 minutes for the elementary level reading task and 36 minutes for the intermediate level reading task.

Regarding the number of prompts that instructors entered, the four instructors wrote 126 prompts in total for reading test development in this study. Instructor 1 wrote 3 out of 11 prompts while creating the elementary level reading task and 8 prompts for the intermediate level reading task. Instructor 2 wrote 4 prompts for elementary and 6 prompts for intermediate. Instructor 3 wrote 86 prompts in total: 37 for the elementary level reading task and 49 for the intermediate level reading task. Instructor 4 wrote 9 prompts for elementary and 10 prompts for intermediate.

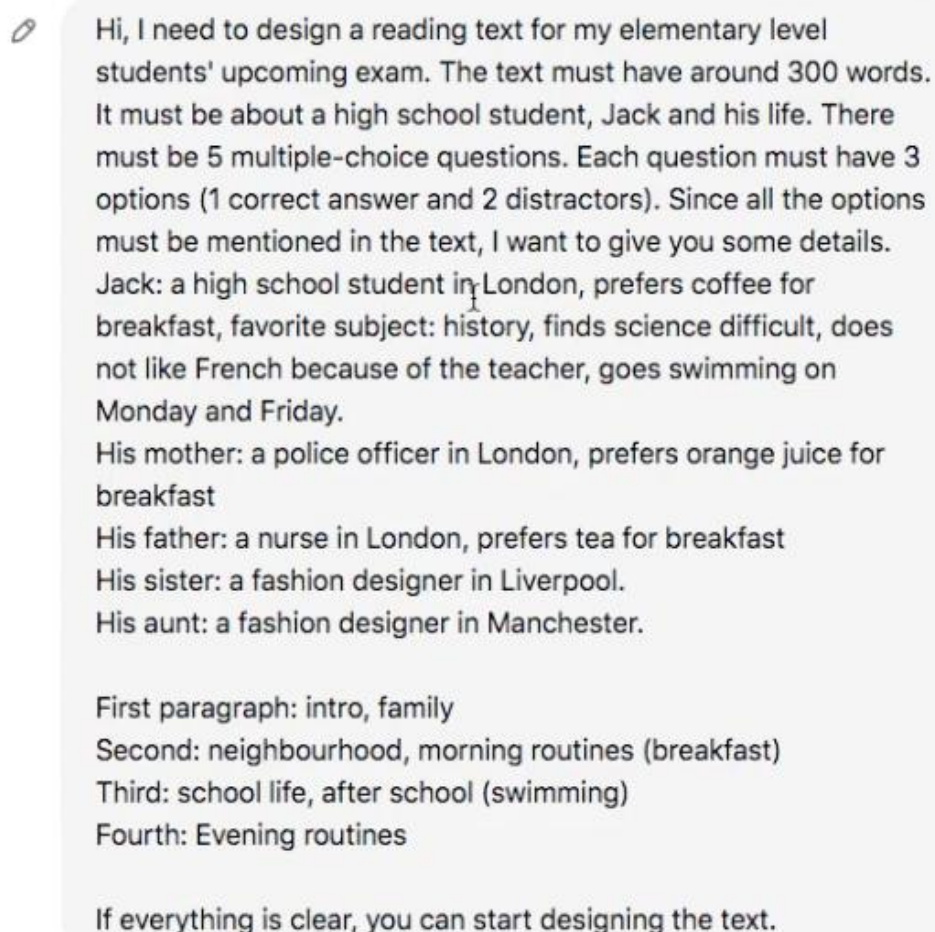
The following sub-sections provide detailed accounts of each instructor's ChatGPT use for each reading task.

Case 1: ChatGPT Use for Elementary-level L2 Reading Task Development

During the pre-ChatGPT use process, Instructor 1 consulted external resources for a total of five times. He consulted the test specifications once, the coursebook three times (twice to select the theme and once to select the text type), and the wordlist once to select vocabulary.

During the ChatGPT use process, Instructor 1 signed into his personal ChatGPT account. He wrote a long, detailed prompt, explaining his need for the text and specifying the word count, topic, characters, item type, text organization, etc. (see Figure 4), stating that he tries to “guide ChatGPT as much as possible with many details in the prompts and giving the criteria at the beginning.” He spent almost half of the time (26 minutes) in the whole video (52 minutes) writing his first prompt.

Figure 4.
Instructor 1's 1st Prompt



Hi, I need to design a reading text for my elementary level students' upcoming exam. The text must have around 300 words. It must be about a high school student, Jack and his life. There must be 5 multiple-choice questions. Each question must have 3 options (1 correct answer and 2 distractors). Since all the options must be mentioned in the text, I want to give you some details.

Jack: a high school student in London, prefers coffee for breakfast, favorite subject: history, finds science difficult, does not like French because of the teacher, goes swimming on Monday and Friday.

His mother: a police officer in London, prefers orange juice for breakfast

His father: a nurse in London, prefers tea for breakfast

His sister: a fashion designer in Liverpool.

His aunt: a fashion designer in Manchester.

First paragraph: intro, family
Second: neighbourhood, morning routines (breakfast)
Third: school life, after school (swimming)
Fourth: Evening routines

If everything is clear, you can start designing the text.

He then engaged in prompt design strategies on seven more occasions, which included providing task details (four times), adjusting word count (once), giving positive feedback to ChatGPT by writing “great” (once), and asking for text revision (once).

He also provided context for the task once by explaining his self-role by writing in his first prompt, “Hi, I need to design a reading text for my elementary level students’ upcoming exam.” External resources were consulted many times, both during and after ChatGPT use — a coursebook and a wordlist for word selection, and a Word document for word count check and editing ChatGPT-generated output. He rationalizes using Word document for word count check by stating, “Why do I use the Word document from the beginning? I don’t trust the word count by ChatGPT; it may not be correct.”

After ChatGPT gave the text (see Figure 5) along with the questions (see Figure 6), instructor 1 performed quality checks three times, which included checking the text twice and checking the word count once. He rationalized checking the text by stating, “I first take a look at the text because sometimes ChatGPT can make mistakes in sentences, even if it is rare.”

Figure 5.
The Initial Text ChatGPT Created

Jack Life Reading



Jack is a high school student in London. He lives with his parents and enjoys spending time with his family. His mother is a police officer, and his father is a nurse. Jack also has a sister, but she lives in Liverpool because she is a fashion designer. His aunt is a fashion designer too, but she works in Manchester.

Jack lives in a busy neighbourhood in London. In the mornings, he wakes up early and has breakfast with his parents. Jack always drinks coffee in the morning, while his mother prefers orange juice and his father likes tea. After breakfast, Jack takes the bus to school.

Jack’s favorite subject at school is history because he enjoys learning about the past. However, he finds science very difficult. He also studies French, but he doesn’t like it because of the teacher. After school, Jack goes swimming twice a week, on Mondays and Fridays. On other days, he goes home and watches TV or listens to music.

In the evenings, Jack helps his parents with dinner. After that, he does his homework and sometimes reads a book. He usually goes to bed around 10 p.m. because he has to wake up early for school the next day.

Figure 6.
The Initial Questions ChatGPT Created

Jack Life Reading ▾

1. Where does Jack live? a) In Liverpool
b) In London
c) In Manchester
(Correct answer: b)
2. What does Jack drink for breakfast? a) Coffee
b) Tea
c) Orange juice
(Correct answer: a)
3. Why doesn't Jack like French? a) He finds it difficult
b) He doesn't like the teacher
c) He prefers history
(Correct answer: b)
4. When does Jack go swimming? a) On Mondays and Fridays
b) On Wednesdays and Sundays
c) On Tuesdays and Thursdays
(Correct answer: a)
5. What does Jack do in the evening? a) Goes to bed early
b) Plays video games all night
c) Helps with dinner and does homework
(Correct answer: c)

A total of eight prompt design strategies he employed included providing task details four times. After ChatGPT created the initial text, instructor 1 wrote in his next prompt "I want you to add more details: his best friend at school, how happy his parents are with their jobs, what does he do on weekends?" and justified providing details by stating "By adding details, we can make it a bit more difficult." Then, ChatGPT created the text (see Figure 7).

Figure 7.
The Revised Text ChatGPT Created

Jack is a high school student in London. He lives with his parents and enjoys spending time with his family. His mother is a police officer, and his father is a nurse. Jack also has a sister, but she lives in Liverpool because she is a fashion designer. His aunt is a fashion designer too, but she works in Manchester.

Jack lives in a busy neighborhood in London. In the mornings, he wakes up early and has breakfast with his parents. Jack always drinks coffee in the morning, while his mother prefers orange juice, and his father likes tea. After breakfast, Jack takes the bus to school.

Jack's best friend at school is Tom. They have many classes together and often sit next to each other. Jack's favorite subject at school is history because he enjoys learning about the past. However, he finds science very difficult. He also studies French, but he doesn't like it because of the teacher. After school, Jack goes swimming twice a week, on Mondays and Fridays. On other days, he goes home and watches TV or listens to music.

Jack's parents are happy with their jobs. His mother enjoys working as a police officer because she likes helping people, and his father loves being a nurse because he finds it rewarding to take care of patients.

On weekends, Jack likes to relax. He often meets his friends at the park or plays football. Sometimes, he goes to the cinema with his family. On Sundays, he helps his parents with housework and studies for his exams.

In the evenings, Jack helps his parents with dinner. After that, he does his homework and sometimes reads a book. He usually goes to bed around 10 p.m. because he has to wake up early for school the next day.

He engaged in editing ChatGPT-generated output during ChatGPT use six times: reorganizing the text (twice), deleting content (once), adding content (once), and paraphrasing (once) to finalize the text (see Figure 8).

Figure 8.
The Text the Instructor Edited

Jack is a high school student in London. He lives with his parents and enjoys spending time with his family. His mother is a police officer, and she is really happy with her job. She likes helping people. Jack's father works as a nurse at a local hospital. He thinks his job is difficult because he has a lot of patients. Jack also has a sister, but she works in Liverpool as a fashion designer. His aunt is a fashion designer too, but she lives in Manchester with her family.

In the mornings, he wakes up early and has breakfast with his parents. Jack always drinks coffee in the morning, while his mother prefers orange juice, and his father likes tea. After breakfast, Jack takes the bus to school.

Jack's best friend at school is Tom. They have many classes together and often sit next to each other. Jack's favorite subject at school is history because he enjoys learning about the past. However, he finds science very difficult. He also studies French, but he doesn't like it because of the teacher. After school, Jack goes swimming on Mondays and Fridays. He also plays football with his friends on Wednesday. On other days, he goes home and watches TV or listens to music.

In the evenings, Jack helps his parents with dinner. After that, he does his homework and sometimes reads a book. He usually goes to bed around 10 p.m. because he has to wake up early for school the next day.

On weekends, Jack likes to relax. He often meets his friends at the park. Sometimes, he goes to the cinema with his family. On Sundays, he helps his parents with housework and studies for his exams.



He also engaged in editing ChatGPT-generated output after ChatGPT use five times by writing items by himself and twice more by deleting two words and adding two words instead (see Figure 9).

Figure 9.

The Final Draft of the Reading Task Instructor 1 Prepared

Jack is a high school student in London. He lives with his parents and enjoys spending time with his family. His mother is a police officer, and she is really happy with her job. She likes helping people. Jack's father works as a nurse at a local hospital. He thinks his job is difficult because he has a lot of patients. Jack also has a sister, but she works in Liverpool as a dentist. His aunt is a dentist too, but she lives in Manchester with her family.

In the mornings, he wakes up early and has breakfast with his parents. Jack always drinks coffee in the morning, while his mother prefers orange juice, and his father likes tea. After breakfast, Jack takes the bus to school.

Jack's best friend at school is Tom. They have many classes together and often sit next to each other. Jack's favorite subject at school is history because he enjoys learning about the past. However, he finds science very difficult. He also studies French, but he doesn't like it because of the teacher. After school, Jack goes swimming on Mondays and Fridays. He also plays football with his friends on Wednesday. On other days, he goes home and watches TV or listens to music.

In the evenings, Jack helps his parents with dinner. After that, he does his homework and sometimes reads a book. He usually goes to bed around 10 p.m. because he has to wake up early for school the next day.

On weekends, Jack likes to relax. He often meets his friends at the park. Sometimes, he goes to the cinema with his family. On Sundays, he helps his parents with housework and studies for his exams.

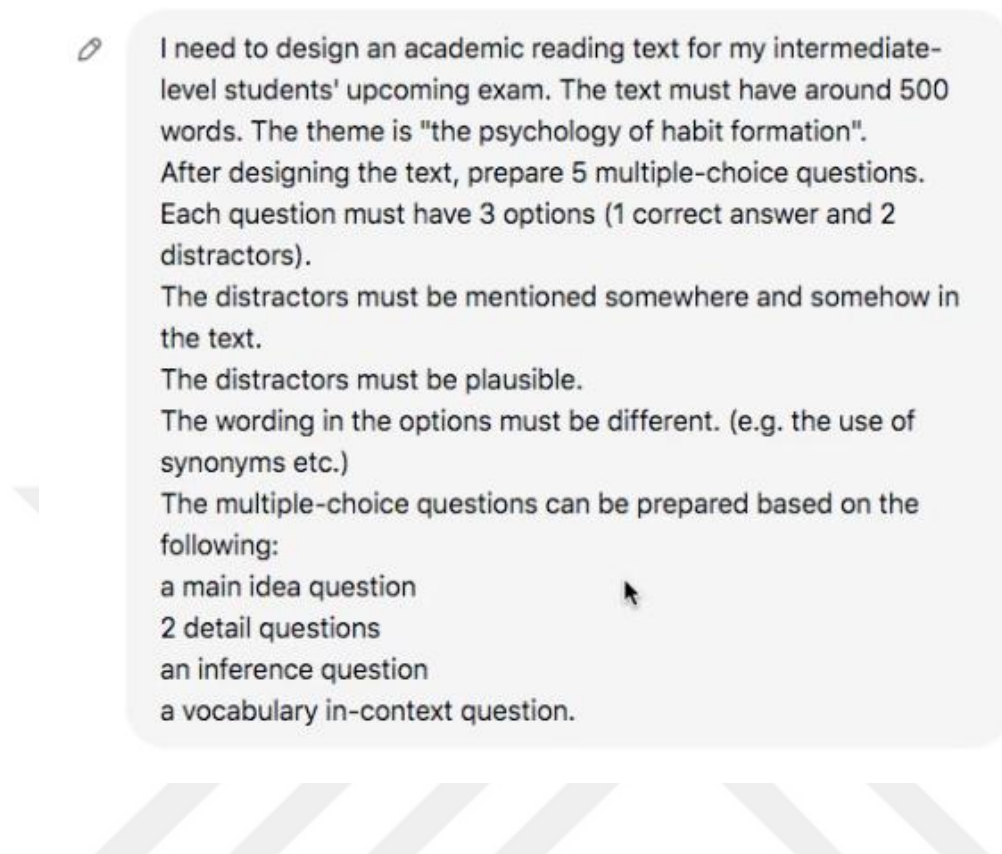
What is Jack's father's job?	A) dentist	B) nurse	C) police office
Where does Jack's sister live?	A) Manchester	B) London	C) Liverpool
What does Jack prefer for breakfast?	A) tea	B) orange juice	C) coffee
Which subject does Jack like the most?	A) history	B) French	C) science
How many times does Jack swim a week?	A) one	B) two	C) three

Case 1: ChatGPT Use for Intermediate-level L2 Reading Task Development

Before using ChatGPT, Instructor 1 consulted external resources twice. This included referring to the test specifications once and consulting the coursebook once for theme selection.

During the ChatGPT use process, Instructor 1 signed into ChatGPT with a different account and stated, "If I need to prepare multiple texts on ChatGPT on the same day, I generally log into ChatGPT with a different account." He started his first prompt by providing the context through which he explained his self-role by writing, "I need to design an academic reading text for my intermediate level students' upcoming exam." He continues providing task details in his prompt (see Figure 10).

Figure 10.
Instructor 1's 1st Prompt



ChatGPT responded by creating the initial text and questions (see Figure 11 and 12), and right above its response, there was a statement “reasoned about academic text and question for 37 seconds.” He copied and pasted the text onto a Word document for both word count check and format and layout adjustment. Another external resource was also consulted three times during the ChatGPT use process, which was the coursebook for text type selection.

Figure 11.
The Initial Text ChatGPT Created

Understanding the Psychology of Habit Formation

Habits are routines or behaviors that become nearly automatic through repetition, allowing individuals to perform them without conscious thought. The psychology of habit formation examines how repeated actions eventually turn into ingrained patterns, often making daily activities more efficient. Researchers in this field investigate not only the benefits of habits but also the challenges they present when one tries to modify or break them.

One widely accepted explanation for habit formation comes from the study of behavioral conditioning. Operant conditioning-where behaviors are shaped by rewards and punishments- plays a central role in how habits are reinforced. Although classical conditioning (the process of associating an environmental stimulus with a response) and observational learning (acquiring behaviors by watching others) are also recognized in behavioral studies, operant conditioning remains the primary framework for understanding why habits persist over time.

A key element in forming a habit is the cycle of cue, routine, and reward. A cue triggers the behavior, the behavior itself is the routine, and the reward reinforces the habit, making it more likely to occur again. This cycle is strengthened with repetition. While some factors like stress and cognitive dissonance- the mental discomfort experienced when holding conflicting beliefs- can influence behavior, they are not the essential drivers of habit formation. Instead, the regular encounter of cues and the immediate gratification of rewards create a stable and predictable pattern.

From a neurological perspective, habit formation is associated with changes in the brain that lead to automaticity. Automaticity is the process by which tasks become almost effortless and are executed with minimal conscious intervention. This transformation occurs as neural pathways are reinforced with every repetition of the habit. Although decision-making processes and emotional responses, such as those provoked by stress, also occur in the brain, they do not directly contribute to the automatic execution of habitual behavior.

Understanding these processes is crucial for those attempting to change unwanted habits. By identifying and modifying the cue-routine-reward cycle, individuals can disrupt old patterns and replace them with healthier behaviors. For instance, rather than relying on sheer willpower, which may be compromised by daily stress or suboptimal decision-making, behavioral interventions often focus on altering environmental cues or finding alternative rewards. This approach has been shown to be more effective than expecting individuals to overcome habits solely through conscious effort.

Moreover, researchers have discovered that even though factors like stress and decision-making play roles in our overall behavior, the deeply embedded nature of habits means that they persist regardless of our fluctuating mental states. In other words, even when we are not under stress or when our decision-making is impaired, the strength of a habit can remain largely unchanged due to its neurological roots. The study of automaticity highlights how repetitive behavior eventually becomes independent of conscious thought, which can be both an asset for efficiency and a hurdle when change is desired.

In summary, the psychology of habit formation provides a detailed look at how our daily routines become automated through repetition and reinforcement. It emphasizes the importance of the cue-routine-reward cycle and the neurological shift towards automaticity while acknowledging that factors like stress, cognitive dissonance, classical conditioning, observational learning, and decision-making processes also influence our behaviors-even if they do not form the core of habit development.

Figure 12.
The Questions of the Initial Text ChatGPT Created

- Multiple-Choice Questions
1. Main Idea Question:
What is the primary focus of the text?
A. The psychology of habit formation
B. The effects of stress on daily routines
C. The influence of decision-making on behavior
 2. Detail Question 1:
According to the text, which element is essential for forming a habit?
A. A cycle of cue, routine, and reward
B. Repetitive cognitive dissonance
C. Constant stress
 3. Detail Question 2:
Which type of conditioning is described as the primary framework for understanding habit reinforcement in the text?
A. Operant conditioning
B. Classical conditioning
C. Observational learning
 4. Inference Question:
Based on the text, why might breaking a habit be particularly challenging?
A. Because repeated behavior strengthens neural pathways, making habits automatic
B. Because individuals do not experience stress
C. Because people lack decision-making ability
 5. Vocabulary in-Context Question:
In the context of the text, what does the term "automaticity" most nearly mean?
A. The ability to perform tasks without conscious thought
B. The tendency to experience cognitive dissonance
C. The result of excessive stress

He used prompt design strategies on ten occasions, which included providing task details (twice), asking for an answer key (once), asking for revision of the text (twice), asking for revision of items (four times), and giving instructions for self-evaluation (once). For example, while asking for revision for both the text and items, he wrote "The text includes a large number of words that might be too advanced for the test takers, such as ingrained, persist, dissonance. Redesign the text using more appropriate vocabulary. Revise the questions accordingly." He also asked ChatGPT for self-evaluation by writing "The correct answer to the 1st question is the same as the title of the text. Do you think it is appropriate?" While waiting for ChatGPT's response, he stated, "Let's see whether ChatGPT can realize its own mistake."

Instructor 1 performed quality checks a total of nine times, including checking the text (three times), items (five times), and options (once). He rationalized checking the text (twice) by stating, "I first take a look at the text because sometimes ChatGPT can make mistakes in sentences, even if it is rare." and "It is useful to look at the text again first. When it comes to

intermediate, ChatGPT can prepare a slightly higher level of vocabulary in the text.” He also engaged in more rationalization, including using the Word document and adjusting the level.

Instructor 1 mentioned editing ChatGPT-generated output on four occasions, which included rewriting the title once, modifying options twice, by rewriting and reordering them.

In the post-ChatGPT use stage, Instructor 1 consulted an external resource (once), which was the Word document, for adjusting the format and layout of the items in the final document (see Figure 13 and 14).



Figure 13.
The Final Draft of the Text Instructor 1 Prepared

Understanding the Psychology of Habit Formation

Habits are actions we do many times, and over time they become easy to do without much thought. Studying how habits form is an important part of psychology. This field helps us understand why we do the same things every day and how we can change habits if needed.

One way to explain how habits form is through reward learning. When you do something and receive a reward, you are more likely to do it again. For example, if you feel happy after finishing your homework, you might want to do it more often. Some researchers also study learning by watching others or by linking ideas together, but reward learning is the main idea in this area.

A simple way to think about a habit is as a loop with three parts. The first part is a cue—a sign or signal that tells you it is time to act. The second part is the action or routine, which is the habit itself. The third part is the reward, which is the good feeling or benefit you get after doing the habit. When this loop happens again and again, the habit becomes strong and happens almost by itself.

The brain plays a big part in making habits automatic. As you repeat the same action, your brain builds strong connections that help you do it without thinking. This means that even when you are tired or under stress, the habit can still occur because your brain is ready for it. Although our feelings, like stress, and careful choices (decision-making) can affect our behavior, the habit loop is the main force that drives our actions.

Changing a habit is not easy because of these strong brain connections. Sometimes, people try to change their habits by using willpower, which means the inner strength or self-control that helps them resist temptations. However, instead of relying solely on willpower, it can help to change one part of the loop. For example, if you have a habit of eating too many sweets when you feel stressed, you might try to change the cue by avoiding situations that cause stress or choose a new reward, like taking a short walk. This method is usually more successful than simply trying to stop the habit without making any changes.

In the end, the psychology of habit formation shows us that repeating an action can turn it into an automatic behavior. It explains that the habit loop of cue, action, and reward is the key idea behind forming habits. Even though other factors—like stress, decision-making, and learning by watching others—also affect how we act, the habit loop is the most important part of this process.

By understanding how habits are made, we can learn how to change habits that are not good for us and build habits that are better. This knowledge can help improve study habits, encourage a healthier lifestyle, or even support learning new skills. The study of habit formation shows why our daily actions can seem automatic and why changing them is sometimes hard.

Figure 14.
The Final Draft of the Questions Instructor 1 Prepared

Main Idea Question:

According to the text, why is it hard to change a habit?

- A. Because habits build strong brain connections through repetition
- B. Because stress always stops any change
- C. Because careful decision-making is too challenging

Detail Question 1:

What are the three parts of the habit loop described in the text?

- A. Cue, action, and reward
- B. Stress, decision, and reward
- C. Cue, observation, and routine

Detail Question 2:

Which method of learning is mentioned as the main way habits are formed?

- A. Reward learning
- B. Learning by watching others
- C. Learning by linking ideas together

Inference Question:

What can be inferred about the role of stress in habit formation?

- A. Stress can affect behavior, but the habit loop is the key factor in forming habits
- B. Stress is the main reason habits are created
- C. Stress completely prevents habits from developing

Vocabulary in-Context Question:

In the text, what does "willpower" mean?

- A. The power to resist quick urges
- B. A signal to start a task
- C. A reward that makes you happy

Case 2: ChatGPT Use for Elementary-level L2 Reading Task Development

During the pre-ChatGPT use process, Instructor 2 consulted an external resource (once), referring to the test specifications before starting the task.

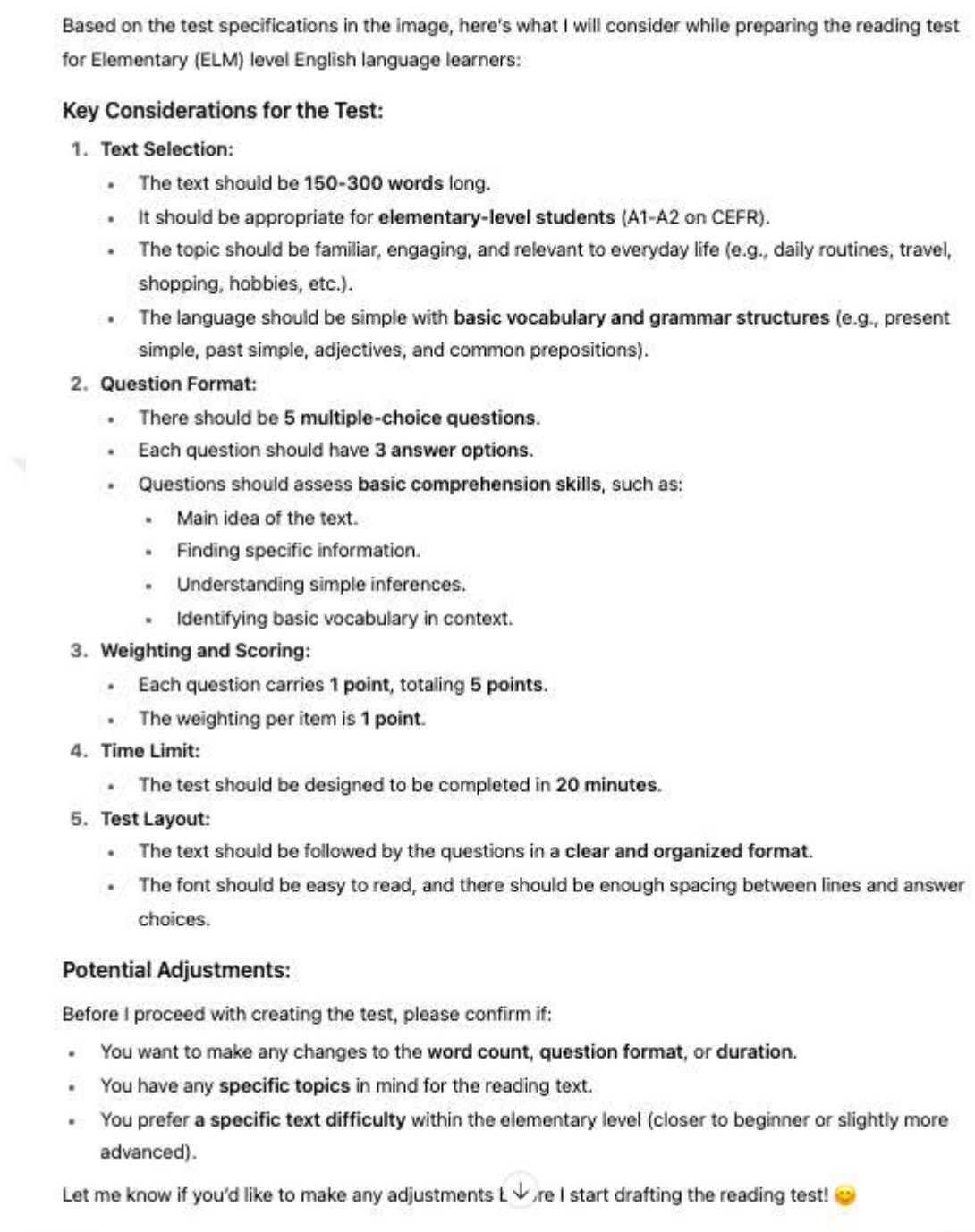
For the reading task development, Instructor 2 signed into his ChatGPT account. During his ChatGPT use, he engaged in a total of ten prompts, including providing task details (three times), asking for suggestions (once), asking ChatGPT for self-evaluation (four times), asking for level adjustment (once), and uploading files (once). He used external resources three times: once for coursebook theme selection and twice for checking the word count in a Word document. He did a total of five quality checks, including two instances of self-evaluation by ChatGPT, and one each for word count, text, and items. He also offered rationalizations for his decisions, for why he checked the text (once), why he provided detailed prompts (twice), and asked ChatGPT for self-evaluation (twice). The following narration illustrates Instructor 2's ChatGPT use, providing a detailed and chronological account of his prompts, interactions with external resources, and reflections.

Regarding task details, Instructor 2 took a screenshot of test specifications and pasted it to ChatGPT by stating, "I give the necessary information to ChatGPT at first, and I build on what I have. I generally move on step by step". In addition to the test specifications, the instructor explained the task with some contextual details and asked ChatGPT for its considerations for test design. The first prompt of Instructor 2 was as follows:

"I attached a test specifications file. Check it out to see what it has for ELE (elementary) group. We are going to prepare a reading test for elementary level English language learners at a university English prep class. Before you prepare the test and the text tell me what you are going to consider for this test so that we can make necessary changes and alterations."

When ChatGPT listed the key considerations for the test, which were text selection, question format, weighting and scoring, time limit, and test layout, the instructor went over ChatGPT's response (see Figure 15).

Figure 15.
ChatGPT's Initial Response



Reading ChatGPT's response that "the text should be appropriate for A1-A2 CEFR level," the instructor wrote a new prompt to lead ChatGPT to focus on A2 level only: "For text selection, remove A1 from the equation and only focus on A2 level. And prepare the text and the items." He also provided context by giving ChatGPT a role twice by writing "When you are writing the items, act like a testing expert and consider key principles of item writing for

language tests.” and “Act like a testing expert who will give this test feedback and give me a report.”

When ChatGPT created the reading text (see Figure 16) along with its items (see Figure 17) , Instructor 2 first copied the text onto a Word document to check the word count, saying that “Sometimes ChatGPT has a problem with word count.” Seeing that the word count was around 170 words, the instructor wrote a new prompt: “the thing is the text looks a bit too short for 5 questions lets expand the text a bit to around 280 words. Also, check to see if all distractors are mentioned in the text somehow but they are still incorrect answers.”

Figure 16.

The Initial Text ChatGPT Created

Got it! I'll prepare a **150-300 word A2-level reading text** along with **5 multiple-choice questions**, ensuring they follow key principles of test item writing (e.g., clarity, plausible distractors, avoiding ambiguity, and testing comprehension effectively).

Here's the reading test:

Reading Passage: A Weekend at the Beach

Last weekend, Emma and her family went to the beach. The weather was sunny and warm, so they decided to spend the whole day there. They arrived early in the morning and found a nice spot near the water. Emma's father set up a big umbrella while her mother prepared some sandwiches for lunch.

Emma and her brother, Jake, played in the water for hours. They swam, collected seashells, and built a large sandcastle. After that, they sat under the umbrella and ate their lunch. The sandwiches were delicious, and they also had fresh fruit and cold lemonade.

In the afternoon, Emma and her family went for a walk along the shore. They saw small fish in the clear water and watched some people surfing. Emma wanted to try surfing one day, but she thought it looked difficult. Before leaving, they took some photos to remember the day.

When they got home, Emma felt tired but happy. It was a perfect weekend, and she couldn't wait to go back to the beach again.

Figure 17.
The Questions of the Initial Text ChatGPT Created

Comprehension Questions

(Each question has three answer options with one correct answer.)

1. Why did Emma and her family decide to spend the day at the beach?
 - a) Because it was a holiday.
 - b) Because the weather was sunny and warm. ✓
 - c) Because they had a new umbrella.
2. What did Emma and Jake do at the beach?
 - a) They learned how to surf.
 - b) They built a sandcastle and swam. ✓
 - c) They went fishing.
3. What did Emma and her family eat for lunch?
 - a) Pizza and ice cream.
 - b) Sandwiches, fruit, and lemonade. ✓
 - c) Hot dogs and cola.
4. What did Emma think about surfing?
 - a) She wanted to try it but thought it looked difficult. ✓
 - b) She was afraid of the water.
 - c) She already knew how to surf.
5. How did Emma feel at the end of the day?
 - a) Tired but happy. ✓
 - b) Angry and disappointed.
 - c) Excited and full of energy.

Key Principles Used in Item Writing:

- ✓ Clear and simple language appropriate for A2-level learners.
- ✓ One correct answer per question with plausible distractors.
- ✓ Avoided grammatical clues in the answer choices.
- ✓ Tested key comprehension skills: main idea, specific details, inference.

Does this match what you had in mind? Let me know if you'd like any refinements! 😊



After receiving the revised text (see Figure 18), the instructor used the Word file again to check the word count, which was around 260. He commented, “Although I wanted a text around 280 words, it gives me a text around 260 words. It is OK. It sometimes does this. It is what it is. Not a big problem.”

Figure 18.
The Revised Text ChatGPT Created

Here's an expanded version of the reading passage, ensuring it's around **280 words** while keeping it at the **A2 level**. I've also checked that **all distractors are mentioned in the text** but remain incorrect answers.

Reading Passage: A Weekend at the Beach

Last weekend, Emma and her family went to the beach. The weather was sunny and warm, so they decided to spend the whole day there. They left home early in the morning and packed everything they needed—towels, sunscreen, and a big blue umbrella. Emma's father carried the cooler, which was full of sandwiches, fresh fruit, and cold lemonade.

When they arrived, they found a perfect spot near the water. Emma and her younger brother, Jake, ran straight to the sea. The water was cool and clear, and they spent hours swimming. Later, they walked along the shore and collected seashells. Jake found a beautiful starfish, but their mother told him to leave it in the water.

After playing in the sand, they built a large sandcastle with tall towers and a moat. They wanted to make it the biggest on the beach, but a big wave destroyed part of it. They laughed and decided to rebuild it.

At lunchtime, they sat under the umbrella and ate their sandwiches. Their parents had also packed cookies, but they saved those for later. While eating, they watched some people surfing. Emma thought surfing looked fun, but she wasn't sure if she could do it.

In the afternoon, they took a long walk. They saw small fish in the water and even a group of people fishing nearby. Before leaving, they took photos to remember the day. When they got home, Emma was tired but happy. It had been a perfect weekend, and she couldn't wait to go back.

For the revised text, the instructor asked ChatGPT to evaluate the new text according to all the given prompts, again instructing it to do this with the role of a testing expert: “Go and check all the prompts I gave you and see how the final version of the text fits all those prompts, and report. Act like a testing expert who will give this test feedback and give me a report.” The instructor justified this request by saying, “because it gets a bit too complicated and it may forget what it has done, it is important for me to read the report in detail as a final step. I will read all the details it has written in the report to check whether it is really so, or whether it is delusionally giving me ideas here or not. Most of the time it doesn't do that.” ChatGPT responded with the details of text length, question format, key principles of item writing, distractors, time appropriateness, and overall assessment (see figure 19).

Figure 19.
ChatGPT's Response

Go and check all the prompts I gave you and see how the final version of the text fits all those prompts and report. Act like a testing expert who will give this test feedback and give me a report.

Test Expert Report on the Reading Test for ELM Level

Evaluation Based on Given Prompts

The final version of the reading test was prepared following the given specifications and item-writing principles. Below is a breakdown of how well the test aligns with each requirement:

1. Text Length & Level (A2 Only) →

- **Requirement:** The passage should be around **280 words** and at the **A2 level** (not A1).
- **Outcome:** The final text contains **280 words**, making it appropriate in length for a 20-minute test. The language is strictly **A2**, using **common verb tenses (past simple)**, **high-frequency vocabulary**, and **simple sentence structures**. There are no complex clauses or CEF B1-level structures that might be too difficult for elementary learners.

2. Question Format →

- **Requirement:** **5 multiple-choice questions**, each with **3 options** (one correct answer + two plausible distractors).
- **Outcome:** The test contains **5 well-structured multiple-choice questions** that effectively test comprehension while adhering to proper test-item writing principles.

3. Key Principles of Item Writing → ✓

Each test item was designed to meet professional testing standards:

- ✓ **Clear and unambiguous stem wording:** Questions are direct and avoid unnecessary complexity.
- ✓ **One correct answer with two plausible distractors:** Each distractor is mentioned in the text but remains incorrect.
- ✓ **No grammatical or structural clues in answer choices:** No option is made obvious due to wording inconsistencies.
- ✓ **Avoidance of subjective or opinion-based questions:** All answers are objectively verifiable from the text.
- ✓ **Balanced distribution of question types:** The test assesses different reading comprehension skills, including:
 - **Main idea** (Q1, Q5)
 - **Specific details** (Q2, Q3)
 - **Inference** (Q4)

4. Distractors → ✓

- **Requirement:** All incorrect answer options should be mentioned in the passage but still be incorrect.
- **Outcome:** Each distractor appears somewhere in the text, making them **plausible but incorrect**.

Example:

- **Q1:** "It was a holiday" → Weekend is mentioned, but no holiday.
- **Q2:** "They went fishing" → They saw people fishing but didn't do it themselves.
- **Q3:** "Pizza and ice cream" → Not mentioned, reinforcing test validity.
- **Q4:** "She was afraid of the water" → She swam, so this is false.
- **Q5:** "Excited and full of energy" → She was excited but also tired.

This ensures **no answer is too obvious** and maintains proper item discrimination.

5. Time Appropriateness (20 Minutes) → ✓

- **Requirement:** The test should be **completed in 20 minutes**.
- **Outcome:** The passage is readable in **2-3 minutes**, and each multiple-choice question should take **2-3 minutes** to answer, making the test duration **reasonable for A2 learners**.

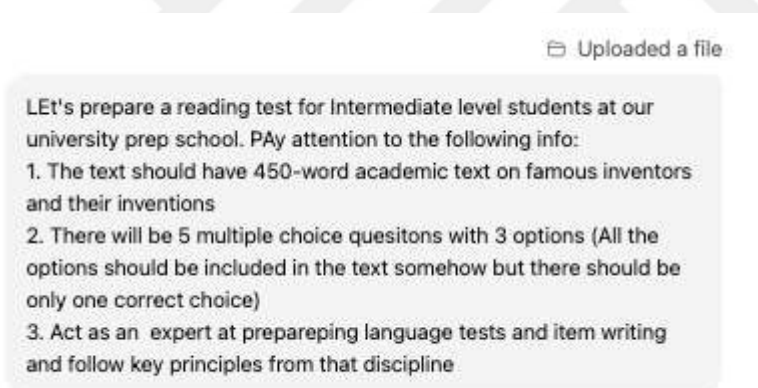
Before ending his elementary level reading task development with ChatGPT, the instructor stated that he would take the final product (i.e., the reading text and the items) and put them on a Word document to go over them himself, saying that the "final touch almost always requires to be done by a human expert." The instructor ended the recording with these comments without editing the final text and items produced by ChatGPT.

Case 2: ChatGPT Use for Intermediate-level L2 Reading Task Development

No pre-ChatGPT use process was observed. Instructor 2 signed into his ChatGPT account and started referring to the wordlist (once), the coursebook for theme selection (once), and the test specifications (once). During the ChatGPT use process, he uploaded the Intermediate wordlist document (which is confidential to the institution) onto ChatGPT and a screenshot of a contents page of one unit of the coursebook. However, as the wordlist is a bit large in size, ChatGPT did not accept the second file. He tried to upload the screenshot of the test specifications, but that did not work either. He stated that typing the test specifications onto ChatGPT is easy, so he kept the wordlist attached there. Thus, he started writing his 1st prompt and while writing the details in his prompt, the instructor kept consulting the external resources, which were checking the test specifications and contents page of the coursebook. The first prompt of Instructor 2 was as in Figure 20:

Figure 20.

Instructor 2's 1st Prompt



During the ChatGPT use process, Instructor 2 employed prompt design strategies a total of twelve times, including providing task details (three times), repeating (once), correcting previous prompts (once), asking for revision of the text (once), asking for revision of items (once), asking for revision of options (once), asking ChatGPT for self-evaluation (twice), uploading files (once) and giving positive feedback (once). He rationalized giving ChatGPT positive feedback by stating "I am writing 'good' because the product that ChatGPT gives is what I wanted. Giving responses like 'Good' is important so that ChatGPT can realize it can understand, and will not change that part." Additionally, regarding self-evaluation, he asked ChatGPT for self-evaluation twice. One of them was in his 2nd prompt as in Figure 21:

Figure 21.
Instructor 2's 2nd Prompt

I forgot to mention the wordlist, see if the text is aligned to the level of vocabulary that is in the list I attached.
See if all the prompts I gave are met with your final prudcut make changes if necessary.
Give feedback as if you were a testing expert in terms of item writing and check if it fits item writing principles.
Report all.

ChatGPT responded with all the details asked (see Figure 22).

Figure 22.
ChatGPT's Response

I've reviewed the text based on the **Language Hub Intermediate Wordlist** and checked its alignment with **key item writing principles** for language testing. Below is my **detailed analysis and revision plan**:

1. Vocabulary Check

I compared the text with the **Language Hub Intermediate Wordlist**. While many words align well with the **Intermediate** level (e.g., "inventor," "develop," "transform," "innovation"), a few words might be **too advanced** for this level, such as:

- **"groundbreaking"** (alternative: *important* or *innovative*)
- **"inefficient"** (alternative: *not effective*)
- **"predecessors"** (alternative: *earlier versions*)
- **"revolutionized"** (alternative: *changed completely*)
- **"concepts"** (alternative: *ideas*)

To ensure alignment, I will **simplify** these words while maintaining the academic tone.

2. Item Writing Principles Check

Each question follows **fundamental principles of good item writing**:

- ✓ **Clear, unambiguous stems:** Each question is direct and avoids unnecessary complexity.
- ✓ **Plausible distractors:** Incorrect answers are **all mentioned in the text**, ensuring they are not **obviously incorrect** (a common test-writing issue).
- ✓ **One correct answer:** There is **only one objectively correct choice**, which aligns with best practices.
- ✓ **Evenly distributed options:** The correct answers are **not in a predictable pattern**, which prevents guessing based on letter distribution.

However, **question 2** could be slightly clearer. The phrase **"He secured the patent before others"** might be misunderstood as "being the first to invent it." I will **reword** it to clarify that Bell was the first to **officially register the patent**, not necessarily the first to invent  voice transmission device.

3. Overall Structure & Content Check

- ✓ **Word Count:** The passage is **450 words**, meeting the length requirement.
- ✓ **Topic Relevance:** The passage is **on famous inventors and their inventions**, as required.
- ✓ **Academic Style:** The passage maintains an **informative and factual** tone.
- ✓ **Multiple-Choice Format:** Exactly **5 MCQs with 3 options each** are included.

4. Revision Plan

To align the vocabulary with the wordlist and enhance clarity, I will:

- Simplify or replace **advanced words** while keeping the academic tone.
- Improve **question 2** for clarity.
- Ensure **concise, direct phrasing** where needed.

I'll now revise the text and questions accordingly.

Edited

Final Report: Test Revision & Improvements

✓ Vocabulary Alignment

- The passage now aligns better with the **Language Hub Intermediate Wordlist** after simplifying **complex words** (e.g., *groundbreaking* → *important*, *inefficient* → *not very effective*).
- The changes ensure **clarity** without losing academic tone.

✓ Item Writing Improvements

- **MCQ Clarity:** Question 2 was revised to specify that Bell was the *first to register a patent* (avoiding confusion about invention vs. patenting).
- **Plausible Distractors:** All incorrect answers are mentioned in the text but are **not misleading or too obvious**.

✓ Structural & Content Review

- The **450-word length** is maintained.
- The topic remains **academic & engaging** for Intermediate learners.
- The **test format** follows good **language assessment principles**.

The test is now **ready for use!** Let me know if you'd like any final tweaks. 🚀

Additionally, he provided context by giving ChatGPT a role twice by writing in his prompts “Act as an expert at preparing language test and item writing and follow key principles from that discipline” and “Give feedback as if you were a testing expert in terms of item writing and check if it fits item writing principles.” He also stated “When I have a problem like this (item generation) with ChatGPT, I try to be as specific as in my instructions, giving my reasoning for the change I want so that it can yield the expected results.”

Quality checks were conducted ten times, including checks for text (three times), items (four times), an answer key (once), and self-evaluation (twice). He asked ChatGPT for revision

of the items by writing “There are some questions that can be answered with world knowledge, we shouldn’t have such questions. Can you fix that?” ChatGPT’s response was as in Figure 23:

Figure 23.
ChatGPT’s Response

You're absolutely right—some questions can be answered with **world knowledge** rather than relying solely on the passage. I will revise them to ensure that **all answers must be found in the text itself** and not assumed from prior knowledge. Here's what I will fix:

Issues & Fixes

1. Question 1 (Edison's invention)

- Someone could answer "**light bulb**" without reading the passage.
- **Fix:** Reword the question to focus on how the invention impacted daily life (information found in the passage).

2. Question 4 (Alan Turing's role)

- Many people already know Turing's role in computing.
- **Fix:** Make the question reference **details specific to the passage**, like his work on early computers **and** codebreaking.

3. Question 5 (Main idea)

- The correct answer ("inventors improve upon earlier ideas") could be guessed using logic.
- **Fix:** Frame the question in a way that requires reference to the **final paragraph** in the text.

Now, I'll make these updates!

Edited

Final Review: World Knowledge Issue Fixed

- **Questions now require information from the passage** rather than general knowledge.
- **Question 1** now asks about Edison's **impact** rather than simply naming his invention.
- **Question 4** ensures students refer to the passage's description of **Turing's role**.
- **Question 5** is explicitly tied to the **final paragraph**, reinforcing reading comprehension.

Your test is now **fully aligned** with **best practices in item writing**! Let me know if you need further tweaks.

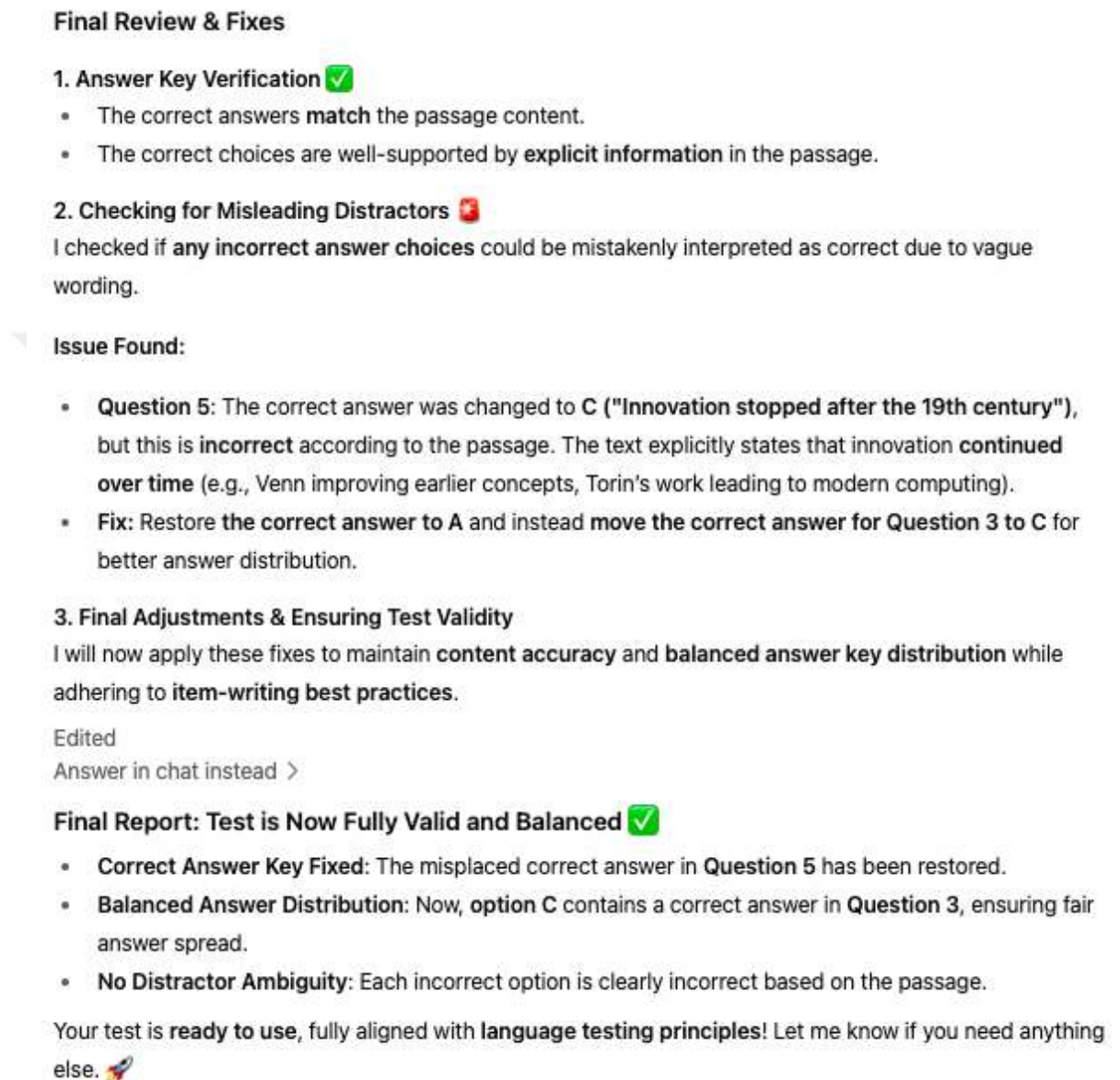


He continued referring to external resources during the ChatGPT process, which included word document twice for editing ChatGPT-generated output. Regarding consulting the Word document, he stated “I don’t use the test ChatGPT gives right away, I put it onto a word document and tailor it to my liking.”

In his last prompt, he asked ChatGPT for self-evaluation as he wrote: “Check the answer key again to see if it is correct. Are there any distractors that might be misinterpreted to be the correct answer? If so, fix it. Report what you have get the final version of the text to match the earliest prompt I gave.”

ChatGPT responded with self-evaluation named as “Final Review & Fixes” as in Figure 24:

Figure 24.
ChatGPT’s Response



Instructor 2 got a warning on the screen: ‘You have reached your free limit for GPT-4o. As there is an attachment in the chat, GPT-4o is required to continue this chat. Your limit resets at 15.52.’ He stated “I used all of the prompts I had for today, which is OK because I am already done with. So I will move the options to spread the correct answers. I will also read the report in detail and make the necessary changes on word document by myself.” and ended the video recording.

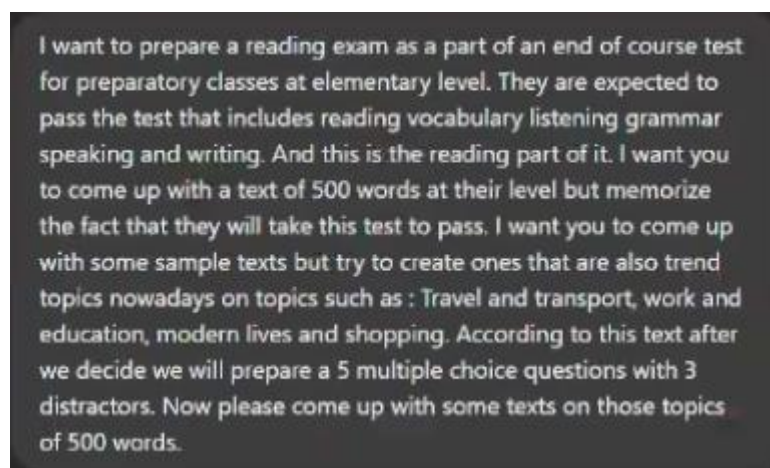
Case 3: ChatGPT Use for Elementary-level L2 Reading Task Development

No pre-ChatGPT use process was observed. Instructor 3 opened chatgpt.com and started using ChatGPT without an account. During the ChatGPT use process, the instructor engaged in multiple activities across several subcategories. Regarding external resource use, Instructor 3 referred to test specifications (once), coursebook (once) and Word document (four times). Regarding prompting behavior, Instructor 3 conducted prompt design strategies (49 times), which included providing task details (once), asking ChatGPT for revisions of the generated text (once), items (18 times) and options (seven times), giving positive feedback (once) and giving negative feedback (twice), requesting alternatives (twice), adjusting the level (twice) and adjusting word count (once), repeating (six times), rationalizing (twice), correcting previous prompts (once), asking for self-evaluation (once), asking for level check (three times) and asking for a new text (once). For quality check, the instructor reviewed the text (twice) and items (eight times). Regarding providing context, Instructor 3 gave ChatGPT no role. For editing ChatGPT-generated output, the instructor engaged in modifications four times, where she rewrote items (once) and options (twice) as well as paraphrasing content in the text. For her actions, the instructor provided one rationalization where she verbalized reasoning for providing details. The following illustrates Instructor 3's ChatGPT use, providing a detailed and chronological account of her prompts, interactions with external resources, and reflections.

Instructor 3 started using ChatGPT without signing into an account. Consulting the contents page of the coursebook and specifications, she posted her first prompt, as follows:

Figure 25.

Instructor 3's 1st Prompt

A screenshot of a text input field with a dark background and light-colored text. The text is a prompt for ChatGPT, detailing the requirements for a reading exam for elementary-level preparatory classes. The prompt asks for a 500-word text on specific topics, followed by five multiple-choice questions with three distractors each.

I want to prepare a reading exam as a part of an end of course test for preparatory classes at elementary level. They are expected to pass the test that includes reading vocabulary listening grammar speaking and writing. And this is the reading part of it. I want you to come up with a text of 500 words at their level but memorize the fact that they will take this test to pass. I want you to come up with some sample texts but try to create ones that are also trend topics nowadays on topics such as : Travel and transport, work and education, modern lives and shopping. According to this text after we decide we will prepare a 5 multiple choice questions with 3 distractors. Now please come up with some texts on those topics of 500 words.

Although ChatGPT had a notification in Turkish (see Figure 26) translated into English as “Welcome back. Sign in or register to get smarter answers, upload files and images, and more,” she ignored it by clicking on ‘Oturum kapalı kalsın (Stay signed out)’

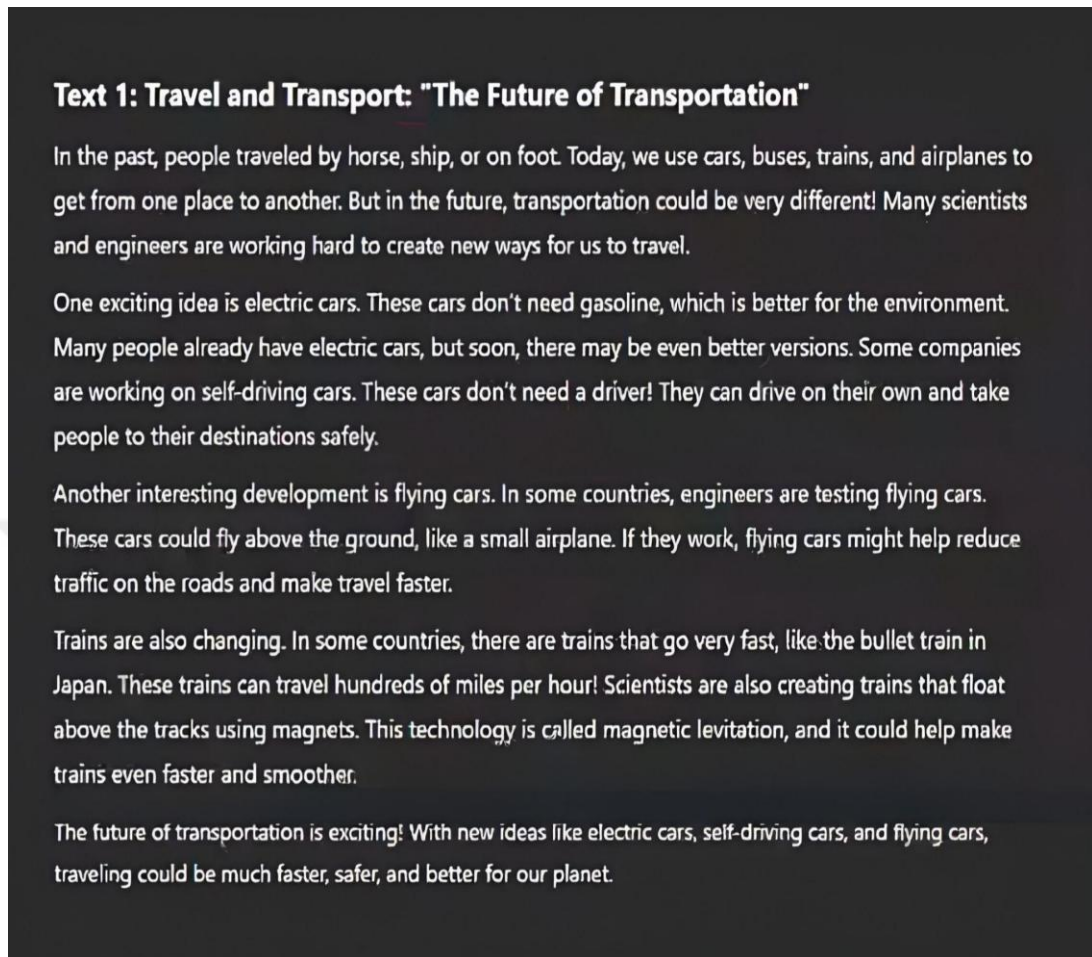
Figure 26.

Turkish Notification on ChatGPT Screen



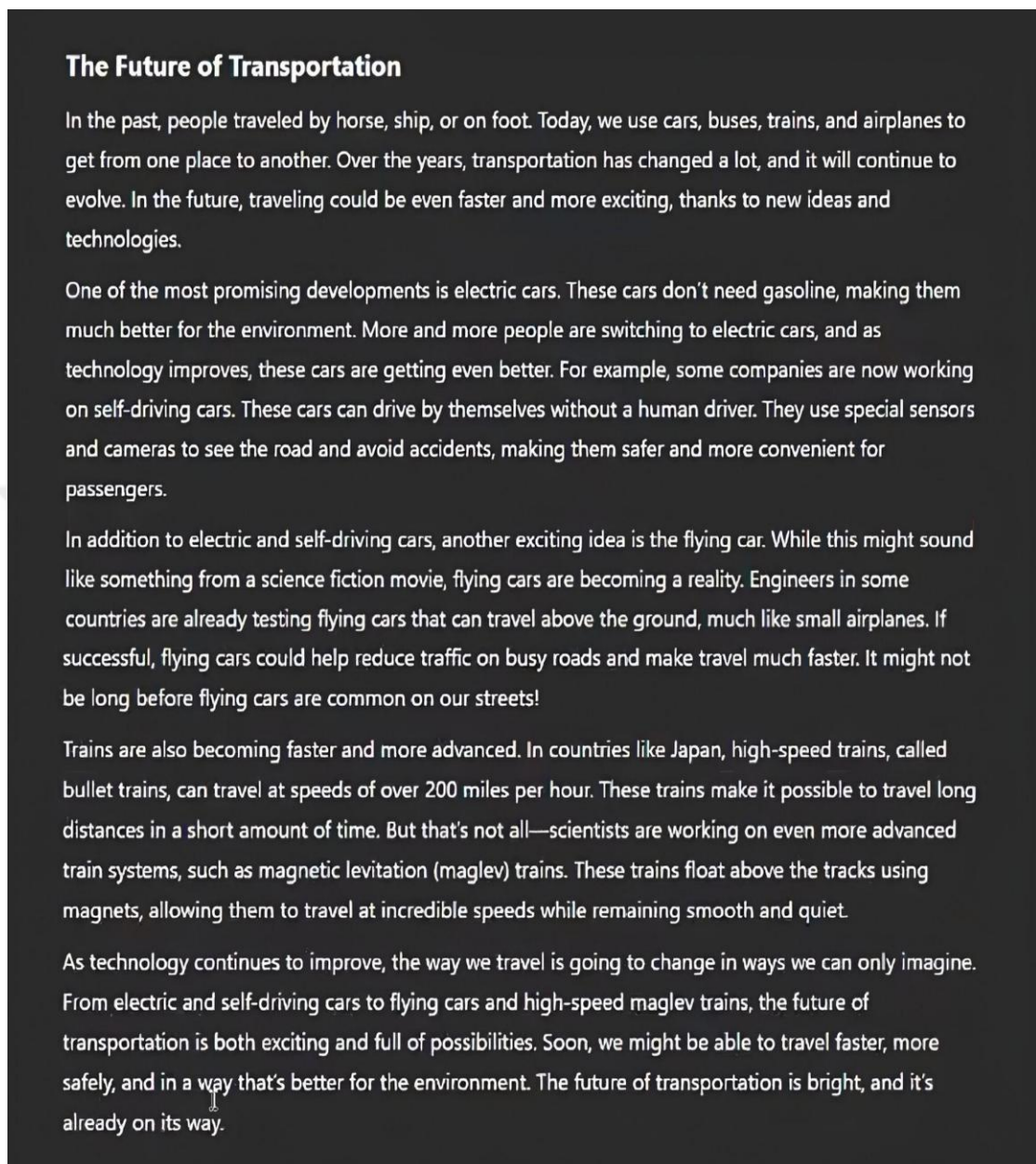
Among the texts ChatGPT produced, the instructor chose the first one (see Figure 27), but was not satisfied with its coherence and cohesion as she wrote a second prompt addressing this: “I liked the first one the most. Please provide coherence, unity and smooth transition between the ideas and paragraphs in order for us to use it as a reading text in the exam.”

Figure 27.
The First Text ChatGPT Created



ChatGPT produced an updated version of the text (see Figure 28), and the instructor seemed to be happy with it as she remarked, "That's great thank you."

Figure 28.
The Revised Text ChatGPT Created



The instructor continued with a new prompt, asking ChatGPT to check whether the text aligned with A2 A2+ level according to CEFR, for which ChatGPT provided the following response:

Figure 29.

ChatGPT's Response for Alignment of the Text Level with CEFR

You're welcome! Let's review the text to see if it aligns with the A2/A2+ level according to the CEFR (Common European Framework of Reference for Languages).

A2 Level Characteristics:

- **Vocabulary:** Simple and familiar words, some basic, concrete terms and phrases.
- **Grammar:** Simple sentence structures with some use of basic tenses (present, past, future).
- **Content:** Texts should be related to familiar everyday situations.
- **Reading Comprehension:** Can understand short texts that deal with familiar topics and simple descriptions.
- **Length:** Around 300-500 words is typical for A2 level.

Review of the Text:

Vocabulary:

- The vocabulary used in the text is relatively simple, with some words that might require clarification (e.g., *self-driving cars*, *maglev trains*, *electric cars*). These are not overly complex terms but could be considered a bit challenging for A2-level students, depending on their prior knowledge. However, they are commonly discussed in the media today, so many students might be familiar with them.

Grammar:

- The text uses basic sentence structures with a clear subject-verb-object order.
- The **present simple** (e.g., "People travel," "It is better") is used correctly, which is common at the A2 level.
- There's also the **future tense** (e.g., "will continue," "could be"), which is appropriate for A2+ students.
- The text doesn't use overly complex grammatical structures, keeping it simple and easy to follow.

Content:

- The content is about **modern technologies**, which is relevant and engaging for young learners, and the theme is familiar (transportation and new technology). While some concepts, like flying cars and self-driving cars, may seem futuristic, they are also popular topics today, making the content appropriate for A2+ learners.
- The text doesn't dive too deeply into technical details, but instead focuses on general descriptions and explanations, which suits A2-level learners who are still building their vocabulary.

Reading Comprehension:

- The text provides clear and easy-to-understand descriptions of future transportation technologies. It's organized logically, starting with familiar modes of transport and moving to new, futuristic ideas.
- The language level is consistent with A2 reading comprehension, as it's not too complex, and the sentences are clear and to the point.

Length:

- At around 450 words, the text is appropriate for A2 students. It's long enough to provide a meaningful context and introduce new concepts, but not so long as to overwhelm elementary learners.

Conclusion:

- **Vocabulary:** Generally simple but with a few terms that may need some explanation. If you include a vocabulary list or some pictures related to the technology, it would help.
- **Grammar:** Mostly A2-level with some future tense and basic structures.
- **Content:** The topic of future transportation is familiar and accessible, but some concepts (like flying cars) may be slightly above A2, although not too complex.
- **Comprehension:** The text is clear, logical, and easy to understand, making it suitable for A2/A2+ students.

Suggestion:

If you're aiming for a **strong A2 level**, you might want to simplify the explanations of futuristic concepts just a bit, especially for students who may not be as familiar with these technologies. For example, "flying cars" could be described as "cars that can fly in the sky, like small airplanes" instead of referring to them as "small airplanes."

Overall: The text is generally well-suited for A2+ students. If you want to aim for A2 specifically, you might consider simplifying the vocabulary and descriptions of some advanced technologies like self-driving and flying cars, or providing a glossary or images to support understanding.

Let me know if you'd like to make any adjustments!

At this point, the instructor received the same warning (i.e., "Tekrar hoşgeldin. Daha zeki yanıtlar almak, dosya ve görsel yüklemek ve daha fazlasını yapmak için oturum aç ya da kaydol" / "Welcome back. Sign in or register to get smarter answers, upload files and images, and more." She ignored it once again, writing her 4th prompt:

I guess this version is OK. Now please use this text below and prepare 5 mc items with 3 distractors. While doing this, please pay attention to creating phrases or sentences at equal length for distractors, also pay attention not to give the correct answers as the same distractors. I mean, the answers cannot A or B, we need to use variety in terms of letters of the correct answers.

Upon receiving the first items, the instructor provided ChatGPT with additional instruction: “I have to remind you of the fact that all distractors should also be included in the text. Also, the questions should be asked in the order they appear in the text also.”

The instructor continued with further requests: “Can you please avoid asking questions from the very beginning of the text, so please get rid of the 1st question and create items starting from the 2nd paragraph. Let’s try once more.”

Upon receiving the revised items, the instructor prompted, “Please prepare the options so that they will be at equal length.”

Going over the revised items, the instructor prompted, “Can we shorten the options as this might increase the cognitive load.”

The instructor wrote a new prompt, this time for the level of the items: “I want to remind you that the questions should also be A2 and A2+ level, can you check them please? They should be that easy either. Let’s try to make them a bit challenging.” She verbalized her action with the following remarks: “In the meantime, of course, I verified both the text and the questions a few times to see if they were appropriate according to CEFR on the basis of level.” When ChatGPT provided the updated version of the items, she opened a word document and copied and pasted the text and the last version of the items to check the word count.

Seeing that the text had 375 words, the instructor gave ChatGPT a new prompt: “Let’s keep the questions this way. But I remember asking you to come up with a text of 500 words but this one is shorter like 375 words. So first can you expand this a little bit, keep all the requirements and specifications I mentioned before in mind.” She verbalized her thinking with the following remarks:

We always have problems especially with the ‘word count’. I said to make the text 500 words, I even started to create the questions from that text, then when I added it to word, I realised that it was not 500 words. I went back to ChatGPT

again, I said you didn't make it 500, I told him to increase it, he made it around 400. Anyway, it's okay.

The instructor then copied the updated text to the word document again, and commented that the word count was OK, which was around 400. When the text was updated once again, she asked ChatGPT to update the items as well: “Let’s try the questions once more. Again keep everything I mentioned in mind while preparing the questions. But this time try to paraphrase the distractors, do not use them as the way they are given in the text.” She checked the updated items by going back to the text and asked for another revision of the items by giving ChatGPT negative feedback: “try not to ask questions that appear one after another. I mean each question might be from one paragraph. For example, the answer of the 1st and 2nd questions are too close. It should not be that way.” When she was commenting on this, she also referred to her own instructions that they were not specific enough:

After ChatGPT asked the questions, I looked and realised that he had asked questions very close to the same place, of course, because I did not say that they should not be close to each other, he had asked questions from very close places. So I said, ‘Change it so that it is not so close.’ He did it again so that it was not so close, but again he asked questions from close places in the same text. I didn't like the questions.

Going over the questions again by looking at the text, she gave negative feedback to ChatGPT regarding the quality of the items: “The answers cannot be predictable. The students should not guess the answer without reading the text. I can answer the 3rd question without reading the text.” ChatGPT provided the updated questions along with its explanations for the adjustments. The instructor asked for a change in one of the items as well as for an additional item: “Please change the last question. Make it an inference question or prepare a question from the last paragraph. Prepare them both. I will choose one of them.” ChatGPT produced two optional items. The instructor chose option 2, but asked ChatGPT for a shorter version: “I like option 2. But please shorten it and also check the level of the question and the distractors as well.”

The interaction between Instructor 3 and ChatGPT continued with many new prompts in which she asked for several new revisions. The following prompts illustrate this process:

Instructor 3: Ok. Except questions 1 and 2, the questions are ok. As for the first and second question, I have doubts because the answers are too close to each

other. Either find another 2 questions because both electric cars and self driving car appear in the same paragraph or change the first question. You might find the first question from the first paragraph especially from the last sentences of the last paragraph. Let's give it a try.

ChatGPT revised the items

Instructor 3: Let me correct myself because I made a mistake. I meant prepare the question from the last part of the first paragraph.

ChatGPT revised the items

Instructor 3: You misunderstood me. Ok then except questions 1 and 2, the questions are ok. As for the first and second question, I have doubts because the answers are too close to each other. Either find another 2 questions because both electric cars and self driving car appear in the same paragraph or change the first question. You might find the first question from the first paragraph especially from the last sentence of the first paragraph. Let's give it a try.

ChatGPT revised the items

Instructor 3: Ok. Let's try once more. Please prepare a question. It will be the first question.

ChatGPT revised the item

Instructor 3: Don't you think it is also predictable?

ChatGPT revised the item

Instructor 3: Can you shorten the question. It is too long. Also the options please.

ChatGPT revised the item

Instructor 3: We do not need to use the phrase according to the last part of the first paragraph. It was a specification that I gave you for a better result.

ChatGPT revised the item

Instructor 3: Can you please write all the questions altogether? Check the level of the questions and the distractors. Keep the options at the same length. Be careful with the order of the questions as well.

ChatGPT wrote all the items together

Instructor 3: Please prepare the first question about electric cars, the second question about flying cars, the third question trains and 4th from maglev trains and the last question from the last paragraph please.

ChatGPT wrote the items

Instructor 3: Don't you memorize what I wrote you about the answers being as in the same distractor like each answer is a here.

ChatGPT revised the items

Instructor 3: What about options being at the same length?

ChatGPT revised the items

Instructor 3: Shorten option A in the last question

ChatGPT shortened option A

Instructor 3: Can you please, lastly, check the level of the questions according to CEFR?

ChatGPT explained the CEFR level of each question

Instructor 3: What is the level of "differ from"?

ChatGPT gave the level of "differ from"

Instructor 3: What about the level of the fourth question

ChatGPT explained the level of the fourth question

Instructor 3: Can you simplify the language in question 4?

ChatGPT provided the simplified version

Instructor 3: Can you make the third question a NOT question?

ChatGPT revised question 3

Instructor 3: Avoid words like only always directly that might make the answer predictable.

ChatGPT revised question 3

Instructor 3: Ok, I didn't like it because the answer is wrong now. try the second question as a NOT question please.

ChatGPT revised question 2

Instructor 3: Still predictable. Do we see any flying cars around?

ChatGPT revised question 2

Instructor 3: Can we add variety while preparing the questions yes they will all be multiple choice but may be different kind of root and question forms. These look similar this way.

ChatGPT revised the questions

Instructor 3: Can we try to find come up with another text about travel and transportation but it should be written in a way that it includes different kinds of information so that asking different kind of questions might be possible.

ChatGPT created a new text titled 'The Changing World of Transportation' with new items

Case 3: ChatGPT Use for Intermediate-level L2 Reading Task Development

No pre-ChatGPT use process was observed as Instructor 3 directly opened chatgpt.com once started the video recording.

During the ChatGPT use process, the instructor demonstrated a wide range of actions and strategies. The subcategory prompt design strategies was the most frequently observed, with a total of 55 instances. These involved providing task details (once), asking for alternatives (once), asking for revision of text (four times), asking for revision of title (once), asking for revision of items (17 times), asking for revision of options (nine times), giving positive feedback (three times), giving negative feedback (nine times), correcting previous prompts (once), asking for an answer key (once), asking for self-evaluation (five times), and rationalizing (three times).

The instructor also engaged in quality check activities 14 times, which included checking items on 9 occasions, checking the text 4 times, and checking the options once. Providing context was not observed as the instructor did not give ChatGPT a role in the task.

Editing ChatGPT-generated output took place five times, which included rewriting items (three times) and options (twice).

The subcategory rationalization was observed twice for choosing familiar texts and tasks as the instructor stated “Our students are used to seeing the thesis statement in the beginning. They would like to see it like that,” and “These topics are covered a lot nowadays, so they may seem too familiar, but it is still nice.”

Instructor 3 consulted external resources 10 times, including the coursebook for theme selection (once) and the Word document for editing output (eight times) and adjusting layout and format (once).

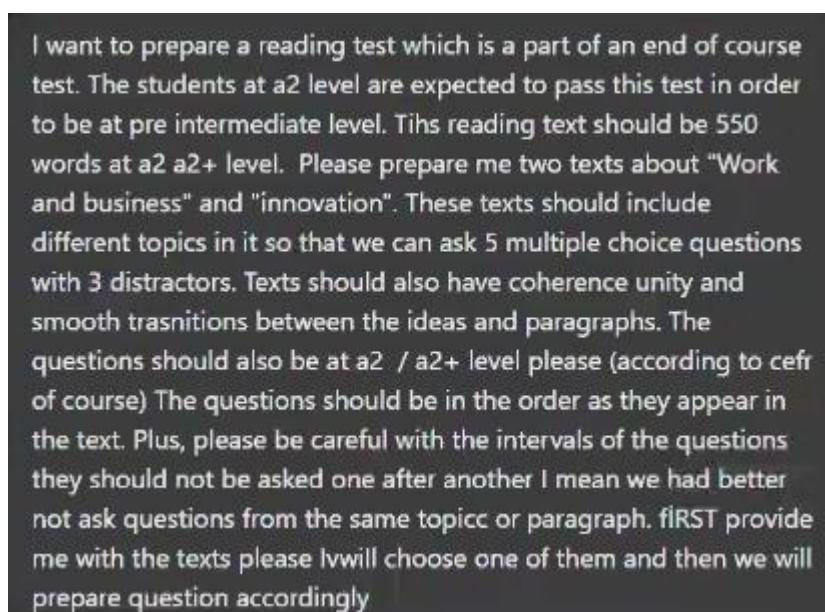
No post-ChatGPT use process was observed.

The following illustrates Instructor 3’s ChatGPT use, providing a detailed and chronological account of her prompts, interactions with external resources, and reflections.

Instructor 3 started using ChatGPT without signing into an account. Consulting the contents page of the coursebook and specifications, she posted her first prompt, as follows:

Figure 30.

Instructor 3’s 1st Prompt

A screenshot of a text input field on a dark background. The text is white and reads: "I want to prepare a reading test which is a part of an end of course test. The students at a2 level are expected to pass this test in order to be at pre intermediate level. Tihs reading text should be 550 words at a2 a2+ level. Please prepare me two texts about "Work and business" and "innovation". These texts should include different topics in it so that we can ask 5 multiple choice questions with 3 distractors. Texts should also have coherence unity and smooth trasnitions between the ideas and paragraphs. The questions should also be at a2 / a2+ level please (according to cefr of course) The questions should be in the order as they appear in the text. Plus, please be careful with the intervals of the questions they should not be asked one after another I mean we had better not ask questions from the same topicc or paragraph. flRST provide me with the texts please lvwll choose one of them and then we will prepare question accordingly". There are several typos in the text: "Tihs", "trasnitions", "topicc", and "flRST".

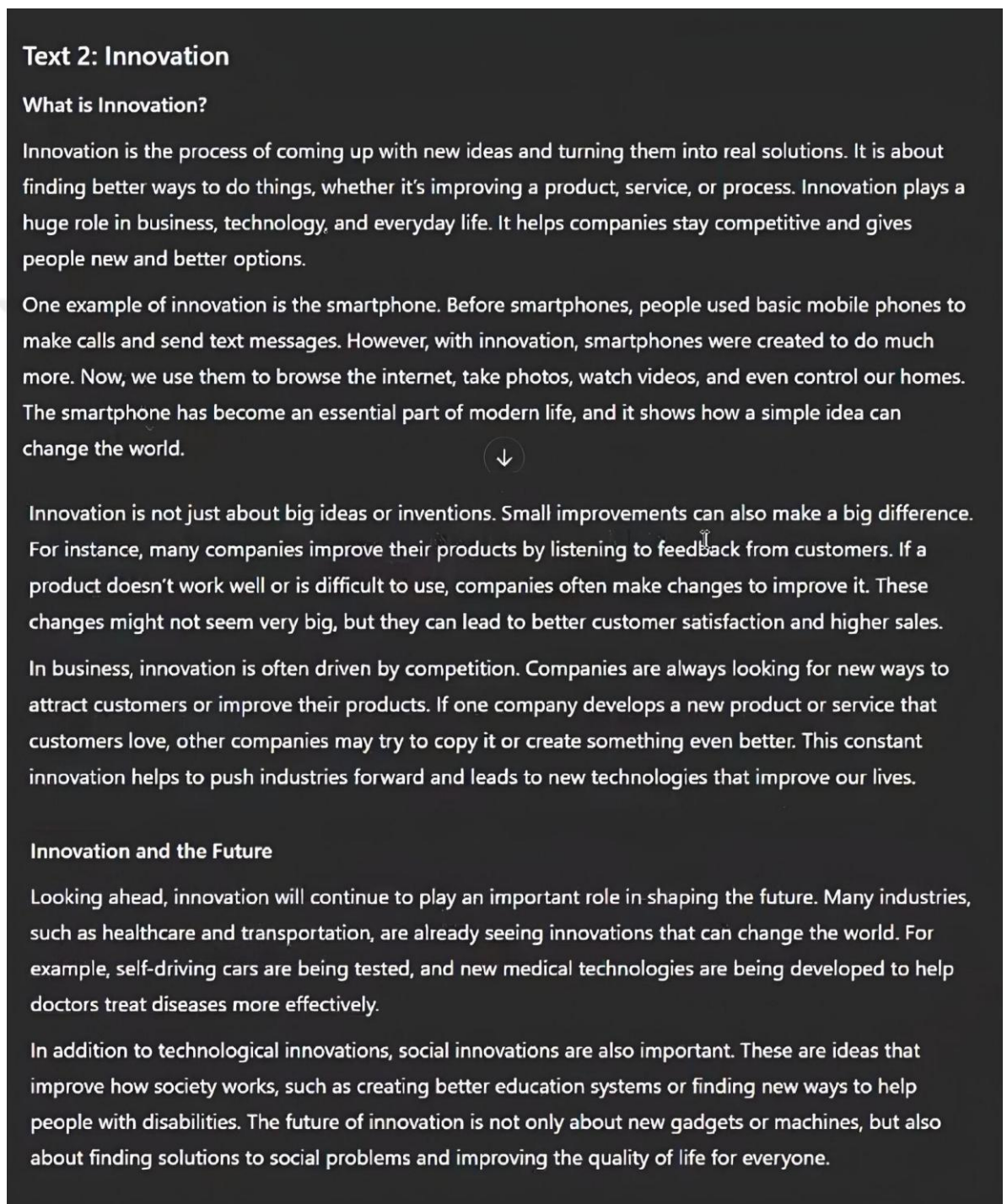
I want to prepare a reading test which is a part of an end of course test. The students at a2 level are expected to pass this test in order to be at pre intermediate level. Tihs reading text should be 550 words at a2 a2+ level. Please prepare me two texts about "Work and business" and "innovation". These texts should include different topics in it so that we can ask 5 multiple choice questions with 3 distractors. Texts should also have coherence unity and smooth trasnitions between the ideas and paragraphs. The questions should also be at a2 / a2+ level please (according to cefr of course) The questions should be in the order as they appear in the text. Plus, please be careful with the intervals of the questions they should not be asked one after another I mean we had better not ask questions from the same topicc or paragraph. flRST provide me with the texts please lvwll choose one of them and then we will prepare question accordingly

On the screen of ChatGPT, a Turkish warning appears which can be translated into English as “Welcome back. Sign in or register to get smarter answers, upload files and images, and more.” The participant ignored it and used ChatGPT without signing in.

ChatGPT created two texts titled “Work and Business” and “Innovation”. The instructor chose the latter (see Figure 31).

Figure 31.

The Initial Text ChatGPT Created

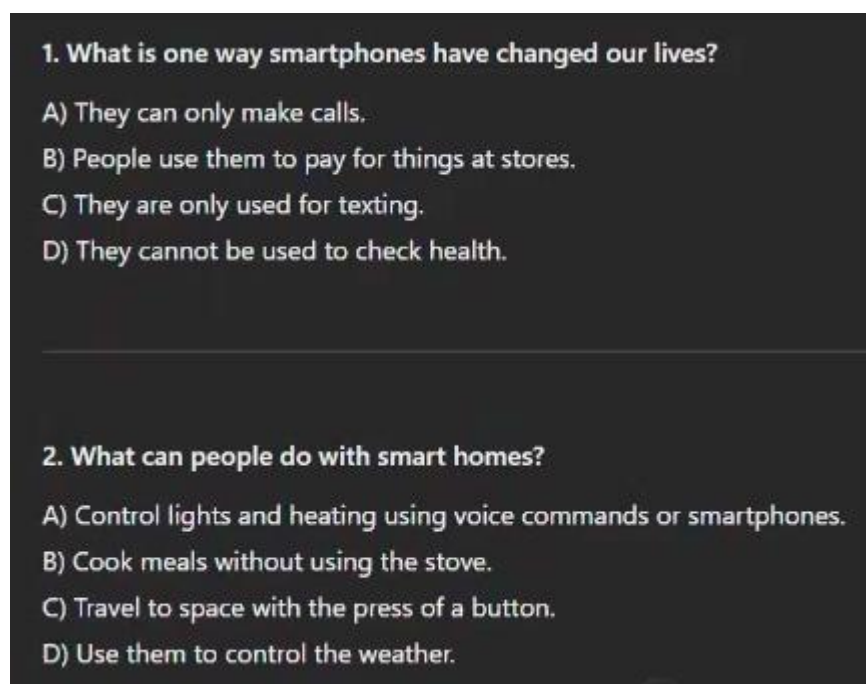


Instructor 3 went over the two texts and wrote her 2nd prompt: “I liked the second text more however I am still not satisfied because it can be more coherent and maybe the topic might be more trendy.” Then, she continued writing a couple of more prompts to ask for revisions of either the text or the title: e.g., “can it be in a different mode”, “instead of dividing it into subheadings can you get rid of the subheadings”, “can you shorten the title and find a more engaging and catchy one” until she wrote: “ok deal.” ChatGPT responded: “Great! I’m glad you like the title. Now that the text and title are set, let me know when you’re ready to move on to creating the multiple-choice questions for the text!”

As noted earlier, Instructor 3 accessed ChatGPT without signing into an account. Thus, no sharable link to the conversation with ChatGPT was available. Additionally, during the screen-recorded test development process, she scrolled rapidly through the generated text and items in the conversation, which limited the visibility of the complete content. As a result, only partial excerpts of the generated text and questions could be retrieved from the recording for illustrative purposes in this section.

When Instructor 3 wrote: “proceed with the questions please”, ChatGPT gave 5 multiple choice questions with 4 options (see Figure 32).

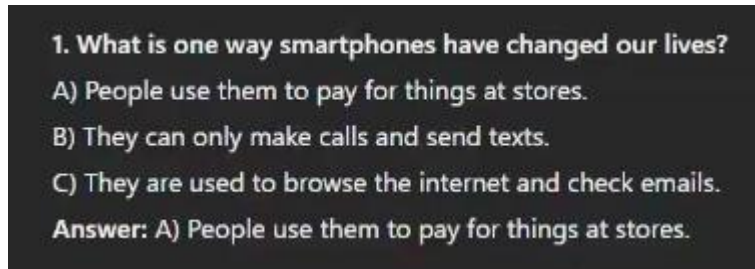
Figure 32.
The Two of the Initial Questions ChatGPT Created



Instructor 3 corrected her previous prompts by writing: “sorry I forgot to tell you that it should be 3 options the options should be at equal length”, ChatGPT gave 5 multiple choice questions with 3 options this time (see Figure 33).

Figure 33.

One of The Revised Questions ChatGPT Created



The instructor continued her prompts by asking ChatGPT for an answer key (once) and for self-evaluation (twice) until ChatGPT gave the questions, options and correct answers. Then, she copied the text onto the Word document (see Figure 34).

Figure 34.
The Revised Text ChatGPT Created

"How Innovation Is Changing Our World"

Innovation is happening all around us, and it's changing the way we live, work, and play. Every day, new ideas and technologies make our lives easier and more exciting. Have you noticed how much technology has changed in just a few years? Things like smartphones and smart homes are examples of how innovation is everywhere, affecting almost every part of our daily lives.

Take smartphones, for example. Not too long ago, phones were only used to make calls. But now, they are so much more. Thanks to innovation, smartphones can do almost anything! You can use them to send emails, play games, find directions, and even check your health. People also use their phones to pay for things at stores! Smartphones have become essential tools, helping us connect with others and stay informed in ways we never imagined.

Smart homes are another exciting example of innovation. Imagine being able to control your lights, heating, and security system just by speaking or tapping your phone. That's what smart homes allow you to do! With smart gadgets like voice assistants (such as Alexa or Google Assistant), smart thermostats, and smart lights, you can make your home more comfortable and energy-efficient. These innovations are making life easier, especially for busy people who want more control over their homes.

Another area of innovation that is changing our world is **Artificial Intelligence (AI)**. AI is when machines are designed to think and learn like humans. You might already know AI from voice assistants like Siri or Alexa, or chatbots that help you with customer service online. But AI is not just about talking to robots. It's also changing many industries in important ways.

For example, in healthcare, doctors are using AI to help them diagnose diseases faster and more accurately. AI can analyze medical images, like X-rays, to help doctors find problems that might be difficult to see. In business, AI helps companies understand customer needs and create better products. AI is making many tasks easier and faster, and it will continue to grow in the future.

While innovation brings many benefits, it also comes with challenges. One concern is that new technologies might take away jobs. Some tasks that people once did are now being done by machines, such as in factories or delivery services. But there is good news too: while some jobs may disappear, new jobs are being created. People will need to learn new skills to work with these technologies.

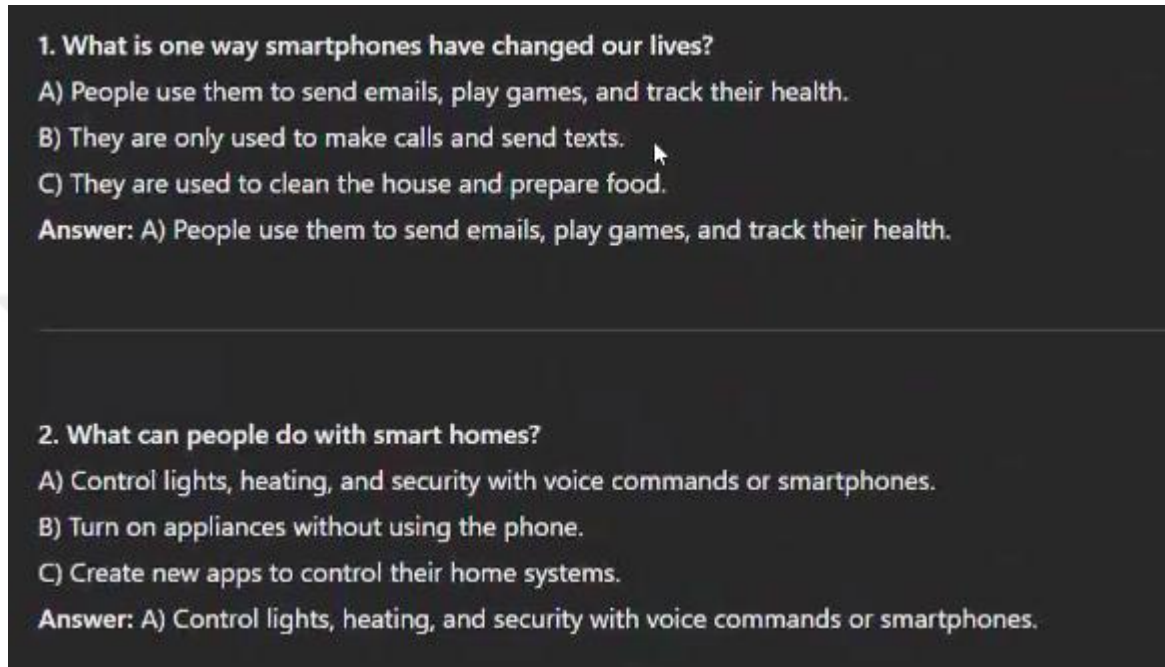
Another challenge is privacy. With so much data being collected by devices and websites, it's important to protect our personal information. As technology gets better, we need to make sure it is used responsibly and in a way that benefits everyone.

The future of innovation is exciting! As new ideas and technologies continue to develop, we can expect even more amazing changes. Who knows what new inventions will make our lives easier or more fun in the future? From healthcare to entertainment, innovation in technology will continue to shape the world in surprising ways.

She conducted a few more quality checks on the text and asked for revisions until she wrote: “now please adjust the questions accordingly.” ChatGPT gave the revised versions of the questions (see Figure 35).

Figure 35.

The Two of the Revised Questions ChatGPT Created



She continued asking ChatGPT for revision a couple of more times by writing:

“the options should be at equal length”

“They are still not at equal length some of them especially the correct answers are longer you should fix it”

“the answers should not be predictable we expect the students to comprehend and answer the questions according to what they understand from the text.”

“all the options should be included in the text”

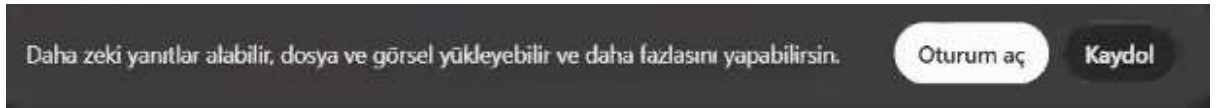
“the length of the options?”

“make one question an inference question and one with NOT”

During the recording, a notification repeatedly appeared on the screen of the conversation stating, “Sign in” or “Sign up” to enable access to more advanced features, such as obtaining smarter responses, uploading files and images, and performing additional functions (Figure 36).

Figure 36.

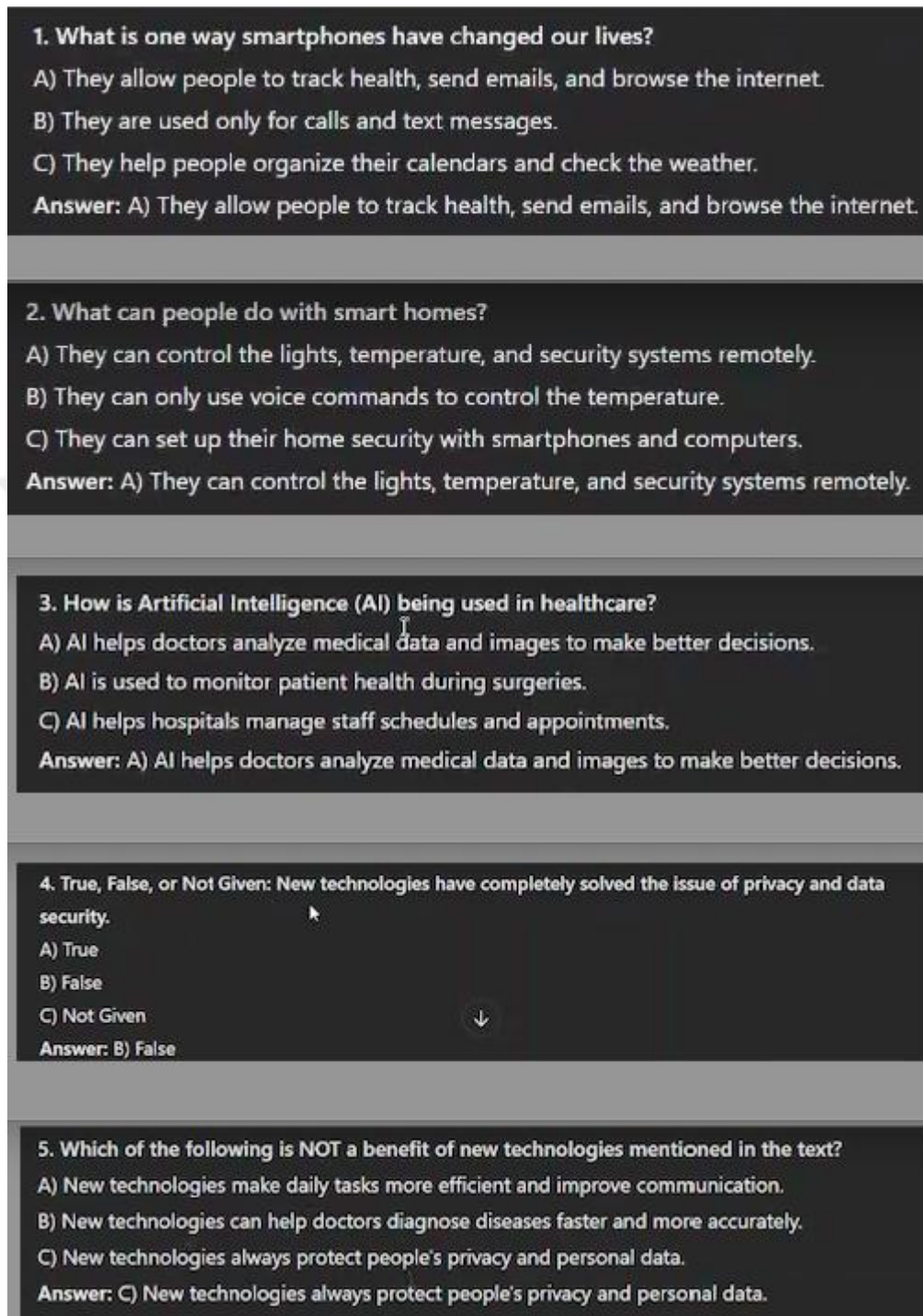
Turkish Notification on ChatGPT Screen



After the instructor wrote: “make one question an inference question and one with NOT”, ChatGPT gave 7 multiple choice questions. She asked for revision by writing: “it will be still 5 questions.” ChatGPT gave 5 questions this time and the instructor wrote: “are they in the order.” ChatGPT gave the revised version of the questions and Instructor 3 wrote: “try to ask different kind of questions one of them even can be true false not given question.” ChatGPT gave the revised version and the instructor went over the questions stating “Okay, I liked the questions.”

She copied the questions and pasted them onto the Word document (see Figure 37), and went over the questions one by one to make some changes in some questions such as rewriting the items and options.

Figure 37.
The Revised Questions ChatGPT Created



1. What is one way smartphones have changed our lives?
A) They allow people to track health, send emails, and browse the internet.
B) They are used only for calls and text messages.
C) They help people organize their calendars and check the weather.
Answer: A) They allow people to track health, send emails, and browse the internet.

2. What can people do with smart homes?
A) They can control the lights, temperature, and security systems remotely.
B) They can only use voice commands to control the temperature.
C) They can set up their home security with smartphones and computers.
Answer: A) They can control the lights, temperature, and security systems remotely.

3. How is Artificial Intelligence (AI) being used in healthcare?
A) AI helps doctors analyze medical data and images to make better decisions.
B) AI is used to monitor patient health during surgeries.
C) AI helps hospitals manage staff schedules and appointments.
Answer: A) AI helps doctors analyze medical data and images to make better decisions.

4. True, False, or Not Given: New technologies have completely solved the issue of privacy and data security.
A) True
B) False
C) Not Given
Answer: B) False

5. Which of the following is NOT a benefit of new technologies mentioned in the text?
A) New technologies make daily tasks more efficient and improve communication.
B) New technologies can help doctors diagnose diseases faster and more accurately.
C) New technologies always protect people's privacy and personal data.
Answer: C) New technologies always protect people's privacy and personal data.

The interaction between Instructor 3 and ChatGPT continued with many new prompts in which she asked for new revisions. The following prompts illustrate this process:

Instructor 3: “1. Please shorten options in question 5. 2. Please make sure all the options are at equal length”

ChatGPT revised the options

Instructor 3: “seçeneklerin uzunluğu aynı ya da yakın boylarda olmalı”

ChatGPT revised the options

Instructor 3: “we cannot use the same words in questions and options as they appear in the text please paraphrase them”

ChatGPT revised the items

Instructor 3: we cannot include words like completely fully only because it makes the answer predictable especially in true false not given questions change it accordingly please

ChatGPT revised the items

Instructor 3: “all has the same function we cannot do that”

ChatGPT revised the items

Instructor 3: “look at question 4 please get rid of all. It gives the answer”

ChatGPT revised the items

Instructor 3: “you cannot recall what I wrote you. Completely fully only all kind of words make it predictable just change the question”

ChatGPT revised the items

Instructor 3: “you did not”

ChatGPT revised the items

Instructor 3: “now we have fully are you kidding me”

ChatGPT revised the items

Instructor 3: “and now is the answer false”

ChatGPT revised the items

Instructor 3: “now you use fully”

ChatGPT revised the items

Instructor 3: “please find another question without any phrases like completely, fully, all only OK”

ChatGPT revised the items

Instructor 3: “but now the answer is problematic not it is not false. Change the question totally”

ChatGPT revised the items

Instructor 3: “it is predictable”

ChatGPT revised the items

Instructor 3: “get rid of true false question find another question from that part”

ChatGPT revised the items

Instructor 3: “go back to first question and shorten the first option”

ChatGPT revised the items

Instructor 3: “the answer of the second question is also predictable because it includes more than one benefit while the other have only one function”

ChatGPT revised the items

Instructor 3: “get rid of only”

ChatGPT revised the items

Instructor 3: “don’t you think all of them plausible”

ChatGPT revised the items

Instructor 3: “the other 2 is also correcy”

ChatGPT revised the items

Instructor 3: “change it into a cant question what cant people control and add one or two sentences into the text so that we can have a better question”

ChatGPT revised the items

Instructor 3: “shorten the first option of the third question a little bit.”

ChatGPT revised the options

Instructor 3: “shorten the first option of the 4th question”

ChatGPT revised the options

Instructor 3: “shorten the 5th question”

ChatGPT revised the items

She went over all the questions on the Word document and made a minor change in one question (question 3) manually by turning the interrogative into a positive sentence with a blank and changed the options as well as adjusting the layout and format.

She wrote ‘thanks’ to ChatGPT and ended the video.

Case 4: ChatGPT Use for Elementary-level L2 Reading Task Development

During the pre-ChatGPT use process, Instructor 4 consulted external resources twice. This included referring to the test specifications once and the coursebook once for theme selection.

Instructor 4 signed into her ChatGPT account, and during the ChatGPT use process, she did not give ChatGPT a role. She engaged in multiple actions across different subcategories. Prompt design strategies were the most frequently employed, occurring 18 times. These included providing task details (four times), asking for revisions for both text (five times) and items (four times), rationalizing (once), asking for alternatives (once), giving positive feedback (twice) and negative feedback (once). She rationalized why she gives details in her prompts by stating “I am trying to write in detail so that ChatGPT does not misunderstand me.”

The instructor also engaged in external resource use eight times, such as test specification (once), coursebook for selecting grammar structures (once), word document for word count check (once) and for editing output (five times), and a readability website for level check (once). Quality check was carried out seven times, which involved checking text (once) and items (five times) generated by ChatGPT. The editing ChatGPT-generated output

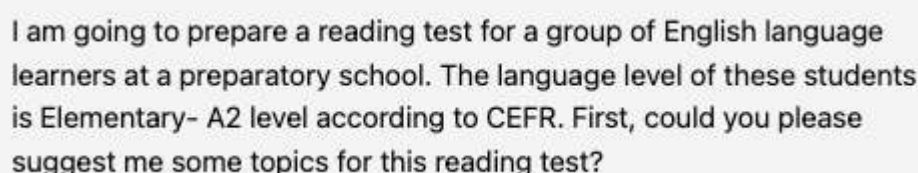
subcategory was also observed nine times, in which the instructor added content (five times), reorganized content (twice), paraphrased content (once), and rewrote items (once).

During the post-ChatGPT use process, Instructor 4 wrote three questions manually by herself and justified this by stating “ChatGPT I guess you won't be able to help me again right now, so I'm continuing so as not to break my concentration. There is nothing I can't do for A2 level anyway.”

The following illustrates Instructor 4's ChatGPT use, providing a detailed and chronological account of her prompts, interactions with external resources, and reflections.

She opened her ChatGPT account and wrote her 1st prompt as follows:

Figure 38.
Instructor 4's 1st Prompt



I am going to prepare a reading test for a group of English language learners at a preparatory school. The language level of these students is Elementary- A2 level according to CEFR. First, could you please suggest me some topics for this reading test?

ChatGPT gave 10 possible topics for the reading text as follows:

Figure 39.
Possible Topics ChatGPT Gave for the Reading Text

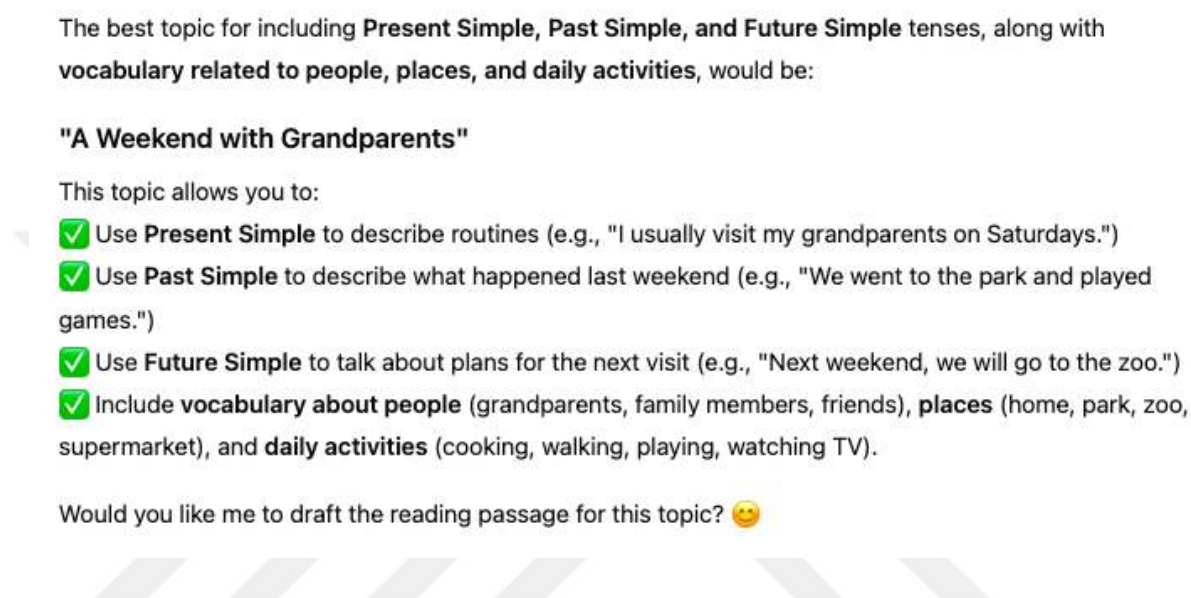
Here are some topic suggestions for your A2-level reading test:

1. **Daily Routines** – A short passage about a person's typical day.
2. **A Trip to the Zoo** – A story about a family's visit to the zoo, describing animals and activities.
3. **A Letter from a Pen Pal** – A letter introducing a new friend from another country.
4. **My Favorite Hobby** – A passage about a person's hobby and why they enjoy it.
5. **At the Supermarket** – A shopping experience, including items and prices.
6. **A Weekend with Grandparents** – A story about spending time with family.
7. **A Day at the Beach** – Describing a fun day by the sea.
8. **Healthy Eating** – A text about different types of food and their benefits.
9. **A School Trip to the Museum** – A passage about a class visiting a museum.
10. **A Famous Person** – A simple biography of a well-known historical or modern figure.

Would you like me to draft a reading passage based on one of these topics?

The instructor wanted ChatGPT to decide on the topic by writing: “Choose me the best topic in which I can include samples of present simple, past simple and future simple tenses. Also, vocabulary related to people, places, daily activities should be in the text. What would you say?” ChatGPT chose “A weekend with grandparents” (see Figure 40).

Figure 40.
ChatGPT’s Response



Instructor 4 stated “I feel like I'm not getting what I want, but I think I gave the wrong instructions.” What she wrote in her 3rd prompt was as follows:

Figure 41.
Instructor 4’s 3rd Prompt

OK, let's start from the beginning. I need a reading text with an appealing topic for the students I mentioned. The text should be at least 300 words so that I can create 5 multiple choice comprehension questions with three alternatives each to test the students' reading performance. Can you create such a text for me please?

When ChatGPT gave a text titled ‘A Weekend with Grandparents’ (see Figure 42) and the questions, the instructor copied the text onto a Word document to check the word count. It was around 220 words, which was ‘less than she wanted, but it was Okay’, as she stated.

Figure 42.
The Initial Text ChatGPT Created

Sure! Here's a **300-word reading passage** on the topic "**A Weekend with Grandparents**", designed for A2-level students.

A Weekend with My Grandparents

Every Saturday, I visit my grandparents. They live in a small house near the countryside. My grandmother loves cooking, and my grandfather enjoys telling stories about his childhood.

Last weekend, I went to their house as usual. When I arrived, my grandmother was baking a delicious apple pie. The house smelled amazing! After lunch, my grandfather and I went for a walk in the park. We saw children playing football, families having picnics, and some people walking their dogs. The weather was sunny, and the birds were singing.

Later in the afternoon, we went to the supermarket to buy some groceries. My grandmother needed vegetables, fruit, and some milk. At the checkout, we met an old friend of my grandfather. They talked about their school days and laughed a lot.

In the evening, we watched a funny movie together. My grandparents love old comedy films. We had popcorn and hot chocolate while we watched. After the movie, I was very tired, so I went to bed early.

Next weekend, I will visit them again. My grandmother says she will teach me how to bake a cake, and my grandfather wants to take me to a museum. I am really excited! Spending time with my grandparents is always special because they make me feel happy and loved.

This passage includes **Present Simple (daily routine)**, **Past Simple (last weekend)**, and **Future Simple (next weekend's plans)** while covering vocabulary related to **people (grandparents, friend)**, **places (park, supermarket, museum)**, and **activities (cooking, walking, watching movies)**.

Would you like me to create **5 multiple-choice comprehension questions** based on this passage? 😊



The instructor wanted to consult an external resource, a readability website, to check the level of the text. However, because of a technical issue, she could not use the tool and stated “Ok, I will check the text. I trust my intuition for the level check. In general, the text seems appropriate for the level. I can make small changes in some vocabulary and grammar structures.” She started making small changes on the Word document such as adding and paraphrasing content. She also checked the contents page from time to time, and had a look at the grammar structures covered so far.

The instructor copied and pasted the text she adapted and wrote:

Figure 43.

Instructor 4's 4th Prompt

This is how I adapted this text: Every Saturday, I visit my grandparents. They live in a small house near the countryside. My grandmother loves cooking, and my grandfather enjoys telling stories about his childhood.

Last weekend, I went to their house as usual. When I arrived, my grandmother was baking a delicious apple pie. The house smelled amazing! After lunch, my grandfather and I went for a walk in the park. We saw a lot of happy people outside. Some children played football, some families had picnics, and some people walked their dogs. The weather was sunny, and the birds were happy, too.

Later in the afternoon, we went to the supermarket and bought some groceries. My grandmother needed vegetables, fruit, and some milk. At the checkout, we met an old friend of my grandfather. They talked about their school days and laughed a lot.

In the evening, we watched a funny movie together. My grandparents love old comedy films. We had popcorn and hot chocolate while we watched. After the movie, I was very tired, so I went to bed early.

Next weekend, I will visit them again. My grandmother says she will teach me how to bake a cake, and my grandfather wants to take me to a museum. I am excited! I always love spending time with my grandparents because they make me feel happy and loved.

Can you now create 5 multiple choice comprehension questions with three alternatives for each. One alternative is the correct answer, while the other two alternatives are distractors. The length of each alternative should be similar and please avoid not in them as well as the questions.

ChatGPT gave 5 multiple choice questions with their correct answers, as follows:

Figure 44.

The Questions ChatGPT Created.

Here are **five multiple-choice comprehension questions** based on your adapted text. Each question has **three answer choices** of similar length:

1. Where do the grandparents live?

- A) In a small house near the countryside. 
- B) In a big apartment in the city.
- C) In a tall building near the park.

2. What was the grandmother doing when the writer arrived?

- A) She was baking an apple pie. 
- B) She was watching a comedy film.
- C) She was walking in the park.

3. Who did they meet at the supermarket?

- A) A friend from school.
- B) An old friend of the grandfather. 
- C) A neighbor from next door.

4. What did the family do in the evening?

- A) They watched a funny movie and had snacks. 
- B) They went to the museum and looked at paintings.
- C) They played football in the park with the children.

5. What does the writer plan to do next weekend?

- A) Bake a cake and go to a museum. 
- B) Play football and have a picnic.
- C) Walk in the park and buy groceries.

Would you like any modifications? 😊



She started going over the questions and wrote: “For the first question, could you add some details to the text to make the question a bit more challenging? This way it is too easy for an MC question.” ChatGPT gave the modified version of the question:

Figure 45.
The Revised Questions ChatGPT Created

Good idea! Here's a slight modification to the text to make the first question more challenging:

Modified Sentence in the Text:

"They live in a small house near the countryside, surrounded by tall trees and colorful flowers. There is also a small lake nearby where my grandfather likes to fish."

Updated Multiple-Choice Question:

1. Where do the grandparents live?

- A) In a small house near the countryside, surrounded by trees and flowers. 
- B) In a small house near the countryside, next to a shopping center.
- C) In a small house near the countryside, close to a train station.

This way, students need to pay closer attention to the details in the text rather than just recognizing a single phrase. Let me know if you'd like any other tweaks! 😊

She checked the 1st question along with the text, made some additions to that part of the text manually, and wrote: "Very good but I need one more sentence about the 'grandmother' so that I can have a better distractor." ChatGPT gave the wanted sentence:

Figure 46.
ChatGPT's Response

Great thinking! Here's the **final modified version** of the text with an added sentence about the grandmother:

Modified Sentence in the Text:

"They live in a small house near the countryside, surrounded by tall trees and colorful flowers. There is also a small lake nearby where my grandfather likes to fish. My grandmother enjoys growing vegetables and flowers in the garden behind the house."

Updated Multiple-Choice Question:

1. Where do the grandparents live?

- A) In a small house near the countryside, surrounded by trees and flowers. 
- B) In a small house near the countryside, next to a shopping center.
- C) In a small house near the countryside, with a big garden full of vegetables.

Now, both distractors sound reasonable, making the question a bit more challenging. Let me know if you'd like any more refinements! 😊

The instructor copied the sentence to the related part in the text and she made some more changes in the text manually by adding some phrases and sentences into the relevant

paragraph. She wrote “instead of nearby, can you suggest an easier word?” ChatGPT gave two words: “close & near.”

She asked for revision by writing: “OK, adapt this paragraph and the question about this paragraph in a way that I have a good comprehension question whose answer is not very obvious because of the distractors. I meana make the distractors related to the information in the paragraph but wrong.” ChatGPT responded:

Figure 47.
ChatGPT’s Response

Great challenge! Here’s the **adapted paragraph** with subtle details to make the question trickier:

Revised Paragraph:

"They live in a small house near the countryside, surrounded by tall trees and colorful flowers. There is also a small lake close to the house where my grandfather likes to fish in the mornings. My grandmother enjoys growing vegetables and flowers in the garden behind the house. She often wakes up early to water her plants before making breakfast."

Updated Multiple-Choice Question:

1. Where do the grandparents live and what do they like doing?

- A) In a small house near the countryside; the grandfather fishes in the mornings, and the grandmother grows vegetables and flowers. 
- B) In a small house near the countryside; the grandfather waters the plants, and the grandmother cooks lunch in the garden.
- C) In a small house near the countryside; the grandfather fishes in the afternoons, and the grandmother grows fruit trees.

Now, the distractors contain **details from the text that are slightly changed**, making the question more challenging while still requiring close reading. Let me know if you need further refinements! 😊

Instructor 4 looked at the revised question and wrote: “No, ask only one thing in the question.” ChatGPT revised the question:

Figure 48.
ChatGPT's Response

Got it! Here's the **adapted version** with a single focus in the question while keeping the distractors closely related but incorrect:

Updated Multiple-Choice Question:

1. What does the grandfather like to do in the mornings?

- A) He likes to fish in the lake. 
- B) He likes to water the plants in the garden.
- C) He likes to cook breakfast with the grandmother.

Now, the distractors are logical but incorrect based on the text, making the question more challenging. Let me know if you'd like any further tweaks! 😊

The instructor went over the question and stated that she liked it. She copied the adapted paragraph and the question onto a Word document. She started making some more changes in that paragraph by changing the place of some sentences and adding some words into the sentences manually. She also shortened the distractors in the 1st question manually as well as playing with the format and layout. She then started adapting the 2nd paragraph of the text manually on the Word document by changing the place of some phrases and adding some words in the sentences. She stated that she did not like the 2nd question ChatGPT created and wrote: "can you do the same for the second paragraph and the second question about that paragraph?" ChatGPT responded:

Figure 49.
ChatGPT's Response

Absolutely! Here's the **adapted second paragraph** with subtle details to make the comprehension question more challenging:

Revised Second Paragraph:

"Last weekend, I went to their house as usual. When I arrived, my grandmother was baking a delicious apple pie in the kitchen. The house smelled amazing! After lunch, my grandfather and I went for a walk in the park. We saw a lot of happy people outside. Some children played football near the big fountain, some families had picnics under the trees, and some people walked their dogs along the river. The weather was sunny, and the birds were happy, too."

Updated Multiple-Choice Question:

2. Where were the children playing football?

- A) Near the big fountain. 
- B) Under the trees with the families.
- C) Along the river with the people walking their dogs.

Now, the **distractors are close to the correct answer** but slightly incorrect, requiring students to pay close attention. Would you like any further refinements? 

The instructor went over the revised item and copied it to the Word document by stating "Hmm, good, ok, it can be used for elementary level, it is quite suitable, I think." She copied it to the Word document and made some changes (see Figure 50):

Figure 50.
The Question the Instructor Adapted

2. Where did the children play football?

- A) Near the big fountain.
- B) Under the tall trees.
- C) Along the river.

Right after ChatGPT gave the revised version of the 2nd question, there appeared a Turkish notification (see Figure 51) that can be translated into English as "You have reached the limit for the Free plan for GPT-4o. Because there is an addition in the conversation, GPT-4o is required to continue this conversation. Your limit will reset as of 15:38."

Figure 51.

Turkish Notification on ChatGPT Screen



As the instructor did not realize there was a notification, she started to write her next prompt “Great job! Can you please the same with the third paragraph and the third questions related to it?” At that moment, she noticed that she reached the limit for the free plan, she clicked ‘New Chat’ to write her prompt. She wrote: “Hello, I am preparing a reading test for a group of students learning English at a preparatory school. Their level of English is Elementary – A2 according to CEFR. The text is ... (she copied the whole text here in the prompt). I am trying to write multiple choice questions for each paragraph. Now, I need to write the 3rd question for the 3rd paragraph. Can you adapt the paragraph so that for the question, I can have one correct answers but two good distractors which are related to the paragraph but wrong?” However, ChatGPT gave a similar warning (see Figure 52), and she could not continue the chat.

Figure 52.

Turkish Notification on ChatGPT Screen



Instructor 4 decided to continue developing the questions without ChatGPT by stating “ChatGPT I guess you won't be able to help me again right now, so I'm continuing so as not to break my concentration. There is nothing I can't do for A2 level anyway.” She started writing the 3rd, 4th and 5th questions manually, and made some changes in the relevant paragraphs (see Figure 53).

Figure 53.

The Final Draft of the Reading Task Instructor 4 Prepared

Every Saturday, I visit my grandparents. They live in a small house near the countryside. There are tall trees and colourful flowers all around the house. There is also a small lake close to the house. My grandfather like fishing there every morning. My grandmother enjoys growing vegetables and flowers in the garden behind the house. She often wakes up early to water her plants before she makes the breakfast.

Last weekend, I went to their house as usual. When I arrived, my grandmother was baking a delicious apple pie in the kitchen. The house smelled amazing! After lunch, my grandfather and I went for a walk in the park. We saw a lot of happy people outside. Some children played football near the big fountain, some families had picnics under the trees, and some people walked their dogs along the river. The weather was sunny, and the birds were happy, too.

Later in the afternoon, we went to the supermarket and bought some groceries. My grandmother needed vegetables, fruit, and some milk. At the checkout, we met an old friend of my grandfather. They talked about their school days and laughed a lot.

In the evening, we watched a funny movie together. My grandparents love old comedy films. We had popcorn and hot chocolate while we watched. After the movie, I was very tired, but I didn't want to go to bed early. I read for half an hour and then slept.

Next weekend, I will visit them again. My grandmother says she will teach me how to bake a cake, and my grandfather wants to take me to a museum. But I want to make a surprise for them and take them to a picnic. I am very excited! I always love spending time with my grandparents because they make me feel happy and loved.

1. What does the grandfather like to do in the mornings?

- a. He goes fishing in the lake.
- b. He waters the plants in the garden.
- c. He cooks breakfast with his wife.

2. Where did the children play football?

- a. Near the big fountain.
- b. Under the tall trees.
- c. Along the river.

3. Who did they meet at the supermarket?

- a. the writer's schoolteacher
- b. the grandmother's neighbour
- c. the grandfather's old friend

4. What did the writer do after the movie?

- a. He went to bed.
- b. He ate some snacks.
- c. He read a book.

5. What is the writer's plan for the next visit?

- a. cooking with the grandmother
- b. having a picnic with the grandparents
- c. taking the grandfather to a museum

Case 4: ChatGPT Use for Intermediate-level L2 Reading Task Development

Before starting to experiment with ChatGPT, Instructor 4 consulted external resources (twice), which included referring to the test specifications (once) and the coursebook (once) for theme selection.

Instructor 4 signed into her ChatGPT account and she started to use prompt design strategies, which were employed 16 times. These strategies were providing task details (once), repeating (once), giving positive feedback (twice) and negative feedback (once), asking for revisions for text (once), items (three times) and options (five times), asking for self-evaluation (once) and rationalizing (twice). In terms of providing context to ChatGPT, she did not give ChatGPT a role.

During the ChatGPT use process, the instructor also benefitted from external resources (five times). These included a Word document to check the number of words in the given text (twice) and to edit the output (twice), and a readability website to check the level of the given text (once). Instructor 4 also conducted quality checks seven times, which involved checking text (three times), items (twice) and options (twice). The instructor was also engaged in editing ChatGPT-generated output six times, in which she rewrote items (three times) and rewrote options (twice).

No post-ChatGPT use process was observed.

The section below outlines Instructor 4's use of ChatGPT, offering a step-by-step narrative of her prompts, engagement with external sources, and personal insights.

Instructor 4 signed into her ChatGPT account and wrote her 1st prompt as follow:

Figure 54.

Instructor 4's 1st Prompt

I am going to prepare a reading test for a group of students learning english at a preparatory school. Their level of English is B1+ Intermediate according to CEFR. The reading text should have a topic appealing for these students. The text should be academic. The length of the text should be minimum 500 words. Also, for this text I need 5 multiple choice comprehension questions. All the questions should have three options, one of which is the correct answer and the others are distractors. The length of all the options in one question should be similar. All the option should be related to the text itself, not out of the context. So, could you please create a reading test considering all these for me?

ChatGPT gave a reading text titled 'The Impact of Technology on Education' (see Figure 55) along with 5 multiple choice questions (see Figure 56).

Figure 55.

The Initial Text ChatGPT Created

Reading Passage: The Impact of Technology on Education

Technology has improved many aspects of modern life, and education is no exception. Over the past few decades, technological advancements have significantly changed how students learn and how teachers instruct. From digital textbooks to virtual classrooms, the impact of technology on education is profound and continues to evolve.

One of the most notable changes is the shift from traditional paper-based learning to digital resources. E-books, online articles, and educational websites provide students with instant access to vast amounts of information. This accessibility allows students to learn at their own pace, revisit materials as needed, and engage with multimedia content such as videos and interactive exercises. Additionally, online platforms offer courses and tutorials that enable learners to acquire new skills without being physically present in a classroom.

Another major transformation is the use of artificial intelligence (AI) in education. AI-driven programs can assess students' performance, provide personalized feedback, and suggest learning paths based on individual progress. For example, adaptive learning software analyzes how students respond to questions and adjusts the difficulty level accordingly. This personalized approach helps students improve their weak areas while reinforcing their strengths.

Furthermore, technology facilitates communication between teachers and students. In the past, students had to wait for scheduled class times or office hours to ask questions. Now, email, messaging apps, and online discussion forums allow for instant communication. This increased interaction enhances the learning experience and ensures that students receive the support they need.

However, despite these advantages, the integration of technology in education also poses challenges. One concern is the digital divide—the gap between individuals who have access to modern technology and those who do not. In some regions, students may lack reliable internet connections or necessary devices, making it difficult for them to benefit from online learning resources. Additionally, excessive screen time and reliance on digital tools can reduce face-to-face social interactions and critical thinking skills.

In conclusion, technology has greatly influenced education by making learning more accessible, personalized, and interactive. While challenges remain, educators and policymakers must work together to ensure that technological advancements benefit all students. As technology continues to evolve, so too will the ways in which knowledge is acquired and shared in educational settings.

Figure 56.
The Initial Questions ChatGPT Created

Comprehension Questions:

1. What is one way technology has changed education? a) It has eliminated the need for teachers in classrooms. b) It has made learning materials more accessible online. c) It has replaced traditional exams with only digital assessments.
2. How does artificial intelligence (AI) help students in education? a) It replaces teachers by providing direct instruction. b) It personalizes learning by adapting to students' progress. c) It limits students' access to different learning materials.
3. What is one challenge of integrating technology in education? a) The digital divide prevents equal access to learning resources. b) Online learning is less engaging than traditional classrooms. c) Technology completely replaces human interaction in education.
4. Why is communication between students and teachers easier now? a) Online platforms provide instant messaging and discussion forums. b) Students do not need to communicate with teachers anymore. c) Teachers have fewer students to interact with due to online learning.
5. What is the overall conclusion of the passage? a) Technology has had a completely negative impact on education. b) While technology improves education, it also presents some challenges. c) Schools should eliminate all traditional learning methods in favor of digital tools.

The instructor said “I like the title and topic” and she copied the text onto a Word document to check the word count. It was around 370. She stated “The word count is less than I expected. I will examine the text and the questions first, then deal with the word count. While preparing the Elementary level exam, I had a problem with quota, so now I do not want to have such a problem in the middle. I would like to prepare for the whole test in one sitting. When there is a quota problem, it distracts my concentration.”

She copied the text onto a readability website to check its appropriateness for the level. She could not use the website for free, so she quickly went over the text stating “I cannot use the website, so I can trust my intuition to decide on the readability by going over the text in more detail.” Then, she stated “Upon reading the rest of the text, I will ask ChatGPT to make

this text easier and longer.” and wrote her 2nd prompt: “Thank you but I said that I needed a text of at least 500 words. So could you please make this text longer and also could you please make the language a bit easier for my students to understand?”

ChatGPT presented the revised text (see Figure 57) along with the revised questions (see Figure 58).

Figure 57.
The Revised Text ChatGPT Created

Reading Passage: The Impact of Technology on Education

Technology has changed many parts of life, including education. Over the past few decades, new technology has transformed how students learn and how teachers teach. From digital textbooks to virtual classrooms, the role of technology in education continues to grow and evolve.

One major change is the shift from traditional paper-based learning to digital resources. Students now have access to e-books, online articles, and educational websites. These materials provide instant information, allowing students to study at their own speed. Many digital resources include videos, quizzes, and interactive exercises that help students understand topics better. Online platforms also offer courses and tutorials, so learners can gain new skills from anywhere in the world.

Another important development is the use of artificial intelligence (AI) in education. AI-driven programs can assess students' performance, give personalized feedback, and recommend learning activities. For example, adaptive learning software adjusts the level of difficulty based on how well a student is performing. If a student struggles with a topic, the software provides extra practice. If a student excels, it introduces more challenging content. This personalized approach helps students strengthen their weak areas while building on their strengths.

Technology has also improved communication between teachers and students. In the past, students had to wait for class or office hours to ask questions. Now, they can send emails, use messaging apps, or participate in online discussion forums. This allows students to get help quickly and stay connected with their teachers and classmates.

Despite these benefits, there are some challenges to using technology in education. One major issue is the digital divide—the gap between students who have access to technology and those who do not. Some students do not have reliable internet connections or devices at home, making it harder for them to participate in online learning. Schools and governments must find ways to provide equal opportunities for all students.

Another challenge is that too much screen time can affect students' health and social skills. Spending long hours in front of a screen can lead to eye strain and reduced physical activity. Additionally, relying too much on digital tools might reduce face-to-face interactions and critical thinking skills. Teachers need to balance digital learning with traditional teaching methods to create the best learning environment.

In conclusion, technology has changed education in many ways. It makes learning more accessible, flexible, and interactive. However, it also presents challenges that educators and policymakers must address. By finding solutions to these challenges, schools can ensure that all students benefit from technological advancements. As technology continues to develop, the way students learn will keep evolving, offering new opportunities for education.



Figure 58.
The Revised Questions ChatGPT Created

Comprehension Questions:

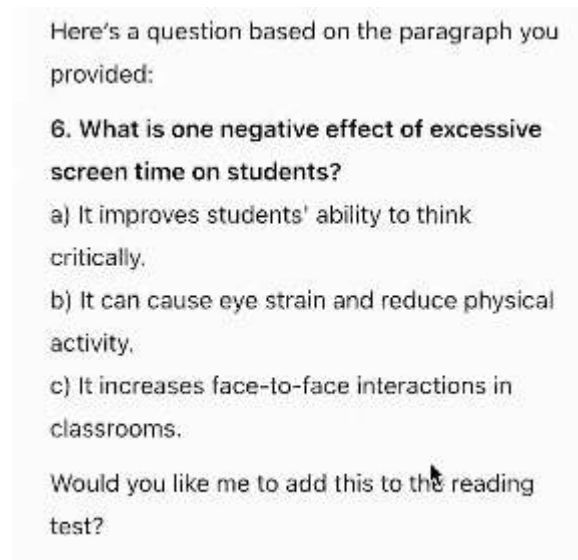
1. What is one way technology has changed education? a) It has eliminated the need for teachers in classrooms. b) It has made learning materials more accessible online. c) It has replaced traditional exams with only digital assessments.
2. How does artificial intelligence (AI) help students in education? a) It replaces teachers by providing direct instruction. b) It personalizes learning by adapting to students' progress. c) It limits students' access to different learning materials.
3. What is one challenge of integrating technology in education? a) The digital divide prevents equal access to learning resources. b) Online learning is less engaging than traditional classrooms. c) Technology completely replaces human interaction in education.
4. Why is communication between students and teachers easier now? a) Online platforms provide instant messaging and discussion forums. b) Students do not need to communicate with teachers anymore. c) Teachers have fewer students to interact with due to online learning.
5. What is the overall conclusion of the passage? a) Technology has had a completely negative impact on education. b) While technology improves education, it also presents some challenges. c) Schools should eliminate all traditional learning methods in favor of digital tools.

She copied both the text and the questions this time onto the Word document first to check the word count, which was 435. She stated “The word count is within the expected limits.” and “The revised version seems a bit better than the 1st one that ChatGPT has given.” She started checking the questions one by one. She started with the 1st question by making a change in the stem by adding the phrase ‘According to paragraph 2. She was satisfied with the adapted question as she said “With this small change, I really like the question 1. It is quite good for the first draft. The length of the options is also very good this time. And of course, I think I am able to give better instructions than the experience with the previous one.”

She continued revising the other questions. She kept the 2nd question as it was, but made some vocabulary changes in the options on her own. regarding the 3rd question, she added the phrase to the stem ‘According to paragraph 5,’ and showed her satisfaction by stating “You created a very good question dear ChatGPT. I like this one.” She continued with the 4th

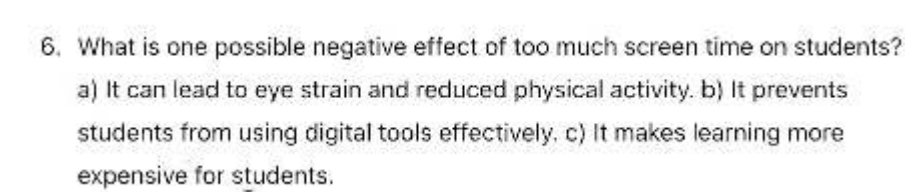
question. She started rewriting the question, then decided to make ChatGPT rewrite the 4th question. She wrote: “Can you write a question for this paragraph? (She copies the related paragraph into the prompt).” ChatGPT responded as follows:

Figure 59.
ChatGPT’s Response



The instructor stated “The question is asking for negative effects, but two options have positive meanings by words ‘improve and increase, etc.’ When the question is asking for negative things, the options cannot have positive meanings.” Therefore, she wrote a new prompt: “If the question is about negative effects, the distractors should also be negative things. Otherwise, it is very easy to figure out the correct answer. Can you adapt the questions accordingly please?” ChatGPT’s response was as follows:

Figure 60.
ChatGPT’s Response



She stated “Ok, I like the new question and its options. There is only a need for a small paraphrase in the related option. I would love to make ChatGPT to paraphrase, but I am afraid of running out of the quota.so I will make it by myself. Actually, I can try it, it is only one sentence.” And she wrote her new prompt: “Can you paraphrase this option (a) in a way that is

not very easy to say it is the correct answer for question 6.” ChatGPT gave the paraphrased version and she used it by deleting some words.

The instructor continued with the last question. She went over the stem and the options in the question. She wrote:

Figure 61.

Instructor 4’s Prompt

For the last paragraph, I want to have a vocabulary question. What is the best word to replace address in the last paragraph? This is the question. Now, I need one correct answer and two options in which I have two other verbs collocating challenge but wrong. Can you create such a question?

ChatGPT created the expected question with the options. She wrote: ‘I am not sure if these are not correct. Can you provide more irrelevant verbs in terms of meaning?’ ChatGPT gave the revised version of the options. ‘solve, ignore, increase’.

She went over the options. She kept ‘solve and ignore’ and changed ‘increase’ into ‘identify’. She wrote: ‘Can you paraphrase solve and solutions with other words?’ ChatGPT gave the revised words. She said ‘No!... because ‘finding solutions’ is already in the text.’” and wrote: ‘Instead of finding solutions’, what else can we say?’ ChatGPT gave the paraphrased version ‘developing strategies to overcome’. She said: “Yes, it is wonderful.” She ended the video by saying that the first draft was now ready.

4.3 Differences in EFL Instructors' Use of ChatGPT for Elementary-level and Intermediate-level L2 Reading Task Development

Across the four cases, instructors engaged in similar overarching processes for both elementary- and intermediate-level L2 reading test development; however, there were differences in the frequency and distribution of codes across categories, subcategories, and specific actions.

Pre-ChatGPT Use Process:

In both levels, the External Resource Use subcategory was present in most cases, though it was absent in Case 3 for both levels and in Case 2 at the intermediate level. At the elementary level, the highest pre-use external resource reference was observed in Case 1 (five times: test specifications, coursebook, wordlist), whereas at the intermediate level, the maximum was two instances (Cases 1 and 4). The types of resources were consistent across levels—test specifications, coursebook, and wordlist—though the elementary level showed more frequent and varied use before ChatGPT was accessed.

ChatGPT Use Process:

For both levels, Prompt Design Strategies was the most frequently employed subcategory. At the elementary level, the highest occurrence was in Case 3 with 49 instances, predominantly focused on asking for revision of items (18 times), asking for revision of options (seven times), and repeating (six times). At the intermediate level, the maximum was observed in Case 3 with 55 instances, again dominated by asking for revision of items (17 times) and asking for revision of options (nine times), along with giving negative feedback (nine times).

While both levels showed a concentration on item- and option-related revisions, intermediate-level development generally included a greater variety of prompt types within the same case, such as higher instances of negative feedback (e.g., nine times in Case 3 intermediate vs. two times in Case 3 elementary). Additionally, providing context—assigning a role to ChatGPT—was more common at the intermediate level (e.g., twice in Case 2) than at the elementary level, where it appeared only in Case 1.

External Resource Use during ChatGPT engagement was frequent in both levels, though elementary-level cases sometimes included resources like a readability website (Case 4), whereas at the intermediate level, such use was less common (only in Case 4). Word documents were widely used at both levels for word count checks and output editing.

Quality Check activities were conducted at both levels, with item checks being the most frequent within this category. For example, in elementary Case 3, items were checked eight

times, while in intermediate Case 3, they were checked nine times. Text checks were also common but less frequent than item checks.

Editing ChatGPT-Generated Output occurred in all cases at both levels, with common edits including rewriting items and options, reorganizing text, adding or deleting content, and paraphrasing. At the elementary level, the highest number of edits was nine (Case 4), compared to a maximum of six in intermediate Case 4. While both levels saw items and options as the primary focus of edits, elementary-level cases also included more frequent text reorganizations and content additions.

Post-ChatGPT Use Process:

Post-use activity was observed in fewer cases. At the elementary level, Case 4 included manual item writing (three times) after ChatGPT use, while at the intermediate level, post-use activity appeared in Case 1 (format and layout adjustment in Word). This suggests a slightly greater tendency at the elementary level to make substantive content contributions after ChatGPT use.

4.4 EFL Instructors' Perceptions of Using ChatGPT for L2 Reading Test Development

All four instructors and eight video recordings of both levels, a total of 140 perceptions were recorded regarding the use of ChatGPT-4 for L2 reading test development. Among these, the most frequently mentioned aspects were item generation (42 mentions, 30%) and following instructions (31 mentions, 22.1%), which together accounted for over half of the perceptions. Other frequently noted aspects included option generation (15 mentions, 10.7%), word count (14 mentions, 10%), and text generation (13 mentions, 9.3%), while less frequently reported perceptions concerned areas such as theme selection, title generation, text organization, level appropriateness, grammar accuracy, usage limitations, response speed, and file upload.

When the perceptions were categorized according to their orientation, 50 were positive (35.7%) and 90 were negative (64.3%). Positive perceptions most commonly emphasized item generation (22 times, 44.0% of positive comments), followed by text generation (11 times,

22.0%), option generation (6 times, 12.0%), and word count (5 times, 10.0%). Conversely, negative perceptions primarily focused on following instructions (27 times, 30.0% of negative comments) and item generation (20 times, 22.2%), followed by option generation (9 times, 10.0%), word count (9 times, 10.0%), and text generation (2 times, 2.2%). Additional negative feedback highlighted issues with level appropriateness (6 times, 6.7%), usage limitations (5 times, 5.6%), response speed (3 times, 3.3%), as well as isolated concerns with title generation, grammar accuracy, text organization, and file upload (each 1.1%).

4.4.1 Case 1: Perceptions of Using ChatGPT for Reading Test Development

In Case 1, nine positive perceptions were noted, most frequently related to text generation (three times) and item generation (three times), while 24 negative perceptions were expressed, predominantly concerning item generation (six times) and following instructions (four times).

Case 1: Perceptions of Using ChatGPT for Elementary-level L2 Reading Task Development

During his ChatGPT use for developing the elementary-level L2 reading test, Instructor 1 made ten comments that revealed his perceptions regarding different aspects of ChatGPT, which were mostly negative comments (seven out of ten comments). The three positive comments were related to the text that ChatGPT generated and the appropriateness of its level. The negative comments of the instructor referred to text generation, text organization, word count, grammatical accuracy, and appropriateness of the level of the text, items, and vocabulary used by ChatGPT. For example, regarding the appropriateness of the level of the ChatGPT-generated chat, Instructor 1 stated: “The text seems suitable for the elementary level in general.” The following comments from the instructor illustrate their negative perceptions:

“Why do I use the word document from the beginning? I don’t trust the word count by ChatGPT. It may not be correct. For example, it says that it has prepared a text of 300 words, but it happens to be 200 or 450 words. So I check the word count on the Word document.” (Word count)

“ChatGPT placed the details I wanted at the end of the text instead of interspersing them.” (Text organization)

“There are some high-level words such as ‘rewarding’” (Level appropriateness of the vocabulary)

Case 1: Perceptions of Using ChatGPT for Intermediate-level L2 Reading Task Development

During his ChatGPT use for developing the intermediate-level L2 reading test, Instructor 1 made 23 comments that revealed his perceptions regarding different aspects of ChatGPT, which were mostly negative comments (17 out of 23 comments). The six positive comments were related to the text that ChatGPT generated (once), item generation (three times), following instructions (once) and word count (once). For example, regarding the item generation, Instructor 1 stated: “The first question is quite good,” and “Now, the question seems better. At least it used different words.” The negative comments of the instructor were on following instructions, item generation, option generation, title generation, level appropriateness of vocabulary and response speed. The following comments from the instructor illustrate his negative perceptions:

“When it comes to Intermediate, ChatGPT can prepare a slightly higher level of vocabulary in the text.” (Level appropriateness: vocabulary)

“ChatGPT is a bit slow at generating what I want.” (Response speed)

4.4.2 Case 2: Perceptions of Using ChatGPT for Reading Test Development

In Case 2, there were eight positive perceptions, such as text generation (twice) and word count (twice), alongside 12 negative perceptions, with item generation (six times) being the most frequent issue.

Case 2: Perceptions of Using ChatGPT for Elementary-level L2 Reading Task Development

Instructor 2 made ten comments that uncovered his perceptions of the use of ChatGPT for elementary-level L2 reading test development. Whereas the five positive comments were related to theme selection, self-evaluation, word count (twice) and following instructions, the other five negative comments were on following instructions (twice), word count, usage limitation and item generation. For example, regarding the negative side of ChatGPT's following instructions, he stated: "Because it gets a bit too complicated and it may forget what it has done, ...", and regarding the word count function of ChatGPT, he stated "Sometimes ChatGPT has a problem with word count." On the other hand, he made two other positive comments on the word count as follows:

"Although I wanted a text around 280 words, it gives me a text around 260 words. It is OK. It sometimes does this. It is what it is. Not a big problem."

"It is not that bad (word count) in terms of test specifications, but I want a bit longer text."

Regarding theme selection, he also made a positive comment: "For elementary group, it is not a big deal. When you give ChatGPT the level, it gives you a lot of suitable themes as well, most probably."

Case 2: Perceptions of Using ChatGPT for Intermediate-level L2 Reading Task Development

Instructor 2 made a total of ten comments which revealed his perceptions regarding the use of ChatGPT for intermediate-level L2 reading test development. Three out of ten comments were positive, which were related to text generation (twice) and item generation. For instance, regarding the text generation, he stated: "... But of course, it gives you a good starting point, which is the important point." His seven negative comments were on file uploading, item generation (five times) and option generation. The following comments from the instructor illustrate his negative perceptions:

"Sometimes we have this problem when you give a file that is a bit large in file in size and you can't upload the others" (file upload)

“I can see question number 2 has a problem because this question can be answered without reading the text. This is problematic. And another problem with the same question is it has actually 2 possible correct answers.” (item generation)

“The 1st question is also problematic here because the test takers already know what Thomas Edison is famous for. We should come up with a solution to this problem, we should maybe change it, replace it, or change the question all together.” (item generation)

4.4.3 Case 3: Perceptions of Using ChatGPT for Reading Test Development

Case 3 produced the highest number of perceptions, with 18 positive, most often about item generation (nine times) and text generation (five times), and 41 negative, mainly concerning following instructions (21 times) and item generation (nine times).

Case 3: Perceptions of Using ChatGPT for Elementary-level L2 Reading Task Development

During her ChatGPT use for developing the elementary-level L2 reading test, Instructor 3 made 27 comments that revealed her perceptions of different aspects of ChatGPT, 18 of which were negative comments. There were nine positive comments made on aspects such as text generation (three times), item generation (five times) and following instructions. For example, regarding ChatGPT’s following instructions, she positively stated: “We changed the questions completely at the last minute. Since ChatGPT remembered all these specifications, it was relatively easy this time. ”

The seven out of 18 negative comments from the instructor were related to following instructions while the rest were about word count, item generation, option generation and text generation. The following comments from the instructor illustrate her negative perceptions:

“... but sometimes ChatGPT does not understand the instructions.” (following instructions)

“Same instructions again. I have given those many times.” (following instructions)

Case 3: Perceptions of Using ChatGPT for Intermediate-level L2 Reading Task Development

During her ChatGPT use for developing the intermediate-level L2 reading test, Instructor 3 made 32 comments that revealed her perceptions of different aspects of ChatGPT, which were mostly negative comments (23 out of 32 comments). The nine positive comments were related to text generation, item generation, word count, option generation whereas the rest of the comments were negative on word count, text generation, title generation, following instructions and option generation. 13 out 23 negative comments were made regarding ChatGPT's following instructions. The following comments from the instructor illustrate her negative perceptions:

"In general, ChatGPT can't get the word count right." (word count)

"he should remember everything we corresponded." (following instructions)

4.4.4 Case 4: Perceptions of Using ChatGPT for Reading Test Development

In Case 4, 16 positive perceptions were reported, with item generation (six times) and option generation (four times) standing out, while 12 negative perceptions were identified, most frequently related to usage limitations (four times) and item generation (two times).

Case 4: Perceptions of Using ChatGPT for Elementary-level L2 Reading Task Development

Instructor 4 made ten comments that revealed her perceptions of the use of ChatGPT for elementary-level L2 reading test development. Whereas five of her ten comments were the negative ones, which were related to following instructions, word count, item generation and usage limitation (twice), the other five comments were positive. These included word count, level appropriateness of vocabulary, level appropriateness of the text generated, and item generation (twice). The following comments illustrate Instructor 4's positive perceptions:

"Good, ok, it can be used for elementary level, it is quite suitable, I think." (item generation)

“When I have a quick look at the text, it seems that there aren't any advanced level words.” (level appropriateness: vocabulary)

Case 4: Perceptions of Using ChatGPT for Intermediate-level L2 Reading Task Development

During her ChatGPT use for developing the intermediate-level L2 reading test, Instructor 4 made 18 comments that showed her perceptions of different aspects of ChatGPT, 10 of which were positive comments. The eight negative comments were regarding word count, usage limitation (twice), item generation (twice), option generation (twice) and response speed. For example, regarding option generation, she stated: “We would never have such options, with extreme words like ‘completely, all’. We need to change them.” The positive comments from the instructor included theme selection, word count, text generation, item generation (three times), and option generation (four times). The following comments illustrate Instructor 4’s positive perceptions:

“The length of the options is also very good this time.” (option generation)

“I like the new questions and its options” (item and option generation)

4.5 Differences in the Perceptions of the Instructors When Developing Elementary-level and Intermediate-level Reading Tasks

Instructor 1 made more negative comments about his ChatGPT use experience when working on the intermediate-level reading task than when he was working on the elementary-level reading task. In the elementary-level reading task development process, his negative perceptions were about word count issues, grammatical issues in the sentences, text organization, and the level of the vocabulary items. In the intermediate-level reading task development process, his negative perceptions were about item generation, following instructions, the level of the vocabulary items, processing speed, distractor length, and item quality. While Instructor 2 made a similar number of comments revealing his negative perceptions about his ChatGPT use experience across the tasks, his negative perceptions were about different aspects: mainly about ChatGPT’s capacity to follow instructions in the

elementary-level reading task development, but mainly about item generation in the intermediate-level reading task development. Instructor 3 reported equal numbers of positive comments in both tasks whereas she reported more negative perceptions when developing the intermediate-level reading task. Similarly, Instructor 4 reported more negative perceptions when she was developing the intermediate-level reading task.



CHAPTER 5: DISCUSSION

5.1 Introduction

This qualitative case study explored how four English language instructors used ChatGPT-4 for L2 reading test development. Each instructor developed two reading tasks using ChatGPT, one at the elementary level and one at the intermediate level. Their perceptions regarding this experience were also explored. In the following subsections, the findings will be discussed for each research question.

5.2 EFL Instructors' Use of ChatGPT for L2 Reading Test Development

Even though the number of instructors in this study was very small, there was a lot of variation in how they engaged with ChatGPT-4 for L2 reading test development. While all instructors used the same AI tool for test development, there were substantial differences in their prompting behaviors. Instructor 1 wrote a long prompt, pre-loaded with details, and the reading text and multiple-choice items that ChatGPT-4 generated seemed to work for the instructor, meeting his standards and/or expectations. His prompting strategy did not require him to follow up with additional questions. On the other hand, Instructor 3 had an extended engagement with ChatGPT, including several less detailed iterative prompts. While Instructor 1 created his tests using ChatGPT in relatively short sessions, Instructor 3 went through prolonged sessions. From these variations, it is obvious that while the task and the AI tool are the same, different users go through different experiences based on how they communicate with the AI tool, what type of questions they ask, and how often. Instructor 3, for example, seemed to be “trapped in the prompt loop,” as Jung et al. (2025) call it. Jung et al. (2025) classified such dissatisfaction as a result of reprompting into two: “Persistent Dissatisfaction,” in which repeated prompts do not resolve an issue, or “Cascade of New Issues,” in which new prompts or questions resolve an issue, but create new ones. The latter seems to have been experienced by Instructor 3.

The analysis of the instructors' screenshot data revealed that a large proportion of their communication with ChatGPT addressed its item and distractor generation rather than text

generation. This may indicate that for them, ChatGPT performed better in producing reading texts than in producing multiple-choice items and distractors for those texts. The instructors had to write several new prompts to fix the issues they identified in the ChatGPT-generated items and distractors, such as too obvious answers, same words in the text and in the items, contrasting distractors, high-level words, or out-of-text distractors. Instructors' greater efforts in trying to get ChatGPT to fix the or improve the items and distractors may be explained with a few reasons. First, creating effective distractors is not an easy work, even for human item writers, as there are several features that they need to follow (Celce-Murcia et al., 1974; Hoshino, 2013). And, this might be more difficult as the level of the reading tasks increases since texts get longer and more complex with more complex grammar and lexis. Thus, it is understandable that ChatGPT-4 also struggled with distractor generation. Second, and more importantly, while the prompts of the instructors included details addressing the features of the desired text, there was not much specific information about the required characteristics of the distractors that would effectively guide ChatGPT. For example, Instructor 1 simply informed ChatGPT that "Each question must have 3 options (1 correct answer and 2 distractors)" and that "all the options must be mentioned in the text." In a similar manner, Instructor 2 wrote: "There will be 5 multiple choice questions with 3 options (All the options should be included in the text somehow but there should be only one correct choice)." It is possible that these specifications were not effective enough to guide ChatGPT for distractor generation.

One consistent behavior among the language instructors in their ChatGPT-4 use for L2 reading test development was their use of external resources in every stage of test development: pre-ChatGPT use, while-ChatGPT-use, and post-ChatGPT use. The two external resources that they used very frequently in the pre-ChatGPT use stage were test specifications and the coursebook. The external resource that they used most frequently in the while-ChatGPT use and post-ChatGPT use stages was the Word document, mainly for checking the word count and for post-editing the output. Instructors' need to use external resources in every stage of reading test development makes it obvious how AI tools are embedded within the existing testing processes and complement testing practices, rather than replacing them or being used as primary resources. This supports the idea that the development of ChatGPT "has reached a level where it could be a useful supplement to language teaching and learning" (Shin & Lee, 2023), as well as testing.

One notable similarity among the three instructors was about role explanation or assignment to ChatGPT in the prompts. For example, Instructor 1 did not assign a professional

role to ChatGPT but explained his need with his own professional role when requesting it to create a text and its items: “I need to design an academic reading text for my intermediate level students’ upcoming exam.” A similar prompt was also used by Instructor 3 and Instructor 4, who assigned no professional roles to ChatGPT: “I want to prepare a reading exam as a part of an end of course test for preparatory classes at elementary level” (Instructor 3) and “I am going to prepare a reading test for a group of English language learners at a preparatory school” (Instructor 4). In these prompts, the instructors positioned themselves as the experts using ChatGPT as a tool that served their needs. Instructor 2, on the other hand, explicitly asked ChatGPT to “act like a testing expert and consider key principles of item writing for language tests,” making it the testing expert. While this difference was observed in role assignment, given the limited number of participants in this study and the limited number of reading tasks developed, it is impossible to discuss how ChatGPT operates with or without professional roles assigned.

The way the instructors in this study carefully checked the ChatGPT output (texts and items) for several aspects, such as difficulty level, length, accuracy, and distractor relevance and characteristics, revealed the crucial role of human agency and expertise for ensuring that the AI-generated material met the required standards. As Instructor 2 stated at the end of his reading task development process that the “final touch almost always requires to be done by a human expert,” humans remain central to producing quality language tests, at least for now.

Lastly, the instructors’ utilization of ChatGPT to varying degrees throughout the study serves as evidence of actual system use (ASU), one of the core constructs of the Technology Acceptance Model (TAM) proposed by Davis (1985). In TAM, actual system use is considered a direct consequence of users’ perceptions of usefulness and ease of use. In this study, the consistent reliance on ChatGPT for generating texts, items, and distractors reveals that instructors perceived it as sufficiently useful to integrate into their test development practices although ChatGPT was not employed as a standalone tool - considering the fact that the instructors consistently consulted external resources, and engaged in post-editing. This demonstrates that perceived usefulness was compelling enough to encourage adoption, even in the presence of limitations.

5.3 EFL Instructors' Perceptions of Using ChatGPT for L2 Reading Test Development

The instructors' comments regarding the texts that ChatGPT-4 created for the reading tasks revealed mostly positive perceptions, which aligns with previous research. Earlier studies also reported ChatGPT's capacity to create L2 reading texts that were found to be natural by pre- and in-service English teachers (Shin & Lee, 2023). However, instructors had more negative perceptions regarding the distractors that ChatGPT-4 generated for items. Similar problems in ChatGPT's distractor generation were also reported in previous research, such as creating options that are not mentioned in the text or having more than one answer (Shin & Lee, 2023).

While instructors expressed a relatively high number of positive perceptions regarding ChatGPT's text generation, their evaluations of item generation were more mixed, comprising both positive and negative comments. This difference may be associated with the divergent expectations attached to each function. Instructors typically asked for only one text for each test, which meant that when an acceptable passage was generated, their needs were met. Consequently, their evaluations were more consistently positive and less in number. On the other hand, item generation necessitated the creation of five multiple questions with three options each, which in turn naturally required more interactions and refinements. As a result, item generation was mentioned more frequently overall and garnered a greater proportion of negative perceptions. This might be indicative of the complexity associated with developing test items that satisfy the testing standards.

Given the important role of prompt characteristics in human-AI interactions as they determine the effectiveness of responses and directly affect users' satisfaction (Ekin, 2023; Hwang et al., 2025), it is possible that the instructors' prompting behavior might have influenced their perceptions and/or satisfaction of using ChatGPT-4 when developing L2 reading tasks. As explained in the previous section, the lack of sufficient specifications about item and distractor generation in the instructor prompts to ChatGPT might have negatively contributed to the instructors' satisfaction using ChatGPT.

The analysis of the instructors' perceptions of their ChatGPT use demonstrated differences in their perceptions across the two reading tasks (i.e., elementary-level and intermediate-level), with more negative comments about ChatGPT when developing the intermediate-level reading task particularly for Instructor 3 and Instructor 4. For example, item quality and distractor quality were bigger issues in the intermediate-level reading task, as the instructors commented

on this negatively several times. The fact that the instructors perceived their intermediate-level reading task development experience more negatively than their elementary-level reading task development experience is probably related with the higher demands of the intermediate-level reading task: longer and more complex text, more complex items, and more nuanced distractors.

Lastly, the findings of this study can be interpreted through the TAM proposed by Davis (1985). Instructors' positive comments regarding ChatGPT-4's capacity to create appropriate texts in a short time is indicative of a strong sense of perceived usefulness (PU), as the tool was perceived to reduce workload and enhance efficiency. However, the recurrent negative perceptions regarding distractor generation, item quality, and word count accuracy underscore issues with perceived ease of use (PEOU), as instructors frequently had to re-prompt ChatGPT, revise outputs, or refer to external resources. These mixed perceptions shaped instructors' attitudes (A) toward using ChatGPT, with some expressing cautious optimism while others emphasized the need for persistent human intervention.

5.4 Implications

Our preliminary answer to the more general inquiry of whether ChatGPT is required in L2 instruction was positive. That is, ChatGPT could exhibit significant potential to help EFL instructors to conveniently generate L2 reading texts along with their items in a shorter period of time, ultimately contributing to less workload. Meanwhile, teacher participation in the revision and refinement of the generated test items is still inevitable because of the current limitation of ChatGPT's capabilities. This finding of the study is consistent with the emphasis placed by Lee et al. (2024) on the need for language instructors to refine ChatGPT's output. The study also revealed the practical constraints of ChatGPT such as the unreliability of word count function, which required external resources for accuracy check. Thus, whereas it is not yet feasible to delegate the whole L2 test development process to ChatGPT, EFL instructors are firmly encouraged to explore its affordances, given its strong ability to guide them.

Additionally, although all four instructors utilized the same set of test specifications, their experiences and perceptions varied significantly, largely due to differences in how they prompted ChatGPT and managed their interaction with this AI-based tool. This emphasizes the

need for new, standardized test specifications in testing offices that are not only written for human test developers but also directly compatible with ChatGPT. Such specifications could be produced in a format that can be copied into ChatGPT, thereby ensuring that critical details such as task type, text length, vocabulary level, and item requirements are consistently communicated. In this way, instructors would not have to repeatedly remind ChatGPT of every criterion one by one in iterative prompts, and they could instead concentrate on refining and evaluating the generated output. This would facilitate the establishment of more systematic prompting strategies across testing units, ensuring efficiency and consistency in AI-assisted test development.



CHAPTER 6: CONCLUSIONS

This qualitative study explored how four English language instructors working in the preparatory program of a state university in Türkiye engaged with ChatGPT-4 to develop two L2 reading tasks, one at the elementary level and the other at the intermediate level. Their ChatGPT use was captured through screen recordings and their perceptions were gathered through think-aloud protocols. The findings revealed how the instructors approached L2 reading test development with ChatGPT, with commonalities and differences in their prompting behavior, the characteristics of their prompts and how their prompting affected the workflow and their perceptions, and how their processes and perceptions were different across different levels of reading tasks.

The value of this study lies in its exploratory and qualitative nature. Existing research on the use of ChatGPT or other AI tools for L2 reading test development has been product-based, examining the quality of the texts and items generated by the tools and users' (teachers or learners) perceptions about them. This study is one of the first process-oriented attempts looking into how language instructors actually use ChatGPT when developing an L2 reading test. Going beyond ChatGPT-products or human perceptions of the products and capturing the interactive process offers insights into the real-time decisions that instructors make when developing tests with ChatGPT and makes apparent its limitations as an assistive tool for language test development.

However, this study is limited in several ways: small sample size, focus on only two proficiency levels, and focus on only multiple-choice item format. ChatGPT may perform better in other item types used for measuring reading comprehension, such as open-ended, yes/no, or cloze questions (e.g., Lee et al., 2024), where distractors are not needed, which is currently an obvious challenge for ChatGPT. Future research should address a wider range of item types, proficiency levels (e.g., high-intermediate), and language skills (e.g., listening).

Additionally, the final test products developed by the instructors using ChatGPT-4 were not assessed for quality in this study. We do not know whether the instructors would consider the outputs ready for test use or would refine them any further. We also did not involve other testing experts or language teachers to gather their perceptions about the quality of ChatGPT-generated L2 reading tasks. These two aspects are important venues for future research.

Only a few weeks prior to the completion of this thesis, OpenAI released ChatGPT-5. Since the study was already in its final phases, it was not possible to integrate the newly launched version into the research design through a comparative analysis. Given the rapid pace of AI development, it is possible that some of the challenges observed in this study, such as distractor generation, item quality, and word count accuracy, may not apply to ChatGPT-5. Comparative studies that examine instructors' experiences and perceptions of ChatGPT-4 and ChatGPT-5, or subsequent versions, would be valuable for evaluating the progress of AI tools in language test development and for identifying whether newer versions are more practical, accurate, and pedagogically appropriate.



REFERENCES

- Abdelatif, K., & Siddiqui, A. (2021). Incorporating artificial intelligence (AI) tools in EFL classes at King Khalid University (KKU). *Journal of Tianjin University Science and Technology*, 54(10), 197-221.
- Abdi, A. N. M., Omar, A. M., Ahmed, M. H., & Ahmed, A. A. (2025). The predictors of behavioral intention to use ChatGPT for academic purposes: evidence from higher education in Somalia. *Cogent Education*, 12(1), 2460250.
- Ajzen, I. (1985). From intentions to actions: A theory of planned behavior. In *Action control: From cognition to behavior* (pp. 11-39). Berlin, Heidelberg: Springer Berlin Heidelberg.
- Al-Mamary, Y. H., Abubakar, A. A., & Abdulrab, M. (2024). The effects of the expectation confirmation model (ECM) and the technology acceptance model (TAM) on learning management systems (LMS) in sub-Saharan Africa. *Interactive Learning Environments*, 32(7), 3875-3891.
- Alarie, B., Niblett, A., and Yoon, A. H. (2018). How artificial intelligence will affect the practice of law. *University of Toronto Law Journal*, 68(1): 106-204.
- Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge University Press.
- Ali, J. K. M. (2023). Benefits and challenges of using ChatGPT: An exploratory study on English language program. *University of Bisha Journal for Humanities*, 2(2), 629-641.
- Ali, Z. (2020). Artificial intelligence (AI): A review of its uses in language teaching and learning. In *IOP Conference Series: Materials Science and Engineering* (Vol. 769, No. 1, p. 012043). IOP Publishing.
- Aljohani, N. J. (2024). Technology-mediated task-based language teaching: a CALL evaluation framework and its pedagogical implications. *Computer Assisted Language Learning*, 1–30. <https://doi.org/10.1080/09588221.2024.2317840>
- Aljohani, R. A. (2021). Teachers and students' perceptions on the impact of artificial intelligence on English language learning in Saudi Arabia. *Journal of Applied Linguistics and Language Research*, 8(1), 36-47.

- Almanea, M. (2024). Instructors' and learners' perspectives on using ChatGPT in English as a foreign language courses and its effect on academic integrity. *Computer Assisted Language Learning*, 1–26. <https://doi.org/10.1080/09588221.2024.2410158>
- Álvarez, M. F. (2016). Language testing in the digital era. In *Technology-Enhanced Language Learning for Specialized Domains* (pp. 83-94). Routledge.
- Alzubi, A. A. (2019). Teachers' perceptions on using smartphones in English as a foreign language context. *Research in Social Sciences and Technology*, 4(1), 92–104.
- Annamalai, N., Eltahir, M. E., Zyoud, S. H., Soundrarajan, D., Zakarneh, B., & Al Salhi, N. R. (2023). Exploring English language learning via Chatbot: A case study from a self-determination theory perspective. *Computers and Education: Artificial Intelligence*, 100148.
- Ansari, M., Waris, S., & Zara, C. (2024). Barriers to Educational Technology Adoption: Navigating Challenges in Integration. *Qlantic Journal of Social Sciences*, 5(3), 240-247.
- Argan, M., Gürbüz, B., Dursun, M. T., Dinç, H., & Tokay Argan, M. (2024). Usage Intention of ChatGPT for Health in Turkey: An Extended Technology Acceptance Model. *International Journal of Human-Computer Interaction*, 1–13. <https://doi.org/10.1080/10447318.2024.2426045>
- Arnold, N., & Ducate, L. (2015). Contextualized views of practices and competencies in call teacher education research. *Language, Learning and Technology*, 19(1), 1–9.
- Arslan, K. (2020). Eğitimde yapay zekâ ve uygulamaları. *Batı Anadolu Eğitim Bilimleri Dergisi*, 11(1), 71-88.
- Ashrafi, A., Zareravasan, A., Rabiee Savoji, S., & Amani, M. (2022). Exploring factors influencing students' continuance intention to use the learning management system (LMS): A multi-perspective framework. *Interactive Learning Environments*, 30(8), 1475-1497.

- Aydın, E. (2004). *Testers' Perceptions of the Test Development Process and Teachers' and Testers' Attitudes Towards the Resulting Achievement Tests at Muğla University School of Foreign Languages* [Master's thesis, Bilkent University, Türkiye].
- Bachman, L. & Palmer, A. (1996). *Language testing in practice: designing and developing useful language test*. New York: Oxford University Press.
- Bachman, L. F. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language testing*, 17(1), 1-42.
- Barrett, N. E., Liu, G. Z., & Wang, H. C. (2022). Student perceptions of a mobile learning application for English Oral Presentations: the case of EOPA. *Computer Assisted Language Learning*, 35(9), 2476-2501.
- Bax, S. (2011). Normalisation revisited: The effective use of technology in language education. *International Journal of Computer-Assisted Language Learning and Teaching (IJCALLT)*, 1(2), 1-15.
- Bekou, A., Mhamed, M. B., & Assissou, K. (2024). Exploring opportunities and challenges of using ChatGPT in English language teaching (ELT) in Morocco. *Focus on ELT Journal*, 6(1), 87-106.
- Belda-Medina, J., & Calvo-Ferrer, J. R. (2022). Using chatbots as AI conversational partners in language learning. *Applied Sciences*, 12(17), 8427.
- Belda-Medina, J., & Kokošková, V. (2024). ChatGPT for language learning: assessing teacher candidates' skills and perceptions using the Technology Acceptance Model (TAM). *Innovation in Language Learning and Teaching*, 1–16. <https://doi.org/10.1080/17501229.2024.2435900>
- Benko, A., & Lányi, C. S. (2009). History of artificial intelligence. In *Encyclopedia of Information Science and Technology, Second Edition*. IGI Global.
- Bernhardt, E. B. (1991). *Reading development in a second language*. Norwood, NJ: Ablex.
- Brown, D. (2010). *Language Assessment: Principle and Classroom Practices*. New York: Pearson Education.

- Brown, J. D. (1996). *Testing in language programs*. Prentice Hall Regents.
- Brown, J. D., & Hudson, T. (2002). *Criterion-referenced language testing*. Cambridge University Press.
- Cain, K. & Oakhill, J. (2006). Assessment matters: Issues in the measurement of reading comprehension. *British Journal of Educational Psychology*, 76, 697–708.
- Cárdenas-Claros, M. S., Roldán-Mora, N., Cárcamo, B., & Rodríguez-Arias, P. (2024). Design, validation, and application of the technology-enhanced listening task acceptance questionnaire. *Computer Assisted Language Learning*, 1-29.
- Carroll, J. B. (1961). Fundamental considerations in testing for English language proficiency of foreign students. In A. Campbell (Ed.), *Teaching English as a second language. A book of readings* (pp. 311–321) (2nd ed., 1965). New York: McGraw-Hill.
- Cavanagh, S. (1997). Content analysis: Concepts, methods and applications. *Nurse Researcher*, 4, 5-16.
- Celce-Murcia, M, Kooshian, GB, Jr, & Gosak, AJ. (1974). Goal: Good multiple-choice language-test items. *English Language Teaching*, 28, 257–262.
- Chapelle, C. (2001). *Computer applications in second language acquisition*. Cambridge: Cambridge University Press.
- Chatterjee, A. (2022). Art in an age of artificial intelligence. *Frontiers in Psychology*, 1-9.
- Chen, X., Xie, H., & Hwang, G. J. (2020). A multi-perspective study on artificial intelligence in education: Grants, conferences, journals, software tools, institutions, and researchers. *Computers and Education: Artificial Intelligence*, 100005.
- Chen, X., Zou, D., Cheng, G., & Xie, H. (2021). Artificial intelligence-assisted personalized language learning: systematic review and co-citation analysis. In *2021 International Conference on Advanced Learning Technologies (ICALT)* (pp. 241-245). IEEE.

- Chen, Y. J., Hsu, L., & Lu, S. W. (2024). How does emoji feedback affect the learning effectiveness of EFL learners? Neuroscientific insights for CALL research. *Computer Assisted Language Learning*, 37(7), 1857-1880.
- Chocarro, R., Cortiñas, M., & Marcos-Matás, G. (2023). Teachers' attitudes towards chatbots in education: a technology acceptance model approach considering the effect of social language, bot proactiveness, and users' characteristics. *Educational studies*, 49(2), 295-313.
- Chomphooyod, P., Suchato, A., Tuaycharoen, N., & Punyabukkana, P. (2023). English grammar multiple-choice question generation using Text-to-Text Transfer Transformer. *Computers and Education: Artificial Intelligence*, 5, 100158.
- Claudio, D., Velázquez, M. A., Bravo-Llerena, W., Okudan, G. E., & Freivalds, A. (2015). Perceived Usefulness and Ease of Use of Wearable Sensor-Based Systems in Emergency Departments. *IIE Transactions on Occupational Ergonomics and Human Factors*, 3(3-4), 177-187. <https://doi.org/10.1080/21577323.2015.1040559>
- Creely, E., Barnes, M., Tour, E., Henderson, M., Waterhouse, P., Pena, M. A., & Patel, S. V. (2025). Exploring attitudes to generative AI in education for English as an additional language (EAL) adult learners. *ReCALL*, 1-17.
- Creswell, J. W., & Creswell, J. D. (2017). *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage Publications.
- Cummings, A. L., Murray, H. G., & Martin, J. (1989). Protocol analysis of the social problem solving of teachers. *American Educational Research Journal*, 26(1), 25-43.
- Davidson, F., & Lynch, B. K. (2002). *Testcraft: A teacher's guide to writing and using language test specifications*. Yale University Press.
- Davis, F. D. (1985). *A technology acceptance model for empirically testing new end-user information systems: Theory and results*. [Doctoral dissertation, Massachusetts Institute of Technology].
- Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 319-340.

- Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1989). User acceptance of computer technology: A comparison of two theoretical models. *Management science*, 35(8), 982-1003.
- De la Vall, R. R. F., & Araya, F. G. (2023). Exploring the benefits and challenges of AI-language learning tools. *International Journal of Social Sciences and Humanities Invention*, 10(01), 7569-7576.
- Debski, R. (2003). Analysis of research in CALL (1980–2000) with a reflection on CALL as an academic discipline. *ReCALL*, 15(2), 177-188.
- Dilzhan, B. (2024). Teaching English and Artificial Intelligence: EFL Teachers' Perceptions and Use of ChatGPT. (*Dissertation, SDU University*).
- Ding, A. C. E., Ottenbreit-Leftwich, A., Lu, Y. H., & Glazewski, K. (2019). EFL teachers' pedagogical beliefs and practices with regard to using technology. *Journal of Digital Learning in Teacher Education*, 35(1), 20–39.
- Dirican, C. (2015). The impacts of robotics, artificial intelligence on business and economics. *Procedia-Social and Behavioral Sciences*, 195, 564-573.
- Djiwandono, P. I. (2019). How language teachers perceive information and communication technology. *Indonesian Journal of Applied Linguistics*, 8(3), 608–616.
- Dooley, P. (2008). Language testing and technology: Problems of transition to a new era. *ReCALL*, 20(1), 21-34.
- Doomen, J. (2023). The artificial intelligence entity as a legal person. *Information & Communications Technology Law*, 32(3), 277–287. <https://doi.org/10.1080/13600834.2023.2196827>
- Dörnyei, Z. (2007). *Research methods in applied linguistics*. Oxford University Press.
- Ekin, S. (2023). Prompt engineering for ChatGPT: A quick guide to techniques, tips, and best practices. *Authorea Preprints*. <https://doi.org/10.36227/techrxiv.22683919.v1>.

- Elo, S., Kääriäinen, M., Kanste, O., Pölkki, T., Utriainen, K., & Kyngäs, H. (2014). Qualitative content analysis: A focus on trustworthiness. *SAGE open*, 4(1), 2158244014522633.
- Elo, S., & Kyngäs, H. (2008). The qualitative content analysis process. *Journal of Advanced Nursing*, 62, 107-115.
- Ericsson, K. A., and Simon, H. A. (1993). *Protocol Analysis: Verbal Reports as Data*. Cambridge, MA: MIT Press.
- Esiyok, E., Gokcearslan, S., & Kucukergin, K. G. (2024). Acceptance of Educational Use of AI Chatbots in the Context of Self-Directed Learning with Technology and ICT Self-Efficacy of Undergraduate Students. *International Journal of Human-Computer Interaction*, 41(1), 641–650. <https://doi.org/10.1080/10447318.2024.2303557>
- Esteban-Millat, I., Martínez-López, F. J., Pujol-Jover, M., Gázquez-Abad, J. C., & Alegret, A. (2018). An extension of the technology acceptance model for online learning environments. *Interactive Learning Environments*, 26(7), 895–910. <https://doi.org/10.1080/10494820.2017.1421560>
- Eze, N. U., Obichukwu, P. U., & Kesharwani, S. (2021). Perceived Usefulness, Perceived Ease of Use in ICT Support and Use for Teachers. *IETE Journal of Education*, 62(1), 12–20. <https://doi.org/10.1080/09747338.2021.1908177>
- Farhady, H., & Shakery, S. (2000). Number of options and economy of multiple-choice tests. *Roshd Foreign Language Teaching Journal*, 14(1), 57.
- Felix, U. (2005). Analysing recent CALL effectiveness research—towards a common agenda. *Computer assisted language learning*, 18(1-2), 1-32.
- Fishbein, M., & Ajzen, I. (1975). *Belief, Attitude, Intention, and Behavior: An Introduction to Theory and Research*. Reading, MA: Addison-Wesley.
- Fleming, M. L., Hatfield, M. D., Wattana, M. K., & Todd, K. H. (2014). Exploratory Study of Emergency Physicians' Use of a Prescription Monitoring Program Using a Framework of Technology Acceptance. *Journal of Pain & Palliative Care Pharmacotherapy*, 28(1), 19–27. <https://doi.org/10.3109/15360288.2013.879242>

- Folkinshteyn, D., & Lennon, M. (2016). Braving Bitcoin: A technology acceptance model (TAM) analysis. *Journal of Information Technology Case and Application Research*, 18(4), 220–249. <https://doi.org/10.1080/15228053.2016.1275242>
- Fryer, L. K., Coniam, D., Carpenter, R., & Lăpuşneanu, D. (2020). Bots for language learning now: Current and future directions. *Language Learning & Technology*, 24(2), 8–22.
- Fulcher, G., & Davidson, F. (2007). *Language testing and assessment*. London and New York: Routledge.
- Fussell, S. G., & Truong, D. (2021). Accepting virtual reality for dynamic learning: an extension of the technology acceptance model. *Interactive Learning Environments*, 31(9), 5442–5459. <https://doi.org/10.1080/10494820.2021.2009880>
- Gabriel, F., Marrone, R., Van Seville, Y., Kovanovic, V., & de Laat, M. (2022). Digital education strategies around the world: practices and policies. *Irish Educational Studies*, 41(1), 85-106.
- Ganji, M., & Esfandiari, R. (2020). Attitudes of language teachers toward multiple-choice item writing guidelines: An exploratory factor analysis. *Journal of Modern Research in English Language Studies*, 7(3), 115-140.
- Genesee, F., & Upshur, J. A. (1996). *Classroom-based evaluation in second language education*. Cambridge University Press.
- Genesereth, M. R., & Nilsson, N. J. (1987). *Logical foundations of artificial intelligence*. Morgan Kaufmann.
- Gibson, S. G., Harris, M. L., & Colaric, S. M. (2008). Technology acceptance in an academic context: Faculty acceptance of online education. *Journal of Education for Business*, 83(6), 355–359. <https://doi.org/10.3200/JOEB.83.6.355-359>
- Godwin-Jones, R., O'Neill, E., & Ranalli, J. (2024). Integrating AI tools into instructed second language acquisition. In C. A. Chapelle, G. H. Beckett, & J. Ranalli (Eds.), *Exploring artificial intelligence in applied linguistics* (pp. 9-23). Iowa State University Digital Press.

- Gruba, P. (2004). Computer assisted language learning (CALL). *The handbook of applied linguistics*, 623-648.
- Gücük, G. (2022). *Perception of English language learners and teachers towards the use of artificial intelligence in the language classroom* (Master's thesis, İstanbul Aydın University, Türkiye).
- Habes, M., Elareshi, M., Salloum, S. A., Ali, S., Alfaisal, R., Ziani, A., & Alsridi, H. (2022). Students' perceptions of mobile learning technology acceptance during Covid-19: WhatsApp in focus. *Educational Media International*, 59(4), 288–306. <https://doi.org/10.1080/09523987.2022.2153990>
- Hamet, P., & Tremblay, J. (2017). Artificial intelligence in medicine. *Metabolism*, 69, S36-S40.
- Han, F. (2024). Factors affecting Chinese undergraduate medical students' behavioural intention and actual use of intelligent tutoring systems. *Australasian Journal of Educational Technology*, 40(2), 76-93.
- Haristiani, N. (2019). Artificial Intelligence (AI) chatbot as language learning medium: An inquiry. In *Journal of Physics: Conference Series* (Vol. 1387, No. 1, p. 012020). IOP Publishing.
- Helm, J. M., Swiergosz, A. M., Haeberle, H. S., Karnuta, J. M., Schaffer, J. L., Krebs, V. E., ... & Ramkumar, P. N. (2020). Machine learning and artificial intelligence: definitions, applications, and future directions. *Current reviews in musculoskeletal medicine*, 13, 69-76.
- Hendriani, S. & Suzanne, N. (2013). *Language Testing*. STAIN Batusangkar Press.
- Hendrickson, A. R., & Latta, P. D. (1996). An Evaluation of the Reliability and Validity of Davis' Perceived Usefulness and Perceived Ease of Use Instrument. *Journal of Computer Information Systems*, 36(3), 77–82. <https://doi.org/10.1080/08874417.1996.11647210>
- Henning, G. (1987). *A guide to language testing: Development, evaluation, research*.
- Hockly, N. (2023). Artificial Intelligence in English Language Teaching: The Good, the Bad and the Ugly. *RELC Journal*, 00336882231168504.

- Hoshino, Y. (2013). Relationship between types of distractor and difficulty of multiple-choice vocabulary tests in sentential context. *Language Testing in Asia* 3, 16. <https://doi.org/10.1186/2229-0443-3-16>
- Hsiao, C. H., & Tang, K. Y. (2015). Examining a model of mobile healthcare technology acceptance by the elderly in Taiwan. *Journal of Global Information Technology Management*, 18(4), 292–311. <https://doi.org/10.1080/1097198X.2015.1108099>
- Hsieh, H. F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative health research*, 15(9), 1277-1288.
- Hsu, L. (2016). An empirical examination of EFL learners' perceptual learning styles and acceptance of ASR-based computer-assisted pronunciation training. *Computer Assisted Language Learning*, 29(5), 881-900.
- Hsu, H. W. (2024). An examination of automatic speech recognition (ASR)-based computer-assisted pronunciation training (CAPT) for less-proficient EFL students using the technology acceptance model. *International Journal of Technology in Education*, 7(3), 456-473.
- Huang, F., Teo, T., & Zhou, M. (2019). Factors affecting Chinese English as a foreign language teachers' technology acceptance: A qualitative study. *Journal of Educational Computing Research*, 57(1), 83–105.
- Huang, H., Chen, Y., & Rau, P. L. P. (2022). Exploring acceptance of intelligent tutoring system with pedagogical agent among high school students. *Universal Access in the Information Society*, 21(2), 381-392.
- Huang, W., Hew, K. F., & Fryer, L. K. (2021). Chatbots for language learning—Are they really useful? A systematic review of chatbot-supported language learning. *Journal of Computer Assisted Learning*, 38(1), 237-257.
- Huang, X., Zou, D., Cheng, G., Chen, X., & Xie, H. (2023). Trends, research issues and applications of artificial intelligence in language education. *Educational Technology & Society*, 26(1), 112-131.
- Hubbard, P. (2003). A survey of unanswered questions in CALL. *Computer Assisted Language Learning*, 16(2-3), 141-154.

- Hughes, A. (2002). Kinds of tests and testing. *Testing for language teachers*, 9-21.
- Hughes, K. L., & Scott-Clayton, J. (2011). Assessing developmental assessment in community colleges. *Community College Review*, 39(4), 327-351.
- Hwang, M., Lee, K-H., & Lee, H-K. (2025). A word to the wise: Crafting impactful prompts for ChatGPT. *System*, 133, Article 103756. <https://doi.org/10.1016/j.system.2025.103756>
- Hz, B. I. R., Laiya, R. E., & Sarumaha, M. S. (2023). Navigating English Writing Proficiency Tests in the Era of Artificial Intelligence. *Journal of English Education and Teaching*, 7(3), 480-498.
- Iftikhar, R., Khan, M. S., & Pasanchay, K. (2022). Virtual reality tourism and technology acceptance: a disability perspective. *Leisure Studies*, 42(6), 849–865. <https://doi.org/10.1080/02614367.2022.2153903>
- Jiang, R. (2022). How does artificial intelligence empower EFL teaching and learning nowadays? A review on artificial intelligence in the EFL context. *Frontiers in Psychology*, 13, 1049401.
- Jimenez, L., & Boser, U. (2021). Future of testing in education: Artificial intelligence. *Center for American Progress*.
- Johnson, C. (2021). *Language learners' perceptions of automatic speech recognition as a writing tool: A Technology Acceptance Model analysis* (Doctoral dissertation, Concordia University).
- Jung, H., Jeong, Y., & Lee, J. (2025). Trapped in the Prompt Loop: Reprompt Behavior in Writing with ChatGPT. *International Journal of Human–Computer Interaction*, 1–29. <https://doi.org/10.1080/10447318.2025.2530100>
- Kalra, R. (2024). Exploring Teachers' Perceptions Toward the Integration of AI Tools in the Language Classroom. *NIDA Journal of Language and Communication*.
- Kamble, A., Golhar, D., & Kalkar, P. (2022). Technology acceptance for online teaching-learning: perspectives of teachers from higher education in India. *Educational Media International*, 59(4), 324–340. <https://doi.org/10.1080/09523987.2022.2153989>

- Kaplan, A., & Haenlein, M. (2019). Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business horizons*, 62(1), 15-25.
- Karabulut-İlğü, A. (2013). *Factors impacting university-level language teachers' technology use and integration*. [Unpublished doctoral dissertation. Iowa State University (ISU), Iowa, USA].
- Karaci, A., Piri, Z., Akyüz, H. İ., & Bilgiçci, G. (2018). Student perceptions of an intelligent tutoring system: A technology acceptance model perspective. *Online Submission*, 182(22), 31-36.
- Kessler, G. (2018). Technology and the future of language teaching. *Foreign language annals*, 51(1), 205-218.
- Khasawneh, O. Y. (2022). Technophobia: How Students' Technophobia Impacts Their Technology Acceptance in an Online Class. *International Journal of Human-Computer Interaction*, 39(13), 2714–2723. <https://doi.org/10.1080/10447318.2022.2085398>
- Khlaisang, J., Teo, T., & Huang, F. (2021). Acceptance of a flipped smart application for learning: a study among Thai university students. *Interactive Learning Environments*, 29(5), 772-789.
- Kietzmann, J., Paschen, J., & Treen, E. (2018). Artificial Intelligence in Advertising: How Marketers Can Leverage Artificial Intelligence Along the Consumer Journey. *Journal of Advertising Research*, 58(3), 263–267. <https://doi.org/10.2501/JAR-2018-035>
- Kocabas, A. D., Calik, E., & Cetinguc, B. (2024). Examining cryptocurrency usage by technology acceptance model in the Turkish context. *African Journal of Science, Technology, Innovation and Development*, 16(5), 630–640. <https://doi.org/10.1080/20421338.2024.2364436>
- Kopplin, C. S. (2022). Chatbots in the workplace: A technology acceptance study applying uses and gratifications in coworking spaces. *Journal of Organizational Computing and Electronic Commerce*, 32(3–4), 232–257. <https://doi.org/10.1080/10919392.2023.2215666>

- Krippendorff, K. (2004). *Content analysis: an introduction to its methodology*, SAGE Publications, the United States of America, ISBN: 0-7619-1544-3.
- Lachman, R., & Joffe, M. (2021). Applications of artificial intelligence in media and entertainment. In *Analyzing future applications of AI, sensors, and robotics in society* (pp. 201-220). IGI Global.
- Landrum, R. E., Cashin, J. R., & Theis, K. S. (1993). More evidence in favor of three-option multiple-choice tests. *Educational and Psychological Measurement*, 53(3), 771-778.
- Lane, S., Raymond, M. R., Haladyna, T. M., & Downing, S. M. (2015). Test development process. In *Handbook of test development* (pp. 3-18). Routledge.
- Law, L. (2024). Application of generative artificial intelligence (GenAI) in language teaching and learning: A scoping literature review. *Computers and Education Open*, 100174.
- Lee, H., & Winke, P. (2013). The differences among three-, four-, and five-option-item formats in the context of a high-stakes English-language listening test. *Language Testing*, 30(1), 99-123.
- Lee, U., Jung, H., Jeon, Y., Sohn, Y., Hwang, W., Moon, J., & Kim, H. (2024). Few-shot is enough: exploring ChatGPT prompt engineering method for automatic question generation in english education. *Education and Information Technologies*, 29(9), 11483-11515. <https://doi.org/10.1007/s10639-023-12249-8>
- Levy, M. (1997). *Computer-assisted language learning: Context and conceptualisation*. Oxford University Press.
- Levy, M. (2002). CALL by design: Products, processes and methods. *ReCALL*, 14(1), 129-142.
- Levy, M., & Hubbard, P. (2005). Why call CALL “CALL”? *Computer Assisted Language Learning*, 18(3), 143-149.
- Levy, M., & Stockwell, G. (2013). *CALL dimensions: Options and issues in computer-assisted language learning*. Routledge.

- Li, R., Meng, Z., Tian, M., Zhang, Z., Ni, C., & Xiao, W. (2019). Examining EFL learners' individual antecedents on the adoption of automated writing evaluation in China. *Computer Assisted Language Learning*, 32(7), 784-804.
- Lin, Y., & Yu, Z. (2024). Learner Perceptions of Artificial Intelligence-Generated Pedagogical Agents in Language Learning Videos: Embodiment Effects on Technology Acceptance. *International Journal of Human-Computer Interaction*, 41(2), 1606–1627. <https://doi.org/10.1080/10447318.2024.2359222>
- Liu, C. L., Wang, C. H. and Gao, Z. M. (2005) 'Using lexical constraints to enhance the quality of computer-generated multiple-choice cloze items'. *Computational Linguistics and Chinese Language Processing*, 10 (3), 303–28.
- Liu, G. L., Darvin, R., & Ma, C. (2024). Exploring AI-mediated informal digital learning of English (AI-IDLE): A mixed-method investigation of Chinese EFL learners' AI adoption and experiences. *Computer Assisted Language Learning*, 1-29.
- Liu, G., & Ma, C. (2024). Measuring EFL learners' use of ChatGPT in informal digital learning of English based on the technology acceptance model. *Innovation in Language Learning and Teaching*, 18(2), 125-138.
- Liu, H., Lin, C. H., & Zhang, D. (2017). Pedagogical beliefs and attitudes toward information and communication technology: A survey of teachers of English as a foreign language in China. *Computer Assisted Language Learning*, 30(8), 745-765.
- Liu, J., Chang, H., Forrest, J. Y. L., & Yang, B. (2020). Influence of artificial intelligence on technological innovation: Evidence from the panel data of China's manufacturing sectors. *Technological Forecasting and Social Change*, 158, 120142.
- Lo, S. (2023). Neural machine translation in EFL classrooms: learners' vocabulary improvement, immediate vocabulary retention and delayed vocabulary retention. *Computer Assisted Language Learning*, 1–20. <https://doi.org/10.1080/09588221.2023.2207603>

- Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2002). Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human communication research*, 28(4), 587-604.
- Luckin, R., & Cukurova, M. (2019). Designing educational technologies in the age of AI: A learning sciences-driven approach. *British Journal of Educational Technology*, 50(6), 2824–2838
- Luckin, R., Holmes, W., Griffiths, M., & Forcier, L. B. (2016). *Intelligence unleashed: An argument for AI in education*. Pearson Education.
- Lund, B. D., & Wang, T. (2023). Chatting about ChatGPT: how may AI and GPT impact academia and libraries?. *Library Hi Tech News*.
- Luo, Z., & Cao, L. (2024). Understanding factors influencing ESL student teachers' adoption of classroom response systems: An integration of TAM and AOI theory. *Interactive Learning Environments*, 1-19.
- Ma, J., Wang, P., Li, B., Wang, T., Pang, X. S., & Wang, D. (2025). Exploring User Adoption of ChatGPT: A Technology Acceptance Model Perspective. *International Journal of Human–Computer Interaction*, 41(2), 1431–1445. <https://doi.org/10.1080/10447318.2024.2314358>
- Magliano, J.P., Millis, K., Ozuru, Y. & McNamara, D.S. (2007). A multidimensional framework to evaluate reading assessment tools. In D.S. McNamara (Ed.), *Reading comprehension strategies: Theories, interventions, and technologies*. (pp. 107–134). New York: Erlbaum.
- Marangunić, N., & Granić, A. (2015). Technology acceptance model: a literature review from 1986 to 2013. *Universal access in the information society*, 14, 81-95.
- McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (1955). A proposal for the Dartmouth summer research project on artificial intelligence. [http:// www-formal.stanford.edu/jmc/history/dartmouth/ dartmouth.html](http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html)
- McCloskey, D. (2016). Evaluating electronic commerce acceptance with the technology acceptance model. *Journal of Computer Information Systems*, 44(2), 49-57.

- McNamara, F. & Roever, C. (2006). *Language Testing: The Social Dimension*. Oxford: Blackwell.
- McNamara, T. F. (2000). *Language testing*. Oxford University Press.
- Memarzadeh, F., Blum, S. C., & Adams, C. (2016). Business Travelers' Intention to Purchase: The Application of Technology Acceptance Model (TAM). *Journal of Quality Assurance in Hospitality & Tourism*, 17(4), 412–424. <https://doi.org/10.1080/1528008X.2015.1096755>
- Mijwil, M. M., Aggarwal, K., Mutar, D. S., Mansour, N., & Singh, R. (2022). The position of artificial intelligence in the future of education: an overview. *Asian Journal of Applied Sciences*, 10(2).
- Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook*. Sage.
- Minsky, M. (Ed.). (1968). *Semantic information processing*. MIT Press.
- Mintz, Y., & Brodie, R. (2019). Introduction to artificial intelligence in medicine. *Minimally Invasive Therapy & Allied Technologies*, 28(2), 73-81.
- Mohamed, A. M. (2023). Exploring the potential of an AI-based Chatbot (ChatGPT) in enhancing English as a Foreign Language (EFL) teaching: perceptions of EFL Faculty Members. *Education and Information Technologies*, 29(3), 3195-3217.
- Moor, J. H. (2003). *The Turing test: the elusive standard of artificial intelligence* (Vol. 30). Springer Science & Business Media.
- Morgan, D. L. (1993). Qualitative content analysis: A guide to paths not taken. *Qualitative Health Research*, 1, 112-121.
- Mori, S. (2018). US defense innovation and artificial intelligence. *Asia-Pacific Review*, 25(2), 16-44.
- Naidoo, S. (2022). Adapting the technology acceptance model to investigate student responses to e-learning. *South African Journal of African Languages*, 42(1), 1–8. <https://doi.org/10.1080/02572117.2021.2015111>

- Nannelli, M., Capone, F., & Lazzeretti, L. (2023). Artificial intelligence in hospitality and tourism. State of the art and future research avenues. *European Planning Studies*, 31(7), 1325–1344. <https://doi.org/10.1080/09654313.2023.2180321>
- Nguyen, T. C. (2024). University Teachers' Perceptions of Using ChatGPT in Language Teaching and Assessment. *Proceedings of the AsiaCALL International Conference*, 4, 116-128. ISSN: 2833-6836; ISBN: 979-8-9870112-4-9. DOI: <https://doi.org/10.54855/paic.2349>
- Nguyen, T. L., Nguyen, H. L., & Le, D. H. (2024). Teachers' Perspectives on AI-Driven Quillionz for Generating EFL Reading Comprehension Quizzes. In *Proceedings of the AsiaCALL International Conference* (Vol. 6, pp. 20-34).
- Nguyen, T. P. T. (2023). The Application of ChatGPT in Language Test Design – The What and How. *Proceedings of the AsiaCALL International Conference*, 4, 104-115. ISSN: 2833-6836; ISBN: 979-8-9870112-4-9. DOI: <https://doi.org/10.54855/paic.2348>
- Niehaves, B., & Plattfaut, R. (2014). Internet adoption by the elderly: employing IS technology acceptance theories for understanding the age-related digital divide. *European Journal of Information Systems*, 23(6), 708–726. <https://doi.org/10.1057/ejis.2013.19>
- O'Loughlin, K. (2006). Learning about second language assessment: Insights from a postgraduate student on-line subject forum. *University of Sydney Papers in TESOL*, 1, 71–85.
- OpenAI. (2022, November 30). *ChatGPT: Optimizing language models for dialogue*. <https://openai.com/blog/chatgpt/>
- Ozuru, Y., Best, R., Bell, C., Witherspoon, A. & McNamara, D.S. (2007). Influence of question format and text availability on the assessment of expository text comprehension. *Cognition and Instruction*, 25, 399–438.
- Park, E., Kim, H., & Ohm, J. Y. (2014). Understanding driver adoption of car navigation systems using the extended technology acceptance model. *Behaviour & Information Technology*, 34(7), 741–751. <https://doi.org/10.1080/0144929X.2014.963672>

- Pavlou, P. A. (2003). Consumer Acceptance of Electronic Commerce: Integrating Trust and Risk with the Technology Acceptance Model. *International Journal of Electronic Commerce*, 7(3), 101–134. <https://doi.org/10.1080/10864415.2003.11044275>
- Peng, J. E., & Liang, W. (2025). Willingness to communicate with artificial intelligence (AI)? insights from tracking EFL learners' perceived acceptance and chat output. *Computer Assisted Language Learning*, 1-27.
- Pérez-Paredes, P. (2022). A systematic review of the uses and spread of corpora and data-driven learning in CALL research during 2011–2015. *Computer Assisted Language Learning*, 35(1-2), 36-61.
- Popenici, S. A. D., & Kerr, S. (2017). Exploring the impact of artificial intelligence on teaching and learning in higher education. *Research and Practice in Technology Enhanced Learning*, 12(22), 1–13.
- Ofori, D., & Appiah-Nimo, C. (2019). Determinants of online shopping among tertiary students in Ghana: An extended technology acceptance model. *Cogent Business & Management*, 6(1). <https://doi.org/10.1080/23311975.2019.1644715>
- Qin, F., Li, K., & Yan, J. (2020). Understanding user trust in artificial intelligence-based educational systems: Evidence from China. *British Journal of Educational Technology*, 51(5), 1693-1710.
- Rafiee, M., & Abbasian-Naghneh, S. (2021). E-learning: development of a model to assess the acceptance and readiness of technology among language learners. *Computer Assisted Language Learning*, 34(5-6), 730-750.
- Renz, A., Krishnaraja, S., & Gronau, E. (2020). Demystification of artificial intelligence in education: How much AI is really in the educational technology? *International Journal of Artificial Intelligence*, 2(1), 14–30.
- Richardson, M. & Clesham, R. (2021). 'Rise of the machines? The evolving role of AI technologies in high-stakes assessment'. *London Review of Education*, 19 (1), 9, 1–13. <https://doi.org/10.14324/LRE.19.1.09>
- Rienties, B., Giesbers, B., Lygo-Baker, S., Ma, H. W. S., & Rees, R. (2014). Why some teachers easily learn to use a new virtual learning environment: a technology

- acceptance perspective. *Interactive Learning Environments*, 24(3), 539–552.
<https://doi.org/10.1080/10494820.2014.881394>
- Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational measurement: issues and practice*, 24(2), 3-13.
- Roumeliotis, K. I., & Tselikas, N. D. (2023). ChatGPT and open-ai models: A preliminary review. *Future Internet*, 15(6), 192.
- Rudolph, J., Tan, S., & Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education?. *Journal of Applied Learning and Teaching*, 6(1).
- Russell, S. J., & Norvig, P. (2010). *Artificial intelligence: A modern approach* (3rd ed.). Prentice Hall.
- Sagnier, C., Loup-Escande, E., Lourdeaux, D., Thouvenin, I., & Valléry, G. (2020). User Acceptance of Virtual Reality: An Extended Technology Acceptance Model. *International Journal of Human–Computer Interaction*, 36(11), 993–1007. <https://doi.org/10.1080/10447318.2019.1708612>
- Saricaoglu, A., & Bilki, Z. (2021). Voluntary use of automated writing evaluation by content course students. *ReCALL*, 33(3), 265–277.
doi:10.1017/S0958344021000021
- Saroia, A. I., & Gao, S. (2019). Investigating university students' intention to use mobile learning management systems in Sweden. *Innovations in Education and Teaching International*, 56(5), 569-580.
- Savitskie, K., Royne, M. B., Persinger, E. S., Grunhagen, M., & Witte, C. L. (2007). Norwegian Internet Shopping Sites: An Application & Extension of the Technology Acceptance Model. *Journal of Global Information Technology Management*, 10(4), 54–73. <https://doi.org/10.1080/1097198X.2007.10856455>
- Saxena, D., Lamest, M., & Bansal, V. (2021). Responsible machine learning for ethical artificial intelligence in business and industry. In *Handbook of Research on Applied Data Science and Artificial Intelligence in Business and Industry* (pp. 639-653). IGI global.

- Schreier, M. (2012). *Qualitative content analysis in practice*. Thousand Oaks, CA: Sage.
- Searle, J. R. (1980). Minds, Brains, and Programs. *Behavioral and Brain Sciences*, 3(3), 417-457.
- Self, J. (2016). The birth of IJAIED. *International Journal of Artificial Intelligence in Education*, 26, 4–12.
- Seufert, S., Guggemos, J., & Sailer, M. (2021). Technology-related knowledge, skills, and attitudes of pre-and in-service teachers: The current situation and emerging trends. *Computers in Human Behavior*, 115, 106552.
- Sharadgah, T. A., & Sa'di, R. A. (2022). A Systematic Review of Research on the Use of Artificial Intelligence in English Language Teaching and Learning (2015-2021): What are the Current Effects?. *Journal of Information Technology Education: Research*, 21.
- Sharma, S., & Saini, J. R. (2022). On the role of teachers' acceptance, continuance intention and self-efficacy in the use of digital technologies in teaching practices. *Journal of Further and Higher Education*, 46(6), 721-736.
- Shen, Y., & Chen, L. (2025). 'Critical chatting' or 'casual cheating': how graduate EFL students utilize ChatGPT for academic writing. *Computer Assisted Language Learning*, 1–29. <https://doi.org/10.1080/09588221.2025.2479141>
- Shin, D., & Lee, J. H. (2023). Can ChatGPT make reading comprehension testing items on par with human experts? *Language Learning & Technology*, 27(3), 27–40. <https://hdl.handle.net/10125/73530>
- Shin, J., Guo, Q. & Gierl, M. J. (2019) 'Multiple-choice item distractor development using topic modeling approaches'. *Frontiers in Psychology*, 10, pp.825.
- Shoufan, A. (2023). Exploring students' perceptions of ChatGPT: Thematic analysis and follow-up survey. *IEEE access*, 11, 38805-38818.
- Sidick, J. T., Barrett, G. V., & Doverspike, D. (1994). Three-alternative multiple choice tests: An attractive option. *Personnel Psychology*, 47(4), 829-835.

- Singh, P. (2024). Transforming healthcare through AI: Enhancing patient outcomes and bridging accessibility gaps. *Journal of Artificial Intelligence Research*, 4(1), 220–232.
- Soegoto, E. S., Utami, R. D., & Hermawan, Y. A. (2019). Influence of artificial intelligence in automotive industry. *Journal of Physics: Conference Series*, 1402(6): 1-6.
- Song, P., & Wang, X. (2020). A bibliometric analysis of worldwide educational artificial intelligence research development in recent twenty years. *Asia Pacific Education Review*, 21(3), 473-486.
- Sumakul, D. T., Hamied, F. A., & Sukyadi, D. (2022). Artificial intelligence in EFL classrooms: Friend or foe? *LEARN Journal: Language Education and Acquisition Research Network*, 15(1), 232-256.
- Sun, W., & Zou, B. (2022). A study of pre-service EFL teachers' acceptance of online teaching and the influencing factors. *Language Learning & Technology*, 26(2), 38–49. <https://doi.org/10.1257/73476>
- Suvorov, R., & Hegelheimer, V. (2013). Computer-assisted language testing. *The companion to language assessment*, 2, 594-613.
- Taherdoost, H. (2018). Development of an adoption model to assess user acceptance of e-service technology: E-Service Technology Acceptance Model. *Behaviour & Information Technology*, 37(2), 173–197. <https://doi.org/10.1080/0144929X.2018.1427793>
- Tasci, G., & Celebi, M. (2020). A new paradigm in education: "Artificial intelligence in higher education". *International Journal of Society Researches*, 16(29), 2346-2370.
- Taylor, T. (2019). Artificial Intelligence in Defence: When AI Meets Defence Acquisition Processes and Behaviours. *The RUSI Journal*, 164(5–6), 72–81. <https://doi.org/10.1080/03071847.2019.1694229>
- Teo, T. (2016). Modelling Facebook usage among university students in Thailand: the role of emotional attachment in an extended technology acceptance model. *Interactive Learning Environments*, 24(4), 745-757.

- Tian, C. (2024). The Influence of Generative AI Technologies On Academic Writing in EFL Education. *Journal of Education, Humanities and Social Sciences*, 28, 575-584.
- 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 learning environments*, 10(1), 15.
- Tubaishat, A. (2017). Perceived usefulness and perceived ease of use of electronic health records among nurses: Application of Technology Acceptance Model. *Informatics for Health and Social Care*, 43(4), 379–389. <https://doi.org/10.1080/17538157.2017.1363761>
- Turing, A. M. (1950). Mind. *Mind*, 59(236), 433-460.
- Üretmen, S. (2024). *Turkish EFL teachers' awareness and perspectives on artificial intelligence incorporation into language instruction* (Master's thesis, Necmettin Erbakan University, Türkiye).
- Vallade, J. I., Kaufmann, R., Frisby, B. N., & Martin, J. C. (2020). Technology acceptance model: investigating students' intentions toward adoption of immersive 360° videos for public speaking rehearsals. *Communication Education*, 70(2), 127–145. <https://doi.org/10.1080/03634523.2020.1791351>
- Varsamidou, A. (2024). Integration of Artificial Intelligence in Foreign Language Teaching: Exploring Views and Beliefs of Greek Foreign Language Teachers. *International Journal of Language and Literary Studies*. 6(3).299-320. <http://doi.org/10.36892/ijlls.v6i3>.
- Vigoroso, L., Roccato, M., Cavallo, E., & Caffaro, F. (2024). Intention to Adopt Digital Games for Safety Training in Young Farm Operators: The Role of Ease of Use, Perceived Usefulness, and Game Design Characteristics. *International Journal of Human-Computer Interaction*, 41(2), 1054–1063. <https://doi.org/10.1080/10447318.2024.2310353>
- Vyas, R., & Supe, A. (2008). Multiple choice questions: a literature review on the optimal number of options. *Natl Med J India*, 21(3), 130-3.

- Weber, R. P. (1990). *Basic content analysis*. Newbury Park, CA: Sage.
- Wei, J., Kim, S., Jung, H., & Kim, Y. H. (2024). Leveraging large language models to power chatbots for collecting user self-reported data. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1), 1-35.
- Wilson, J., Ahrendt, C., Fudge, E. A., Raiche, A., Beard, G., & MacArthur, C. (2021). Elementary teachers' perceptions of automated feedback and automated scoring: Transforming the teaching and learning of writing using automated writing evaluation. *Computers & Education*, 168, 104208.
- Wolf, D. F. (1993). A comparison of assessment tasks used to measure FL reading comprehension. *The Modern Language Journal*, 77(4), 473-489.
- Wu, R., & Yu, Z. (2023). Investigating Users' Acceptance of the Metaverse with an Extended Technology Acceptance Model. *International Journal of Human-Computer Interaction*, 40(19), 5810–5826.
<https://doi.org/10.1080/10447318.2023.2241295>
- Wu, T., He, S., Liu, J., Sun, S., Liu, K., Han, Q. L., & Tang, Y. (2023). A brief overview of ChatGPT: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica*, 10(5), 1122-1136.
- Yang, S., & Appleget, C. (2024). An exploration of preservice teachers' perceptions of Generative AI: Applying the technological Acceptance Model. *Journal of Digital Learning in Teacher Education*, 40(3), 159–172.
<https://doi.org/10.1080/21532974.2024.2367573>
- Yanhua, Z. (2020). The application of artificial intelligence in foreign language teaching. In *2020 International Conference on Artificial Intelligence and Education (ICAIE)* (pp. 40-42). IEEE.
- Yaşar, D. I. (2020). *An investigation of the relationship between EFL instructors' perceptions on technology use and their possible language teacher selves* (Master's thesis, Bilkent University, Türkiye).
- Yunjiu, L., Wei, W., & Zheng, Y. (2022). Artificial intelligence-generated and human expert-designed vocabulary tests: A comparative study. *SAGE Open*, 12(1), 21582440221082130.

- Yunus, M. M. (2007). Malaysian ESL teachers' use of ICT in their classrooms: expectations and realities. *ReCALL*, 19(1), 79-95.
- Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – where are the educators? *International Journal of Educational Technology in Higher Education*, 16(39), 1–27.
- Zhai, N., & Ma, X. (2022). Automated writing evaluation (AWE) feedback: A systematic investigation of college students' acceptance. *Computer Assisted Language Learning*, 35(9), 2817-2842.
- Zhang, D., Hennessy, S., & Pérez-Paredes, P. (2023). An investigation of Chinese EFL learners' acceptance of mobile dictionaries in English language learning. *Computer Assisted Language Learning*, 1-25.
- Zhang, D., Hoang, T., Pan, S., Hu, Y., Xing, Z., Staples, M., ... & Quigley, A. (2023). Test-takers have a say: understanding the implications of the use of AI in language tests. *arXiv preprint arXiv:2307.09885*.
- Zhang, D., & Pérez-Paredes, P. (2021). Chinese postgraduate EFL learners' self-directed use of mobile English learning resources. *Computer Assisted Language Learning*, 34(8), 1128-1153.
- Zhao, Y. (2003). Recent developments in technology and language learning: A literature review and meta-analysis. *CALICO journal*, 7-27.
- Zhou, S., & Bacanin, N. (2024). Artificial intelligence in advanced manufacturing. *International Journal of Computer Integrated Manufacturing*, 37(4), 383–384. <https://doi.org/10.1080/0951192X.2024.2327226>
- Ziqi, C., Xinhua, Z., Qi, L., & Wei, W. (2024). L2 students' barriers in engaging with form and content-focused AI-generated feedback in revising their compositions. *Computer Assisted Language Learning*, 1-21.
- Zou, B., Lyu, Q., Han, Y., Li, Z., & Zhang, W. (2023). Exploring students' acceptance of an artificial intelligence speech evaluation program for EFL speaking practice: an application of the Integrated Model of Technology Acceptance. *Computer Assisted Language Learning*, 1-26.

APPENDICES

APPENDIX I

INFORMED CONSENT FORM

Dear Colleague,

My name is Gülçin Akyol. I am an instructor at the School of Foreign Languages, at Social Sciences University of Ankara (ASBU), Türkiye. As an MA student in the English Language Teaching program at ASBU, I am conducting a study entitled “The Use of Artificial Intelligence in L2 Reading Test Development: A Study of Language Instructors' Perceptions and Experiences in an English Preparatory School.” As you are working or have previously worked as a test developer at the same school, I would like to invite you to participate in my research study. This form includes details of the study to help you decide whether or not to become a participant. Participation is on a voluntary basis, and you may wish to withdraw from the study at any time without any negative consequences.

The purpose of this study is to explore your perceptions and practices of using ChatGPT for test development purposes. If you agree to participate, you will be asked to develop two reading tests at two different proficiency levels (i.e, elementary and intermediate) with five multiple-choice questions in each test. Your ChatGPT use will be screen recorded. The data collection process will be completed online. All the information you provide will be kept confidential.

If you have any questions about the study, you can contact me.

Thank you,

1. I understand the nature and purpose of the study and my involvement in it.
2. My participation is voluntary and I am free to withdraw at any time without giving any reasons. This will not affect my status now or in the future.
3. I will use a screen recording extension to record my computer screen and my voice.
4. My personal information will be treated in strict confidence.
5. There are no incentives or rewards for participation.
6. The results obtained from this study can be reported and published scientifically, provided that anonymity and confidentiality are maintained.

Please tick below if you want to volunteer to participate in the study.

☐ I have been informed about this study and I agree to voluntarily participate in this study.

SIGNATURE

DATE

APPENDIX II

GENERAL OVERVIEW OF ChatGPT FOR TEST DEVELOPMENT

I. Introduction to ChatGPT

ChatGPT is a language model developed by OpenAI. It is designed to understand and generate human-like text based on the input provided. It has the potential to assist in various tasks, including the development of reading tests with multiple-choice questions (MCQs). ChatGPT has been fine-tuned on a diverse range of text data to understand and respond to user prompts effectively. It can be applied in a wide array of contexts, including content generation, customer support, language translation, and, in your case, test development.

II. Key Concepts and Terminology

Large Language models like ChatGPT are computer programs trained to understand and generate human language. They learn patterns/structures from large datasets to generate coherent text.

Fine-Tuning is the process of customizing a pre-trained language model like ChatGPT for specific tasks. It allows you to adapt the model to generate content relevant to your tests.

Prompt Engineering involves providing specific instructions or questions to ChatGPT to generate desired responses. Effective prompts are essential for obtaining accurate and relevant content.

III. How ChatGPT Works

ChatGPT is built on a neural network architecture known as the Transformer. This architecture enables the model to capture intricate language patterns and generate coherent text. ChatGPT is trained on vast amounts of text data from the internet. It learns from this data to understand and generate text in a variety of styles and topics. When you input a prompt or question, ChatGPT generates responses based on the patterns and information it has learned during training. It aims to provide coherent and contextually relevant text.

IV. Using ChatGPT for Test Development

Generating Test Content: You can use ChatGPT to generate reading test items and MCQs. By providing clear and precise prompts, you can obtain content aligned with your objectives.

Customization: ChatGPT can be fine-tuned for specific test requirements. This customization process ensures that the generated content is tailored to your needs and maintains consistency.

Review and Quality Control: While ChatGPT is a valuable tool, it's essential to review and edit the content it generates. Human oversight is crucial to ensure accuracy and relevance.

V. Practical Tips

Effective Prompts: Craft clear and specific prompts to obtain the desired content. Experiment with different phrasings to optimize results.

Collaboration: Encourage collaboration and feedback among test developers. Collaborative efforts can enhance the quality of test items.

Data Management: Implement data management practices to organize and store generated content securely. Back up data regularly.

VI. How to Register and Log into ChatGPT

Visit the OpenAI Platform: To access ChatGPT, you need to visit the OpenAI Platform.

Create an Account: If you don't already have an OpenAI account, you will need to create one. Click on the "Sign Up" or "Register" button to get started.

Provide Your Information: Fill out the required information for registration. This typically includes your name, email address, and a secure password. Ensure that you agree to OpenAI's terms and conditions.

Verification: You may be required to verify your email address by clicking on a verification link sent to your email. Follow the provided instructions for verification.

Payment Information: Depending on the usage plan and pricing structure offered by OpenAI, you may need to provide payment information. This may include credit card details or other payment methods. Ensure that you understand the pricing and billing terms.

Account Confirmation: Once your registration is complete and payment information is verified, you will receive confirmation that your account is active.

Logging into ChatGPT

Visit the OpenAI Platform: To log into ChatGPT, go to the OpenAI Platform.

Enter Your Credentials: Click on the "Log In" or "Sign In" button on the platform's homepage. Enter the email address and password associated with your OpenAI account.

Two-Factor Authentication (if enabled): If you have two-factor authentication (2FA) enabled for added security, you will need to provide the verification code sent to your registered mobile device or email.

Accessing ChatGPT: After successfully logging in, you will have access to ChatGPT and its features. You can start using ChatGPT by providing prompts and interacting with the model.

Please note that the registration and login process may vary depending on OpenAI's policies and any updates they make to their platform. Be sure to review the latest information on their website for the most accurate guidance.

VII. Support

OpenAI Community and Forums

<https://community.openai.com/>

For technical support or inquiries related to ChatGPT, you can contact OpenAI Support.

APPENDIX III

GENERAL OVERVIEW OF SCREEN RECORDING

HOW TO INSTALL AND USE A SCREEN RECORDER EXTENSION ON GOOGLE CHROME

This overview will help you install and use a screen recorder extension to record your screen and voice while preparing reading tests on ChatGPT.

Step 1: Open Google Chrome

Make sure you have Google Chrome open on your computer.

Step 2: Go to the Chrome Web Store

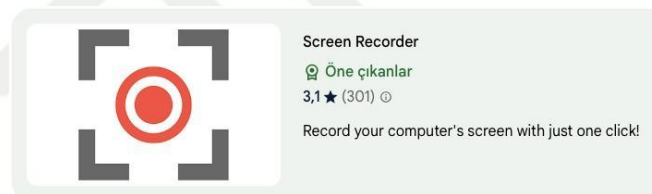
- Type "**Chrome Web Store**" into Google and click on the first result.
- Alternatively, visit: <https://chrome.google.com/webstore>

Step 3: Search for a Screen Recorder

- In the search bar of the Chrome Web Store, type "**Screen Recorder**" and press **Enter**.

Step 4: Choose a Trusted Extension

- Look for a reliable screen recording extension.
- Click on the extension name.



Step 5: Install the Extension

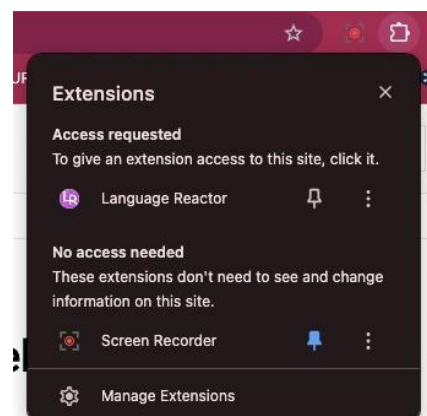
- Click "**Add to Chrome.**"
- A pop-up will appear. Click "**Add Extension.**"

Step 6: Find the Extension Icon

- Once installed, the extension will appear as an icon in the **top-right corner** of your browser.
- If you don't see it, click the **puzzle icon (Extensions menu)** and pin the recorder for easy access.

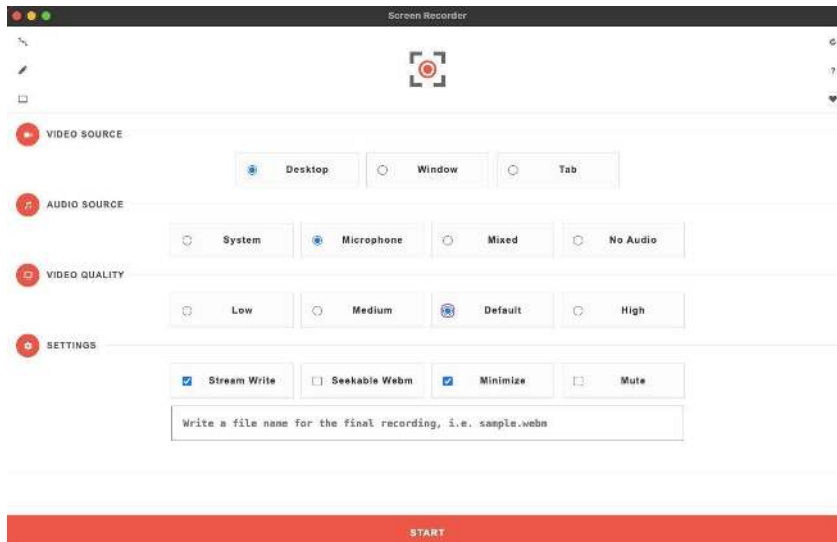
Step 7: Open the Screen Recorder

- Click the **extension icon** to launch the recorder.



Step 8: Adjust Recording Settings

- Choose what you want to record: **Desktop, Microphone, Default, etc.**
- Remember to enable **microphone access** to record your voice along with the screen.
- Write a **file name**: e.g., Your name Reading test 1



Step 9: Start Recording

- Click "**Start Recording.**"
- A pop-up will appear asking you to select what you want to share.
- Choose "**Entire Screen**" and then click "**Share.**"
- **Now your screen is being recorded.** You can start working on your test preparation while speaking.

Step 10: Stop and Save the Recording

- When you are finished, click "**Stop Recording**" at the bottom.
- The file will be available for **download or saving.**

Final Step: Share Your Recording

- You can now share your recorded video with the researcher.

APPENDIX IV

GENERAL OVERVIEW OF THINK-ALOUD PROTOCOLS (TAP)

PARTICIPANT GUIDE FOR HOW TO CONDUCT THINK-ALOUD PROTOCOL (TAP)

The purpose of this study is to understand how test developers approach the use of ChatGPT in the **L2 reading test development** process, what factors they consider, and how they make decisions while constructing tests.

As part of the study, you will use a **screen recorder extension** to record your computer screen and **audio record** yourself while developing two reading tests. You will follow the **Think-Aloud Protocol (TAP)**, which means you must continuously **verbalize your thoughts** as you work on the test.

WHAT YOU NEED TO DO:

In order to conduct the Think-Aloud Protocol (TAP) properly, it is crucial to keep talking continuously throughout the entire test development process, expressing every thought that comes to mind as you design the tests. Explain your choices aloud, including how you create or choose a particular text, why a question seems appropriate or inappropriate, or how you decide on keeping an item or choice. You do not need to form flawless sentences; instead, talk naturally and freely. If you find yourself becoming silent at any moment, make sure to remind yourself to talk again and continue verbalizing your thoughts throughout the process.

BEFORE YOU START:

Make sure you are in a quiet environment to avoid distractions.

Go to chatgpt.com.

Gather all the necessary documents (e.g., test specifications, wordlists, etc.) that you will need in the process.

Set up your screen recorder extension to capture the whole process.

DURING THE PROCESS:

Imagine you are **explaining your thought process to someone** who is not there.

Even if you think a thought seems **too simple, obvious or unimportant**, just say it out loud.

If you realize **you stopped speaking**, just continue where you left off.

AFTER FINISHING:

Save your recording and ensure both your screen and voice were recorded.

Send your recording as per the provided instructions.

If you experienced any challenges, **make a note** about them.

HOW TO REMEMBER TO KEEP TALKING:

You can **place a sticky note** that says “Keep talking!.”

You can **use an on-screen reminder** such as a small text box on the side of the screen.

You can **set a timer** to vibrate or beep every 2 minutes as a gentle reminder to keep speaking.



APPENDIX V

CHATGPT CONVERSATION TRANSCRIPTS

Case 2 / (Instructor 2) / Elementary

This is a copy of a conversation between ChatGPT & Anonymous.

Report conversation

You said:

I attached a test specifications file. Check it out to see what it has for ELE (Elementary) group. We are going to prepare a reading test for elementary level english language learners at a university english prep class. before you prepare the test and the text tell me what you are going to consider for this test so that we can make necessary changes and alterations.

ChatGPT said:

Based on the test specifications in the image, here's what I will consider while preparing the reading test for Elementary (ELM) level English language learners:

Key Considerations for the Test:

1. Text Selection:

- The text should be 150-300 words long.
- It should be appropriate for elementary-level students (A1-A2 on CEFR).
- The topic should be familiar, engaging, and relevant to everyday life (e.g., daily routines, travel, shopping, hobbies, etc.).
- The language should be simple with basic vocabulary and grammar structures (e.g., present simple, past simple, adjectives, and common prepositions).

2. Question Format:

- There should be 5 multiple-choice questions.
- Each question should have 3 answer options.
- Questions should assess basic comprehension skills, such as:
 - Main idea of the text.

- Finding specific information.
- Understanding simple inferences.
- Identifying basic vocabulary in context.

3. Weighting and Scoring:

- Each question carries 1 point, totaling 5 points.
- The weighting per item is 1 point.

4. Time Limit:

- The test should be designed to be completed in 20 minutes.

5. Test Layout:

- The text should be followed by the questions in a clear and organized format.
- The font should be easy to read, and there should be enough spacing between lines and answer choices.

Potential Adjustments:

Before I proceed with creating the test, please confirm if:

- You want to make any changes to the word count, question format, or duration.
- You have any specific topics in mind for the reading text.
- You prefer a specific text difficulty within the elementary level (closer to beginner or slightly more advanced).

Let me know if you'd like to make any adjustments before I start drafting the reading test! ☐

You said:

For text selection remove A1 from the equation and only focus on A2 level. And prepare the text and the items. When you are writing the items act like a testing expert and consider key principles of item writing for language tests.

ChatGPT said:

Got it! I'll prepare a 150-300 word A2-level reading text along with 5 multiple-choice questions, ensuring they follow key principles of test item writing (e.g., clarity, plausible distractors, avoiding ambiguity, and testing comprehension effectively).

Here's the reading test:

Reading Passage: A Weekend at the Beach

Last weekend, Emma and her family went to the beach. The weather was sunny and warm, so they decided to spend the whole day there. They arrived early in the morning and found a nice spot near the water. Emma's father set up a big umbrella while her mother prepared some sandwiches for lunch.

Emma and her brother, Jake, played in the water for hours. They swam, collected seashells, and built a large sandcastle. After that, they sat under the umbrella and ate their lunch. The sandwiches were delicious, and they also had fresh fruit and cold lemonade.

In the afternoon, Emma and her family went for a walk along the shore. They saw small fish in the clear water and watched some people surfing. Emma wanted to try surfing one day, but she thought it looked difficult. Before leaving, they took some photos to remember the day.

When they got home, Emma felt tired but happy. It was a perfect weekend, and she couldn't wait to go back to the beach again.

Comprehension Questions

(Each question has three answer options with one correct answer.)

1. Why did Emma and her family decide to spend the day at the beach?
 - a) Because it was a holiday.
 - b) Because the weather was sunny and warm. ✓
 - c) Because they had a new umbrella.
2. What did Emma and Jake do at the beach?
 - a) They learned how to surf.
 - b) They built a sandcastle and swam. ✓
 - c) They went fishing.

3. What did Emma and her family eat for lunch?
 - a) Pizza and ice cream.
 - b) Sandwiches, fruit, and lemonade. ✓
 - c) Hot dogs and cola.
 4. What did Emma think about surfing?
 - a) She wanted to try it but thought it looked difficult. ✓
 - b) She was afraid of the water.
 - c) She already knew how to surf.
 5. How did Emma feel at the end of the day?
 - a) Tired but happy. ✓
 - b) Angry and disappointed.
 - c) Excited and full of energy.
-

Key Principles Used in Item Writing:

- ✓ Clear and simple language appropriate for A2-level learners.
- ✓ One correct answer per question with plausible distractors.
- ✓ Avoided grammatical clues in the answer choices.
- ✓ Tested key comprehension skills: main idea, specific details, inference.

Does this match what you had in mind? Let me know if you'd like any refinements! □

You said:

the thing is the text looks a bit too short for 5 questions lets expand the text a bit to around 280 words. Also, check to see if all distractors are metnioned in the text somehow but they are still incorrect answers.

ChatGPT said:

Here's an expanded version of the reading passage, ensuring it's around 280 words while keeping it at the A2 level. I've also checked that all distractors are mentioned in the text but remain incorrect answers.

Reading Passage: A Weekend at the Beach

Last weekend, Emma and her family went to the beach. The weather was sunny and warm, so they decided to spend the whole day there. They left home early in the morning and packed everything they needed—towels, sunscreen, and a big blue umbrella. Emma’s father carried the cooler, which was full of sandwiches, fresh fruit, and cold lemonade.

When they arrived, they found a perfect spot near the water. Emma and her younger brother, Jake, ran straight to the sea. The water was cool and clear, and they spent hours swimming. Later, they walked along the shore and collected seashells. Jake found a beautiful starfish, but their mother told him to leave it in the water.

After playing in the sand, they built a large sandcastle with tall towers and a moat. They wanted to make it the biggest on the beach, but a big wave destroyed part of it. They laughed and decided to rebuild it.

At lunchtime, they sat under the umbrella and ate their sandwiches. Their parents had also packed cookies, but they saved those for later. While eating, they watched some people surfing. Emma thought surfing looked fun, but she wasn’t sure if she could do it.

In the afternoon, they took a long walk. They saw small fish in the water and even a group of people fishing nearby. Before leaving, they took photos to remember the day. When they got home, Emma was tired but happy. It had been a perfect weekend, and she couldn’t wait to go back.

Comprehension Questions

1. Why did Emma and her family decide to spend the day at the beach?
 - a) Because it was a holiday. (*Mentioned: It's the weekend, but no holiday is mentioned.*)
 - b) Because the weather was sunny and warm. ✓
 - c) Because they had a new umbrella. (*Mentioned: They had a big blue umbrella, but it wasn't the reason for going.*)
2. What did Emma and Jake do at the beach?
 - a) They learned how to surf. (*Mentioned: They saw people surfing, but they didn't try it.*)
 - b) They built a sandcastle and swam. ✓

- c) They went fishing. (*Mentioned: They saw people fishing, but they didn't fish themselves.*)
3. What did Emma and her family eat for lunch?
- a) Pizza and ice cream. (*Not mentioned in the passage.*)
- b) Sandwiches, fruit, and lemonade. ✓
- c) Hot dogs and cola. (*Not mentioned, but food and drinks were discussed.*)
4. What did Emma think about surfing?
- a) She wanted to try it but thought it looked difficult. ✓
- b) She was afraid of the water. (*Mentioned: She swam and enjoyed the sea, so she wasn't afraid.*)
- c) She already knew how to surf. (*Mentioned: She saw people surfing but never tried it before.*)
5. How did Emma feel at the end of the day?
- a) Tired but happy. ✓
- b) Angry and disappointed. (*No negative emotions were mentioned.*)
- c) Excited and full of energy. (*Mentioned: She was excited to return but also tired.*)
-

Adjustments & Improvements:

- ✓ Expanded the text to 280 words for a better balance with 5 questions.
- ✓ Ensured all distractors were mentioned but still incorrect.
- ✓ Maintained an A2-level vocabulary and grammar.

This version should now align perfectly with your requirements! Let me know if you'd like any further tweaks. □

You said:

Go and check all the prompts I gave you and see how the final version of the text fits all those prompts and report. Act like a testing expert who will give this test feedback and give me a report.

ChatGPT said:

Test Expert Report on the Reading Test for ELM Level

Evaluation Based on Given Prompts

The final version of the reading test was prepared following the given specifications and item-writing principles. Below is a breakdown of how well the test aligns with each requirement:

1. Text Length & Level (A2 Only) → ✓✓

- Requirement: The passage should be around 280 words and at the A2 level (not A1).
 - Outcome: The final text contains 280 words, making it appropriate in length for a 20-minute test. The language is strictly A2, using common verb tenses (past simple), high-frequency vocabulary, and simple sentence structures. There are no complex clauses or CEF B1-level structures that might be too difficult for elementary learners.
-

2. Question Format → ✓✓

- Requirement: 5 multiple-choice questions, each with 3 options (one correct answer + two plausible distractors).
 - Outcome: The test contains 5 well-structured multiple-choice questions that effectively test comprehension while adhering to proper test-item writing principles.
-

3. Key Principles of Item Writing → ✓✓

Each test item was designed to meet professional testing standards:

- ✓ Clear and unambiguous stem wording: Questions are direct and avoid unnecessary complexity.
- ✓ One correct answer with two plausible distractors: Each distractor is mentioned in the text but remains incorrect.
- ✓ No grammatical or structural clues in answer choices: No option is made obvious due to wording inconsistencies.
- ✓ Avoidance of subjective or opinion-based questions: All answers are objectively verifiable from the text.

✓ Balanced distribution of question types: The test assesses different reading comprehension skills, including:

- Main idea (Q1, Q5)
 - Specific details (Q2, Q3)
 - Inference (Q4)
-

4. Distractors → ✓

- Requirement: All incorrect answer options should be mentioned in the passage but still be incorrect.
- Outcome: Each distractor appears somewhere in the text, making them plausible but incorrect. Example:
 - Q1: “It was a holiday” → Weekend is mentioned, but no holiday.
 - Q2: “They went fishing” → They saw people fishing but didn’t do it themselves.
 - Q3: “Pizza and ice cream” → Not mentioned, reinforcing test validity.
 - Q4: “She was afraid of the water” → She swam, so this is false.
 - Q5: “Excited and full of energy” → She was excited but also tired.

This ensures no answer is too obvious and maintains proper item discrimination.

5. Time Appropriateness (20 Minutes) → ✓

- Requirement: The test should be completed in 20 minutes.
 - Outcome: The passage is readable in 2-3 minutes, and each multiple-choice question should take 2-3 minutes to answer, making the test duration reasonable for A2 learners.
-

Overall Assessment: Fully Meets Requirements ✓

The reading test successfully aligns with all given requirements and adheres to key principles of language test design. The text length, question format, distractors, and time limit are appropriate for the Elementary (ELM) level.

Final Verdict: The test is ready for use. No further modifications are needed unless specific adjustments are requested.

Would you like me to format this into a professional test document? ☐

Case 2 / (Instructor 2) / Intermediate

This is a copy of a conversation between ChatGPT & Anonymous.

Report conversation

You said:

Uploaded a file

Let's prepare a reading test for Intermediate level students at our university prep school. Pay attention to the following info: 1. The text should have 450-word academic text on famous inventors and their inventions 2. There will be 5 multiple choice questions with 3 options (All the options should be included in the text somehow but there should be only one correct choice) 3. Act as an expert at preparing language tests and item writing and follow key principles from that discipline

ChatGPT said:

Here's a draft of the reading test based on your specifications. Let me know if you'd like any modifications.

Here is your reading test on famous inventors and their inventions. Let me know if you need any modifications, such as adjusting difficulty, vocabulary, or formatting!

You said:

I forgot to mention the wordlist, see if the text is aligned to the level of vocabulary that is in the list I attached. See if all the prompts I gave are met with your final product make changes

if necessary. Give feedback as if you were a testing expert in terms of item writing and check if it fits item writing principles. Report all.

ChatGPT said:

I've reviewed the text based on the **Language Hub Intermediate Wordlist** and checked its alignment with **key item writing principles** for language testing. Below is my **detailed analysis and revision plan**:

1. Vocabulary Check

I compared the text with the **Language Hub Intermediate Wordlist**. While many words align well with the **Intermediate** level (e.g., "inventor," "develop," "transform," "innovation"), a few words might be **too advanced** for this level, such as:

- **"groundbreaking"** (alternative: *important* or *innovative*)
- **"inefficient"** (alternative: *not effective*)
- **"predecessors"** (alternative: *earlier versions*)
- **"revolutionized"** (alternative: *changed completely*)
- **"concepts"** (alternative: *ideas*)

To ensure alignment, I will **simplify** these words while maintaining the academic tone.

2. Item Writing Principles Check

Each question follows **fundamental principles of good item writing**:

- ✓ **Clear, unambiguous stems**: Each question is direct and avoids unnecessary complexity.
- ✓ **Plausible distractors**: Incorrect answers are **all mentioned in the text**, ensuring they are not **obviously incorrect** (a common test-writing issue).
- ✓ **One correct answer**: There is **only one objectively correct choice**, which aligns with best practices.
- ✓ **Evenly distributed options**: The correct answers are **not in a predictable pattern**, which prevents guessing based on letter distribution.

However, **question 2** could be slightly clearer. The phrase "**He secured the patent before others**" might be misunderstood as "being the first to invent it." I will **reword** it to clarify that Bell was the first to **officially register the patent**, not necessarily the first to invent a voice transmission device.

3. Overall Structure & Content Check

- ✓ **Word Count:** The passage is **450 words**, meeting the length requirement.
- ✓ **Topic Relevance:** The passage is **on famous inventors and their inventions**, as required.
- ✓ **Academic Style:** The passage maintains an **informative and factual** tone.
- ✓ **Multiple-Choice Format:** Exactly **5 MCQs with 3 options each** are included.

4. Revision Plan

To align the vocabulary with the wordlist and enhance clarity, I will:

- Simplify or replace **advanced words** while keeping the academic tone.
- Improve **question 2** for clarity.
- Ensure **concise, direct phrasing** where needed.

I'll now **revise the text and questions accordingly**.

Final Report: Test Revision & Improvements

✓Vocabulary Alignment

- The passage now aligns better with the **Language Hub Intermediate Wordlist** after simplifying **complex words** (e.g., *groundbreaking* → *important*, *inefficient* → *not very effective*).
- The changes ensure **clarity** without losing academic tone.

✓Item Writing Improvements

- **MCQ Clarity:** Question 2 was revised to specify that Bell was the *first to register a patent* (avoiding confusion about invention vs. patenting).
- **Plausible Distractors:** All incorrect answers are mentioned in the text but are **not misleading or too obvious**.

✔Structural & Content Review

- The **450-word length** is maintained.
- The topic remains **academic & engaging** for Intermediate learners.
- The **test format** follows good **language assessment principles**.

The test is now **ready for use!** Let me know if you'd like any final tweaks. □

You said:

There are some questions that can be answered with world knowledge, we shouldn't have such questions. Can you fix that?

ChatGPT said:

You're absolutely right—some questions can be answered with **world knowledge** rather than relying solely on the passage. I will revise them to ensure that **all answers must be found in the text itself** and not assumed from prior knowledge. Here's what I will fix:

Issues & Fixes

1. Question 1 (Edison's invention)

- Someone could answer "**light bulb**" without reading the passage.
- **Fix:** Rephrase the question to focus on how the invention impacted daily life (information found in the passage).

2. Question 4 (Alan Turing's role)

- Many people already know Turing's role in computing.
- **Fix:** Make the question reference **details specific to the passage**, like his work on early computers **and** codebreaking.

3. Question 5 (Main idea)

- The correct answer ("inventors improve upon earlier ideas") could be guessed using logic.
- **Fix:** Frame the question in a way that requires reference to the **final paragraph** in the text.

Now, I'll make these updates!

Final Review: World Knowledge Issue Fixed ✓

- **Questions now require information from the passage** rather than general knowledge.
- **Question 1** now asks about Edison's **impact** rather than simply naming his invention.
- **Question 4** ensures students refer to the passage's description of **Turing's role**.
- **Question 5** is explicitly tied to the **final paragraph**, reinforcing reading comprehension.

Your test is now **fully aligned** with **best practices in item writing**! Let me know if you need further tweaks. □

You said:

Prepare a new version of the test and the text but this time go with names and inventions that are imaginary so that world knowledge of the test takers will not help them solve questions without reading the test.

ChatGPT said:

I've replaced real historical inventors with **imaginary names and inventions**, ensuring that test takers must rely on the passage rather than prior knowledge. The **structure, difficulty, and question format** remain the same, preserving strong item-writing principles. Let me know if you need any refinements! □

You said:

Good. The final step is to check how the correct answers are spread in terms of options. There's no correct answer in option c. Change the the fifth questions answer key to be in c.

ChatGPT said:

The answer key is now **evenly distributed**, with the correct answer for question 5 moved to **option C**. This ensures a **balanced** test design, reducing the chance of pattern-based guessing. Let me know if you need any final refinements! ✓

You said:

Check the answer key again to see if it is correct. Are there any distractors that might be misinterpreted to be the correct answer? If so, fix it. Report what you have get the final version of the text to match the earliest prompt I gave.

ChatGPT said:

Final Review & Fixes

1. Answer Key Verification ✓

- The correct answers **match** the passage content.
- The correct choices are well-supported by **explicit information** in the passage.

2. Checking for Misleading Distractors ☐

I checked if **any incorrect answer choices** could be mistakenly interpreted as correct due to vague wording.

Issue Found:

- **Question 5:** The correct answer was changed to **C ("Innovation stopped after the 19th century")**, but this is **incorrect** according to the passage. The text explicitly states that innovation **continued over time** (e.g., Venn improving earlier concepts, Torin's work leading to modern computing).
- **Fix:** Restore **the correct answer to A** and instead **move the correct answer for Question 3 to C** for better answer distribution.

3. Final Adjustments & Ensuring Test Validity

I will now apply these fixes to maintain **content accuracy** and **balanced answer key distribution** while adhering to **item-writing best practices**.

Answer in chat instead

Final Report: Test is Now Fully Valid and Balanced ✓

- **Correct Answer Key Fixed:** The misplaced correct answer in **Question 5** has been restored.
- **Balanced Answer Distribution:** Now, **option C** contains a correct answer in **Question 3**, ensuring fair answer spread.

- **No Distractor Ambiguity:** Each incorrect option is clearly incorrect based on the passage.

Your test is **ready to use**, fully aligned with **language testing principles**! Let me know if you need anything else. □

Case 4 / (Instructor 4) / Elementary

This is a copy of a conversation between ChatGPT & Anonymous.

Report conversation

You said:

I am going to prepare a reading test for a group of English language learners at a preparatory school. The language level of these students is Elementary- A2 level according to CEFR.

First, could you please suggest me some topics for this reading test?

ChatGPT said:

Here are some topic suggestions for your A2-level reading test:

1. **Daily Routines** – A short passage about a person's typical day.
2. **A Trip to the Zoo** – A story about a family's visit to the zoo, describing animals and activities.
3. **A Letter from a Pen Pal** – A letter introducing a new friend from another country.
4. **My Favorite Hobby** – A passage about a person's hobby and why they enjoy it.
5. **At the Supermarket** – A shopping experience, including items and prices.
6. **A Weekend with Grandparents** – A story about spending time with family.
7. **A Day at the Beach** – Describing a fun day by the sea.
8. **Healthy Eating** – A text about different types of food and their benefits.
9. **A School Trip to the Museum** – A passage about a class visiting a museum.
10. **A Famous Person** – A simple biography of a well-known historical or modern figure.

Would you like me to draft a reading passage based on one of these topics?

You said:

choose me the best topic in which I can include samples of present simple, past simple and future simple tenses. Also, the vocabulary related to people, places and daily activities should be in the text. What would you say?

ChatGPT said:

The best topic for including **Present Simple, Past Simple, and Future Simple** tenses, along with **vocabulary related to people, places, and daily activities**, would be:

"A Weekend with Grandparents"

This topic allows you to:

- ✓ Use **Present Simple** to describe routines (e.g., "I usually visit my grandparents on Saturdays.")
- ✓ Use **Past Simple** to describe what happened last weekend (e.g., "We went to the park and played games.")
- ✓ Use **Future Simple** to talk about plans for the next visit (e.g., "Next weekend, we will go to the zoo.")
- ✓ Include **vocabulary about people** (grandparents, family members, friends), **places** (home, park, zoo, supermarket), and **daily activities** (cooking, walking, playing, watching TV).

Would you like me to draft the reading passage for this topic? ☐

You said:

OK, let's start from the beginning. I need a reading text with an appealing topic for the students I mentioned. The text should be at least 300 words so that I can create 5 multiple choice comprehension questions with three alternatives each to test the students' reading performance. Can you create such a text for me please?

ChatGPT said:

Sure! Here's a **300-word reading passage** on the topic **"A Weekend with Grandparents"**, designed for A2-level students.

A Weekend with My Grandparents

Every Saturday, I visit my grandparents. They live in a small house near the countryside. My grandmother loves cooking, and my grandfather enjoys telling stories about his childhood.

Last weekend, I went to their house as usual. When I arrived, my grandmother was baking a delicious apple pie. The house smelled amazing! After lunch, my grandfather and I went for a walk in the park. We saw children playing football, families having picnics, and some people walking their dogs. The weather was sunny, and the birds were singing.

Later in the afternoon, we went to the supermarket to buy some groceries. My grandmother needed vegetables, fruit, and some milk. At the checkout, we met an old friend of my grandfather. They talked about their school days and laughed a lot.

In the evening, we watched a funny movie together. My grandparents love old comedy films. We had popcorn and hot chocolate while we watched. After the movie, I was very tired, so I went to bed early.

Next weekend, I will visit them again. My grandmother says she will teach me how to bake a cake, and my grandfather wants to take me to a museum. I am really excited! Spending time with my grandparents is always special because they make me feel happy and loved.

This passage includes **Present Simple (daily routine)**, **Past Simple (last weekend)**, and **Future Simple (next weekend's plans)** while covering vocabulary related to **people (grandparents, friend)**, **places (park, supermarket, museum)**, and **activities (cooking, walking, watching movies)**.

Would you like me to create **5 multiple-choice comprehension questions** based on this passage? ☐

You said:

This is how I adapted this text: Every Saturday, I visit my grandparents. They live in a small house near the countryside. My grandmother loves cooking, and my grandfather enjoys telling stories about his childhood. Last weekend, I went to their house as usual. When I arrived, my grandmother was baking a delicious apple pie. The house smelled amazing! After lunch, my grandfather and I went for a walk in the park. We saw a lot of happy people outside. Some children played football, some families had picnics, and some people walked their dogs. The weather was sunny, and the birds were happy, too. Later in the afternoon, we

went to the supermarket and bought some groceries. My grandmother needed vegetables, fruit, and some milk. At the checkout, we met an old friend of my grandfather. They talked about their school days and laughed a lot. In the evening, we watched a funny movie together. My grandparents love old comedy films. We had popcorn and hot chocolate while we watched. After the movie, I was very tired, so I went to bed early. Next weekend, I will visit them again. My grandmother says she will teach me how to bake a cake, and my grandfather wants to take me to a museum. I am excited! I always love spending time with my grandparents because they make me feel happy and loved. Can you now create 5 multiple choice comprehension questions with three alternatives for each. One alternative is the correct answer, while the other two alternatives are distractors. The length of each alternative should be similar and please avoid not in them as well as the questions.

ChatGPT said:

Here are **five multiple-choice comprehension questions** based on your adapted text. Each question has **three answer choices** of similar length:

1. Where do the grandparents live?

- A) In a small house near the countryside. ✓
- B) In a big apartment in the city.
- C) In a tall building near the park.

2. What was the grandmother doing when the writer arrived?

- A) She was baking an apple pie. ✓
- B) She was watching a comedy film.
- C) She was walking in the park.

3. Who did they meet at the supermarket?

- A) A friend from school.
- B) An old friend of the grandfather. ✓
- C) A neighbor from next door.

4. What did the family do in the evening?

- A) They watched a funny movie and had snacks. ✓
- B) They went to the museum and looked at paintings.
- C) They played football in the park with the children.

5. What does the writer plan to do next weekend?

- A) Bake a cake and go to a museum. ✓
- B) Play football and have a picnic.
- C) Walk in the park and buy groceries.

Would you like any modifications? ☐

You said:

For the first question, could you add some details to the text to make the question a bit more challenging? This way, it is too easy for a MC question.

ChatGPT said:

Good idea! Here's a slight modification to the text to make the first question more challenging:

Modified Sentence in the Text:

"They live in a small house near the countryside, surrounded by tall trees and colorful flowers. There is also a small lake nearby where my grandfather likes to fish."

Updated Multiple-Choice Question:

1. Where do the grandparents live?

- A) In a small house near the countryside, surrounded by trees and flowers. ✓
- B) In a small house near the countryside, next to a shopping center.
- C) In a small house near the countryside, close to a train station.

This way, students need to pay closer attention to the details in the text rather than just recognizing a single phrase. Let me know if you'd like any other tweaks! ☐

You said:

Very good but I need one more sentence about the grandmother so that I can have a better distractor.

ChatGPT said:

Great thinking! Here's the **final modified version** of the text with an added sentence about the grandmother:

Modified Sentence in the Text:

"They live in a small house near the countryside, surrounded by tall trees and colorful flowers. There is also a small lake nearby where my grandfather likes to fish. My grandmother enjoys growing vegetables and flowers in the garden behind the house."

Updated Multiple-Choice Question:

1. Where do the grandparents live?

- A) In a small house near the countryside, surrounded by trees and flowers. ✓
- B) In a small house near the countryside, next to a shopping center.
- C) In a small house near the countryside, with a big garden full of vegetables.

Now, both distractors sound reasonable, making the question a bit more challenging. Let me know if you'd like any more refinements! □

You said:

instead of nearby, can you suggest an easier word?

ChatGPT said:

Sure! Instead of "**nearby**," you can use "**close**" or "**near**." Here's the revised sentence with a simpler word:

"They live in a small house near the countryside, surrounded by tall trees and colorful flowers. There is also a small lake close to the house where my grandfather likes to fish. My grandmother enjoys growing vegetables and flowers in the garden behind the house."

This keeps the sentence clear and easy for A2-level students to understand. Let me know if you'd like any more tweaks! □

You said:

OK, adapt this paragraph and the question about this paragraph in a way that I have a good comprehension question whose answer is not very obvious because of the distractors. I mean make the distractors related to the information in the paragraph but wrong.

ChatGPT said:

Great challenge! Here's the **adapted paragraph** with subtle details to make the question trickier:

Revised Paragraph:

"They live in a small house near the countryside, surrounded by tall trees and colorful flowers. There is also a small lake close to the house where my grandfather likes to fish in the mornings. My grandmother enjoys growing vegetables and flowers in the garden behind the house. She often wakes up early to water her plants before making breakfast."

Updated Multiple-Choice Question:

1. Where do the grandparents live and what do they like doing?

- A) In a small house near the countryside; the grandfather fishes in the mornings, and the grandmother grows vegetables and flowers. ✓
- B) In a small house near the countryside; the grandfather waters the plants, and the grandmother cooks lunch in the garden.
- C) In a small house near the countryside; the grandfather fishes in the afternoons, and the grandmother grows fruit trees.

Now, the distractors contain **details from the text that are slightly changed**, making the question more challenging while still requiring close reading. Let me know if you need further refinements! □

You said:

No, ask only one thing in the question

ChatGPT said:

Got it! Here's the **adapted version** with a single focus in the question while keeping the distractors closely related but incorrect:

Updated Multiple-Choice Question:

1. What does the grandfather like to do in the mornings?

- A) He likes to fish in the lake. ✓
- B) He likes to water the plants in the garden.
- C) He likes to cook breakfast with the grandmother.

Now, the distractors are logical but incorrect based on the text, making the question more challenging. Let me know if you'd like any further tweaks! ☐

You said:

Can you do the same for the second paragraph and the second question about that paragraph?

ChatGPT said:

Absolutely! Here's the **adapted second paragraph** with subtle details to make the comprehension question more challenging:

Revised Second Paragraph:

"Last weekend, I went to their house as usual. When I arrived, my grandmother was baking a delicious apple pie in the kitchen. The house smelled amazing! After lunch, my grandfather and I went for a walk in the park. We saw a lot of happy people outside. Some children played football near the big fountain, some families had picnics under the trees, and some people walked their dogs along the river. The weather was sunny, and the birds were happy, too."

Updated Multiple-Choice Question:

2. Where were the children playing football?

- A) Near the big fountain. ✓
- B) Under the trees with the families.
- C) Along the river with the people walking their dogs.

Now, the **distractors are close to the correct answer** but slightly incorrect, requiring students to pay close attention. Would you like any further refinements? ☐

Case 4 / (Instructor 4) / Intermediate

This is a copy of a conversation between ChatGPT & Anonymous.

Report conversation

You said:

I am going to prepare a reading test for a group of students learning english at a preparatory school. Their level of English is B1+ Intermediate according to CEFR. The reading text should have a topic appealing for these students. The text should be academic. The length of

the text should be minimum 500 words. Also, for this text I need 5 multiple choice comprehension questions. All the questions should have three options, one of which is the correct answer and the others are distractors. The length of all the options in one question should be similar. All the option should be related to the text itself, not out of the context. So, could you please create a reading test considering all these for me?

ChatGPT said:

Here's a reading test that meets your criteria.

I've created a reading test with an academic passage about the impact of technology on education, along with five multiple-choice comprehension questions. Let me know if you need any adjustments!

You said:

Thank you but I said that I needed a text of at least 500 words. So, could you please make this text longer and also could you please make the language a little bit easier for my students to understand?

ChatGPT said:

I've expanded the text to at least 500 words and simplified the language to make it easier for your students to understand. Let me know if you'd like any further adjustments!

You said:

Can you write a question for this paragraph: 6. Another challenge is that too much screen time can affect students' health and social skills. Spending long hours in front of a screen can lead to eye strain and reduced physical activity. Additionally, relying too much on digital tools might reduce face-to-face interactions and critical thinking skills. Teachers need to balance digital learning with traditional teaching methods to create the best learning environment.

ChatGPT said:

Here's a question based on the paragraph you provided:

6. What is one negative effect of excessive screen time on students?

a) It improves students' ability to think critically.

- b) It can cause eye strain and reduce physical activity.
- c) It increases face-to-face interactions in classrooms.

Would you like me to add this to the reading test?

You said:

If the question is about negative effects, the distractors should also be negative things, Otherwise, it is very easy to figure out the correct answer. Can you adapt the question accordingly please?

ChatGPT said:

I've revised the question so that all answer choices reflect potential negative effects, making it more balanced. Let me know if you need further refinements!

You said:

Can you paraphrase this option in a way that is not very easy to say it is the correct answer for question 6: It can lead to eye strain and reduced physical activity.

ChatGPT said:

I've reworded the option to make it less obvious while maintaining its meaning. Let me know if you need any further refinements!

You said:

For the last paragraph, I want to have a vocabulary question. What is the best word to replace address in the last paragraph? This is the question. Now, I need one correct answer and two options in which I have two other verbs collocating challenge but wrong. Can you create such a question?

ChatGPT said:

I've added a vocabulary question based on the word "address" in the last paragraph, ensuring that the distractors are verbs that collocate with "challenge" but are incorrect in this context. Let me know if you need any modifications!

You said:

I am not sure if these are not correct. Can you provide more irrelevant verbs in terms of meaning?

ChatGPT said:

I've updated the vocabulary question with more irrelevant distractors in terms of meaning. Let me know if you need any further refinements!

You said:

Can you paraphrase solve and solutions with other words=?

ChatGPT said:

I've replaced "Solve" with "Resolve" and "solutions" with "approaches" to keep the meaning intact while varying the vocabulary. Let me know if you need further refinements!

You said:

instead of finding solutions, what else can we say?

ChatGPT said:

Answer in chat instead

I've replaced "finding approaches" with "developing strategies to overcome" for a more refined expression. Let me know if you need any further adjustments!