İSTANBUL UNIVERSITY-CERRAHPAŞA
INSTITUTE OF GRADUATE STUDIES

M.Sc. THESIS

THE EFFECT OF VIRTUAL REALITY ON STUDENTS' SPEAKING PERFORMANCE IN EFL CLASSES

Sümeyra Mürvet AYDIN

SUPERVISOR
Assoc. Prof. Dr. Tuncer CAN

Department of Foreign Language Education

English Language Teaching Programme

January, 2026

THESIS ACCEPTANCE AND APPROVAL

This thesis titled **"The Effect of Virtual Reality on Students' Speaking Performance in EFL Classes"**, prepared by **Sümeyra Mürvet AYDIN** under the supervision of **Assoc. Prof. Dr. Tuncer CAN** , was accepted as a **M.Sc. Thesis** after being **consensus** approved by our jury as a result of the exam held on **30/01/2026**.

Examining Committee Members

		Signature	Decision
SUPERVISOR	Assoc. Prof. Dr. Tuncer CAN İstanbul University-Cerrahpaşa Foreign Language Education Department		☐ Accept ☐ Reject
MEMBER	Assist. Prof. Dr. Gülay KIRAY İstanbul University-Cerrahpaşa Foreign Language Education Department		☐ Accept ☐ Reject
MEMBER	Assist. Prof. Dr. Cemil Gökhan KARACAN Beykent University English Language and Literature Department		☐ Accept ☐ Reject

To my loving family, for their endless support and encouragement …

FUNDING

THE EFFECT OF VIRTUAL REALITY ON STUDENTS' SPEAKING PERFORMANCE IN EFL CLASSES

No financial support was received from any institution for this thesis.

ACKNOWLEDGEMENT

First of all, I would like to express my deepest gratitude to my thesis supervisor, Assoc. Prof. Dr. Tuncer CAN, who has guided and supported me throughout this journey and far beyond the scope of this thesis. Having known him since my undergraduate years, I have continuously benefitted from his guidance as a teacher, researcher, and most importantly, as a human being. He taught me to trust myself and approach research with confidence and integrity, while always offering support when needed. His belief in me and his mentorship shaped both this thesis and my academic growth.

I would also like to sincerely thank my thesis committee members, Assist. Prof. Dr. Gülay KIRAY and Assist. Prof. Dr. Cemil Gökhan KARACAN, for their valuable contributions to this study. I am especially grateful to Assist. Prof. Dr. Gülay KIRAY, whom I have also known since my undergraduate years, for her insightful feedback and ability to provide the most accurate and timely advice that have been invaluable. I would also like to specifically thank Assist. Prof. Dr. Cemil Gökhan KARACAN for offering crucial support and guidance at moments when it was most needed.

I would like to express my sincere gratitude to Scott Cardwell, the co-founder of ImmerseMe, for providing complimentary licensing support that enabled this MA thesis to be conducted within a high-quality virtual reality environment.

I extend my most sincere thanks to Dr. Halil KARAGÜLLE for his assistance with the quantitative data analysis, which constituted a major component of this study. His expertise and guidance were essential to the rigor and clarity of the quantitative findings.

My heartfelt thanks go to my dearest friend, Dilşad KAPLAN, without whom I truly could not have completed my thesis. Her constant presence, encouragement and unwavering support throughout this process made even the most challenging moments manageable.

Finally, I would like to thank my family and friends for their endless support, patience and understanding throughout this demanding journey. Their belief in me has been a constant source of strength and motivation.

January, 2026 Sümeyra Mürvet AYDIN

CONTENTS

Page

THESIS ACCEPTANCE AND APPROVAL ..... **ii**
DECLARATION ..... **iii**
FUNDING ..... **v**
ACKNOWLEDGEMENT ..... **vi**
CONTENTS ..... **vii**
LIST OF FIGURES ..... **xii**
LIST OF TABLES ..... **xiv**
LIST OF SYMBOLS AND ABBREVIATIONS ..... **xvi**
ÖZET ..... **xvii**
ABSTRACT ..... **xix**
1. INTRODUCTION **1**
1.1. BACKGROUND OF THE STUDY 1
1.2. STATEMENT OF THE PROBLEM 3
1.3. PURPOSE OF THE STUDY 4
1.4. RESEARCH QUESTIONS 4
1.5. SIGNIFICANCE OF THE STUDY 5
1.6. DEFINITIONS OF KEY TERMS 6
2. CONCEPTUAL FRAMEWORK **8**
2.1. SPEAKING SKILLS IN EFL CONTEXTS 8
2.1.1. Importance of Speaking in Language Learning 8
2.1.2. Obstacles Affecting EFL Learner's Speaking Development 9
2.1.3. Proposed Solutions to Developing Speaking Skills 10
2.1.4. Instructional Approaches to Teaching Speaking 11
2.2. ASSESSMENT OF SPEAKING SKILLS 23
2.2.1. Theoretical Framework of Speaking Assessment 23
2.2.2. Assessment and Testing in Language Teaching 23
2.2.3. Types of Assessment 24

2.2.4. Approaches to Language Testing 25
2.2.5. Core Principles of Language Assessment 26
2.2.6. Purposes of Assessment 31
2.2.7. Types of Scoring 31
2.2.8. Types of Scales and Rubrics 32
2.2.9. Conceptual Models of L2 Speaking Assessment 33
2.2.10. Challenges in Assessing Spoken Production 40
2.3. SPEAKING RUBRICS AND CRITERIA 49
2.3.1. Conceptual Base of Spoken Language 49
2.3.2. Common European Framework for Assessing Speaking Skills 50
2.3.3. IELTS Speaking Rubric 55
2.4. VIRTUAL REALITY IN LANGUAGE EDUCATION 59
2.4.1. Definitions of VR and Related Technologies 59
2.4.2. Implications of Using VR in Language Learning 63
2.4.3. Theoretical and System-Level Perspectives on VR in Language Learning 64
2.4.4. Empirical Research on Virtual Reality in Language Learning 67
2.4.5. Validated Instrument on Perceptions of VR for Speaking Skills 72
2.4.6. ImmerseMe 74
2.4.7. Empirical Research Conducted on ImmerseMe 78
2.5. PHILOSOPHICAL FRAMEWORK 81
2.5.1. Post-positivist Approach 81
2.5.2. Constructivist Approach 83
2.5.3. Situated Learning Theory and VR 87
2.6. RESEARCH GAP AND RATIONALE 87
2.6.1. Summary of Key Findings from Literature 87
2.6.2. Identified Gaps in Research 88
2.6.3. Relevance of the Current Study 89
2.7. CONCLUSION 89
3. METHOD 90
3.1. RESEARCH DESIGN 92
3.2. RESEARCHER POSITIONALITY 94
3.3. SETTING AND PARTICIPANTS 95
3.3.1. BILET Exam Procedures 98
3.3.2. Description of Level 4 101

3.3.3. Research Setting 102
3.3.4. Sampling 102
3.3.5. Participant Profile 103
3.3.6. Variables 105
3.4. DATA COLLECTION INSTRUMENTS 106
3.4.1. Demographic Information Survey 106
3.4.2. Pre- and Post-Tests 107
3.4.3. Quantitative Perception Scale 108
3.4.4. Semi-Structured Questionnaire 108
3.5. DATA COLLECTION PROCEDURE 112
3.5.1. Raters 114
3.5.2. Week 1 - Inter-Rater Reliability 116
3.5.3. Week 2 – Calibration Process 117
3.5.4. Weeks 3 and 4 – Pre-Test and ImmerseMe Accounts 117
3.5.5. Weeks 4 to 8 – Intervention 118
3.5.6. Week 9 – Last Week of Intervention and Post-Test 120
3.5.7. Week 10 – Ironsi's (2023) Perception Scale and Quantitative Data Analysis 121
3.5.8. Week 11 – Semi-Structured Questionnaire and Qualitative Data Analysis 121
3.6. RELIABILITY AND VALIDITY 122
3.6.1. Reliability 122
3.6.2. Validity 123
3.6.3. Threats to Internal Validity 125
3.6.4. Trustworthiness 127
3.7. DATA ANALYSIS PROCEDURES 129
3.7.1. Interrater Reliability and Calibration Process 129
3.7.2. Pre- and Post- Test Results 134
3.7.3. Quantitative Perception Scale 136
3.7.4. Semi-Structured Questionnaire 141
3.8. ETHICAL CONSIDERATIONS 147
3.9. LIMITATIONS 148
3.10. SUMMARY 148
4. RESULTS 150
4.1. THE EFFECT OF VR ON STUDENTS' SPEAKING PERFORMANCE 150
4.1.1. Speaking Performance Results of Pre- and Post-Tests 151

4.1.2. Inferential Analysis of Speaking Performance in Pre- and Post-Tests 151
4.1.3. Speaking Subskill and Overall Individual Rater Mean Results of Pre- and Post-Tests 155
4.2. STUDENTS' PERCEPTIONS OF USING VR FOR LEARNING AND SPEAKING ENGLISH 158
4.2.1. Quantitative Results of Students' Perceptions Toward Using VR..... 158
4.2.2. Qualitative Results of Students' Perceptions on the Use of VR for Learning and Speaking English..... 170
4.3. COMPARISON OF STUDENTS' PERCEIVED AND ACTUAL SPEAKING PERFORMANCES..... 180
4.3.1. Quantitative Results of Pre- and Post-Test Scores and Ironsi's (2023) Perception Scale 181
4.3.2. Qualitative Results from Semi-Structured Questionnaires: Students' Descriptions of Their Speaking Performance..... 184
4.4. SUMMARY OF RESULTS 185
5. DISCUSSION..... 186
5.1. DISCUSSION OF FINDINGS IN RELATION TO RQ I: WHAT IS THE EFFECT OF VIRTUAL REALITY ON STUDENTS' SPEAKING PERFORMANCE?..... 187
5.1.1. Pre-test Comparisons Between Groups..... 187
5.1.2. Post-Test Comparison Between Groups 188
5.2. DISCUSSION OF FINDINGS IN RELATION TO RQ II: WHAT ARE THE PERCEPTIONS OF STUDENTS TOWARD USING VR FOR LEARNING AND SPEAKING ENGLISH?..... 193
5.2.1. Discussion of Descriptive Statistics from Ironsi's (2023) Perception Scale..... 193
5.2.2. Discussion of Findings in Relation to RQ ii.i: Are There Any Correlated Items in the Students' Perceptions Toward Using VR for Learning and Speaking? 195
5.2.3. Interpretation of Thematic Analysis in Relation to RQ ii.ii: How Do Students Describe Their Perception toward Using VR for Learning and Speaking? 202
5.2.4. Integrated Interpretation of Quantitative and Qualitative Findings in Relation to RQ ii. 211
5.3. DISCUSSION OF FINDINGS IN RELATION TO RQ III: IS THERE A DIFFERENCE BETWEEN STUDENTS' PERCEPTIONS OF THE EFFECT OF VIRTUAL REALITY ON THEIR SPEAKING PERFORMANCE AND THEIR ACTUAL SPEAKING PERFORMANCE? 213
5.3.1. Interpretation of Students' Individual Perceptions of VR on Their Speaking Performance and Their Actual Speaking Performance Outcomes..... 213
5.4. SUMMARY OF THE DISCUSSION 222
6. CONCLUSION AND RECOMMENDATIONS..... 225

6.1. SUMMARY AND OVERALL INTERPRETATION OF KEY FINDINGS.....225
6.2. PEDAGOGICAL IMPLICATIONS.....228
6.3. LIMITATIONS.....231
6.4. SUGGESTIONS FOR FUTURE RESEARCH.....231
6.5. CONCLUDING REMARKS.....232
REFERENCES.....233
APPENDICES.....255
FIRST PAGE OF THE PLAGIARISM REPORT.....310
ETHICS COMMITTEE PERMISSION.....311
INSTITUTIONAL PERMISSIONS.....313
CURRICULUM VITAE.....315

LIST OF FIGURES

Page

Figure 2.1: The Relationship Between Teaching, Assessment, Measurement, and Evaluation 24

Figure 2.2: The Tri-Layered Concentric Ring Model of The Core Principles of Assessment. 30

Figure 2.3: Skehan's Model of The Assessment of Speaking Performance 34

Figure 2.4: Fulcher's Expanded Model Based on Skehan's Model 35

Figure 2.5: Weir's Socio-Cognitive Framework for Assessing and Validating Speaking Performance 36

Figure 2.6: The Peculiarly Entangled Machinery of Speaking Assessment 39

Figure 2.7: Interaction of Cognitive, Affective, and Task-Related Factors in Speaking Assessment 48

Figure 2.8: Hierarchical Organization of Linguistic Units in Spoken Language 50

Figure 2.9: Conversion Chart of IELTS to CEFR 59

Figure 2.10: Reality-Virtuality Continuum 60

Figure 2.11: The Reality-Virtuality Continuum with Visuals 61

Figure 2.12: Comparative Levels of Immersion, Perception and Telepresence across VR Modalities 62

Figure 2.13: The Interface of ImmerseMe 76

Figure 2.14: The B1.1 Level-Specific Interface of ImmerseMe 77

Figure 2.15: A Role-Play in B1.1 Level of ImmerseMe 77

Figure 3.1: Mapping of Research Questions to Data Collection Instruments 91

Figure 3.2: Phase 1 and Phase 2 Illustration of This Study 93

Figure 3.3: CEFR-Based Assessment Framework for B1 Level Speaking Performance 111

Figure 3.4: The Outline of Data Collection Procedure 112

Figure 3.5: Overview of the Data Analysis Procedure 129

Figure 3.6: A Flowchart of Selecting an ICC 131
Figure 3.7: Qualitative Data Analysis Procedure 141
Figure 3.8: Integration Levels of Methods, and Interpretation and Reporting 146

LIST OF TABLES

Page

Table 2.1: The Chronology of the Conceptual Evolution of Communicative Competence.....16

Table 2.2: The Difference Between Traditional and Alternative Assessment 26

Table 2.3: The Core Principles of Assessment.....27

Table 2.4: Subscales of Sustained Monologue.....52

Table 2.5: Linguistic and Pragmatic Competences of Communicative Language Competence 53

Table 2.6: Band Descriptors of IELTS Speaking Rubric 56

Table 2.7: The Comparison of Post-Positivism and Constructivism; Ontological, Epistemological and Axiological Aspects.....86

Table 2.8: The Comparison of Post-Positivism and Constructivism: Research Designs and Methods 86

Table 3.1: BILET 1 Exam Placement Levels.....100

Table 3.2: Recognized Examinations and Minimum Exemption Scores for BEPP.....101

Table 3.3: The Summary of The Data Collection Procedure 113

Table 3.4: Speaking Activity Content for the Control Group 120

Table 3.5: The Results of ICC for Interrater Reliability and Calibration Process.....123

Table 3.6: Interpretation of ICC values.....132

Table 3.7: Pre- and Post-Test Individual Score Differences Interpretation 136

Table 3.8: Interpretation of the Descriptive Statistics of Ironsi's (2023) Perception Scale Based on Alkharusi's (2022) Class Intervals 137

Table 3.9: Interpretation of SD Ranges of Ironsi's (2023) Perception scale.....138

Table 3.10: Interpretation of Pearson Correlation Coefficient Strength 139

Table 3.11: Categorization of Items in Ironsi's (2023) Perception Scale 140

Table 3.12: Analytic Categorization of Items in Ironsi's (2023) Perception Scale.....140

Table 3.13: Integration Levels within Mixed-Methods Research 142

Table 4.1: Descriptive Statistics for Experiment and Control Groups on Pre- and Post-Test Speaking Scores 151

Table 4.2: Normality Assessment and Selection of Statistical Tests for Pre-Test Scores 152

Table 4.3: Normality Assessment and Selection of Statistical Tests for Post-Test Scores 153

Table 4.4: Independent Samples t-Test Results of the Overall Scores for Pre- and Post-Tests 154

Table 4.5: Independent Samples t-Test for Pre-Test Speaking Measures 155

Table 4.6: Mann-Whitney U Test Results for Pre-Test Speaking Measures 156

Table 4.7: Independent Samples t-Test for Post-Test Speaking Measures 157

Table 4.8: Mann-Whitney U Test Results for Post-Test Speaking Measures 157

Table 4.9: The Overall Descriptive Statistics of Ironsi's (2023) Perception Scale 159

Table 4.10: Mean Scores and Standard Deviation of Ironsi's (2023) Perception Scale 159

Table 4.11: Summary of Pearson Correlation Results for Ironsi's (2023) Perception Scale 162

Table 4.12: Positive Correlation Patterns Across Analytic Categories in Ironsi's (2023) Perception Scale 165

Table 4.13: No Linear Correlation Patterns Across Analytic Categories in Ironsi's (2023) Perception Scale 167

Table 4.14: Negative Correlation Patterns Across Analytic Categories in Ironsi's (2023) Perception Scale 168

Table 4.15: Thematic Analysis of Semi-Structured Questionnaire 170

Table 4.16: Pre- and Post-Test Speaking Scores and Score Differences for the Experiment Group 181

Table 4.17: Individual Students' Mean Score and Standard Deviation of Ironsi's (2023) Perception Scale 182

Table 4.18: Alignment Between Students' Speaking Performance Change and Perception Levels of VR-Based Speaking Instruction 183

LIST OF SYMBOLS AND ABBREVIATIONS

Symbols	Description
p	: Probability value
t	: t-test statistic
df	: Degrees of Freedom
U	: Mann-Whitney U test
z	: Standardized test statistic
M	: Mean
SD	: Standard Deviation
N	: Number of participants
n	: Sample size
<	: Less than
≥	: Greater than or equal to

Abbreviations	Description
VR	: Virtual Reality
LiVR	: Low-Immersion Virtual Reality
HiVR	: High-Immersion Virtual Reality
BEPP	: Bilgi English Preparatory Program
B1/ B1	: Common European Reference of Proficiency Levels
CEFR	: Common European Framework of References for Languages
EFL	: English as a Foreign Language
IELTS	: International English Language Testing System
ICC	: Intraclass Correlation Coefficient
IRR	: Interrater Reliability
RQ	: Research Question
SLT	: Situated Learning Theory

ÖZET

YÜKSEK LİSANS TEZİ

YABANCI DİL OLARAK İNGİLİZCE SINIFLARINDA SANAL GERÇEKLİĞİN ÖĞRENCİLERİN KONUŞMA PERFORMANSI ÜZERİNDEKİ ETKİSİ

Sümeyra Mürvet AYDIN

İstanbul Üniversitesi-Cerrahpaşa

Lisansüstü Eğitim Enstitüsü

Yabancı Diller Eğitimi Anabilim Dalı

İngiliz Dili Eğitimi Programı

Danışman : Doç. Dr. Tuncer CAN

Bu çalışma, sanal gerçekliğin (VR) İngilizceyi yabancı dil olarak öğrenen öğrencilerin konuşma performansı üzerindeki etkisini incelemekte ve öğrencilerin VR kullanımına yönelik algılarını araştırmaktadır. Geleneksel sınıf ortamlarında öğrencilerin gerçek hayata benzer iletişimsel durumlara sınırlı ölçüde maruz kalması, konuşma becerisinin gelişimini zorlaştırmaktadır. Bu bağlamda VR, öğrencilerin hedef dili gerçekçi ve bağlamsallaştırılmış ortamlarda kullanmalarına olanak tanıyan yenilikçi bir öğrenme aracı olarak öne çıkmaktadır. Araştırmada, nicel ve nitel verilerin bütünleştirildiği açıklayıcı sıralı karma yöntem deseni benimsenmiştir. Çalışmaya orta düzey İngilizce yeterliğine sahip toplam 24 öğrenci katılmış; öğrenciler deney (n=12) ve kontrol (n=12) gruplarına ayrılmıştır. Örneklem, gönüllü yanıt ve amaçlı örnekleme yöntemleriyle belirlenmiştir. VR uygulaması altı hafta sürmüş; deney grubu ImmerseMe platformu aracılığıyla günlük yaşama dayalı konuşma senaryolarında bireysel olarak pratik yaparken, kontrol grubu geleneksel konuşma etkinliklerine devam etmiştir. Her iki grubun ele aldığı konuşma konuları aynı tutulmuştur.Nicel veriler, IELTS konuşma rubriğine dayalı ön test ve son testler yoluyla toplanmış ve iki bağımsız değerlendirici tarafından puanlanmıştır. Öğrencilerin VR'ye yönelik algıları, geçerliliği sağlanmış bir algı ölçeği ve yarı yapılandırılmış anket aracılığıyla incelenmiştir. Bulgular, her iki grubun da konuşma becerilerinde gelişim gösterdiğini; ancak deney grubunun özellikle akıcılık ve telaffuz alt becerilerinde daha yüksek puanlar elde ettiğini ortaya koymuştur. Bununla birlikte, gruplar arasındaki başlangıç düzeyi farklılıkları sonuçların temkinli yorumlanmasını gerekli kılmıştır.Algı bulguları, öğrencilerin VR'yi ilgi çekici, gerçekçi ve motivasyonu artırıcı bir öğrenme ortamı olarak

değerlendirdiklerini; ancak pratik sınırlılıkların da farkında olduklarını göstermiştir. Nitel bulgular, VR'nin kaygıyı azaltan, anonimlik hissi sağlayan ve özgüveni destekleyen bir ortam sunduğunu ortaya koymuştur. Ancak algılar ile ölçülen konuşma kazanımları arasında her zaman güçlü bir ilişki bulunmamıştır. Sonuç olarak, VR'nin konuşma becerisini doğrudan belirleyen bir unsurdan ziyade, öğrenme sürecini destekleyen ve yapılandıran tamamlayıcı bir araç olduğu söylenebilir.

Ocak 2026 , 309 sayfa.

Anahtar kelimeler: Sanal gerçeklik, Yabancı dil olarak İngilizce konuşma becerisi gelişimi, öğrenci algıları, İngilizce dil öğrenimi, karma yöntem araştırması

ABSTRACT

M.Sc. THESIS

THE EFFECT OF VIRTUAL REALITY ON STUDENTS' SPEAKING PERFORMANCE IN EFL CLASSES

Sümeyra Mürvet AYDIN

İstanbul University-Cerrahpaşa

Institute of Graduate Studies

Department of Foreign Language Education

English Language Teaching Programme

Supervisor : Assoc. Prof. Dr. Tuncer CAN

This study investigated the effect of virtual reality (VR) on EFL students' speaking performance and explored their perceptions of using VR for learning and speaking English. It adopted an explanatory sequential mixed-methods design, integrating quantitative speaking performance data with learners' perceptions gathered through a validated perception scale and a semi-structured questionnaire. It aims to provide a comprehensive understanding of VR as a speaking practice tool within an English preparatory program. The participants were 24 EFL students, divided into an experimental group (n = 12) and a comparison group (n = 12). Over a six-week period, the experimental group engaged in VR-based speaking practice using the ImmerseMe platform, while the comparison group followed conventional speaking activities. Speaking performance was assessed through pre- and post-tests using the IELTS speaking rubric and rated by two independent raters. Descriptive statistics and inferential analyses indicated that both groups improved over time; however, the experimental group consistently obtained higher scores at both testing stages. Due to baseline differences between groups, post-test findings were interpreted cautiously. Subskill analyses suggested that group differences were particularly evident in fluency and coherence, and pronunciation, indicating that VR may have shaped the nature of speaking development rather than uniformly increasing performance. Perception scale results showed that students generally viewed VR positively, describing it as engaging, authentic, effective, and supportive of learning, while also acknowledging practical constraints. Correlation analyses revealed that positive perceptions did not always correspond to measurable speaking gains, suggesting a complex relationship influenced by affective and

individual factors. Qualitative findings supported this interpretation, highlighting VR as a low-anxiety, immersive environment that enhanced learners' confidence, motivation, and willingness to speak. Overall, the findings suggest that VR functions best as a supportive instructional tool rather than a sole determinant of speaking development, contributing valuable insights to VR-based language learning research.]

January 2026, [309] pages.

Keywords: [Virtual reality, EFL speaking development, ImmerseMe, Learner perceptions, English language learning, mixed-methods research]

1. INTRODUCTION

This chapter introduces the focus of the present thesis, which explores the use of the virtual reality application ImmerseMe in English as a Foreign Language (EFL) classrooms in relation to its effect on learners' speaking skills. The study examines whether engagement with virtual reality (VR) enhanced speaking tasks lead to measurable differences in learners' speaking performance when compared to traditional instructional practices in an English preparatory program. By adopting an explanatory sequential mixed-methods design, the research addresses both performance-based outcomes and learners' perceptions of their VR experiences for learning and speaking. In this chapter, the overall design of the study is delineated through a presentation of the background of the study, the research problems, the purpose, the research questions, the significance of the study and the definitions of commonly used terms in this study.

1.1. BACKGROUND OF THE STUDY

Among the four language skills, speaking is widely considered to be particularly demanding as it involves the simultaneous management of linguistic accuracy, fluency, pronunciation, pragmatic competence and interactional skills. In EFL settings, learners' opportunities for authentic spoken interaction are mostly restricted by structural constraints such as large class sizes, exam-oriented curricula, time limitations, among many others. Additionally, this skill also includes the influence of affective factors such as anxiety and low confidence. As a result, many learners fail to achieve communicative competence despite years of classroom-based instruction.

The salience of this skill isn't merely in relation to achieving academic success as beyond its communicative function, language embodies how individuals think, who they are and the cultural contexts they belong to. Especially reflecting back on a discussion with a Turkish opera singer, it

was highlighted in that moment that words, sounds and flow of speech can be understood as a form of music through which identity, culture, ideology and lived experience are expressed in every note, rhythm and intonation. Moreover, when considered broadly, humans rely predominantly on speaking to convey meaning, express identity and interact with others. Humans are different from other living beings as they use spoken language to convey complex ideas, emotions and intentions. This is why speaking is a central part of human interaction and social life. This fundamental role of speaking can be seen in English as a Foreign Language (EFL) context, where oral communication is widely recognized as a core skill. Hence, speaking proficiency and communicative competence are heavily and consistently highlighted in the literature for communicative competence.

To address the ongoing challenges in language learning, technological advancements have been explored in recent years. Among these, VR has gained prominence for its capacity to create immersive, contextualized and interactive learning environments that are alike real-world communicative situations. Thus, learners are able to participate in these VR-based situated tasks and role-play activities that are difficult to achieve in conventional classroom settings.

Hence, VR environments are considered to facilitate communicative language use as they provide visual contextualization, authentic scenarios and opportunities for sustained practice.

Empirical research on the use of VR in language education has expanded substantially with many studies reporting positive learner perceptions, increased motivation and enjoyment, and reduced speaking anxiety. Despite these favourable affective outcomes, evidence concerning VR's effect on assessing speaking performance remains inconsistent. A significant number of studies rely heavily on learners' perceived language gains, small samples, short term interventions and they lack rigorous assessment procedure or appropriate comparison groups. Furthermore, the speaking assessment procedures are often insufficiently reported with limited detail regarding rating criteria, rater training or the reliability of scoring.

Especially in tertiary-level English preparatory programs, where learners are expected to achieve functional communicative competence within a limited timeframe, there is an increasing need for empirically grounded evidence to clarify whether and how VR can support speaking development. In particular, high-quality controlled studies examining VR-enhanced speaking

practice through transparent, reliable and performance-based assessment procedures in realistic institutional contexts remain scarce.

1.2. STATEMENT OF THE PROBLEM

Despite the growing body of research on VR in language learning, significant methodological concerns continue to limit firm conclusions of its effectiveness. Systematic reviews have demonstrated that even though the field has grown quickly, high-quality experimental and quasi-experimental studies are limited as many lack control groups, transparent methodological reporting and verifiable learning outcomes (Butorova et al., 2024; Mallek et al., 2024). In response to these shortcomings, it has repeatedly been emphasized that there is a need for rigorous quantitative, qualitative and mixed-methods research to strengthen the empirical foundations of VR-based language learning research (İnalöz and Yılmaz, 2025; Raman et al., 2024). Consequently, the limited number of experimental and quasi-experimental designs restricts the extent to which causal relationships between VR use and language developments can be established (Ironsi, 2023).

Although positive outcomes of immersive VR are frequently reported, these results are often based on the learners' self-reported perceptions rather than direct assessments of language performance, which restricts conclusions about speaking development. In particular, research that examines VR-enhanced speaking practice often omits control groups, so it isn’t possible to assign the observed improvements in speaking performance specifically to VR use.

Beyond the issues of research design, recent reviews have also indicated the insufficient consideration of cognitive processes, such as perception, memory and attention within VR-based learning environments, despite their relevance to effective instructional design (Butorova et al., 2024). These gaps point to the need for methodologically reliable research that combines objective, performance-based speaking assessment with learners’ experiential accounts to provide a clearer evidence of how VR-enhanced language learning contribute to EFL learners’ speaking development (Raman et al., 2024; Soto-Lucena, 2024). In short, further research is required to address these methodological and conceptual concerns, which incorporates exploring VR-enhanced speaking practice that have a controlled design and include both performance-based outcomes and learners’ experiences.

1.3. PURPOSE OF THE STUDY

The present study examines the effect of VR on students' speaking performance. This includes a desktop version of VR through a platform called ImmerseMe. It compares the experiment group and control groups' speaking performance for students who are at the intermediate level of English in an English preparatory program. In other words, it compares students who had experience practicing speaking with VR with those who had traditional speaking practices. It employed an explanatory sequential mixed-methods design to understand the learners' perceptions of using VR for speaking practice, which was to help assist the interpretation of the quantitative findings. This study aims to generate empirically grounded evidence of the effectiveness of VR on students' speaking skills under controlled conditions in order to contribute to the existing literature that is currently persisted in methodological and empirical limitations.

1.4. RESEARCH QUESTIONS

Building on the issues outlined in the previous section, the present study focuses on the implementation of ImmerseMe as a VR tool for supporting EFL learners' speaking performance. It intends to determine the extent to which VR-enhanced speaking practice contributes to speaking development within a controlled educational environment. It also inspects learners' perceptions of the use of VR on their learning and speaking skills. Therefore, the study is guided by the following research questions:

i. What is the effect of virtual reality on students' speaking performance?

ii. What are the perceptions of students toward using VR for learning and speaking English?

ii.i Are there any correlated items in the students' perceptions toward using VR for learning and speaking?

ii.ii How do students describe their perception toward using VR for learning and speaking?

iii. Is there a difference between students' perceptions of the effect of virtual reality on their speaking performance and their actual speaking performance?

1.5. SIGNIFICANCE OF THE STUDY

At the core of language learning is that humans need to speak to convey meaning, develop identity, and engage in social interaction. In EFL contexts, opportunities for meaningful spoken communication are limited. This means that learners don't have enough space to turn their linguistic and pragmatic knowledge into communicative ability. Thus, the current study is significant because it examines whether VR-based learning can provide results that are in line with the pedagogy to develop speaking skills in a realistic educational setting.

This study contributes to the field as it displays speaking as a construct that can be measured to see VR's effect on it and not just provide its outcomes on learners' affective factors. Previous studies have reported learners' engagement, motivation and enjoyment in VR environments; however, there are only a few studies that are transparent and controlled which offer VR's influence on speaking skills. By addressing this gap, the study strengthens the empirical foundation to assess VR's impact on learners' speaking skills in EFL contexts. This is especially thanks to its research design being explanatory sequential mixed-methods, which brings numerical value and lived experiences together. Thus, quantitative data of pre- and post-tests provide insight into the observable learning outcomes regarding speaking skills, the quantitative scale of perception assists to systematically measure learners' perceived impact of VR on their speaking performance and learning experience, and the qualitative data of learners' perceptions helps to understand the processes, challenges and possibilities of the quantitative results. Therefore, this study not only assesses VR's effect on students' speaking skills but also how learners experience and interact while speaking in a VR environment.

Moreover, this study can be valuable for tertiary-level English language programs as they expect learners to develop their communicative competence in a short amount of time. By using a desktop version of VR, the study depicts that it can actually be accessible and aligns with the reality that high-immersion VR is not feasible in all educational contexts. Hence, it may be beneficial for instructors, curriculum designers and even program administrators who want to incorporate novel technology into their EFL lessons.

Finally, as VR learning involves situated learning, this study makes a contribution to theoretical discussions that revolve around how language development is shaped by context, interaction and authenticity. In doing so, it stresses the idea that speaking isn't solely a skill to be

practiced in an isolated manner, but rather a socially situated act that unfolds through meaning engagement with people, tasks and environments, which in this case doesn't differentiate between whether it's physical or virtual.

To conclude, the study is salient as it provides methodologically rigorous, pedagogically grounded and theoretically informed evidence on the role of VR for developing speaking skills in EFL contexts since it facilitates the more complex understand of how technology may meaningfully support the speaking, which is the language skill that is most central in human communication.

1.6. DEFINITIONS OF KEY TERMS

VR: virtual reality refers to computer-generated environments that allow users to interact with simulated scenarios that resemble real-life or imagined environments. In language education, VR is used to create contextualized and interactive environments that support communicative practice.

Desktop-Based VR: Refers to non-immersive VR environments accessed through a standard computer screen using a keyboard, mouse and audio devices.

iVR: Immersive VR refers to VR environments that provide a sense of immersion and presence through technologies such as interactive virtual environments, spatial audio, head-mounted displays and others.

LiVR: Low-immersion VR describes environments that offer limited sensory immersion such as those accessed through desktop computers or mobile devices.

HiVR: High-immersion VR describes VR environments that provide a high degree of sensory engagement through advanced technological hardware such as head-mounted display, spatial audio, and motion tracking.

Speaking Performance: refers to learners' observable and measurable spoken language output that is produced during a communicative task.

Performance-Based Speaking Assessment: Involve the evaluation of learners' spoken production through authentic speaking tasks that require active language use. Its criteria reflect fluency and

coherence, lexical resource, grammatical accuracy and pronunciation to assess the learners' communicative performance.

CEFR: The Common European Framework of Reference for Languages (CEFR) is an internationally recognized framework that is used to describe language proficiency levels.

2. CONCEPTUAL FRAMEWORK

This chapter structures the theoretical foundations and empirical research that have informed the design of this study. It examines the notions and development of speaking skills in EFL classrooms. This is also carried out with a specific focus on the role of virtual reality on improving students' speaking performance, especially since it is a salient technological advancement on which this research is conducted. Therefore, emphasis regarding the research on learners' perceptions towards the integration of VR to enhance their speaking skills will be incorporated as well to provide further insight on this matter in relation to its literature.

2.1. SPEAKING SKILLS IN EFL CONTEXTS

2.1.1. Importance of Speaking in Language Learning

"Language, at its core, encompasses more than just communication; it manifests one's identity, cognition, and culture" (Al-khresheh, 2024, p.2). Thus, when taken into consideration, humans predominantly rely on the language skill of speaking to do so, which highlights its criticality in human lives. This naturally leads to it being a central part in EFL contexts as the significance of oral proficiency in English is widely acknowledged in the literature. This is attributed to it being a crucial component of functional communicative engagement in an interwovenly connected world (Al-khresheh, 2024; Alzubi et al., 2025; Caesar, 2023; Oktaviani et al., 2025; Rambe and Daulay, 2025; Sasirekha et al., 2024; Wang et al., 2024). As the renowned linguist, Crystal (2003), had once indicated that the continuation of globalization and the advancement of technology have placed English as a global language especially thanks to it enabling interpersonal communication between individuals with varied backgrounds, across various locations and even being able to transcend the geographical limitations. Consequently, it can be noted that Crystal's argument still remains to be relevant even though slightly more than 20 years have passed since they are mentioned in current literature. Thus, this further strengthens

the view that the need to still speak and learn English is required nowadays to have the capability of engaging with others on a global scale in this globalized world. This requirement has become so essential that it has evolved into a pressurizing demand to better English proficiency (Almutairi and Alfarwan, 2024; Caesar, 2023). In particular, the speaking proficiency is a necessity to cultivate communicative competence and engage in real-life interactions (Alzubi et al., 2025; Santos and Luque-Agulló, 2025) that enables to deliberate on the societal and global expectations to use and speak English effectively in diverse domains that ranges from professional to personal areas of life, which encompasses academic, business and social (Al-khresheh, 2024; Rambe and Daulay, 2025; Segar and Asmawi, 2024; Wang et al., 2024). These points exceptionally emphasize the vital role of speaking skills not only in EFL, but also in the globalized world. Thus, clearly the enhancement of this skill would eventually enable the learners to navigate real-life communication across various areas of life.

2.1.2. Obstacles Affecting EFL Learner's Speaking Development

In alignment with the CLT approach of language teaching, it is paramount that all four of the language skills be instructed and assessed (Dinçer and Bal Gezegin, 2024). Despite this, the greatest challenges that are the cause of major stress and tension are related to developing fluency and accuracy in speaking in a foreign language, especially in EFL contexts (Ni'mah et al., 2025; Oktaviani et al., 2025; Pituxcoosuvarn et al., 2025; Segar and Asmawi, 2024). There are a number of reasons behind this being the foremost challenge, which will be explained in the following sentences. The rote learning of grammar and vocabulary (Al-khresheh, 2024; Caesar, 2023; Mrabet-Abid, 2024; Sasirekha et al., 2024; Segvan and Saeed, 2025), limited classroom time to foster speaking skills (Almutairi and Alfarwan, 2024; Ay Kaya and Yangın Ersanlı, 2025; Fathi et al., 2024; Ni'mah et al., 2025), limited exposure to the target language speaking settings (Al-khresheh, 2024; Almutairi and Alfarwan, 2024; Ni'mah et al., 2025), the misalignment between the curricular goals and classroom practices regarding speaking skills (Al-khresheh, 2024; Almutairi and Alfarwan, 2024; Ay Kaya and Yangın Ersanlı, 2025; Dinçer and Bal Gezegin, 2024; Segvan and Saeed, 2025), which leads to a causal chain as speaking becomes absent in assessments, so it also becomes further neglected in teaching (Dinçer and Bal Gezegin, 2024; Somé, 2025). Noticeably, these factors are all in relation to education that consists of a wide

spectrum of elements that also seem to be interrelated. However, it is important to highlight that the challenges concerning speaking skills aren't solely connected to education, but also affective factors pertaining to learners' themselves. The impact of negative peer feedback causing reluctance to speak (Fathi et al., 2024; Sasirekha et al., 2024; Suratin and Sribayak, 2025), having low self-confidence (Ni'mah et al., 2025; Oktaviani et al., 2025; Phisutthangkoon, 2024; Sasirekha et al., 2024; Suratin and Sribayak, 2025; Tao et al., 2025), lack of motivation (Oktaviani et al., 2025; Sasirekha et al., 2024; Suratin and Sribayak, 2025; Tao et al., 2025), and other factors that have a negative influence on the learner's socio-emotional and psychological state such as communicative incompetence, language barriers, and insufficient fluency (Oktaviani et al., 2025; Phisutthangkoon, 2024; Tao et al., 2025), which can all also be a cause of speaking anxiety (Suratin and Sribayak, 2025; Tao et al., 2025). As a result, the immense influence of learner-related and external conditions forms the development of the learner's speaking skills because components such as psychological processing, educational factors and the influence of the language elements themselves holistically affect the learner's oral proficiency.

2.1.3. Proposed Solutions to Developing Speaking Skills

In spite of the wide acknowledgement of these interrelated challenges, the conventional teaching methods and theoretical approaches are inadequate in successfully addressing them (Al-khresheh, 2024; Sasirekha et al., 2024; Segar and Asmawi, 2024). Therefore, these methods must be supplemented by digital solutions in order to efficiently tackle these difficulties regarding speaking skills (Karagöl et al., 2025). In other words, these classical methods shouldn't be completely discarded but rather be strengthened with innovative approaches. Certain ideas such as enhanced teacher education (Al-khresheh, 2024; Somé, 2025), and curriculum alignment or adjustment are presented as potential solutions for this issue. However, innovative approaches that incorporate both experiential and interactive learning should be adopted to properly address this matter (Ho-Minh and Suppasetseree, 2025; Segvan and Saeed, 2025), especially ones with a focus on the adoption and incorporation of novel technology resources (Sasirekha et al., 2024; Segvan and Saeed, 2025). The advanced technological tools and devices provide opportunities to augment learners' motivation, which leads to active participation (Sasirekha et al., 2024). This indicates that the use of such technology can potentially foster more positive emotional and psychological

states among students. Hence, learning environments that stimulate more dynamic and learner-centred spaces are required as they encourage active participation that provide a better chance for learners to advance their speaking skills (Al-khresheh, 2024; Oktaviani et al., 2025; Segar and Asmawi, 2024; Segvan and Saeed, 2025). Thus, it is important to mention VR here as it plays a key role in this matter by providing such instructional environments that facilitate speaking production and in turn speaking development in realistic and interactive situations (Çetin Köroğlu, 2025). This argument is further strengthened by Lee (2025), as VR displays contextualized environments in which learners are involved in authentic immersive communicative scenarios that are rarely achievable to recreate in the physical context of a conventional classroom. Moreover, VR alleviates students' speaking anxiety thanks to its nature (Lee, 2025). Therefore, this reinforces Sasirekha et al.'s (2024) assertion that the inclusion of innovative approaches to develop learners' speaking skills also have a positive outcome on their emotional and psychological well-being. Evidently, the fortification of conventional teaching methods with the convergence of innovative approaches that supply interactive and context-based dynamic environments, such as the VR learning environments, will truly equip learners with the desired level of oral proficiency that's essential for their success in academic, professional and social domains. Further elaboration on the utilization of VR in language learning is provided in Section 2.4.

2.1.4. Instructional Approaches to Teaching Speaking

The historical foundations of language education in terms of teaching and learning expands far back in time. As delineated by Richards and Rogers (2001), the categorization of language teaching methods and approaches can be divided into three sections; major twentieth century trends, alternative methods, and contemporary communicative approaches. Furthermore, in relation to this, the theory of language itself can be conceptualized under three categories as well; structural, functional and interactional views (Richards and Rogers, 2001). Each section includes a variety of approaches and methods; however, elaborating on all of them here would go beyond the scope of this section. Therefore, merely the approaches and methods that have guided the evolution of the instruction of speaking skills will be addressed.

It is worth mentioning that especially Richards and Rogers (2001) account for the lion's share of information in this section of the literature review. This is due to them providing one of the most authoritative and prominent contributions in relation to language teaching methods. Thus, they were predominantly referenced to ground this review in a well-established source alongside being supported with other canon studies published by eminent scholars such as Howatt and Widdowson (2004), Larsen-Freeman (2000), Nunan (1989), Willis and Willis (2001) and many other notable figures in language teaching in regards to approaches, techniques and methods to establish foundational knowledge, demonstrate scholarly awareness and increase the credibility of the review.

Without delving into this issue any further, it is highly important to make the theoretical differentiation between the concepts of approach, method and technique in language teaching to avoid any confusion at a later point. It must be written that throughout time there have been disputes on this issue. However, one of the most widely accepted definitions that's relevant even until now was published by Edward Anthony in 1963. In short, the main organizational point is that techniques serve to carry out methods that correspond with an overall approach (Anthony, 1963). The association of the three is hierarchical; approach is the broader scope that has presuppositions, theories and principles which can incorporate multiple methods with various techniques (Anthony, 1963). Method is the procedure and techniques constitute the practical aspect of the method being utilized that is in an alignment with a specific approach (Anthony, 1963). Though as highlighted previously in regards to disputes on this matter, certain scholars such as Richards and Rogers (2001) do argue that there is no hierarchy and that methods can be generated from any of the categories of approaches, designs or procedures. For instance, a successful procedure can lead to later being justified by a theoretical approach and not vice versa (Richards and Rogers, 2001).

It is not possible to proceed without initially mentioning the earliest formal method of language teaching, which is known as the Grammar Translation Method (GTM). The era between the 1840-1940's witnessed the dominance of GTM in European and foreign language education as it greatly structured them (Richards and Rogers, 2001). The primary aim of this method was to

boost the literacy skills of students to gain an understanding of literature in addition to facilitating mental discipline and intellectual growth (Larsen-Freeman, 2000; Richards and Rogers, 2001). Hence, as this was the case, speaking and listening received little to no attention (Larsen-Freeman, 2000; Richards and Rogers, 2001), which is why it was profoundly criticized (Richards and Rogers, 2001).

With the contribution of the Reform Movement members (Henry Sweet, Wilhelm Vietor, and Paul Passy), in addition to the invention of the International Phonetic Alphabet (IPA), naturalistic language learning became a priority and this became the inspiration for approaches that mimicked first language acquisition (Howatt and Widdowson, 2004; Richards and Rogers, 2001). This increased the number of natural methods and eventually led to what became known as the Direct Method, which became popularized by Maximillian Berlitz (Howatt and Widdowson, 2004; Richards and Rogers, 2001). Despite being called a method, it lacked scientific foundation (Richards and Rogers, 2001). Nonetheless, its significance is viewed by some as it lies in paving the way to the "methods era", and being the first "method" to focus on the value of the speaking skills (Howatt and Widdowson, 2004; Richards and Rogers, 2001). The Direct Method drew parallels with first language acquisition by ignoring the realities in the classroom and lacking a thorough linguistic basis (Howatt and Widdowson, 2004; Larsen-Freeman, 2000; Richards and Rogers, 2001). Additionally, it relied heavily on native teacher fluency and avoided the use of mother tongue or translation at all costs (Howatt and Widdowson, 2004; Larsen-Freeman, 2000; Richards and Rogers, 2001), which made it impractical and counterproductive to use (Richards and Rogers, 2001). Thus, it also received heavy backlash from critics (Richards and Rogers, 2001).

By building on the notions of the Reform Movement, the applied linguists operated to systematize its shortcomings by generating principles for it in the 1920s and 1930s. The contributions led to the establishment of Audiolingualism in the U.S and the Oral Approach/ Situational Language Teaching (SLT) in Britain (Richards and Rogers, 2001). These methods have certain resemblances to one another such as both of these methods are speaking oriented, teacher-led and use repetition (Johnson and Johnson, 1998; Smith and Loewen, 2018). Furthermore, they share features such as the order of skills to teach, and the inclusion of drills is used to ensure

accuracy (Johnson and Johnson, 1998; Smith and Loewen, 2018). However, as SLT evolved from the Direct Method, it is insufficient in terms of the linguistic and psychological foundation that is evident in the basis of Audiolingualism (Richards and Rogers, 2001). Despite their similarities, their traditions were developed independently from one another; SLT started in the 1930s and Audiolingualism began in the 1950s (Richards and Rogers, 2001). The two primarily differ as SLT viewed language as a purposeful activity that was connected to real-world goals and natural contexts, in which learners learned inductively so explanations were discouraged as presented by their advocates Palmer and Hornby, who aided in its formation and foundation (Smith and Loewen, 2018). Additionally, the well-known "P-P-P" lesson model is a reflection of the essential elements that are embedded in SLT (Smith and Loewen, 2018). In contrast, as the Audiolingual method was grounded in behaviourism it viewed language learning through stimulus-response mechanisms that incorporated conditioning and reinforcement to produce error-free performance (Johnson and Johnson, 1998; Richards and Rogers, 2001). That's why, even though it was an oral-focused approach, it stressed immediate error correction and accurate speech (Johnson and Johnson, 1998; Richards and Rogers, 2001). Thus, little attention was given to grammatical explanations and discussion of the language itself as the sole focus was on repetition and memorization through the use of dialogues and drills to strengthen stimulus-response connections (Johnson and Johnson, 1998; Richards and Rogers, 2001). Both of these methods were particularly influential in their era, yet Audiolingualism had a more considerable impact (Richards and Rogers, 2001). However, the notable linguist, Chomsky, challenged the very basis that Audiolingualism was built upon as he asserted that humans create language creatively from abstract rules and not through imitation (Johnson and Johnson, 1998; Richards and Rogers, 2001). Thereby, casting doubt on the behaviorist approaches and the whole of the Audiolingual method itself by weakening the heavy reliance on drills, repetition and memorization (Johnson and Johnson, 1998; Richards and Rogers, 2001).

This entailed the beginning to the alternative approaches and methods to emerge from the 1970s-1980s, such methods include the Total Physical Response, Silent Way, Neurolinguistic Programming, Suggestopedia and others (Richards and Rogers, 2001). Nunan (1989) referred to the methods formed in this era as the "designer" methods, while Brown (1997) called them the

"spiritual seventies". This was due to this era being a result of a rapid development and innovation of methods (Brown, 1997). Though they hold historical value, as they don't bear much significance in language education today, they will not be discussed in this section.

Although it's also considered as an alternative method, Communicative Language Teaching (CLT) is mentioned here as it has managed to stay relevant in the language education of today. The emergence of CLT can be associated with traditions of the British language teaching in the late 1960s that came from the changes made to SLT (Johnson and Johnson, 1998). This was especially a result of Audiolingualism having lost its credibility in the United States during the mid-1960s period, which caused a chain reaction that made the British linguists question the assumptions of SLT (Richards and Rogers, 2001). Thus, by doing so, they managed to highlight a critical component of language learning that had been previously overlooked by the previous methods; the emphasis on the functional and communicative dimensions of language to support communicative competence (Larsen-Freeman, 2000; Richards and Rogers, 2001). This concept was first proposed by Wilkins (1972) as a perspective of emphasizing the functional and communicative nature of language that formed the basis of the communicative syllabus design by providing an analysis of the meaning learners are required to learn for comprehension and expression. The two types of meaning he distinguished between were notional and communicative function. The former addresses aspects such as time, order, amount, place, and frequency, while the latter involves functions such as requests, refusals, offers, and complaints (Richards and Rodgers, 2001). Advocates on this view, including Christopher Candlin and Henry Widdowson, extended theoretical insights originating from the British functional linguists such as those associated with Firth and Halliday (Richards and Rodgers, 2001; Roberts, 2004) along with those of American sociolinguistic researchers such as Hymes, Gumperz, and Labov (Richards and Rodgers, 2001), and even philosophical work such as from Austin and Searle (Richards and Rodgers, 2001; Roberts, 2004). Thus, it encompasses work from the Council of Europe and contributions from Wilkins along with other British applied linguists to inform its theoretical foundations of communicative and functional language teaching (Roberts, 2004). At present, scholars in both the United States and Britain regard it as an approach that seeks to foster communicative competence by adopting an integrated view of the four skills, which stresses the

interdependent nature of language and communication (Richards and Rodgers, 2001). Language teaching within this approach is primarily focused toward fostering communicative competence, as defined by Hymes (1972).

According to Hymes (1972), communicative competence involves possessing the necessary knowledge and skills to function effectively within a speech community. Achieving communicative competence involves both language comprehension and contexutally appropriate language use (Hymes, 1972). Hyme's view offers a more thorough an account of what is means to possess knowledge of a language when compared with that of Chomsky's, which stressed abstract grammatical knowledge (Celce-Murcia, 2007). Halliday constructed a comprehensive theory of language functions that supplements Hyme's model of communicative competence, which had been noted by many CLT scholars such as Brumfit, Johnson and Savignon (Richards and Rodgers, 2001). Seven primary functions of language that occur in children's first language acquisition were identified, which ranges from basic instrumental and regulatory uses of language to interpersonal, cognitive, imaginative, and eventually representational uses that allow speakers to express abstract information and construct meaning (Halliday, 1975). As the years unravelled, the evolution of Hymes' communicative competence commenced by Canale and Swain building a model of their own on communicative competence (Celce-Murcia, 2007; Richards and Rodgers, 2001). This model was further developed by Bachman, and later extended by Celce-Murcia, Dörnyei and Thurrell (Richards and Rodgers, 2001), and the latest version was provided by Littlewood (Taş and Khan, 2020).

Table 2.1: The Chronology of the Conceptual Evolution of Communicative Competence

The Chronology of the Conceptual Evolution of Communicative Competence	Features
Chomsky (1957, 1965)	Linguistic competence
Hymes (1967, 1972)	Linguistic competence, Sociolinguistic competence
Canale and Swain (1980)	Grammatical competence, Strategic competence, Sociolinguistic competence
Canale (1983)	Grammatical competence, Strategic competence, Sociolinguistic competence, Discourse competence

Table 2.1 (continued): The Chronology of the Conceptual Evolution of Communicative Competence

Bachman (1990)	Strategic competence, Language competence (Organizational competence – grammatical competence and textual competence; Pragmatic competence – Illocutionary competence; Sociolinguistic competence), Psycho-physiological Factors
Celce-Murcia, Dörnyei and Thurrell (1995)	Linguistic competence, Strategic competence, Sociolinguistic competence, Actional competence, Discourse competence
Celce-Murcia (2007)	Sociocultural Competence, Discourse Competence, Linguistic Competence, Formulaic Competence, Interactional Competence, Strategic Competence
Littlewood (2011)	Linguistic competence, discourse competence, pragmatic competence, sociolinguistic competence, sociocultural competence

Note. Developed based on information from Celce-Murcia (2007), and Taş and Khan (2020).

Prior to proceeding any further, it is essential to differentiate between competence and performance as this study examines students' speaking performance. There have been many disputes on the definition of competence and performance and whether they refer to the same construct. (Canale and Swain, 1980). According to Chomsky's (1965) weaker claim of the definition of the two, competence refers to an individual's knowledge of grammar and other linguistic aspects, whereas performance involves the actual use of language. This distinction is one which most linguists agree with. However, theoretical assumptions of what's relevant or not of a data could be a potential threat (Canale and Swain, 1980). Chomsky's (1965) stronger claim is that competence is equivalent to a theory of grammar, which focuses on the linguistic rules to create and describe the syntax of a language. Whereas performance addresses the acceptability of sentences in actual speech that encompasses both perception and production, which reflects the interaction of grammatical and non-grammatical psychological elements that have an impact on language use. However, Campbell and Wales (1970), and Hymes (1972) were one of the earliest scholars to note that there was an exclusion of sociocultural appropriateness and the contextual significance of utterances in this stronger version of Chomsky's definition. Thus, it can be said that whether implicitly or explicitly, the distinction between communicative competence and performance were made by scholars (Canale and Swain, 1980). As defined by Canale and Swain

(1980), communicative competence incorporates the interrelationship between grammatical competence and sociolinguistic competence. On the other hand, communicative performance denotes the realization and interplay of communicative competence during real-time language production and comprehension. Therefore, the present study employs the term performance as defined by Canale and Swain (1980) due to this definition accounting for the full range of elements involved in assessing speaking skills.

An important consideration to be made is that CLT is not a method in itself, but an approach as it is extremely comprehensive as it consists of communicative principles that assists flexibility and varied classroom practices (Harmer, 2007; Littlewood, 2014; Richards and Rodgers, 2001). Consequently, there has been an extensive discussion about syllabus planning and models in CLT (Littlewood, 2014; Richards and Rodgers, 2001). The communicative approach promotes varied activities that enhance curricular goals through interaction where information exchange and negotiation of meaning take place in pairs or groups (Littlewood, 2014; Richards and Rodgers, 2001). Central part of CLT is the belief that language learning in the classroom is fostered through active interaction that incorporates activities with authentic and meaningful exchanges for real-life communication (Littlewood, 1981, Richards and Rodgers, 2001). They highlight the value of fluency, accuracy, holistic approach towards the skills, and sees language as part of a creative process inside the learner that naturally holds trial and error (Littlewood, 1981, Richards and Rodgers, 2001). By placing communication and interaction at the heart of teaching, CLT succeeded in resonating with those who appreciated a humanistic alternative (Littlewood, 1981, Richards and Rodgers, 2001).

Task-Based Language Teaching (TBLT) conceptualizes tasks as the fundamental unit of curriculum and design and instructional practice as an approach (Nunan, 1989; Willis and Willis, 2001). It can be considered to be an extension of CLT since it relies on the principles set by the communicative movements of the 1980s (Willis, 1996). Though many definitions of the term “task” are provided in the context of TBLT, the general consensus is on it being an activity or objective that is accomplished through the meaningful use of language (Nunan, 1989; Willis and Willis, 2001). Krashen (1981) prioritized comprehensible input for being sufficient for language

acquisition, yet others argue that output is also an indispensable part of this process. However, this notion was especially underscored by Swain (1985) as the need for opportunities to use language actively as output. This is especially significant as Swain's assertion was also strongly supported by other scholars who have had a substantial influence on the development of TBLT such as Nunan (1989), and Willis and Willis (2001). Thus, it can be comprehended that TBLT highly values the essential contribution of output or, in other words production, in language learning. Therefore, drawing on findings from SLA research focusing on interaction and negotiation, TBLT situates tasks as the focal point for encouraging input and output practice, meaning negotiation, and transactional interaction (Nunan, 1989; Willis and Willis, 2001). The procedure of this approach consists of pre-task, task cycle (planning, task, report) along with emphasis on language focus upon completion of task cycle (Willis, 1996; Willis and Willis, 2001), and post-task (optional) (Richards and Rodgers, 2001). Tasks entail 6 components; goals, input, activity, roles of teachers and learner, and settings (Nunan, 1989). Though there have been a few categorizations of task types created by scholars such as Prabu and Stern (Willis and Willis, 2001), Willis (1996) proposed a classification which subsumes the task types that were produced earlier and was intended to function as a generative pedagogical instrument. "The operations are: listing; ordering and sorting; comparing; problem-solving; sharing personal experiences; creative tasks" (Willis and Willis, 2001, p.177).

In this aspect, role-plays are considered by some to be seen as a task which can have some of the following operations previously mentioned. These well-known scholars are ones such as Harmer (1989), Nunan (2003), and Stern (1992), who advocate that role-plays indeed support real language use and thus should be utilized as a task in classrooms to improve speaking skills. However, others such as Willis (1990) assert that role-plays have no real value as learners are displaying instead of using the target language if they are asked to only act out a situation without involving any problem-solving skills to generate real language use.

"Role playing allows participants to immerse themselves in a learning environment by acting out the role of a character or part in a particular situation" (Hale Feinstein et al., 2002, p.735). Additionally, role-plays are generally divided into three forms; fully scripted, semi-scripted (SSRP) and non-scripted (NSRP) (Harper-Whalen and Morris, 2005). Generally, the first is

suitable for lower levels, the second being more appropriate for elementary to intermediate levels and the third can be practical for intermediate to advanced levels (Krebt, 2017). Hence, role-plays have been recognized to have consistently been identified as an effective approach for enhancing speaking skills in EFL contexts (Altun, 2015; Alwehebi, 2025; Ayuningtias et al., 2019; Islam and Islam, 2012; Krebt, 2017; Neupane, 2019; Oradee, 2012; Sukmara et al., 2025; Tran et al., 2025). This could be attributed to role-plays providing ideal situations for students to practice functions that highlight verbal communication serves a purpose such as apologizing, giving advice etc. (Lackman, 2010) and that they offer a safe environment in which learners practise speaking based on specific contexts before they perform in real world (Altun, 2015; Alwehebi, 2025; Bailey, 2003; Islam and Islam, 2012).

This is in terms of pertaining to factors such as fluency, accuracy, pronunciation, (Islam and Islam, 2012; Neupane, 2019; Sukmara et al., 2025; Tran et al., 2025) lexicon (Altun, 2015; Krebt, 2017; Neupane, 2019; Sukmara et al., 2025), and communicative confidence (Altun, 2015; Ayuningtias et al., 2019; Islam and Islam, 2012; Neupane, 2019; Oradee, 2012; Sukmara et al., 2025; Tran et al., 2025). These outcomes also provide evidence for the socio-cognitive view of language development being enhanced through social interaction as role-plays allow learners to engage and interact with others while negotiating meaning and adapting to different communicative roles (Krebt, 2017; Oradee, 2012; Sukmara et al., 2025). It also increases active participation and engagement (Altun, 2015; Alwehebi, 2025; Neupane, 2019; Oradee, 2012; Tran et al., 2025). However, some findings indicate that even though role-plays positively impact students' speaking skills, the effect might not be so significant for between the control and the experiment groups (Cho, 2015; Druckman and Ebner, 2008; Krebt, 2017), while others have revealed the difference between the control and experiment groups to be major (Alwehebi, 2025; Neupane, 2019; Sukmara et al., 2025). Furthermore, regardless of the type of role-play, findings of studies have shown that when compared to one another they present no significant difference and they all do foster students' speaking skills in the target language (Cho, 2015; Mahboob-Eshghdoost and Mall-Amiri, 2019; Rodpradit and Sinwongsuwat, 2012). Nonetheless, others also claim that the type of role-play has an influence on the certain aspects of students' speaking skills, such as NSRP being better than scripted role-plays for fluency (Rodpradit and Sinwongsuwat, 2012; Romadhona et al., 2023). A further question remains about whether skills gained through role-play can be applied in authentic communication (Alexander and LeBaron, 2009). In short,

similar to other studies relating to humans, there is not a single answer to these issues; the variation can be attributed to humans having a dynamic nature and the absence of a universally accepted approach. Thus, factors such as learners' demographic and educational backgrounds should be considered when interpreting the findings and their possible impact on the results.

Building upon the earlier discussion of language methodology, it can be said that in contrast to approaches, methods have enjoyed a brief moment of popularity when glanced upon in language teaching history; having their peak until the late 1980s (Richards and Rodgers, 2001). The reasons for this are presumed to be that they are closely knitted to particular claims and teaching practices, so they fade out once these become outdated (Richards and Rodgers, 2001). Also, that they were overly rigid, became indistinguishable from one another in their later stages and they couldn't be fully supported by empirical evidence (Brown, 1997). Moreover, Phillipson (1992) underlined that they reflected the main interests of powerful countries and excluded the local needs of those in less influential nations.

The early 2000s signalled the origins of the post-method era, which has been characterized by the move away from rigid methods and towards the acknowledgement of the complex nature of language learning (Galante, 2014). This era challenged the misconception regarding methodologies that functioned as a universal understanding that claimed to be designed to suit every learner and every situation (Brown, 1997; Galante, 2014). As Brown (1997) has perfectly described this era through the use of a metaphor "We have emerged well beyond the dark ages of language teaching when a handful of pre-packaged elixirs filled up a small shelf of options" (p.10).

The post-method era, which encompasses Technology Enhanced Language Learning (TELL), Computer-assisted Language Learning (CALL), Mobile-assisted Language Learning (MALL) and Internet-assisted Language Learning (IALL), has extended the language learning opportunities beyond the classroom walls and rigid methodologies and thereby created new opportunities for learning (Almaktary and Al-Kadi, 2017; Hasumi and Chiu, 2024). This transformation initially fostered key ideas such as autonomy, informal learning and mobile learning (Almaktary and Al-Kadi, 2017; Hasumi and Chiu, 2024). It can be noted that the more

technology is involved in language learning, the less dominant traditional methods become (Almaktary and Al-Kadi, 2017). However, for effective teaching to take place, theoretical principles to attain learning goals are still required (Prabhu, 1990). This effective teaching needs to include an eclectic approach that relies on multiple methods to sustain learner engagement and motivation (Almaktary and Al-Kadi, 2017; Bhandari and Bhandari, 2025; Daskalovska et al., 2025; Prabhu, 1990). Thus, this is the core understanding and representation of the post-methods era, because in addition to there being a no best method, teachers are advised to take action in accordance with the need of their students as studies linked to humans always tend to contain the element of relativity (Bhandari and Bhandari, 2025; Daskalovska et al., 2025; Ellis, 2005; Kumaravadivelu, 2001; Prabhu, 1990). This can also be concisely comprehended through Kumaravadivelu's (2001) conceptualization of the post-method pedagogy by the categorization of it around a three-dimensional model that integrates particularity, practicality and possibility. The first is concerned with rejecting universally accepted methods and objectives, and promotes context-sensitive approaches with the inclusion of linguistic, sociocultural and political contexts, the second places an emphasis on teachers constructing their own theories of practice, the third involves extending the language teaching beyond the classroom by including socio-political awareness for identity and societal transformation (Kumaravadivelu, 2001).

As a result, with the beginning of the post-methods era, the integration of technology is necessary to enable this eclectic approach as it also functions as a central component in many areas of contemporary life (Almaktary and Al-Kadi, 2017; Hasumi and Chiu, 2024). In line with the principles of the post-method era, novel technology such as the adoption and integration of artificial intelligence (AI) has begun to become popular in language education in recent times (Dennis, 2024; Grab, 2025; Karagöl et al., 2025; Zhang et al., 2023). Furthermore, innovative technologies such as extended reality (XR), which also encompasses VR, augmented reality (AR) and mixed reality (MR), have become a major focus of research and gave rise to immersive learning instruments (Kaplan-Rakowski et al., 2023; Ma et al., 2025; Schwaiger et al., 2024; Wang and Li, 2024). In relation to the previous discussion on role-plays, it must be emphasized that especially employing VR for role-play tasks enables learners to engage in lifelike scenarios that cultivate their ability to become more effective interlocutors (Kaplan-Rakowski et al., 2023). Thus, it can be argued that VR based simulations are the modern extensions that are grounded in the

same fundamental pedagogical principles as role-plays. However, further information on the use of VR in language education will be discussed in detail under section 2.4.

2.2. ASSESSMENT OF SPEAKING SKILLS

2.2.1. Theoretical Framework of Speaking Assessment

The assessment of speaking skills is grounded in the notion of communicative competence, which has been explained thoroughly above in section 2.1.4. The Common European Framework (CEFR) (Council of Europe [CoE], 2020) adopts this communicative orientation and defines the overall communicative language competence as consisting of three dimensions; linguistic, sociolinguistic and pragmatic components (CoE, 2020). This provides the conceptual framework for contemporary assessment of speaking skills, which seek to capture both linguistic accuracy and communicative effectiveness in authentic tasks. In accordance with this, modern rubrics have been designed to be able to implement these competences through observable performance criteria such as the IELTS speaking rubric.

2.2.2. Assessment and Testing in Language Teaching

A test is a method, a measure of an individual's competence, knowledge or performance within a defined domain (Bachman, 1990; Brown, 2004). The terms 'test' and 'assessment' aren't synonymous with one another; test is a scheduled and formal procedure where the learners demonstrate their abilities for evaluation, whereas assessment is a continuous process that accounts for a wider scope of learning (Brown, 2004; Brown and Abeywickrama, 2019). Thus, tests compose one component of the wider process of assessment (Brown, 2004; Brown and Abeywickrama, 2019) and although tests evaluate actual performance, the results are understood to represent the learner's underlying competence in the language (Brown and Abeywickrama, 2019).

Measurement refers to assigning numerical values to describe students' observed performance, while evaluation is when the results of tests or other assessments are interpreted and used to make educational decisions (Bachman, 1990; Brown and Abeywickrama, 2019). The figure below is adapted from Brown (2004), which depicts teaching as the outer layer, assessment

within it and tests as the innermost circle. This current design of assessment and testing integrates both of these key concepts.

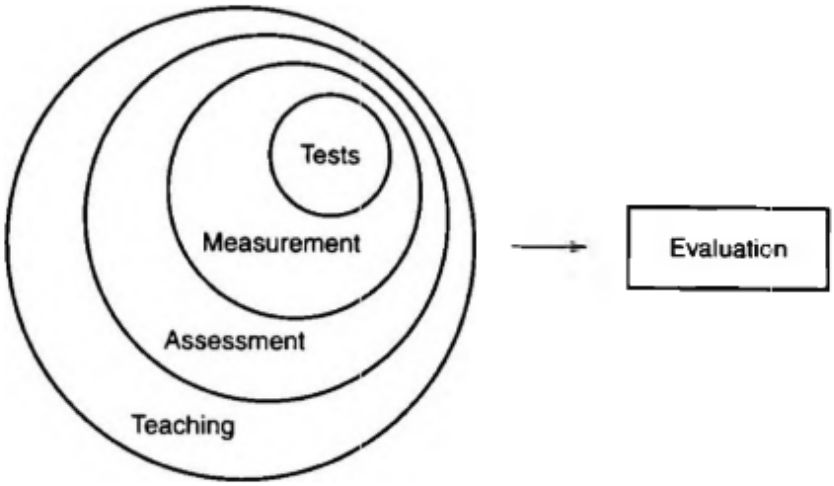


Figure 2.1: The Relationship Between Teaching, Assessment, Measurement, and Evaluation

Note. Reprinted from Brown and Abeywickrama (2019, p.7).

2.2.3. Types of Assessment

2.2.3.1. Informal and Formal Assessment

Informal assessment encompasses unplanned interactions such as spontaneous comments, scaffolding or giving feedback in class (Brown, 2004; Brown and Abeywickrama, 2019). On the other hand, formal assessment refers to structured and systematically designed tasks intended to measure learners' proficiency and achievement (Brown, 2004; Brown and Abeywickrama, 2019). Tests represent one type of formal assessment, but they aren't the sole means of formal evaluation (Brown, 2004; Brown and Abeywickrama, 2019).

2.2.3.2. Formative and Summative Assessment

Formative assessments aim to foster learning by evaluating students' progress during the development of skills, while summative assessment provides an overview of their attained knowledge and performance at the completion of a course or a unit (Brown, 2004; Brown and Abeywickrama, 2019; Bachman, 1990).

2.2.3.3. Norm-Referenced and Criterion-Referenced Tests

Norm-referenced testing evaluates individuals by comparing their results to the performance of a wider group, whereas criterion-referenced testing focuses on providing feedback on the individual learner's achievement of specific learning objectives rather than comparing learners with one another (Bachman, 1990; Brown, 2004; Brown and Abeywickrama, 2019). In language teaching, compared to norm-referenced testing, criterion-referenced testing places greater emphasis on the achievement of curriculum-defined learning outcomes (Bachman, 1990; Brown, 2004; Brown and Abeywickrama, 2019; Hughes, 2003).

2.2.4. Approaches to Language Testing

2.2.4.1. Discrete-Point Testing and Integrative Testing

Discrete-point testing assesses individual elements of language separately in isolation such as only grammar or vocabulary; however, integrative testing measures the learners' ability through a holistic manner within real communication rather than as separate elements (Bachman, 1990; Brown, 2004; Hughes, 2003; Oller, 1979).

2.2.4.2. Communicative Language Testing

Incorporates the models of communicative competence and "to measure language proficiency ... account must now be taken of: where, when, how, with whom, and why language is to be used, and on what topics, and with what effect" (Weir, 1990, p.1); Moreover, the authenticity of the task is a must in this context (Brown, 2004; Brown and Abeywickrama, 2019; Hughes, 2003).

2.2.4.3. Performance-Based Assessment

In language education, performance-based assessment evaluates learners through authentic tasks like speaking, writing, integrative performances (Brown, 2004; Brown and Abeywickrama, 2019). Despite being time consuming and requiring more effort, it enhances greater content validity by directly measuring learners' ability to perform authentic language tasks (Brown, 2004; Brown and Abeywickrama, 2019).

2.2.4.4. Direct Testing and Indirect Testing

Direct testing measures a learner's ability by having them perform the intended skill to be assessed by the test (Bachman, 1990; Brown and Abeywickrama, 2019; Hughes, 2003). In

contrast, indirect testing assesses the underlying knowledge and abilities that support language skills rather than the skills themselves such as measuring linguistic components in a traditional test (Bachman, 1990; Brown and Abeywickrama, 2019; Hughes, 2003). The best way to achieve content validity of a test is through the direct assessment of students' performance (Brown, 2004; Brown and Abeywickrama, 2019).

2.2.4.5. Traditional and Alternative Assessment

The contrast between the two is illustrated in the table below by Brown (2004), who devised the table based on descriptions provided by Armstrong (1994) and Bailey (1998). However, it should be highlighted that distinguishing between traditional and alternative assessment in the manner that Armstrong (1994) and Bailey (1998) have done is challenging since many assessment types naturally blend features of both (Brown, 2004; Brown and Abeywickrama, 2019).

Table 2.2: The Difference Between Traditional and Alternative Assessment

Traditional Assessment	Alternative Assessment
One-shot, standardized exams	Continuous long-term assessment
Timed, multiple-choice format	Untimed, free-response format
Decontextualized test items	Contextualized communicative tasks
Scores suffice for feedback	Individualized feedback and washback
Norm-referenced scores	Criterion-referenced scores
Focus on the "right" answer	Open-ended, creative answers
Summative	Formative
Oriented to product	Oriented to process
Non-interactive performance	Interactive performance
Fosters extrinsic motivation	Fosters intrinsic motivation

Note. Reprinted from Brown (2004, p.13).

2.2.5. Core Principles of Language Assessment

Principles of language assessment include five profoundly well-known components; "practicality, reliability, validity, authenticity, and washback" (Brown, 2004, p.19). Nonetheless, there is an ultimate component that is often disregarded, which is fairness (McNamara and Roever, 2006).

Table 2.3: The Core Principles of Assessment

Principle	Definition	Subcomponent
Practicality	A practical test is one that remains affordable, administered within reasonable time limits, easy to manage, and has a straightforward scoring procedure (Bachman, 1990; Brown, 2004; Weir, 2005).	
Reliability	Refers to the extent to which results remain consistent across different test administrations, raters, and versions; A reliable test produces stable and dependable outcomes regardless of external variables (Bachman, 1990; Brown, 2004; Hughes, 2003; Weir, 2005).	Student-related reliability: Learner related reliability issues that stem from temporary factors such as fatigue, illness or anxiety (Brown, 2004; Weir, 2005). These factors can cause a difference between a learner's observed and true performance (Brown, 2004; Weir, 2005). Rater reliability/ Scoring validity: Bias, error and subjectivity are rater related factors that can affect the reliability of the scoring (Bachman, 1990; Brown, 2004; Weir, 2005). Thus, interrater reliability is needed as it is concerned with the consistency of scores assigned by different raters (Bachman, 1990; Brown, 2004; Weir, 2005). However, inexperience or disregard for scoring criteria can reduce this consistency. Reliability isn't a matter solely between two raters. Intra-rater reliability relates to factors such as an individual teacher's fatigue, bias, inattentive scoring or unclear assessment criteria that may influence scoring consistenct (Bachman, 1990; Brown, 2004; Weir, 2005). Test Administration Reliability: Refers to the physical aspects of the test taken such as environment, seating conditions, timing, equipment and distractions (Brown, 2004; Cohen, 2001; Weir, 2005). Along with the procedural aspects such as task familiarity, instructions, examiner behavior (Bachman, 1990; Brown, 2004; Cohen, 2001; Hughes, 2003; Weir, 2005). Test reliability: The structure of a test can cause reliability issues (Bachman, 1990; Brown, 2004; Cohen, 2001; Hughes, 2003; Weir, 2005). Extremely long tests may cause fatigue, and time—constraint tests may disadvantage students who have anxiety to perform under pressure (Brown, 2004) Moreover, ambiguous or poorly written items can contribute to measurement error and unreliability (Bachman, 1990; Brown, 2004; Cohen, 2001; Hughes, 2003).

Table 2.3 (continued): The Core Principles of Assessment

Validity	Refers to the extent to which a test correctly measures the intended construct and whether the results are meaningful for their intended purpose (Brown, 2004; Weir, 2005).	Content validity - when a test accurately reflects the material it aims to assess and requires learners to perform the required skills that are to be measured (Bachman, 1990; Brown, 2004; Hughes, 2003; Weir, 2005). Criterion-related validity – the degree to which a test's results correspond to those of an established instrument measuring the same construct (Bachman, 1990; Brown, 2004; Hughes, 2003; Weir, 2005). There are two versions of this validity: concurrent validity and predictive validity. The former refers to how closely a test aligns with another established test of the same skill, whereas the latter examines how well test results can anticipate future performance outcomes of the same skills (Brown, 2004; Weir, 2005). Construct validity – This term describes how effectively a test measures the theoretical construct it is designed to assess (Bachman, 1990; Brown, 2004; Hughes, 2003; Weir, 2005). In language learning, constructs are ones such as proficiency, communicative competence, motivation, self-esteem among many others (Brown, 2004). Consequential validity – Refers to the effect a test has such as measurement accuracy, test preparation, the impact on the learner in addition to its influence on social outcomes (Brown, 2004; Weir, 2005). Face Validity- This term denotes how appropriate and credible a test appears to test-takers in terms of content and appearance (Bachman, 1990; Brown, 2004; Hughes, 2003; Mosier, 1947; Weir, 2005). A test tends to have higher face familiarity when its format and tasks are familiar to test-takers, clearly presented, manageable within the given time, connected to course content, and set at a suitable level of difficulty (Brown, 2004; Weir, 2005). However, because face validity is grounded in the test-taker's perceptions, they are best positioned to judge it (Allen et al., 2023).

Table 2.3 (continued): The Core Principles of Assessment

Authenticity	Refers to how accurately a test task represents real-world communicative language use. A test is considered authentic when it uses natural language, presents items in meaningful contexts, includes relevant and engaging topics, follows a clear theme or storyline, and resembles real life communication tasks (Bachman, 1990; Bachman and Palmer, 1996; Brown, 2004; Weir, 2005). The principle of authenticity is also connected to the manner in which the test is conducted (Bachman, 1990).	
Fairness	Refers to assessments being designed, implemented and scored in ways that ensure equitable treatment of all candidates and that the test offers an accurate and valid representation of the construct it aims to assess (McNamara and Roever, 2006).	Justice – Is referred to as being an extension of fairness because it refers to the broader social, ethical, and policy implications of test use (McNamara and Ryan, 2011). While fairness focuses on the internal technical qualities of assessments, justice is concerned with whether the purposes and consequences of the test are socially justifiable (McNamara and Ryan, 2011). Both of these are overlooked in assessment; however, the validation of assessment must also consider the social roles and policy implications of language tests (McNamara and Ryan, 2011).
Washback	Refers to the influence a test has on both teaching and learning (Alderson and Wall, 1993; Bachman, 1990; Brown, 2004; Hughes, 2003; Weir, 2005).	Positive washback – Refers to the beneficial influence of testing on teaching (Alderson and Wall, 1993; Bachman, 1990; Hughes, 2003). It occurs when test formats align with and accurately reflect the skills and abilities that were emphasized in the course (Bachman, 1990). Direct testing, making assessment criterion-referenced, and test familiarity to learners result in positive washback (Hughes, 2003). Negative washback – this term denotes the potentially harmful experience of a test on instructional practices and learner development such as test-induced anxiety on learners due to performing abnormally under pressure, teacher anxiety due to poor results and focusing instruction on the content of the test, which would lead to reducing the curriculum (Alderson and Wall, 1993).

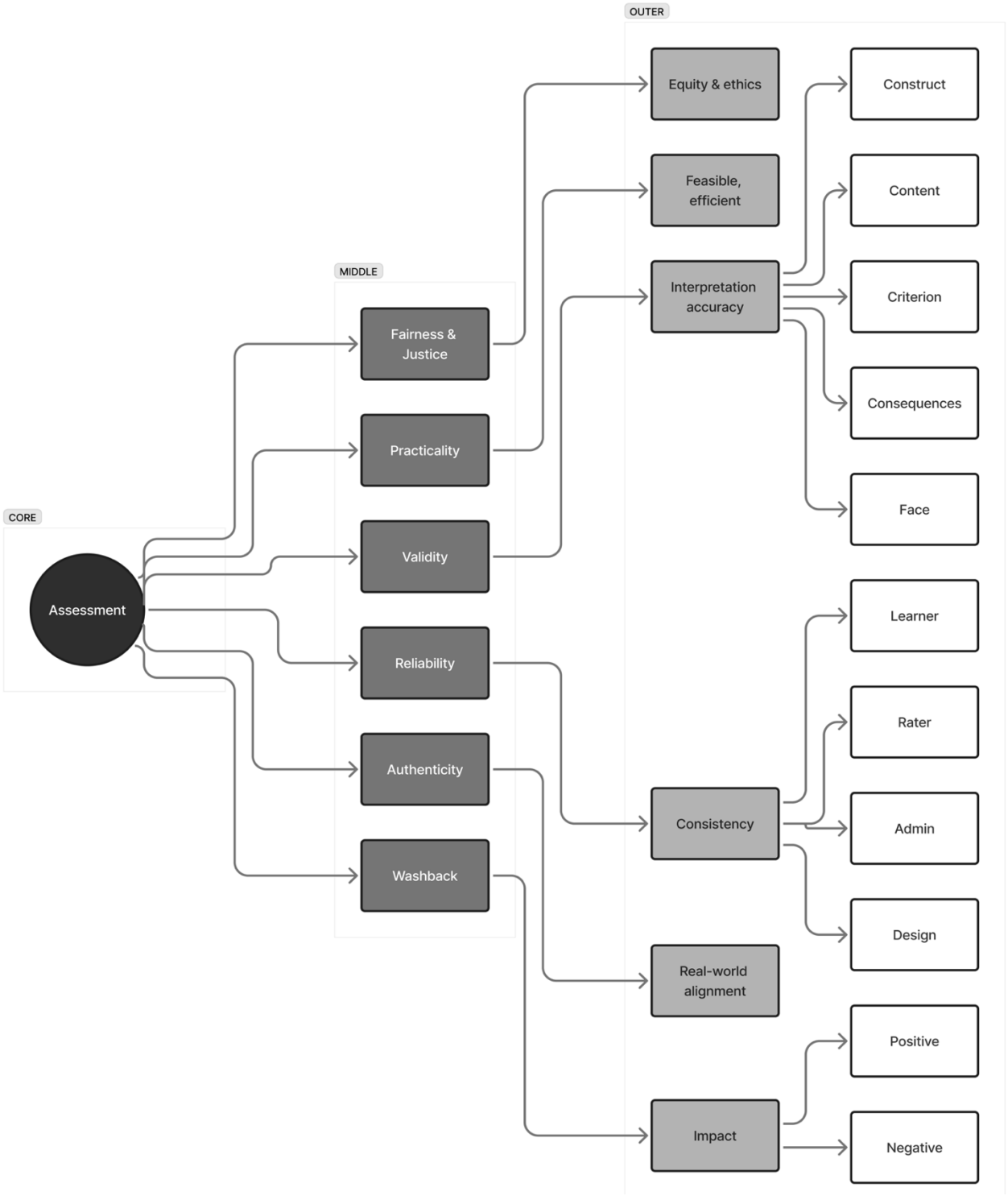


Figure 2.2: The Tri-Layered Concentric Ring Model of The Core Principles of Assessment

2.2.6. Purposes of Assessment

2.2.6.1. Achievement Test

The primary function of this testing is to measure students' learning outcomes aligned with a particular lesson, unit, or overall curriculum, so it is in line with course objectives and administered after relevant instruction has taken place (Bachman, 1990; Brown and Abeywickrama, 2019; Hughes, 2003).

2.2.6.2. Diagnostic Test

This test seeks to highlight learners' strengths and weaknesses connected to specific language areas that necessitates further development (Bachman, 1990; Brown and Abeywickrama, 2019; Hughes, 2003).

2.2.6.3. Placement Tests

This test is utilized to determine the most suitable level for a learner within a language program. It typically includes sample content from different courses to assist in this determination so that the student can be placed at an appropriately challenging level (Bachman, 1990; Brown and Abeywickrama, 2019; Hughes, 2003).

2.2.6.4. Proficiency Tests

This test measures a learner's overall ability rather than focusing on a specific course, curriculum or skill. These tests usually incorporate all of the four main skills and the two subskills of grammar and vocabulary (Bachman, 1990; Brown and Abeywickrama, 2019; Hughes, 2003).

2.2.7. Types of Scoring

2.2.7.1. Analytic Scoring

Raters score each criterion separately (Fulcher, 2003; Kuiken and Vedder, 2021; Weir, 2005).

2.2.7.2. Holistic Scoring

Raters assign one score to a production to represent its overall quality (Cooper, 1977; Fulcher, 2003; Kuiken and Vedder, 2021; Weir, 2005).

2.2.7.3. Primary Trait Scoring

Raters specifically assign an overall score to the requirements of an individual task (Fulcher, 2003; Kuiken and Vedder, 2021).

2.2.7.4. Multiple Trait Scoring

Raters score multiple aspects of performance separately (Fulcher, 2003; Kuiken and Vedder, 2021). Thus, this scoring is situated between analytic and holistic scoring as it is alike both of them (Kuiken and Vedder, 2021), especially as it includes task specific performance criteria and communicative competence. The distinction between analytic scoring and multiple trait scoring seems to be a debate topic among scholars as some refer to both as being the same, whereas others separate the two (Kuiken and Vedder, 2021).

It is significant to mention that all of these scoring types have their own rating scales.

2.2.8. Types of Scales and Rubrics

While the term "rating scale" had been used more commonly and consistently until now, the term "rubric" is equally common and even more familiar to educators and researchers nowadays (Baker and Turner, 2022). Rubrics are defined as descriptive statements that specify learners' performance at particular levels on a rating scale and are also known as rating scales or band descriptors (Brown and Abeywickrama, 2019). However, there isn't a consensus on this issue as some differentiate between the two terms. Some refer to them as meaning the same (Baker and Turner, 2022; Brown and Abeywickrama, 2019; CoE, 2001; Fulcher, 2003; Kuiken and Vedder, 2021). In contrast, some view rating scales to assess students' performance and their corresponding success level, whereas rubrics provide explicit performance descriptions for each criterion at every achievement level for them (Brookhart, 2018; Brookhart and Chen, 2015; Doğan and Uluman, 2017). Thus, in the context of this study, the two terms refer to the same definition.

2.2.8.1. Rating Scales or Rubrics

They are a sequence of increasing proficiency levels that are used as reference points to judge language performance (CoE, 2001).

2.2.8.2. Rating Criteria

Particular components of the construct that form the basis to evaluate the performance such as linguistic accuracy, fluency, organization, and others (Brown and Abeywickrama, 2019; Fulcher, 2003; Kuiken and Vedder, 2021; Luoma, 2004).

2.2.8.3. Bands

A scale level that details what a learner can do in the language and linguistic features of the performance; a rating scale incorporates several of these levels (CoE, 2020; IELTS, 2025; Kuiken and Vedder, 2021).

2.2.8.4. Descriptor

A statement that specifies what a learner's performance looks like at a specific level; each band contains one or more descriptors for each criterion (CoE, 2020; Kuiken and Vedder, 2021).

2.2.8.5. Functions of Scales

As defined by Alderson (1991) the functions of scales are listed as user oriented, assessor oriented, or construct orients. User oriented scales describes what test takers can typically do at a certain level such as "Can do..." statements. Assessor oriented scales guide raters in evaluating performances. Construct oriented scales supports test designers in choosing suitable tasks for the assessment

2.2.9. Conceptual Models of L2 Speaking Assessment

Below are three figures of conceptual models that portray the assessment of speaking performance. All of which are well-recognized. The first is Skehan's (2001) model that incorporates internal factors, task variables, external factors, and the outcome of the performance. All of these are key concepts in spoken production as they display how speaking performances happen. However, it lacks major notions regarding speaking assessment such as rating scale, decisions and consequences, along with other missing components, which is why Fulcher's (2003) revised model proves to be beneficial.

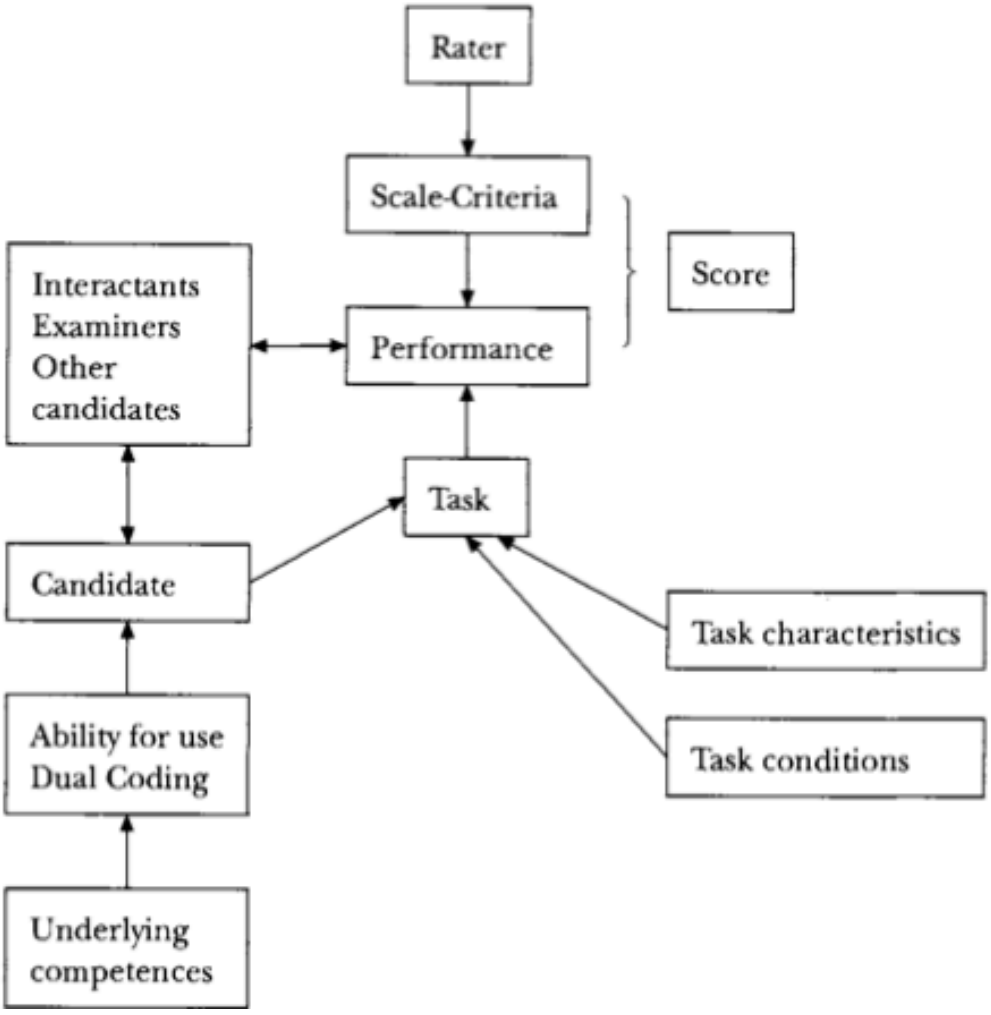


Figure 2.3: Skehan's Model of The Assessment of Speaking Performance

Note. Reprinted from Skehan (2001, p.169)

The second is Fulcher's (2003) model. It is built upon Skehan's (2001) but it is much more assessment oriented. It encompasses entire layers missing from Skehan's (2001) model, such as rater characteristics, rater training, scoring philosophy, rating scale/ band descriptors, local test conditions, and construct definition. Thus, this version elaborates on the full assessment and interpretation of the spoken production process.

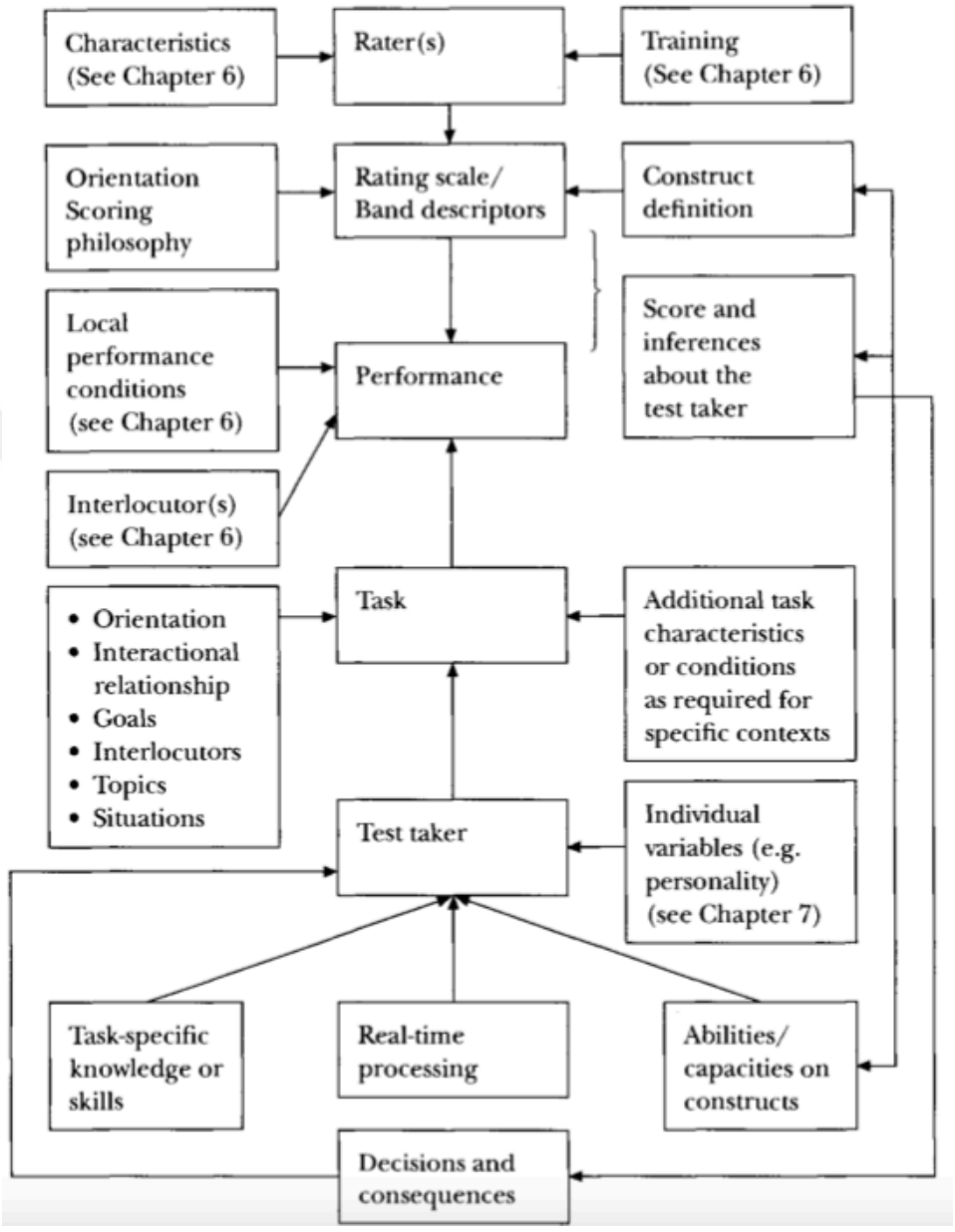


Figure 2.4: Fulcher's Expanded Model Based on Skehan's Model

Note. Reprinted from Fulcher (2003, p. 115).

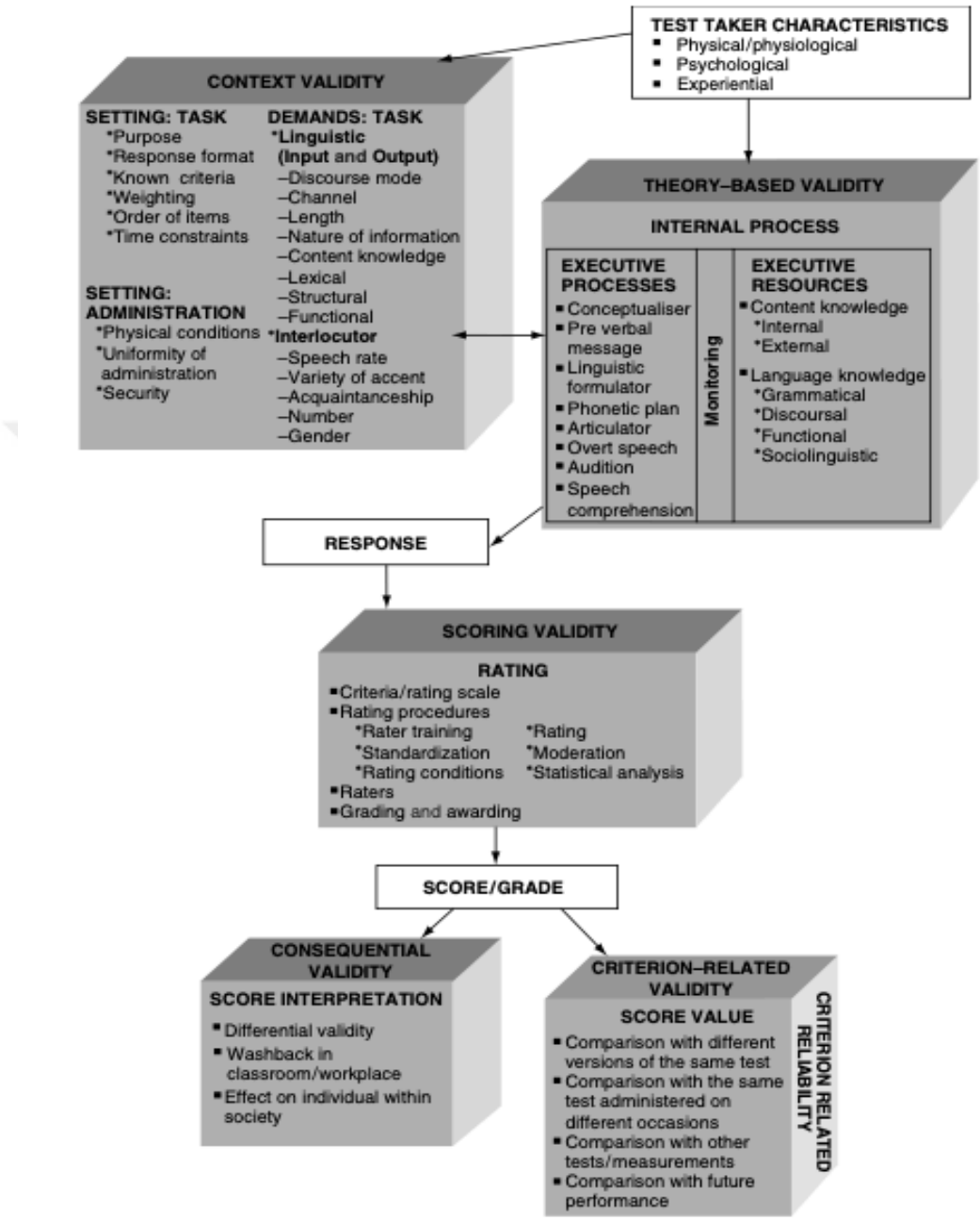


Figure 2.5: Weir's Socio-Cognitive Framework for Assessing and Validating Speaking Performance

Note. Reprinted from Weir (2005, p. 46).

The third is Weir's (2005) socio-cognitive framework that displays a comprehensive approach to validating speaking assessments by examining the dynamic relationship between the

test taker, the task itself and assessment procedures. This model highlights task authenticity, the cognitive and linguistic processes involved, the reliability of scoring procedures, the alignment between test scores and other measures of proficiency, and the wider educational and social effects of test use. All together, these components ensure that the evaluation of speaking is accurately and fairly determined.

Fulcher's (2003) model and Weir's (2005) socio-cognitive framework address speaking assessment from complementary but distinct perspectives. Fulcher (2003) conceptualizes speaking assessment as an integrated system in which performance emerges from the interplay among the test-taker, the task, interlocutors, and contextual performance conditions. All of these are interpreted through rater behaviour, rating scales and scoring philosophies. His model highlights sources of score variation by explicitly incorporating rater training, band descriptors, interactional dynamics and individual test-taker variables. Thus, it emphasizes how speaking ability is operationalized and interpreted within a specific assessment context. In contrast, Weir's (2005) socio-cognitive framework is structured as a validity-oriented model that follows the development of test performance and score meaning through successive stages of context validity, theory validity, scoring validity, criterion-related validity and consequential validity. Rather than focusing primarily on the procedures of rating, he stresses the alignment between task demands, cognitive processes, scoring procedures, and the purposes for which the test scores are intended to be interpreted and used. Moreover, Weir's (2005) socio-cognitive model addresses fairness more explicitly and thoroughly as its embedded ethical considerations across multiple dimensions of validity, which includes context, cognitive processing, scoring, and consequential validity. In this model, fairness is examined through the appropriateness and accessibility of tasks, the alignment between task demands and the cognitive processes they prompt, the consistency and transparency of scoring procedures and the implications of score interpretation and use for learners and institutions. In contrast, Fulcher's (2003) model contributes to fairness in a more indirect manner by highlighting potential sources of score variation such as rater behavior, task characteristics and interactional variables without placing fairness at the centre of the validity argument. In brief, while Fulcher's (2003) model provides a detailed account of how speaking performance is produced and rated within an assessment system, Weir's (2005) framework offers a broader

validity argument that justifies why scores can be meaningfully interpreted and used. Both of these models provide a comprehensive foundation for speaking assessment by linking construct realization and rating practices with evidence-based validity claims.

The figure on the following page is a conceptual model of the interconnected ecosystem of speaking assessment, in which multiple components dynamically interact to shape learners' spoken performance and the outcomes of assessment. In this model, the construct definition provides an interpretive framework for understanding speaking ability rather than functioning as a prescriptive starting point for task design or rating procedures. Task characteristics, performance conditions, and test-taker characteristics interact with one another by displaying how cognitive, contextual, affective, physiological and background collectively affect students' spoken output. This interaction results in the production of the observed performance, which is observed through fluency, accuracy and complexity. This production is then interpreted by raters whose judgments are carved by rating scale descriptors, scale use, potential bias, scoring severity or leniency, and rater training and calibration. The final assessment outcomes in terms of validity, reliability and fairness are further examined through criterion-related validity (i.e., external alignment with others), and consequential validity (i.e., washback, ethics, and appropriate score interpretation) and also give feedback to the construct itself and the broader testing context. This feedback loop suggests that construct interpretations may be redefined by empirical evidence in addition to emphasizing that the test's impact must be continuously monitored. Therefore, clearly, this model does not propose a new theoretical approach, rather it synthesises central elements from Skehan's (2001), Fulcher's (2003) and Weir's (2005) established frameworks into a single visual representation that is straightforward, conceptually coherent and doesn't cause cognitive strain since it offers a holistic overview of factors influencing L2 speaking assessment. Hence, its contribution merely lies in its integration and clarity of presentation. In brief, this model is a clear indicator that speaking assessment is not linear. In fact, it is a highly fluid and interrelated process where each component mutually moulds one another. Nevertheless, the critical consideration to be made here is the task of isolating, measuring and evaluating speaking performance proving to be a challenge due to this complex interdependence. Therefore, referring to this process as merely a challenge would be an understatement, which can be especially understood from the information provided below that elaborates on the impediments of this process.

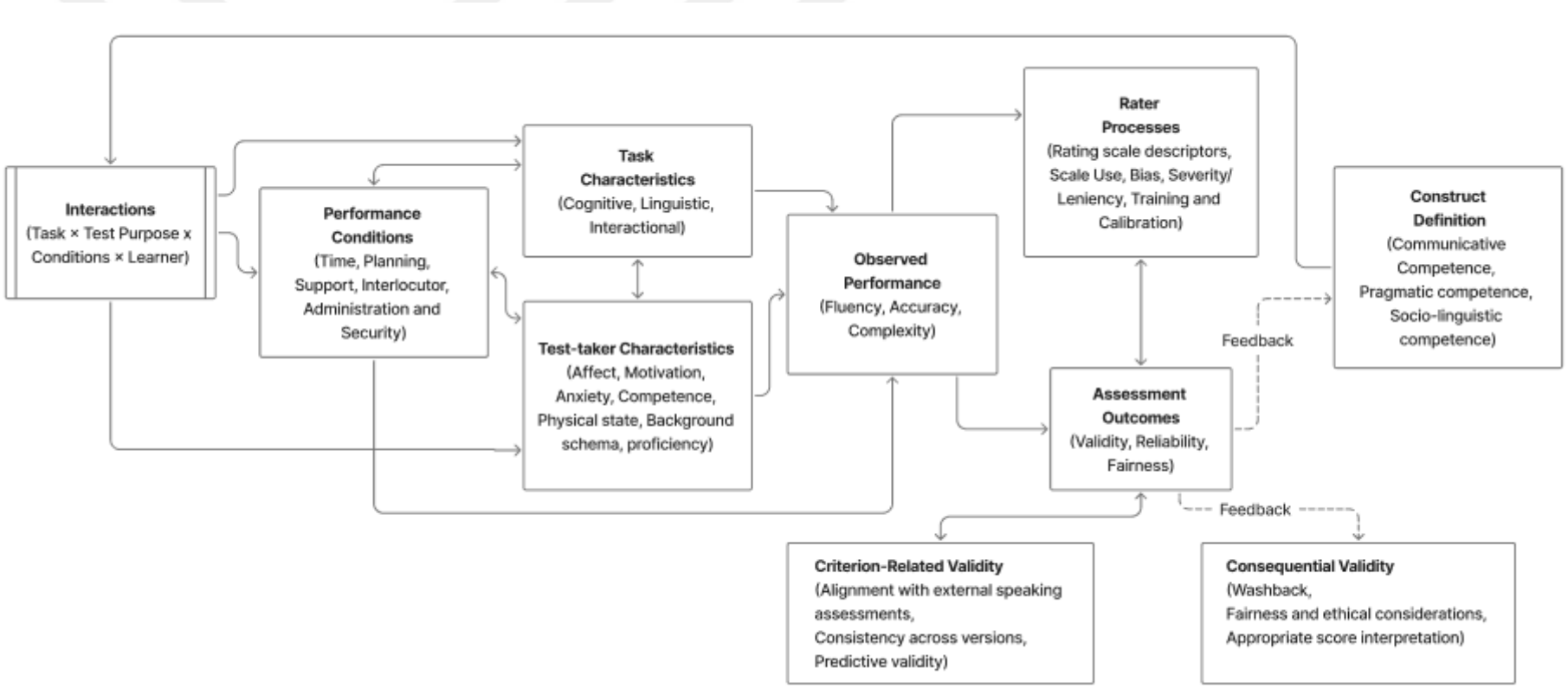


Figure 2.6: The Peculiarly Entangled Machinery of Speaking Assessment

2.2.10. Challenges in Assessing Spoken Production

There are three core components of L2 speaking assessments and they can be classified under the representation of constructs, the design of test formats and tasks and the procedures used for scoring and providing feedback (Nakatsuhara et al., 2020). The main challenges pertaining to assessing oral abilities is similar to that of testing writing skills (Hughes, 2003; Isaacs, 2016). However, Isaacs (2016) asserts spoken language, because it is less tangible, is significantly harder to evaluate than written language. The procedure of which, Lundeberg (1929) illustrates as time-consuming and labour-intensive, which can still be said to remain valid for the assessment of speaking skills in the current times. Nonetheless, assessment involving spoken production is a matter to be dealt with by using direct or semi-direct formats as they are acknowledged as the only valid means of evaluation for L2 oral production (O'Loughlin 2001), especially as indirect test items were found to produce extremely low reliability and related validity concerns (Buck, 1989; Lado, 1961). Task design represents one of the most demanding aspects of speaking assessment as it requires to represent a balanced and realistic sample of the range of oral communication skills learners are expected to perform (Hughes, 2003). The second, which is initially linked to the first, is ensuring that these tasks elicit authentic and representative language use that accurately reflect learners' communicative competence (Hughes, 2003), especially since speaking assessment relies on clearly defined constructs and test tasks are essential for eliciting language in performances in order to measure them (Nakatsuhara et al., 2021). Therefore, prior to the assessment of speaking, the definition of precisely what speaking ability consists of such as communicative competence, pragmatic competence etc. is required. Once the construct is defined, the test tasks should be structured to elicit performances that actually reveal those abilities so that the rubric could meaningfully measure them, which corresponds to the last issue mentioned by Hughes (2003). The last one is scoring the performances in a valid and reliable manner to ensure consistency and fairness (Hughes, 2003). The last statement is applicable for cases such as assessing spoken production when assessment is dependent on the listener's judgment as providing solely objective scoring cannot be verified (Isaacs, 2016). This challenge is more apparent in oral assessment since open-ended tasks are inherently harder to score consistently since test-takers have more freedom in self-expression and response (Brown and Abeywickrama, 2019). This brings about the issues regarding interrater reliability and rater training to ensure consistent scoring in agreement that doesn't jeopardize the factor of fairness. Thus, interrater reliability and the standardization of

scoring of these raters is needed to be able to offer reliable and valid findings (Weir, 1990; Wilkinson, 1968).

Fulcher (2003) agrees with this view and elaborates by mentioning that standardization is a process that ensures raters interpret and use the rating scale as intended. Studies have presented that standardization training is especially helpful for novice raters or those who rate too harshly or too leniently (known as rater severity) (Bijani et al., 2022; Davis, 2015; Elder et al., 2005; Weigle, 1998) as it aids their scoring to become more consistent with that of other raters (Davis, 2015) or due to it improving intra-rater reliability as the rater gives more consistent scores for themselves rather than interrater reliability (Bijani et al, 2022; Weigle, 1998) and become less biased (Elder et al., 2005). Moreover, outcomes have indicated the potential of brief and user-friendly training sessions that may assist in reducing rater bias at least in the short term (Bijani et al, 2022; Kang et al., 2019). Nonetheless, there have been many other studies that have presented contrary results to this such as additional training for experienced raters often producing only minimal changes in scoring due to rater severity persisting even after standardization (Bijani et al, 2022; Eckes, 2011). In fact, Fulcher (2003) makes a firm assertion that relying solely on scores from trained rater may not provide strong validity evidence as their judgments are shaped by the their prior experience rather than by the inherent clarity of the rating scale; however, if untrained raters are able to implement the descriptors of a new scale with limited external guidance or naturally pay attention to the similar performance features on the rating scale, this would indicate the scale is interpretable and grounded in observable qualities of spoken performance. Thus, when rater scores are used to support a validity argument, it is important to consider whether those scores truly reflect the quality of the scale rather than the effects of the rater's prior experience and training (Fulcher, 2003). In addition to these, other studies have resulted in no significant improvements in inter-rater consistency, rater bias or intra-rater reliability after standardization as well (Knoch, 2011; O'Sullivan and Rignall, 2007). The difference in results could be due to the rater's familiarity with a particular test, background, skills or their general experience engaging with language learners (Davis, 2015). Overall, significant improvements in inter-rater consistency, rater bias or intra-rater reliability are not consistently observed across most assessment related studies. That being said, it is widely acknowledged in the literature that high rater reliability in speaking assessments is

relatively easy to attain (Fulcher, 2003). Especially, reviewing and discussing the scoring rubric is expected to direct the rater's attention to the features of a performance the assessment intends to evaluate (Fulcher, 2003) as standardizations don't guarantee that raters will use the rubric appropriately to achieve full consensus (Bijani et al, 2022; Eckes, 2011). Hence, the salience of interrater reliability cannot be denied in assessing spoken production; however, the main strength lies in being able to attain the high rater reliability relatively easily, which proves it to be not a major issue.

In addition to the scoring rubric, using exemplary performances or practice scoring may be crucial for helping raters align their perceptions to the scale, particularly when the descriptors lack clarity (Davis, 2015). This claim is strengthened by psychological and behavioral economics, which underline that individuals assess an item by referencing other comparable items either present in the immediate environment or recalled from memory in a comparative process (Pollitt, 2012; Stewart et al., 2006). This is similar to the manner in which raters assess the different levels in which a learner's speaking performance should be placed even when using a rubric.

Among the different formats for assessing spoken production, the interview is the most frequently used one (Brown and Abeywickrama, 2019; Hughes, 2003; Luoma, 2004). Despite this, it is regarded as one of the most challenging independent task types as it entails more structured and coherent speech (Nakatsuhara et al., 2021). Additionally, it has drawbacks regarding the power dynamic between the interviewer and the candidate since it often inhibits natural communication, which results in limited speech functions being elicited (Brown and Abeywickrama, 2019; Hughes, 2003). This issue can be overcome by incorporating varied elicitation techniques that can help produce more representative and interactive language samples such as the utilization of asking questions and requesting further elaboration, pictures, role-plays, prepared monologue and other techniques as the operations in such representative tasks of speaking include expressing, directing, describing, eliciting, narrating and reporting (Hughes, 2003). However, the problem which cannot be overlooked in paired speaking activities is that it encompasses socio-interactional tasks and its problems pertain to how test takers should be assessed such as individually or jointly, which brings about further issues such as defining the speaking construct along with interactional competence

and the notion of fairness of scoring (Isaacs, 2016). As paired activities result in further complications, it can be said that spoken production is best assessed through tasks that require performances of monologues.

Furthermore, the complete isolation of speaking from listening is difficult for assessing spoken production as they are interrelated skills (Brown and Abeywickrama, 2019). Only specific forms of speech (such as monologues, telling a story, speeches, reading aloud and others) allow speaking to be assessed independently of interaction with a listener (Brown and Abeywickrama, 2019). Even though speaking is an observable productive skill, its assessment is often affected by the test-taker's listening abilities and this interdependence may reduce the reliability and validity of oral production assessment (Brown and Abeywickrama, 2019). This further strengthens the claim to favour formats such as interviews in which there is extended speech due to monologue so as to not diminish the reliable and valid results of oral skills assessment. The extensive speaking tasks involve little or no listener interaction because they are planned and formal in nature; however, they can encompass elements of informal speech such as personal anecdotes or explanations (Brown and Abeywickrama, 2019), which relates back to Hughes' (2003) techniques that require certain functions of spoken production to elicit more language from the test taker in an interview format.

The interlocutor effect refers to the influence an interlocutor could possibly have on a test-taker's spoken performance and assessment results (Brown, 2004; Brown and Abeywickrama, 2019; Hughes, 2003; Nakatsuhara et al., 2021). In relation to speaking test anxiety, it has been indicated that interlocutor behaviors can either be the cause or the remedy for test-taker's stress depending on the interlocutor's approach (Fulcher, 2003). As mentioned above, interlocutor related variables that cause power dynamics are due to the interlocutor's status (higher, lower, or equal to the test-taker), and the degree of familiarity can have an impact on the outcome of the assessment, which may affect the performance by changing the formality or informality of the interaction (Brown and Abeywickrama, 2019; Fulcher, 2003; Hughes, 2003). In addition, in speaking exams, some interlocutors try to assist the student's performance by providing accommodation strategies such as lexical simplification or reducing their speech rater, whereas

others try to make it more difficult for the student by restricting their support by asking more closed questions, accepting shorter response and permitting longer silences in the interaction (Brown, 2004). The core issue is whether supportive or non-supportive interlocutor behavior creates a fairer and more valid speaking test (Brown, 2004). Initially it was accepted that coming to the conclusion that the variations in interlocutor behaviors can cause a difference in the outcome of the speaking performance is merely theoretical as there was no available research that demonstrates that these variations of interlocutor behavior certainly led to unfair or biased rating (Lazaraton, 1996); however, this view was later reconsidered as interaction and meaning focused research gained prominence that caused researchers to question the dynamics between the interviewers and candidates in speaking interviews, which led to the view of recognizing spoken performance as socially situated and acknowledging that interlocutors in speaking assessments influence one another within co-constructed interaction (Ducasse and Brown, 2009; He and Young, 1998; McNamara, 2001; Seedhouse and Nakatsuhara, 2018). The interlocutor's effect on the outcome of the speaking assessment revolved around accent (Alderson, 1991; Weir, 1990), speech rate, degree of familiarity, number of participants in the interaction, and gender (Weir, 1990). In speaking exams that require monologues, the interlocutor is the examiner. Therefore, these factors are also in relation to the examiner in such assessment formats. The problems stemming from the interlocutor's effect on the performance outcome is due to interlocutor training not being developed or prioritised to the same extent as rater training. (Fulcher, 2003). Even though there is no one set way of interlocutors to conduct the speaking assessment, there have been suggestions or guidelines provided to aid with the testing. According to Buck (1989), interlocutors should refrain from slowing their speech, echoing or correcting candidates, providing vocabulary or assisting with their syntax, pressuring response times, asking display questions (questions whose answers are already known) or overusing yes/no questions, and changing topics abruptly. Moreover, recommendations outlined in the VOTE: Oral Testing Video of British Council (1983) suggest that interlocutors shouldn't correct errors, speak hastily, mumble, dominate the exchange, make condescending or culturally inappropriate comments or ask about unrelated tasks during the test, but they should maintain eye contact throughout assessment; all of which is provided as advice to ensure fair assessment conditions.

Moreover, the rating scale used to assess the spoken production can also contribute to creating issues in evaluating speaking due to it affecting the differences in scoring (Isaacs, 2016). Therefore, scales that are practical and user friendly should be used (Isaacs, 2016; Weir, 1990). Nevertheless, this practical type of rating scale can make it difficult to fully represent complex or wide-ranging constructs (Davis, 2015; Fulcher, 2003; Isaacs, 2016; Isaacs and Thomson, 2013; Isaacs and Trofimovich, 2012), such as the descriptors in the IELTS bands for the speaking rubric.

Regardless, as part of the standardized language proficiency tests outlined by Brown (2004), IELTS is included as one of the four principal examples. He argued that standardized tests offer efficiency, consistency, and valuable insights into learners' language skills; despite these findings, careful interpretation is required as the tests may overlook important contextual and interactive aspects of the language performance (Brown, 2004).

In brief, the effectiveness of an oral interview relies on several key factors; establishing clear administrative procedures to ensure practicality, designing questions that align with the assessment's aims to maintain validity, eliciting sufficient and high-quality language output from the test taker to promote optimal performance, minimizing the negative influence of the interviewer and interviewee power dynamics and implementing a consistent and reliable scoring system (Brown and Abeywickrama, 2019).

A key construct issue regarding pronunciation in speaking assessments is the notion of what constitutes intelligibility to be counted as acceptable since it varies widely among raters (Harding, 2012; Issacs, 2016; Issacs and Harding, 2017; Issacs and Trofimovich, 2012). This is due to intelligibility of pronunciation being judged by listeners' subjective rating so the scales appear to reflect comprehensibility rather than true intelligibility (Isaacs and Thomson, 2013; Issacs and Harding, 2017; Issacs and Trofimovich, 2012; Saito and Plonsky, 2019). The subjectivity of accessing pronunciation has the potential to cause systemic bias and raise issues regarding fairness as this scoring can be influenced by the rater's linguistic background, experience, or sensitivity towards accents (Carey et al., 2011; Isaacs and Harding, 2017; Winke et al., 2013). In the past, there was a distinction between the terms of comprehensibility and

intelligibility (Isaacs and Harding, 2017; Isaacs and Trofimovich, 2012). Intelligibility describes the learner's ability to comprehend spoken language and so it's treated as being interchangeable with comprehensibility (Levis, 2006). In contrast, comprehensibility relates to listeners' evaluations of the ease of understanding speech and is commonly measured using rating scales focusing on perceived pronunciation ease (Munro and Derwing, 1999). Nonetheless, the distinction between the two is not applicable to current times as rating scales present intelligibility as comprehensibility (Isaacs and Trofimovich, 2012). Thus, nowadays these two terms are used interchangeably. A shortcoming concerning L2 oral proficiency scales is the inconsistent or even absent representation of pronunciation (Issacs, 2016; Issacs and Thomson, 2013; Issacs and Trofimovich, 2012). Even if pronunciation is incorporated in the rating scale, the descriptions tend to be too general to articulate a well-defined construct even in large-scale assessments such as TOEFL and IELTS (Isaacs and Thomson, 2013; Issacs and Trofimovich, 2012; Saito and Plonsky, 2019). Furthermore, it is unclear whether pronunciation refers to only segmental errors (individual sounds) or if it includes suprasegmental features such as intonation, rhythm, word stress and others (Isaacs and Trofimovich, 2012; Saito and Plonsky, 2019). A further limitation in relation to this issue is the terms of comprehensibility and accentedness being merged into a single construct (Harding, 2012; Saito and Plonsky, 2019). The focal point to reach a conclusion for this issue is that having an accent doesn't automatically equate to making speech strenuous to understand; however, listeners often focus on it because it is extremely noticeable (Derwing and Munro, 2009). Hence, even if raters assign the same score, they may do so for entirely different reasons (Isaacs and Thomson, 2013). The overlap of these two terms in rating scales can lead to misinterpretation of the rating scale and result in invalid scoring as it's unclear what exactly is being measured in terms of pronunciation. This is due to the inadequate research of empirical work (apart from accent-focused studies) that outlines the defining features of comprehensible speech for use in scale developments (Isaacs and Trofimovich, 2012).

In relation to the rubrics, there are issues regarding both of the major scoring types, which are analytical and holistic as mentioned above. The utilization of analytical rubrics can result in the halo effect, which is when a high score is assigned to one criterion and that may unintentionally increase the scores of the other criteria (Weir, 2005). The limitation of holistic rubrics is that it

overlooks the various components that contribute to the production skill such as speaking, which is in fact composed of several distinct elements (Fulcher, 2003). Therefore, assessment through the use of the appropriate construct for the right purpose is necessary to ensure the inclusion of the required subskills to establish criterion validity (Fulcher, 2003; Weir, 2005). Nonetheless, as explicit descriptors in analytic rubrics aid in the assessment of performance for specific strengths and weaknesses, it provides richer diagnostic information (Kuiken and Vedder, 2021). Hence, since such detail cannot be obtained from holistic rubrics, it can be said that analytic rubrics are more valuable as they provide better insight and more accurate results regarding the outcome of the speaking performance, which is why they should be used instead.

Test-taker affective variables are also a challenge not only in the EFL classrooms but also in spoken production assessments (Bao and Liu, 2021; Ellis, 2008; Fulcher, 2003; Horwitz et al., 1986; Kormos and Préfontaine, 2016; Sutarsyah, 2017; Waluyo and Bakoko, 2022). Affective variables pertaining to the test-taker's individual differences encompass factors such as self-esteem, hesitation, anxiety, personality, motivation, attitude and a range of other factors (Bao and Liu, 2021). As delineated by Bachman and Palmer (1996) affective factors are impacted by personal characteristics, familiarity with topics, language knowledge, strategic competence and characteristics of the task itself and its setting; all of which are also connected to the test candidate's past experience and prior success. To put simply, affective factors are influenced by the learner (traits, knowledge, and strategies) and the task context (difficulty, conditions, interlocutor, etc.). In the figure below, it can be observed that Bachman and Palmer (1996) place affect at the centre to be depicted as the focal point as emotional reactions mediate how knowledge and strategies emerge in accordance with the context during an assessment performance.

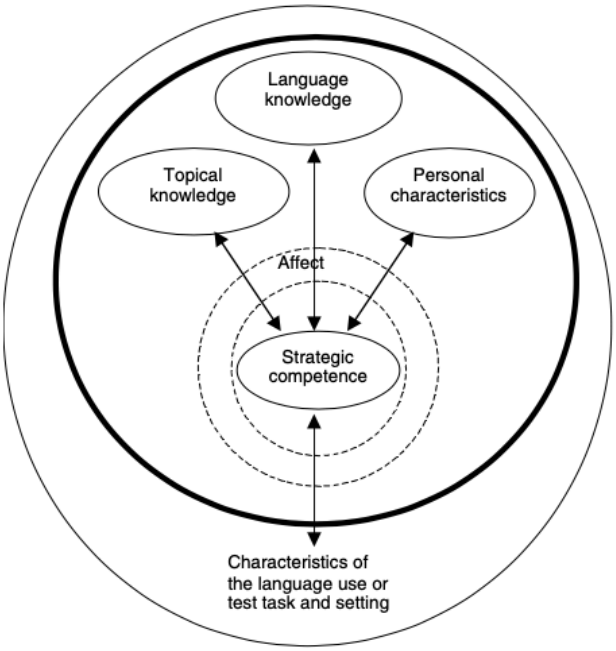


Figure 2.7: Interaction of Cognitive, Affective, and Task-Related Factors in Speaking Assessment

Note. Reprinted from Bachman and Palmer (1996 p.63).

Moreover, test-takers believe that individual personality traits such as shyness or low confidence, may negatively influence the outcome of their speaking performance (Fulcher, 2003). The task cognitive load impacted the test-takers' affective reactions and also their measurable and perceived fluency (Kormos and Préfontaine, 2016). While task anxiety and perceived success were the strongest affective predictors of test-taker's fluency, interest and motivation played a greater role in native speaker's fluency judgments (Kormos and Préfontaine, 2016). The results have demonstrated that speaking anxiety adversely affects the learner's speaking performance as learners with lower anxiety level achieved higher scores than those with greater anxiety (Sutarsyah, 2017). Nervousness was identified as the primary source of anxiety followed by worry and tension (Sutarsyah, 2017). However, a separate study indicated that self-confidence, motivation, and perseverance had positive correlation with speaking assessments, while these factors themselves were negatively correlated (Waluyo and Bakoko, 2022). Despite these relationships, motivation was the sole salient predictor of speaking assessment outcomes in this study unlike the other one mentioned previously (Waluyo and Bakoko, 2022). Another study's findings displayed that regardless of gender, trait anxiety and language-related anxiety directly

contributed to learners' anxiety during the assessment, whereas test anxiety indirectly had an influence on this issue (Huang, 2018). Nonetheless, Ellis (2008) claims that instructional practices need to focus on cultivating a balanced and productive form of anxiety so that it motivates learners rather than interfering with their performance. It is observed that most componential models are beneficial but too rigid and linguistically focused (Luoma, 2004). Therefore, they omit the internal, affective and interactional aspects of communication (unlike the figure mentioned above), so assessment developers should also take into consideration these elements too (Luoma, 2004). Putting all of this aside, there was a case that portrayed there to be no correlation between the test taker's personality and motivation to speaking performance but implied other factors such as learners' linguistic competence, their anxiety levels, or the nature of classroom interaction that may influence speaking performance outcome (Rohima and Pitaloka, 2025).

2.3. SPEAKING RUBRICS AND CRITERIA

2.3.1. Conceptual Base of Spoken Language

Prior to describing the theoretical framework of assessing speaking skills, to present the conceptual base in relation to clearly identifying the basis of the theoretical framework is essential, as it encompasses all linguistic components relevant to assessing speaking skills; the figure below demonstrates the hierarchical nature of linguistic analysis and how successive levels build upon one another from sound to meaning in communicative context (Bailey, 2003; Van Lier, 1966). This figure also portrays the connection between all of the linguistic components as communication itself is multifaceted and it can be said that the competence of higher levels is dependent on the competence of lower levels since interpretation of meaning in the discourse is based on the comprehension of clauses, phrases, words, morphemes and phonemes (Bailey, 2003).

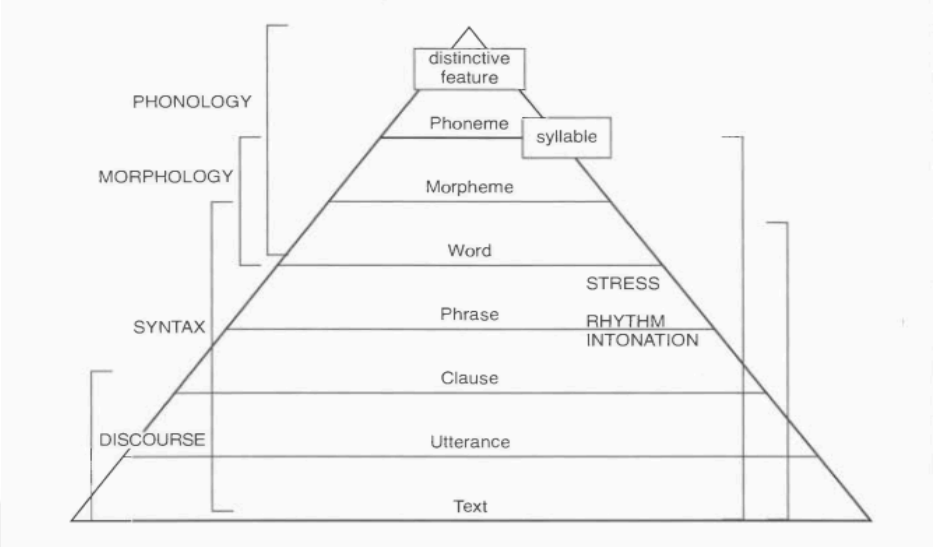


Figure 2.8: Hierarchical Organization of Linguistic Units in Spoken Language

Note. Reprinted from Bailey (2003, p. 51) who developed this figure based on the conceptual information provided by Van Lier (1966).

2.3.2. Common European Framework for Assessing Speaking Skills

The CEFR (CoE) published in 2020 builds on the Council of Europe's earlier educational reform and aims to enhance the language learning, teaching and assessment in language education. Therefore, in this section of the literature review, only this version has been utilized as it is the latest one that has been adapted and/ or adjusted based on the previous editions so it avoids inconsistencies or construct mismatches associated with earlier editions. In short, the 2020 edition ensures a greater accuracy, consistency, detail and alignment with current assessment practices. The CEFR (CoE, 2020) provides illustrative descriptor scales for communicative language competences. The descriptors place the first level that a learner is capable of performing a task, and are used as reference tools; not for assessment. There are three subheadings under it; linguistic competence, sociolinguistic competence and pragmatic competence. However, only those that are relevant to this study will be included here. The general linguistic range, grammatical accuracy, vocabulary range, vocabulary control and phonological control all pertain to the linguistic competence under communicative language competences, whereas fluency, propositional precision, thematic development, and coherence and cohesion are under pragmatic competence for communicative language competences (CoE, 2020).

For overall spoken production of the B1 level, the CEFR (CoE, 2020) presents it as learners being capable of delivering reasonably fluent descriptions of topics within their field of interest, with ideas presented in a linear and organized progression (CoE, 2020).

It is essential to note that the CEFR (CoE, 2020) designates B1 as the threshold level and the B1+ level to be the strong threshold level. As a result, the overlap of descriptors from these two levels within level 4 appears to be consistent with this framework; both of which refer to learners being independent users of the target language. The evidence for this is provided in the methodology section. Furthermore, not all segments have a distinction between B1 and B1+, which is another reason for the incorporation of B1 descriptors, otherwise there wouldn't be enough information to discuss the foundations of this study; making it inadequate in terms of grounding it in a solid framework. Hence, for the purpose of clarification, when B1 level is referenced in this study, it also covers the descriptors and content for B1 + level.

Some descriptors in the CEFR (CoE, 2020) relate to the speaking task itself such as describing experiences, giving information and putting a case for sustained monologue in spoke production, whereas other descriptors relate to the rubric used for assessment such as linguistic and pragmatic competence for communicative competence as they can only be assessed through the students' performance. Under production activities for spoken production, the CEFR (CoE, 2020) description for B1 level's overall oral production is defined as being able to maintain a fairly fluent and coherent description of familiar topics linked to personal interest by arranging ideas in a structured and logical order. This study has used the means of interviews to assess the students' speaking performance, so this is a means that requires a sustained monologue regarding describing experience, giving information and putting a case, which is about since interviews are used to assess learners' ability to produce extended speech without interactional support. Thus, the three subscales under spoken production are both included in the CEFR (CoE, 2020).

Sustained monologue of describing experience is about describing the learners' ability to narrate and describe, which moves from basic personal information to more complex and structured accounts. It emphasizes communicative functions.

Sustained monologue of giving information refers to the learners' ability to deliver information in an extended turn. It ranges from simple factual descriptions to complex explanations. The focus is on precision and the ability to distinguish between closely related ideas.

Sustained monologue of putting a case provides how learners advance from giving simple opinions to developing well-structured arguments on complex issues. It includes the development in presenting, supporting and structuring arguments effectively while considering the interlocutor's perspective.

The following table summarizes the B1 level descriptors of each subscale in a sustained monologue.

Table 2.4: Subscales of Sustained Monologue

Oral Production Sub-scale	B1 level descriptors
Describing experience in a sustained monologue	Learners can clearly articulate their emotions regarding personal experience and provide justifications for them. They can offer straightforward descriptions on a range of familiar topics concerning their area of interest and narrate simple stories or accounts with reasonable fluency and coherence. They can provide elaborations for experiences by explaining their emotions and reactions in both predictable and unpredictable situations. They can summarize the storyline of a book or a film and express their reactions. They can describe their dreams, aspirations and ambitions. This level enables them to narrate both real and imagined events.
Giving information in a sustained monologue	Learners can explain the central points of an idea or an issue with a fair degree of accuracy. They can clearly describe step-by-step processes to do something by providing detailed instructions. They can report facts on familiar subjects such as describing a problem or giving directions when given some time for preparation.
Putting a case in a sustained monologue	Learners can express opinions on familiar topics using simple language, provide basic reasons to support their viewpoints, and justify their approval or disapproval of others' actions.

Note. Adapted from CEFR (CoE, 2020).

The table below illustrates the communicative language competence incorporating scales of linguistic competence and pragmatic competence along with the descriptors of their subskills specifically for B1 level. This research exclusively focuses on the linguistic and pragmatic aspects

of communicative competence and excludes the sociolinguistic competence mentioned in the CEFR (CoE, 2020) due to its irrelevance to the study's specific objectives.

Table 2.5: Linguistic and Pragmatic Competences of Communicative Language Competence

Communicative language competence	Sub-skills of the specified communicative language competence	Descriptors of the sub component specific for B1 level
Linguistic competence	General linguistic competence	Learners demonstrate enough language range to unpredictable situations, convey main ideas with precision and discuss cultural or abstract content. They can communicate on familiar topics such as family, hobbies, jobs, etc., but limited vocabulary can cause hesitation or repetition.
	Vocabulary range	Learners have a strong vocabulary range for familiar subjects and everyday contexts such as family, hobbies, work, travel, etc., and may resort to using circumlocution when necessary.
	Grammatical accuracy	Learners communicate accurately in familiar contexts with some errors and some mother-tongue influence that doesn't hinder the meaning. They use common patterns or 'routines' in predictable situations.
	Vocabulary control	Learners have a good control of basic vocabulary but make errors with complex or unfamiliar topics. They utilize a variety of simple vocabulary when discussing familiar subjects.

Table 2.5 (continued): Linguistic and Pragmatic Competences of Communicative Language Competence

	Phonological control	Overall phonological control - Learners' pronunciation is mostly intelligible, and their intonation and stress don't hinder meaning. The speaker's accent is often affected by the other languages they speak.
		Sound articulation - Learners are usually intelligible even though they regularly mispronounce less familiar words and sounds.
		Prosodic features – Learners can communicate their message intelligibly even when stress, intonation or rhythm is strongly influenced by other language(s) they speak.
Pragmatic competence	Flexibility	Learners can adjust their language to handle less familiar or challenging situations. They can flexibly employ a broad range of simple language to express themselves effectively.
	Thematic development	Learners effectively indicate the chronological order in narratives and develop arguments that are generally easy to follow. They display awareness of the conventional structure pertaining to the text type while narrating or describing events in a coherent sequence.
	Coherence and cohesion	Learners can present simple counter-arguments in discursive texts (e.g with "however"). They can link short discrete points into coherent linear sequence, form longer sentences using a set of limited cohesive devices and structure longer texts with basic paragraphing.

Table 2.5 (continued): Linguistic and Pragmatic Competences of Communicative Language Competence

	Propositional precision	Learners can present the central points of an idea or issue with acceptable accuracy, communicating basic and relevant information so that their message is understandable.
	Fluency	Learners can communicate relatively easily, and overcome pauses or formulation difficulties, and continue speaking without requiring assistance. They remain comprehensible despite pausing to plan for grammatical and lexical accuracy, and self-corrections, which are evident in free production that necessitates longer stretches of speech.

Note. Adapted from CEFR (CoE, 2020).

2.3.3. IELTS Speaking Rubric

The IELTS rubric for speaking consists of 9 speaking band descriptors and four main categories: fluency and coherence, lexical resource, grammatical range and accuracy, and pronunciation. The bands ranging from 0-9 were given explicit descriptions for each criterion in the rubric. The IELTS test is a test with worldwide recognition of assessing English language abilities (Hashemi and Daneshfar, 2018; IELTS, 2025). It is a prestigious certificate that is acknowledged by institutions relating to both academia and immigration (Hashemi and Daneshfar, 2018). This test is renowned for its fairness, reliability, and rigorous standards for longer than 35 years, which is also co-owned by three prominent organizations in language learning and testing and these organizations are the British Council, IDP IELTS and Cambridge University Press and Assessment (IELTS, 2025). Additionally, it has been demonstrated that despite teachers having varying lengths of professional experience, they still generate consistent and valid IELTS speaking scores due to the use of a standardized rubric (Taman, 2019). Hence, this rubric was chosen to be used in this study since it has an international esteem regarding its well-established reliability and

validity. Below is a table that clearly explains each band score with its corresponding skill level and descriptors.

Table 2.6: Band Descriptors of IELTS Speaking Rubric

IELTS Band Score	Skill Level	Descriptors
9	Expert	Demonstrates complete and highly accurate command of English, using the language naturally, precisely and fluently in all contexts with full comprehension.
8	Very good	Uses English effectively and with strong accuracy, though occasionally isolated slips may occur. Misunderstandings are rare and generally do not impede communication.
7	Good	Shows solid operational control of English by producing clear and detailed language, but some inaccuracies or misinterpretations may arise in less familiar situations.
6	Competent	Handles everyday conversation adequately despite noticeable errors or mistakes. Can manage moderately complex language if the topic is familiar.
5	Modest	Possesses a limited yet functional grasp of English. Can understand and produce general meaning but frequent mistakes and misunderstandings are expected.
4	Limited	Communicates only in routine and familiar situations. The range of language is restricted and difficulties in comprehension and expression occur regularly.
3	Extremely limited	Able to express only very basic ideas and understand simple messages. Communication tends to break down often even in familiar contexts.
2	Intermittent	Displays minimal ability to communicate. Understanding and producing English is highly limited. Has severe challenges with productive skills.
1	Non-user	Shows virtually no functional ability in English apart from isolated words that may be produced or recognized.
0	Did not attempt the test	Provides no usable language, which is either due to not taking the test or leaving all items unanswered.

Note. Adapted from IELTS (2025).

Moreover, the IELTS speaking rubric was selected for this study because it aligns closely with the CEFR (CoE, 2020) model of communicative language competence that encompasses both linguistic and pragmatic dimensions. The IELTS speaking rubric specifies the same construct of subskills under fluency and coherence, lexical resource, grammatical range and accuracy, and pronunciation.

Fluency and coherence in the IELTS rubric correspond to the CEFR (CoE, 2020) pragmatic competence with the subskills of thematic development, coherence and cohesion and fluency. Both the IELTS rubric and pragmatic competence in CEFR (CoE, 2020) describe how well learners can organize their ideas, maintain logical sequencing and sustain extended speech with minimal breakdowns.

Lexical resource, and grammatical range and accuracy in the IELTS rubric align with the CEFR (CoE, 2020) linguistic competence. Both describe a learner's capacity to use a sufficiently broad range of vocabulary and grammatical forms with appropriate control and some flexibility. These two criteria in the IELTS rubric reflect the CEFR's (CoE, 2020) focus on vocabulary range and control, general linguistic competence and grammatical accuracy.

Pronunciation in the IELTS rubric corresponds to phonological control in the CEFR (CoE, 2020). Both emphasize intelligibility, consistency and the ability to employ prosodic features such as stress, rhythm, and intonation to convey meaning.

Although the IELTS speaking rubric doesn't include propositional precision as a separate criterion, this subskill pertaining to pragmatic competence in the CEFR (CoE, 2020) can be seen across the IELTS rubric. This is due to it describing the learner's ability to convey main ideas clearly, maintain relevance and use vocabulary precisely, which are seen under fluency and coherence, lexical resource, and grammatical range and accuracy. Similarly, flexibility under pragmatic competence isn't assessed as an independent criterion but is seen across the rubric in Fluency and coherence, Lexical resource, and grammatical range and accuracy. Thus, they have both been included in this study.

The subskill of turn taking under pragmatic competence hasn't been included in this study as it is irrelevant in the speaking assessment of this study.

In brief, the CEFR (CoE, 2020) provides the competences' descriptors qualitatively, and the IELTS speaking rubric supplies observable and rateable descriptors for these competences.

To clearly depict the relationship between IELTS and CEFR (CoE, 2020), below is a figure that illustrates the relationship between IELTS speaking bands and the CEFR proficiency levels as it is a conversion chart that is used to interpret what an IELTS score means in terms of the CEFR (CoE, 2020). Hence, this chart is useful to show the equivalence and compare IELTS to CEFR (CoE, 2020). The IELTS bands 4.0-5.0 correspond to the B1 level, 5.5-6.5 align with the B2 level, and 7.0-8.0 reflect C1 proficiency. An IELTS score that is below 4.0 represents A1 or A2 level. This chart represents clear alignment with CEFR (CoE, 2020) so it enables clearer scoring and benchmarking of the learner' oral proficiency and supports more consistent interpretations of placement and expected communicative abilities. Last but not least, as the participants in this study have an intermediate proficiency level, their IELTS score is generally expected to range between 4.0 and 5. 0.



Figure 2.9: Conversion Chart of IELTS to CEFR

Note. retrieved from IELTS (2025).

2.4. VIRTUAL REALITY IN LANGUAGE EDUCATION

2.4.1. Definitions of VR and Related Technologies

As stated much earlier on in this literature review, the umbrella term that incorporates VR is referred to as XR and these novel concepts have increasingly captured the attention of scholars, which has contributed to the development of immersive learning tools (Kaplan-Rakowski et al., 2023; Ma et al., 2025; Schwaiger et al., 2024; Wang and Li, 2024). Nevertheless, Virtual Reality as a term originated in the late 1980s with Jaron Lanier (Schroeder, 1993) and though it came later, it denoted earlier technologies with comparable features (Ruscica et al., 2025).

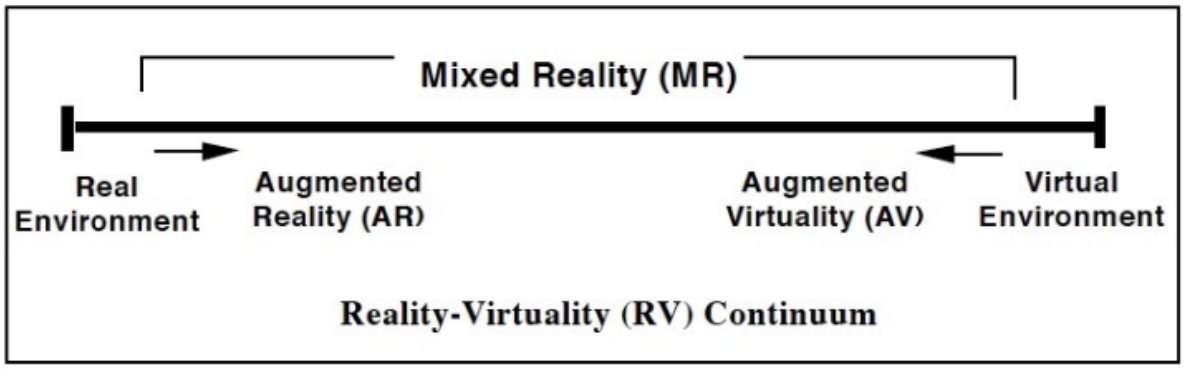


Figure 2.10: Reality-Virtuality Continuum

Note. Reprinted from Milgram and Kishino (1994, p.1322).

The figure above illustrates the Reality-Virtuality Continuum that was proposed by Milgram and Kishino (1994), which is a well-known model that is used to distinguish between different types of mediated environments. It's based along a spectrum for the degree of reality, which ranges from fully real to fully virtual depending on the balance of real and virtual content. This figure indicates that reality and virtuality are part of a continuous spectrum rather than being isolated or independent domains (Milgram and Kishino, 1994).

AR stays closest to the real word because users stay in their physical environment while digital layers are added to what they see (Piumsomboon et al., 2017; Ruscica et al., 2025; Vallarino et al., 2025; Vo et al., 2018). For instance, a HoloLens user being in a real room with real furniture that has virtual layers (Piumsomboon et al., 2017). However, AV leans more toward virtuality as it places users largely inside a virtual space while still integrating elements from the real world (Piumsomboon et al., 2017; Vallarino et al., 2025) such as a user with head-mounted device viewing and interacting with a virtual reconstruction of the real room that they are in (Piumsomboon et al., 2017). Essentially MR is a blend of real and virtual elements so this includes both AR and AV (Milgram and Kishino, 1994; Vo et al., 2018). Whereas VR is at the end of the spectrum compared to the other two since it provides a completely computer-generated environment without having anything from the user's physical environment (Piumsomboon et al., 2017; Ruscica et al., 2025; Vallarino et al., 2025).

Below is a figure adapted from Milgram and Kishino's (1994) model by Vo et al. (2018) that incorporates visuals to clearly demonstrate the difference between the terms.

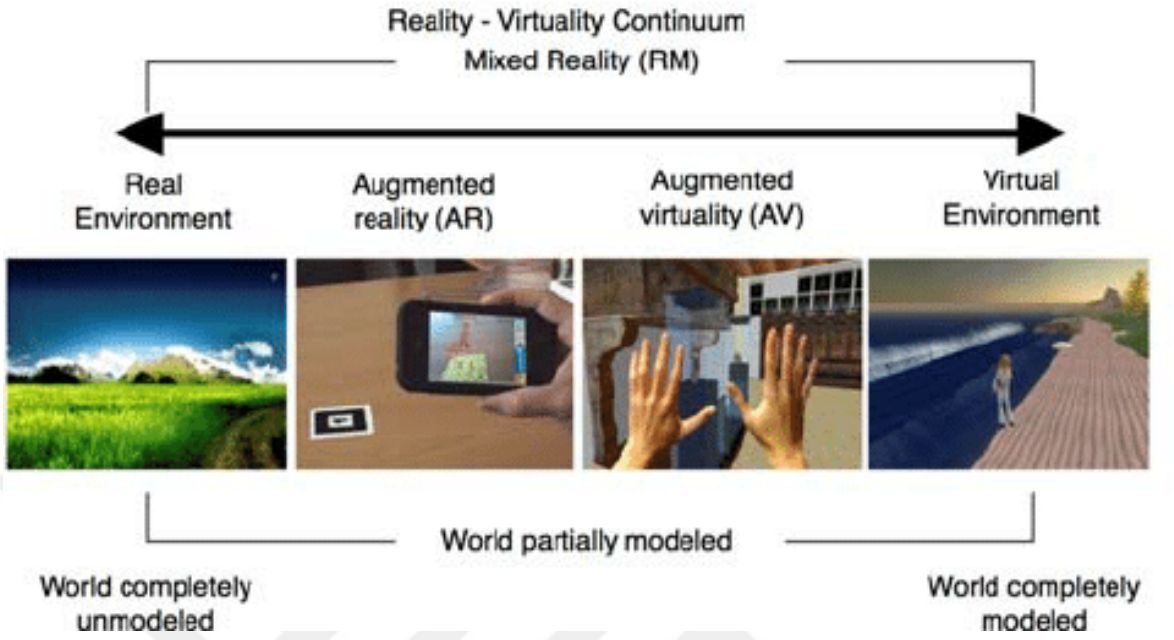


Figure 2.11: The Reality-Virtuality Continuum with Visuals

Note. Reprinted from Vo et al. (2018, p.2), who adapted it from Milgram and Kishino's (1994) RV continuum

However, the challenge of defining and categorizing VR has been present since its emergence due to its rapidly advancing nature (Kaplan-Rakowski and Gruber, 2019). Immersive VR (iVR) can be categorized into two; Low immersion VR (LiVR) involves a 3D virtual space viewed on conventional devices like a desktop monitor, whereas high immersion VR (HiVR) offers a full immersive 360-degree environment through a head-mounted display nature (Kaplan-Rakowski and Gruber, 2019). As a result, LiVR and HiVR differ significantly in the embodiment and virtual body ownership they support in nature (Kaplan-Rakowski and Gruber, 2019).

Nonetheless, alternative classifications divide VR into three main forms as fully immersive, semi-immersive and non-immersive environments (Cowie and Alizadeh, 2025; Çeltek, 2020; Musa et al., 2022). Non- immersive VR refers to standard technologies such as computers, video game consoles, displays, and input devices that offer limited sensory engagement (Çeltek, 2020; Musa et al., 2022). Semi-immersive VR includes the sense of presence through systems like flight simulators, IMAX cinemas with 3D or 4D features, and hands on training simulators (Çeltek, 2020; Musa et al., 2022). Fully immersive VR incorporates the highest level of sensory involvement through equipment such as VR glasses, head-mounted displays, gesture-tracking tools, projection or display walls and specialized application software (Çeltek, 2020; Musa et al.,

2022). Therefore, it is important to clarify that LiVR and non-immersive VR refer to the same VR type used in this study.

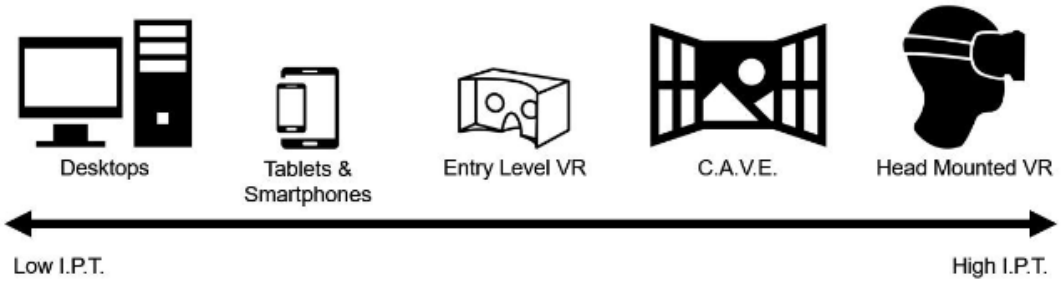


Figure 2.12: Comparative Levels of Immersion, Perception and Telepresence across VR Modalities

Note. Reprinted from Kim and Leathem (2018, p.4).

This figure above proposed by Kim and Leathem (2018) illustrates how much different VR devices vary in the three elements of immersion, perception and telepresence. Immersion reflects how fully virtual reality replaces the real one. Perception is concerned with sensory detail that influences how believable the simulation feels. Telepresence describes the feeling of actually being inside the virtual setting. Just as with the virtual-reality continuum, this figure presents VR devices on a continuous scale as they are connected technologies rather than existing as separate entities. The desktops and mobile screens are placed at the lower end, entry level VR is in the middle and C.A.V.E leans closer to the higher end, but the peak of this continuum belongs to the head-mounted VR as it offers the most convincing and engaging experiences.

A significant observation is that most studies conducted on iVR in language learning are predominantly based on English, which is followed by Mandarin Chinese and Spanish as the next most commonly studied languages in second, foreign or additional language learning contexts involving iVR (Restall et al., 2023).

2.4.2. Implications of Using VR in Language Learning

Despite widespread success across domains including manufacturing and healthcare, VR remains an emerging area language education (Cowie and Alizadeh, 2025; İnalöz and Yılmaz, 2025). Moreover, in relation to other areas in the field of language education, it has been used for developing public speaking skills (Lestiono and Lee, 2024; Lestiono and Setyaningrum, 2025; Patterson et al., 2025), integrated into game-based learning environments (Wu et al., 2025; Zhang, 2025) and even employed as a tool for preserving and teaching endangered languages (Camacho-Fidalgo et al., 2025).

With continual technological process, VR has gained importance as an educational tool, especially for language learning since traditional methods are limited in their ability to offer immersive and interactive qualities, and engaging experiences that resemble authentic communicative situations (Chen and Yuan, 2025; Lee et al., 2024; Lou, 2025; Wu et al., 2025; Yan et al., 2024). A related point is that VR is gradually redefining language learning by providing immersive and interactive environments that develop students' speaking abilities (İnalöz and Yılmaz, 2025; Lee, 2025; Lou, 2025; Wu et al., 2025; Yan et al., 2024; Zhang, 2025). In addition, it facilitates growth in other language skills, namely listening, reading and writing (İnalöz and Yılmaz, 2025; Peixoto et al., 2024; Zhang, 2025). Moreover, VR has also shown to support vocabulary acquisition (Chen and Yuan, 2025; Yan et al., 2024) and grammatical accuracy (Lou, 2025; Lu, 2025; Vallarino et al., 2018; Yan et al., 2024) in addition to enhancing fluency and pronunciation (Yan et al., 2024). Furthermore, VR consistently strengthens affective-cognitive factors such as reducing anxiety, building confidence, and increasing motivation and enjoyment (Gu, 2025; Mallek et al., 2024; Pani and Ray, 2024; Vallarino et al., 2018; Yudintseva, 2024; Zhang, 2025). Thus, studies have clearly displayed that VR can increase motivation, engagement and contextualized learning, which can promote better outcomes in language learning contexts (Chen and Yuan, 2025; Cowie and Alizadeh, 2025; Lee et al., 2024; Lutfi et al., 2025; Tafazoli, 2024; Wang, 2024; Wu et al., 2025; Yan et al., 2024; Yudintseva, 2024). These VR-enhanced language skills can assist students with utilizing them more effectively in real-life situations (Chen and Yuan, 2025; Kaplan-Rakowski et al., 2023). Nevertheless, research has shown that VR-based instruction does not always result in language learning gains (Butorova et al., 2024; Yudintseva, 2024). This indicates that its effectiveness may vary across contexts and implementations.

Despite the heightened interest in the adoption and integration of VR into language learning, several challenges remain. Limited experience with VR (Cowie and Alizadeh, 2025; Yudintseva, 2024), the lack of time, the high cost and insufficiency of VR resources for meaningful experimentation (Chen and Yuan, 2025; Cowie and Alizadeh, 2025; Ironsi, 2023; Mallek et al., 2024), the potential occurrence of cybersickness (Cowie and Alizadeh, 2025), the technical issues and restricted availability of VR devices (Chen and Yuan, 2025; Mallek et al., 2024) are a concern. A further concern is that the effectiveness of VR is highly dependent on instructional design (Restall et al., 2023; Yudintseva, 2024), its cognitive load and task structure (Yudintseva, 2024). This suggests that its success is determined by thoughtful design, sufficient support and appropriate implementation.

2.4.3. Theoretical and System-Level Perspectives on VR in Language Learning

The affective and behavioral patterns identified in prior studies raise questions about the cognitive foundations of VR-enhanced learning, which makes Schütze's (2024) theoretical analysis essential to complement these empirical observations. He articulates a compelling argument that in the discussion of the use of VR that is integrated with AI in language learning, the key point is that strong sensory stimulation by itself is not enough for meaningful purpose. He asserts that what also matters is the level to which various areas of the brain are interconnected and channel information into semantic hubs. Moreover, he claims VR and AI could assist in the creation of immersive experience and neural integration, but whether they truly do so still remain unresolved. Furthermore, he mentions that immersion in a 360-degree VR environment can have the potential to strongly capture learners' attention; however, even though VR is effective for drawing learners in, maintaining cognitive engagement requires the adaptive capabilities of AI. He contends that together, VR and AI may create a more effective environment for sustained learning and knowledge transfer as VR and AI are most beneficial in second language learning when they are able to both attract and hold students' attention. However, he also lists several barriers that can limit their effectiveness, such as unreliable speech recognition, the absence of body language cues, and overly monotone chatbot voices that reduce engagement. Consequently, he emphasizes the need for instructional design that aligns with learners' language levels and motivational needs for these technologies to be used in. Even though Schütze (2024) offers several convincing points

regarding the cognitive conditions required for VR and AI to meaningfully support language learning, important questions persist in the actual feasibility of applying these technologies in the current educational contexts. While his claims about the combined potential of VR and AI are persuasive, the individual utilization of VR and AR still remains relatively novel in language learning contexts nowadays. Thus, this integration of both of them within a single application is a limited practice or adoption in the real-world currently, which makes it difficult to fully evaluate the effects he proposes. Nonetheless, his argument for carefully designed instructions is highly valid as the effectiveness of VR or AI depends on pedagogical design that is sensitive to learners' proficiency levels, cognitive load and motivation.

While Schütze (2024) provides a theoretical foundation for understanding how VR may support cognitive and affective dimensions of L2 learning, studies such as Lou (2025), Lu (2025) and Zhang (2025), move toward technological implementation. Zhang (2024) illustrates how such pedagogical principles can be embedded into a VR platform through real-time speech analysis and automated feedback. Similarly, Lou (2025) and Lu (2025) adopt computational approaches using machine-learning models and large VR datasets to explore how linguistic complexity, engagement and communicative performance develop within iVR. Although these studies were not empirical, they demonstrate how VR technologies are being engineered and analytically modelled in ways that align with cognitive and affective perspectives highlighted by Schütze (2024).

A further study (Zhang, 2025) described the creation of an immersive VR platform for developing English speaking and listening skills. This custom system included Unity3D and Vuforia SDK. This system can be utilized on Android mobile devices and relies on camera image recognition, so it can be classified as low immersion VR or at most semi-immersive VR; however, the author does not specifically state the immersion level. The system recreated every day and workplace interactions that gave learners a safe and realistic setting to practice in. This study used a maximum information coefficient algorithm, which examines learners’ speech in depth to provide immediate and customised feedback. This strengthened the pedagogical relevance of the training tasks and supported the learners’ development faster. The findings depict that the VR environments improve learners’ speaking and listening skills in addition to increasing their

engagement, motivation when compared to traditional methods. It is important to note that Zhang (2025) does not follow a formal research design, rather it is a system development study that describes the construction of an immersive VR platform and illustrates its technical functions without conducting an empirical experiment. The paper does not report participant data, assessment procedures, duration or comparative performance results. Thus, its positive effects on speaking and listening are assertions that are not backed by participant testing or controlled comparisons, which makes it provide only theoretical benefits.

An additional study (Lu, 2025) used a machine-learning technique called RHH-ISVM to examine how students learn English in VR. The system analysed data from 150 students on how their learning progressed week by week along with their speech patterns and complexity of language use. However, this article doesn't disclose the name of the VR nor does it provide the length of the intervention. The results revealed similar findings in the sense that VR contributed to the learners' language development by making connections through experience and increasing motivation while reducing anxiety. The results declared the effectiveness of VR as accuracy, training gain and VR scores all passed 97%. Accuracy represented the learners' proficiency, VR score meant the engagement, comprehension and personalized learning that made it effective, training gain refers to how much improvement learners made after using VR, and detention score measured how well the system detected the learners' progress and ability by analysing speech and behaviours. However, the sample of participants were 15- from a narrow educational context so questions regarding generalizability of the results are raised. Furthermore, the results may be biased due to the higher probability that participants with prior VR experience volunteered for the study than those who weren't specially as it was reported by some to have faced confusion or discomfort while using VR.

A related study conducted by Lou (2025) utilized VR-LearnEng dataset to collect information on learners' interactions, language use, engagement level and AI-generated feedback, which allowed a detailed examination of how communication skills develop in VR-enhanced English learning contexts. The progress was monitored through a learning success scale of low, medium and high. The dataset further included demographic information, session-specific information, linguistic complexity metrics, interaction data and personalized feedback. With 1000

well-organized samples in CSV format, it provided meaningful insights into learner's vocabulary development, fluency, contextual interpretation and overall progress. The findings revealed a nearly identical pattern of results, which reported high accuracy rates, strong VR learning scores, substantial training gains and reliable detention and retention measures. The findings strengthen the view that immersive, interactive and personalized VR environments provide meaningful support for the development of communicative skills. Nonetheless, this study lacks methodological detail, which limits its transparency. It doesn't identify the VR platform or learning environment and it doesn't specify the number of participants, the length of the intervention, session design or the extent of VR exposure. However, it mentions the task types used in VR such as debate, interactive story, simulation and others.

2.4.4. Empirical Research on Virtual Reality in Language Learning

A systematic review conducted by Butorova et al. (2024) highlighted how limited high-quality research exists iVR for adult language learning between the years of 2014 and 2024. Out of 783 papers, only 18 studies met the criteria for experimental analysis. These studies dealt with different types of VR supported learning that included speaking, vocational language, listening, vocabulary, and social interaction. 10 of them presented clear benefits; however, the remaining 8 did not offer enough evidence to judge whether VR actually improved learning. This pattern reveals both the potential and the methodological gaps within the field as many studies still fail to provide sound outcome data. Most studies in the review examined vocabulary learning in iVR but only several reported the effectiveness being more than traditional methods. The findings revealed that in some cases multimodal instruction in iVR hindered vocabulary retention, action modelling did not consistently enhance verb learning, and several studies reported no objective improvement but reported learners' perceived gains. Research on language skills and particularly listening skills did not include control groups or show performance advantages. Nonetheless, learners consistently reported positive emotional experiences. Moreover, social interaction studies rarely assessed learning outcomes. Only one mentioned that it increased accuracy but also noted issues relating to VR sickness and interface-related difficulties. Importantly, the authors acknowledge the review's limitations due to the omission of Scopus and PubMed sources.

Another systematic review conducted by Mallek et al. (2024) examined VR more broadly and emphasized its general pedagogical advantages within higher education, which included personalisation, flexibility, and immersive and interactive experiences. Results indicated that learning gains are strongest when VR is grounded in frameworks such as constructivism, which encompasses experiential and collaborative learning. It also mentioned challenges regarding instructional design, accessibility, high cost in addition to other difficulties. Although the review specifies the databases used in the search such as ACM, IEEE, ScienceDirect, Springer, Wiley, it does not specify the publication years covered by its search, which is a key limitation of this review that decreases the transparency. Nonetheless, both systematic reviews portray that VR consistently enhances engagement and affective responses.

A study performed by Yan et al. (2023) employed an explanatory sequential mixed-methods design to evaluate the impact of Immerse on speaking skills and perceptions of 16 first-year EFL program's students at a university in China. Although students reported a range of previous EFL learning experiences, none had used VR tools before the study. By using the desktop version of Immerse, students took part in role-play speaking tasks in 6 speaking sessions Because the platform could only accommodate a limited number of users at once, the 16 students were randomly divided into two groups of 8 to use Immerse. These learners completed a series of assessments before and after the intervention, which included surveys and interviews. The students' speaking performance was evaluated by two raters who used the Student Oral Language Observation Matrix Revised (SOLOM-R) produced by Wright (2015). This tool measured five components, which consists of comprehension, fluency, vocabulary, pronunciation and grammar. Each is represented by five descriptors that are rated on a 1-5 scale. The higher ratings correspond to stronger proficiency in the corresponding skill's criterion. The pre-survey gathered demographic information while the post-survey measured learners' overall views on Immerse and included two open-ended questions that related to its advantages and limitations for speaking development. All survey items were rated on a 5-point Likert scale from 1 being strongly disagree to 5 being strongly agree. To further explore students' perspectives and experiences with Immerse, semi-structured interviews were conducted with 6 volunteers. The quantitative results portrayed clear gains in speaking ability with the strongest improvements in fluency, vocabulary and grammar. The learners' comments were in alignment with these findings that Immerse improved their understanding, motivation, engagement and ability to apply language in authentic contexts, which

are central elements in situated learning. However, a significant number of participants perceived that their VR experience felt more like play than like participating in a formal lesson. The thematic analysis of the open-ended questions revealed 6 key themes regarding VR-enhanced speaking environments. These themes incorporated that Immerse assisted learners with understanding speaking topics more clearly, helping them remember vocabulary and sentences more easily, the immersive setting kept them focused and curious, it increased greater motivation to improve speaking skills, and it supported applying what they learned to real-life situations.

Although Yan et al. (2023) provide useful insight on how learners interacted with Immerse, several methodological issues limit the strength of their claims about speaking improvement. The study's sample size is very small and there is an absence of a control group. Thus, any gains cannot be confidently attributed to the VR intervention as it isn't possible to isolate VR as the cause. Their use of the SOLOM-R scale also presents certain issues as there is no mention of rater training, calibration or inter-rater reliability, which makes it unclear whether the observed improvements in fluency, vocabulary and grammar show genuine progress or have no rater subjectivity. Moreover, there is a heavy reliance on the learners' self-reported perceptions. While learners' perceptions are valuable, it cannot replace performance measurement because it does not equal the actual performance outcome itself.

A mixed-method quasi experimental study conducted by Yudintseva (2024) explored how low and high immersion VR environments affect ESL learners' willingness to communicate. 20 advanced graduate students completed two speaking tasks; one in each VR modality. Engage XR platform was utilized for low immersion VR, whereas Meta Quest 2 was used for high immersion VR. Participants varied in English proficiency and VR familiarity. The sample included four B1-B2 learners and 16 C1-C2 learners. All participants were ESL learners aged 18 or above. In the findings, no significant differences were found between the two VR modalities. However, task order and action-oriented instruction appeared to play a greater role. That being said, learners reported higher cognitive load and greater enjoyment in high-immersion conditions, and participation in both VR modalities led to reduced speaking anxiety and increased confidence. Furthermore, qualitative results also highlighted how technological, affective-cognitive and

individual factors interact to shape willingness to communicate. Although the study did not measure speaking performance directly, its insights into anxiety reduction and learner engagement provide valuable implications for VR-supported oral communication that also complements the present study's focus on speaking outcomes and learner perceptions.

However, another study conducted by Vallarino et al. (2018) demonstrated that HiVR was more effective than LiVR as the experiment group who used VR achieved higher test scores for students' perceptions of usefulness, learning value, enjoyment and comfort and they also had higher scores for grammar and vocabulary. In this study, researchers conducted an experiment within a B1 English MOOC at the University of Genoa to evaluate iVR versions of the course's grammar and vocabulary activities in particular. 40 students participated; 25 completed the VR tasks and 15 completed equivalent screen-based activities on the university's Moodle platform. For the VR group, learning gains were measured by using a pre and post-test design, whereas the MOOC groups completed only the final test. Thus, the final test scores for the MOOC group were rescaled from a 30-point to a 20-point scale for comparison. The procedure examined both learning outcomes and learner perceptions. VR activities were delivered individually across two sessions using Meta Quest 2 headsets and Unreal Engine 4 tasks adapted from the book's unit. A one-week gap separated the sessions, and the post-test took place one week later. The learners' perceptions further strengthened these results as students in the experiment group reported that VR was effective, enjoyable, helpful and comfortable to use. Thus, these studies' qualitative findings align with those of the previous study as it also presents the affective-cognitive benefits of using VR. This indicates that VR's potential to enhance learning is related to the instructional design and the positive emotional experience as it fosters lower anxiety, higher engagement and greater confidence. Nonetheless, it is important to note that this is an older study and the technology has advanced greatly in 7 years, which implies even though there might still be some differences, the variation in the effects of LiVR and HiVR might not be so great today.

A separate study conducted by Restall et al. (2023) adopted a multiple-methods case-study approach to explore how two Year 10 English classes in Seoul experienced four iVR modules in addition to their teachers' perspectives on this experience. The students' proficiency level were not stated. There were two groups of students. In the first group, students were given cardboard

VR headsets with mobile phones, which they were encouraged to use at home to explore four VR modules. The second group used higher-quality plastic VR goggles compatible with mobile devices within a teacher-led classroom setting. Both groups engaged with the four YouTube VR modules over a two-week period. Case study 1 included 50 students, while case study 2 involved 30 students. Learners were allowed to revisit the VR content freely. However, the home-based group had significantly greater access than the in-class group due to the limited number of devices owned in the school. Upon completion, students completed an online questionnaire and recorded a spoken response to one of the four iVR tasks. Both teachers were interviewed to discuss how iVR might support spoken production that builds on learners' formal education received in class. Students generally linked their enjoyment of the modules to improved confidence and perceived competence, while the teachers emphasized the pedagogical potential of iVR's multimodal resources and the importance of integrating authentic interaction to enhance language learning. This point is especially about the teachers' contrasting views on whether instructional design elements can be effectively incorporated into the immersive experience. In short, iVR was seen as beneficial not only for communicative competence but also for cognitive development since it grounds the context within a real-world environment that is supported by multimodal scaffolding.

Overall, the systematic reviews and empirical studies indicate that VR in language learning reliably enhances affective outcomes such as motivation, engagement, enjoyment and reduced anxiety. Nevertheless, evidence for measurable language gains remains inconsistent and dependent on methodological quality. The systematic reviews highlight that meticulous experimental studies are limited as they often lack control groups or comprehensive reporting. Vocabulary learning shows the strongest support, though the findings present mixed results as iVR sometimes leads to better outcomes, but in other cases the multimodal and actional-based features overwhelm some students and reduce retention. Evidence for listening and broader language use is limited. A crucial point made is that learning gains appear most likely when VR is embedded within established pedagogical frameworks rather than used in isolation, which is reinforced by empirical classroom studies. Considering the empirical research mentioned above, the evidence suggests that VR's strongest contribution is within the affective-cognitive domains, whereas linguistic gains depend heavily on instructional design, target skills and methodological precision. This provides a clear rationale for the present study's focus on speaking performance and learner perceptions within a structured VR-supported environment.

2.4.5. Validated Instrument on Perceptions of VR for Speaking Skills

Learner perceptions are frequently discussed in the literature regarding how learners view themselves and how they interpret the learning environment (Wesely, 2012). This also encompasses self-perceptions of how students interpret and make sense of their own abilities and the learning processes (Liskin-Gasparro, 1998; Williams and Burden, 1999). However, learners' perceptions of the learning environment often relate to how they experience different features of the classroom such as the behavior and actions of the instructor (Brown, 2009). Even though learner beliefs are not often formally separated from perceptions in the literature, they are usually comprehended as broader and long-lasting, whereas perceptions typically relate to particular learning experiences (Wesely, 2012). Learners' beliefs about themselves are frequently linked to the concept of self-efficacy. This is referred to as students' evaluations of their ability to plan and carry out the actions necessary to succeed in academic tasks (Mills et al., 2007). These beliefs include how learners understand the task of learning another language (Barcelos and Kalaja, 2003) and their perspectives on various issues and controversies related to language education (Horwitz, 1988). Perceptions originate from learners' deeper sometimes unconscious emotions, and these perceptions form the attitudes they develop toward learning a language (Despagne, 2010).

It is worth mentioning that since many studies blur the distinction between perceptions and beliefs, the current study treats perception as the broader category that also encompasses learner's beliefs. This perspective is consistent with existing literature, in which both terms tend to overlap and assist in the explanation of learner's experiences. Hence, this approach allows for a more integrated interpretation of learners' affective-cognitive responses in this study.

2.4.5.1. Ironsi's (2023) Questionnaire on Students' Perceptions of VR-Enhanced Speaking Lessons

An important addition to assess students' perceptions on VR-enhanced education introduced by Ironsi (2023), who developed a scale to capture both student and teacher views on the use of VR in English speaking lessons based on the united theory of acceptance and use of technology (UTAUT), the Technology Acceptance Model (TAM) and findings from studies to propose several factors such as performance expectancy, effort expectancy, social influence and

facilitating conditions for how users experience technology. These elements can significantly influence whether users ultimately accept and use a given technology.

This study employed a mixed-methods design using a concurrent triangulation strategy as data were gathered through both structured questionnaires and interviews. This approach enabled the quantitative and qualitative findings to support and validate each other. Since the topic required insights into how students and teachers experienced VR in practice, the combination of methods was more appropriate. The questionnaires highlighted students' self-reported perceptions toward VR use, while the interviews shed light on their and their teachers' actual experience with VR. These complementary data provided meaningful perspectives on VR use in the classroom. An A1 and A2 level course based on CEFR (CoE, 2020) was created and implemented for the study. Thus, the sampling of this study was convenience sampling as the students and teachers were at the English foundation school within the same university in North Cyprus. The instructor group included both native and non-native English teachers. In total, the study involved 75 students and 10 instructors, which resulted in the sample of participants to be 85.

VR was incorporated into the speaking component of the course with learners completing speaking drills using VR goggles. These VR-based sessions placed students in simulated real-life communicative settings. The full course spanned over a 14-week period and covered both productive and receptive skills. However, VR was specifically used during the speaking and listening lessons. The goggles were programmed with a range of activities, which included context-based dialogue simulations. Due to the limited number of devices, learners practised in rotation. This pattern continued until week 7 with VR incorporated into two hours of speaking lessons each day for five days a week. This resulted in the total being 10 hours weekly and 60 hours of VR exposure across 7 weeks. At the end of the programme, both students and teachers used their experiences and learning outcomes to provide feedback on VR as a tool for language learning.

The primary instrument was a 13-item questionnaire. It used a 5-point Likert scale ranging from strongly agree (5) to strongly disagree (1), and it was available in both Turkish and English. A pilot study with 30 students produced a Cronbach's alpha of 0.83, which confirmed that the instrument reliably measured the construct specified within the UTAUT framework. This validated version was then administered to students to determine their level of agreement with each statement, which were later analysed descriptively. To complement the questionnaire, open-ended semi-structured interviews were carried out with 40 participants; 30 students and 10 students who were randomly selected. The interviews were recorded, transcribed and examined using thematic analysis to capture deeper insights aligned with the study's objectives. The results indicated that students perceived VR as helpful in their speaking classes, but the teachers remained neutral. Moreover, participants also reported that VR was difficult to use, expensive to obtain, and that they were still unsure about how effective it actually is. Nonetheless, both students and teachers felt that VR was easy to use and enjoyable, even though the lesson objectives hadn't been fully achieved. This highlighted the notion that implementing VR without sufficient planning may not lead to improvements in students' speaking abilities or their grades.

Unlike studies that focus solely on psychometric scale construction, this instrument was embedded in a structured VR instructional programme, which allowed perceptions to be interpreted on learners' direct experiences. By gathering data based on perception after an extensive and real-world use of VR, the study was able to uncover insights that felt authentic and meaningful to real teaching and learning. This makes the scale a valuable guide for the current study as it investigates how learners make sense of VR-based speaking instruction within an organized classroom context.

2.4.6. ImmerseMe

The organization comprises an international team of more than 50 educators who are committed to VR-based language learning (ImmerseMe, 2025). It was founded online in 2015 and it is grounded in values of sustainability, education and innovation (ImmerseMe, 2025). It currently supports more than 150,000 learners in over 30 countries and it aims to become a globally

influential academic resource for language learning (ImmerseMe, 2025). Moreover, it is designed in accordance with the CEFR (CoE, 2020) and includes content that is organized by each proficiency level. As the participants were intermediate level learners, the platform provided this level appropriate content under “B1.1”.

The following information was presented on the platform’s official website by the co-founder of ImmerseMe Scott Cardwell (2025). ImmerseMe aims to support learners' target language proficiency by placing them in contextualized virtual environments that simulate authentic communication. Thus, it can be seen as a VR platform for role-play. It employs semi-scripted role-play activities that are combined with structured linguistic input that usually provides learners a choice. While interaction paths and response options are pre-designed, learners actively participate in shaping the dialogue, which allows controlled and meaningful speaking practice. The learners can give their own output; however, as ImmerseMe doesn’t have a conversational and generative AI, the platform does not always recognize it. The platform integrates 360-degree video recordings of real-world simulations with voice-recognition technology. This enables learners to recreate interaction with native speakers. The scenarios are organized into small, manageable learning stages that gradually support the development of fluence and learner independence.

Its four instructional models offer sequential learning opportunities that scaffold skill growth by drawing on principles associated with constructivist educational theory. The platform’s design also incorporates principles proposed by Dr. Gianfranco Conti (2025), who identifies several phases in language development. These stages progress from Listening as Modelling (LAM), Thorough Processing, Structured Production, and Autonomy and Spontaneity. Thus, it develops from an awareness-building phase to the deeper understanding stage, which is followed by learners beginning to produce language with support and finally the last phase in which they gradually become independent and free users of the language. Cardwell (2025) highlights that by especially building on Conti’s second phase, ImmerseMe supports learners to strengthen their processing of target language input and begin producing language in a supported way. He states that in relation to speaking skills, the pronunciation mode breaks input into manageable steps by allowing the learners to listen, read and then transform that input into spoken responses. Thus, the

pronunciation mode guides learners from listening and reading a model response to reproducing it orally.

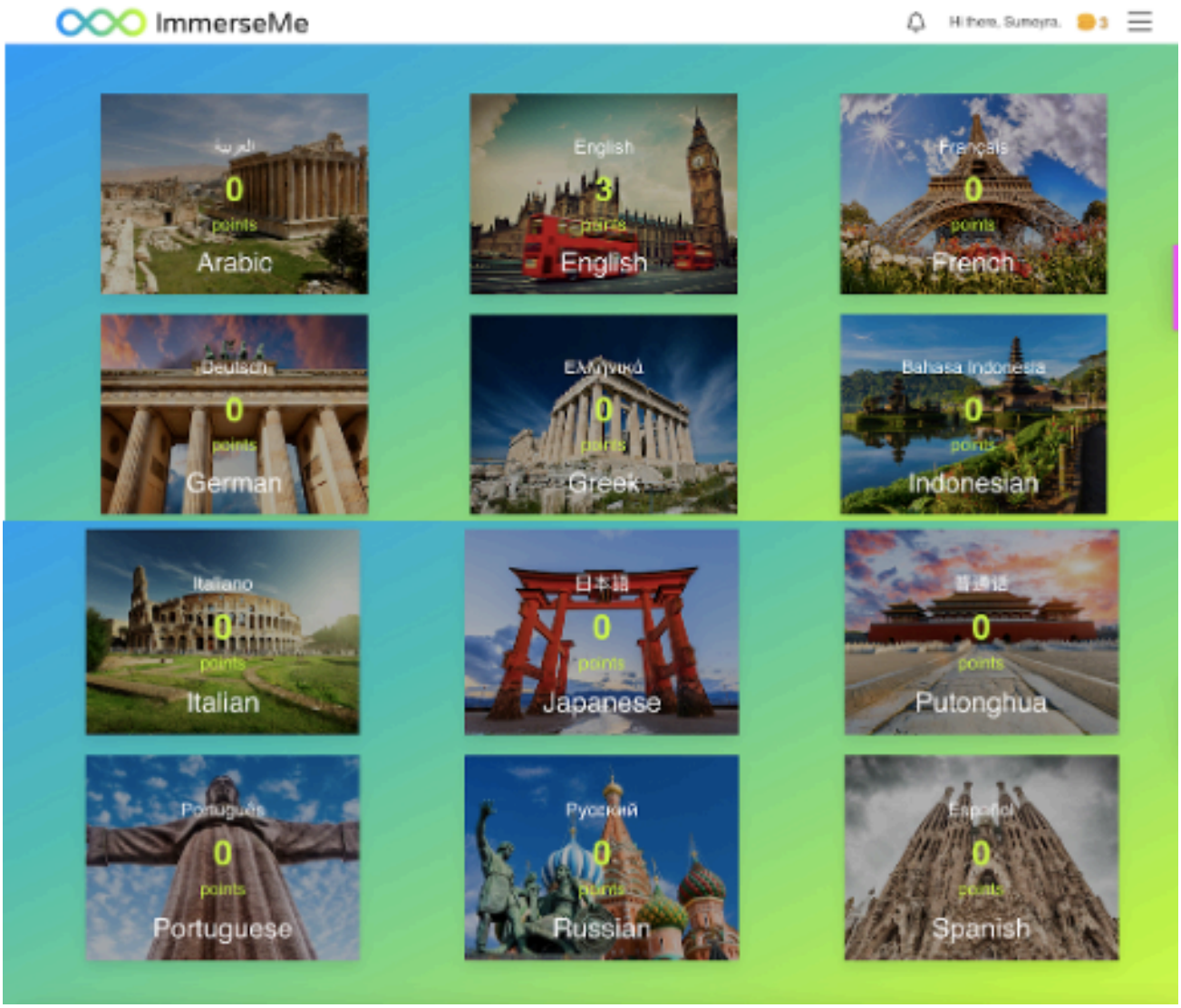


Figure 2.13: The Interface of ImmerseMe

Note. Screenshot was captured by the researcher from the desktop version of ImmerseMe.

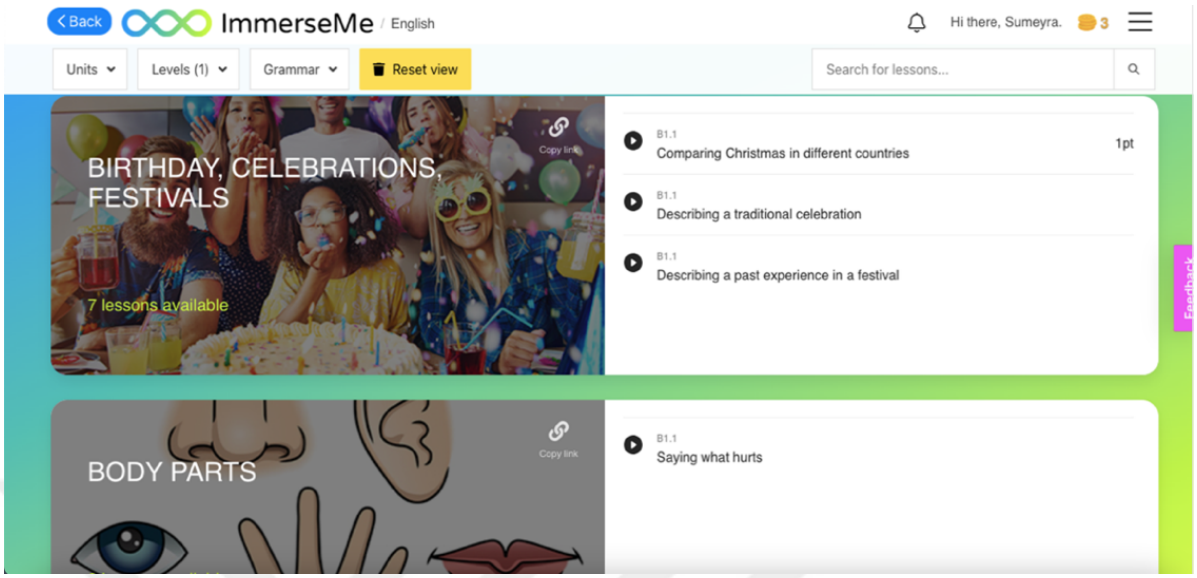


Figure 2.14: The B1.1 Level-Specific Interface of ImmerseMe

Note. Screenshot was captured by the researcher from the desktop version of ImmerseMe.

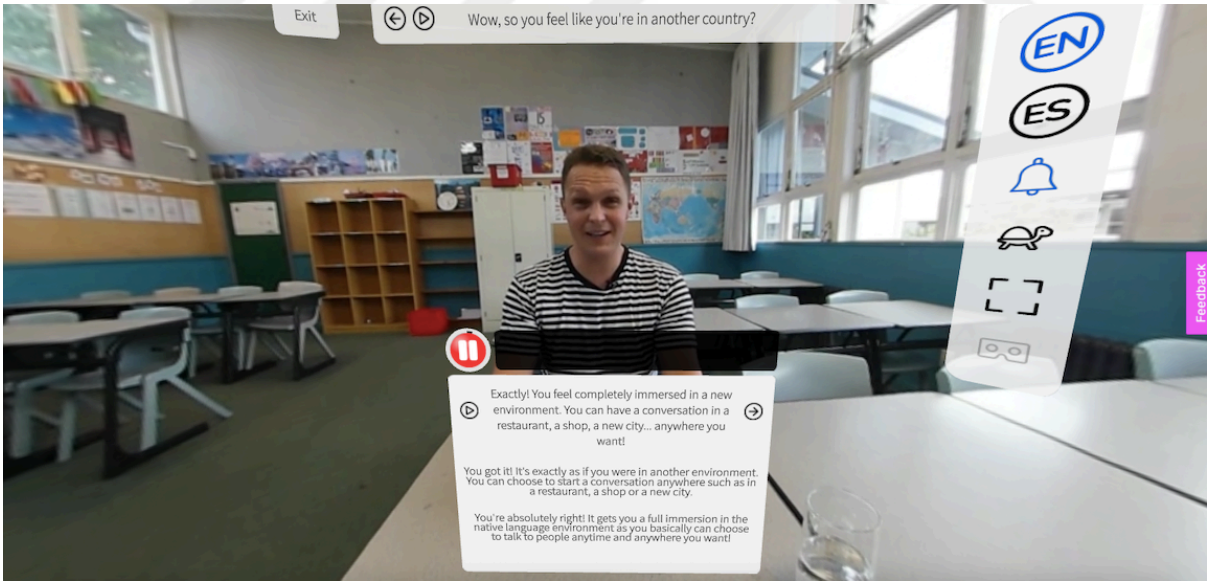


Figure 2.15: A Role-Play in B1.1 Level of ImmerseMe

Note. Screenshot was captured by the researcher from the desktop version of ImmerseMe.

2.4.7. Empirical Research Conducted on ImmerseMe

The primary purpose of Raman et al.'s (2024) study was to examine the effects of incorporating VR technology into the development of English verbal communication skills among ESL learners. The research aimed to provide evidence-based insight into the extent of which VR can support improvements in communicative abilities. Accordingly, the main research question was about how ImmerseMe contributes to ESL learners' fluency and accuracy in addition to how it influences their communicative competence during speaking assessments. This study involved 15 students from a Chinese Independent High School in Malaysia at A2 and B1 proficiency levels. For 8 weeks, learners used mobile phones equipped with VRG (economical VR glasses) together with ImmerseMe. The data was collected through ImmerseMe's automatically generated reports and a speaking assessment conducted by using the national oral assessment, which is known as the Sijil Pelajaran Malaysia (SPM) speaking assessment. In general, the quantitative findings indicated notable gains in learners' performance following the VR intervention.

Overall, the results displayed a considerable variation in participants' fluency and accuracy throughout the 8 weeks of VR intervention. These variations can be due to well-documented factors such as task difficulty, learners' proficiency levels, cognitive load, individual difference, and varying familiarity with vocabulary and precision. Higher performing learners demonstrated stronger vocabulary knowledge, better comprehension of language structures and greater communicative confidence. Whereas low-scoring students struggled with limited vocabulary, complex structures and fluency and accuracy. Furthermore, the differences in task relevance and engagement affected this as more familiar or meaningful scenarios led to a better performance. To sum up, these findings for RQ1 highlight how learner characteristics and task demands interact to determine outcomes in fluency and accuracy during VR-enhanced language learning.

Moreover, this intervention indicated positive gains in English verbal communication with most learners progressing in their CEFR levels as measured by the SPM Oral Test scales. Although one learner remained at A2, which was likely linked to affective and contextual factors such as anxiety, exam pressure, limited preparation and others, rather than to the VR intervention itself. By contrast the other A2 level participants and majority of the B3 showed clear progress. This suggested that sustained exposure to authentic iVR environments supported measurable improvements in speaking skills. In short, these patterns indicate that regular engagement with

ImmerseMe can effectively strengthen ESL learners' fluency, accuracy and overall communicative performance.

While Raman et al. (2024) provide a valuable attempt to explore VR as a tool for enhancing spoken English, there are several key aspects in their study that reduce the clarity and credibility of the conclusions they draw. One of the main issues is the way learner proficiency was reported. The authors refer to progress at so-called "B3" and "B4" levels in the CEFR, which do not exist within the framework. Thus, the mentioning of these non-standard levels creates confusion about what authors actually measured and casts doubt on any claims of CEFR-related improvement. In addition, although the researchers stated that the SPM national speaking assessment was used to evaluate learners' oral performance, they give almost no detail about assessment criteria. Only broad categories such as grammar, vocabulary and communicative competence, which are rated on a 0-6 scale. A holistic rubric is used for the overall performance provided, which ranges from 0-6. However, the type of rubric used for grammar, vocabulary and communicative competence is not specified. Without a clear explanation of how these components were defined, how the scale aligned with CEFR descriptors or how the raters were trained and calibrated, the reliability and interpretability of the speaking scores remain uncertain. As a result, the extent to which the assessment truly reflects gains in communicative ability is unclear. Finally, the study does not include a control group. Without a comparison, the increases reported cannot be attributed to the VR intervention alone, as natural development, test familiarity or additional practice may also explain the observed changes. Consequently, the findings should be viewed cautiously as the methodological shortcomings prevent firm conclusions about the true impact of VR on the learners' communicative performance.

Another study conducted by Akay and Kessler (2024) investigated the contribution of ImmerseMe to the oral pragmatic competence of Turkish university EFL learners. A quasi-experimental mixed-methods design was implemented with two groups of students (30 students in each group) over a 6-week period. In this study, the pragmatic competence was measured specifically in relation to learners' abilities to produce appropriate requests and apologies, which about the 16 prompts written in the Oral Discourse Completion Task. While the experiment group used ImmerseMe as LiVR, the control group used conventional classroom activities. The data was

drawn from pre- and post-test oral discourse completion tasks as well as semi-structured interviews. The results indicated that although the VR group had a higher average than the control group, but the difference was not statistically meaningful. The interview data suggested that students generally valued the VR experience as the majority of the students perceived ImmerseMe to be beneficial for their oral and pragmatic abilities despite facing difficulties.

A similar study implemented by Gracia (2024) explored how ImmerseMe as HiVR can be used to teach English L2 pragmatics to international students in an intensive English program at a U.S. university. There were 77 participants. The research integrated HiVR into explicit pragmatic instruction and employed a quasi-experimental design to compare the pragmatic performance of learners in an experimental and control group during an 8-week course for intermediate level; however, the VR intervention only lasted 4 weeks. The pragmatic tasks included declining an invitation, offering an apology for being late, requesting a favor and politely asking about someone's private life. The results depicted that the students who used HiVR outperformed those in the control group. This indicated that HiVR can strengthen L2 pragmatic development. The qualitative feedback further revealed that learners enjoyed the VR experience and perceived it as realistic, which highlighted HiVR's potential for creating meaningful and real-world conversational practice.

A separate study carried out by Loughlin and Mentzer (2025) examined the use of ImmerseMe to enhance world language learning in middle school Spanish classes; 30 year 7 students and 38 year 8 students. Students practiced conversations through HMD headsets and ImmerseMe. They used an adapted version of the Cognitive Affective Model of Immersive Learning (CAMIL) framework, which emphasizes presence and agency. They tailored the model for this age group by merging six of its constructs into three broader developmental categories. The survey data and the teacher notes revealed that learners felt a strong presence and improved confidence in pronunciation. The study concludes that ImmerseMe can support engagement and perceived learning when instructional design balances immersive features with learner autonomy, regardless of learners encountering problems with technical issues and sometimes having limited independent actions.

A further study performed by Soto-Lucena (2024) as a Ph.D. student involved 18 university students at intermediate level of Spanish. This study aimed to explore the role of iVR in enhancing language learning experiences and learner's willingness to communicate. Most of these students were native English speakers who had varied prior experience in the language. Each participant completed a one-hour session that included an ImmerseMe desktop VR task, a post-task survey, a demographic questionnaire and a LexTALE-Esp proficiency test. The survey results displayed high satisfaction with the immersive practice especially due to its realism, low-pressure environment and support for pronunciation and vocabulary development. Many viewed ImmerseMe as a helpful supplement to limited real-world exposure. However, participants also reported frustrations with automatic speech recognition and expressed mixed feelings about grammar-focused and repetitive activities. The findings largely indicate that ImmerseMe can enhance language-learning experience and help reduce learners' discomfort or anxiety, and most have found it motivating, engaging and supporting their willingness to speak. Nonetheless, there were differences about learners' views pertaining to use of ImmerseMe. The variability in affective measures of motivation, willingness to communicate and anxiety scores suggested individual learner characteristics influenced their perceptions of the use of ImmerseMe.

Other studies on ImmerseMe have examined teachers' perceptions (Ericsson, 2023; Tran, 2021), its use for world language instruction (Facer, 2024; Facer and Uebiyev, 2023), its impact on fostering intercultural development (Facer, 2024), and students' reflections following their experience of using the platform (Ericsson and Sofkova Hashemi, 2024; Facer and Uebiyev, 2025).

2.5. PHILOSOPHICAL FRAMEWORK

2.5.1. Post-positivist Approach

Karl Popper (1959) argued that scientific theories should be subjected to empirical testing with the aim of falsification, which would allow weaker explanations to be replaced by improved theoretical models. Through this process, scientific research moves toward an increasing understanding of phenomena and their causal relationships. Popper's (1959) emphasis on fallibility and revision has since been used as a key intellectual foundation for the post-positivist framework.

Post-positivism promotes a comprehensive view of research that extends beyond specialized methodological boundaries (Panhwar et al., 2017; Phillips and Burbules, 2000; Ryan, 2006). It rejects the separation of theory and practice, which emphasizes their interdependence in generating meaningful findings (Phillips and Burbules, 2000; Ryan, 2006). Moreover, the researcher's values and level of engagement are treated as central to the research aims, which challenges the assumption that meticulous research is defined solely by technical accuracy in data collection and classification (Schratz and Walker, 1995). The shift from positivist to post-positivist has emerged as these movements and groups that have highlighted that knowledge cannot be neutral, have challenged simplistic black-and-white modes of thinking and have stressed the ethical responsibilities involved in conducting research (Fox, 2008; Maksimovic and Evtimov, 2023; Panhwar et al., 2017; Ryan, 2006). Therefore, researchers frame their claims in terms of probabilistic interpretations of truth rather than absolute certainty (Maksimovic and Evtimov, 2023). The most that research can reasonably achieve is an approximation of truth that is sustained through rigorous methodological practices, triangulation of multiple data sources, and detailed analytical procedures that might be generalized with caution (Fox, 2008; Phillips and Burbules, 2000).

This approach both enables and challenges researchers to acknowledge the shared humanity between themselves and research participants since researchers also adopt learning instead of simply just testing (Ryan, 2006). Accordingly, research is viewed as a collaborative process carried out with participants, which fosters learning with them rather than treating individuals as objects of investigation (Wolcott, 1990). Hence, this approach requires the researcher to engage empathetically with the participants being studied, which results in a reflexive awareness of their own positioning and assumptions (Fox, 2008; Ryan, 2006). In contrast to positivist approaches that attempt to uncover an objective truth from participants, post-positivist research aims to move beyond the conventional practices as they encourage and engage participants in the collaboratively constructed interaction (Ryan, 2006), which allows their prior knowledge and lived experiences to emerge more naturally (Ritchie and Rigano, 2001). Thus, a critical post-positivist perspective implies that overall truth cannot be derived merely by combining or accumulating data (Fox, 2008; Maksimovic and Evtimov, 2023; Panhwar et al., 2017; Phillips and Burbules, 2000; Ryan, 2006). This emphasises that findings must be interpreted within their social, contextual, and methodological conditions (Fox, 2008; Panhwar et al., 2017;

Ryan, 2006). Post-positivism does not intend to replace positivism with a more stable foundation (Lather, 1991). Instead, it encourages critical awareness of the complex, historical contextualization, and fragile methods used to produce knowledge (Fox, 2008; Lather, 1991). This perspective requires researchers to remain attentive to the tensions and contradictions that arise throughout the research process, rather than attempting to remove them completely (Ryan, 2006). It also recognizes that research is limited by its reliance on explicit evidence, which may constrain its ability to engage with aspects of human experience that are deeply intuitive or ontological in nature (Ryan, 2006). This statement challenges the assumption that empirical research alone should guide decision-making as it demands further data that can sometimes delay action (Panhwar et al., 2017; Ryan, 2006).

It foregrounds intuition as experience-based knowledge that is particularly valuable in complex human contexts and draws attention to artistic and narrative examination as alternative means of epistemologies that are capable of revealing dimensions of experience that empirical approaches may fail to access (Ryan, 2006). Therefore, in a post-positivist framework, empirical research remains significant but it is only one valuable source of knowledge among several others rather than being the sole authority (Fox, 2008; Maksimovic and Evtimov, 2023; Panhwar et al., 2017; Ryan, 2006). Acknowledging the epistemological limitations of the post-positivist approach, the present study adopts the constructivist approach in the second phase to explore how participants interpret and experience the learning process to gain insight to the quantitative findings.

The quantitative phase of this study is grounded in a post-positivist perspective in that it seeks to measure a clearly defined construct using standardized instruments and through systematic assessment procedures, while acknowledging that the results indicate general patterns rather than definitive outcomes that apply homogeneously across all individuals and contexts.

2.5.2. Constructivist Approach

Constructivism has no single founder; however, in research methodology, it is mostly associated with Lincoln and Guba (1985). It offers an approach for examining social relations in their many forms (Fox, 2008; Guba and Lincoln, 1994; Onuf, 2015). Although it incorporates

elements from other perspectives, it stands independently as a coherent system of concepts and assumptions (Onuf, 2015). Moreover, it doesn't seek to explain why people act in particular ways or why societies differ so it isn't a theory (Onuf, 2015). Instead, it creates space for theorizing connections among seemingly unrelated issues by inspecting how meaning is constructed through shared concepts and practices (Fox, 2008; Guba and Lincoln, 1994; Lincoln and Guba, 1985; Magoon, 1977; Onuf, 2015).

At its core, it treats meaning-making as a process of active knowledge construction (Cobern, 1993; Fox, 2008; Lincoln and Guba, 1985; Magoon, 1977) and interpreted within social, linguistic and historical contexts (Schwandt, 2001). Thus, the meaning of constructivism is embedded in the term itself as it conceptualizes meaning-making as an active process of building conceptual understanding (Cobern, 1993; Lincoln and Guba, 1985). Epistemologically, this approach implies that knowledge is fundamentally context-dependent and that the researcher's role in constructing knowledge must be explicitly acknowledged (Erciyes, 2020; Fox, 2008; Schwandt, 2001). Furthermore, its philosophical foundation lies in epistemological fallibilism, which holds that knowledge is incomplete and uncertain (Cobern, 1993; Guba and Lincoln, 1994; Lincoln and Guba, 1985). To put it simply, it denies the presence of an objective social reality independent of human action and cognition, which means that social phenomena cannot be investigated using the same approaches applied to natural scientific processes (Fox, 2008; Guba and Lincoln, 1994; Lincoln and Guba, 1985; Magoon, 1977). Hence, it places greater emphasis on the fluid and evolving meanings constructed by social actors (Fox, 2008; Guba and Lincoln, 1994; Lincoln and Guba, 1985; Magoon, 1977). From this view, knowledge does not represent an exact reflect of reality but rather a meaningful interpretation of lived experience (Cobern, 1993; Guba and Lincoln, 1994; Lincoln and Guba, 1985), in which systematic and rigorous interpretive procedures can minimise bias in this research approach with methods that focus on essential features that demand and involve extensive analytical work (Pouliot, 2004; Schwandt, 2001). There is no methodological shortcut in constructivism, as grounding research in the idea of socially constructed reality demands sustained theoretical and analytical effort (Pouliot, 2004).

This study draws on constructivism at two complementary levels. At the pedagogical level, constructivist theories developed by Piaget and Vygotsky inform the language learning as both

active and socially mediated. At the methodological level, the qualitative phase is guided by a constructivist epistemology, which conceptualizes knowledge as co-constructed through participants' experiences and interpretations (Guba and Lincoln, 1994; Lincoln and Guba, 1985). By distinguishing between these levels, the study maintains conceptual clarity while ensuring theoretical and methodological coherence.

Although constructivist learning theories are commonly associated with Piaget and Vygotsky, these frameworks were not addressed explicitly mentioned in this section as the present study is not theory-testing and is based on an intervention that is outcome-oriented. Thus, the literature review was structured around empirical studies more often focusing on the application of VR in language learning, particularly its impact on speaking performance, learner engagement, and assessment outcomes rather than foundational cognitive or sociocultural models of learning. While principles underlying constructivist learning are implicitly present in the design of the VR-enhanced speaking tasks, detailed review of these classical learning theories and methodologies go beyond the aims of this study. Therefore, constructivism is more in relation to the methodological aspect to inform qualitative interpretation of learners' experiences instead of being provided as a pedagogical theory.

In brief, the primary distinction of post-positivism and constructivism is not about the methodological consistency and soundness but in their epistemological assumptions. Post-positivism holds that reality exists but can only be understood imperfectly through systematic investigation, while constructivism maintains that reality is constituted through human interpretation, so it is inseparable from it.

The table below provides an overview of the post-positivist and constructivist paradigms in terms of their ontological, epistemological and axiological foundations.

Table 2.7: The Comparison of Post-Positivism and Constructivism; Ontological, Epistemological and Axiological Aspects

Paradigm	Ontology	Epistemology	Axiology
Post-Positivism	**Nuanced Realist**: An objective reality exists, but it can only be imperfectly understood.	**Critical and Fallible**: Knowledge is theory-laden and subject to revision through empirical testing.	**Value-Aware**: Complete objectivity is unattainable, and researchers must be reflexive about biases.
Constructivism	**Individual Relativist**: Reality is subjective, constructed by individuals through personal experiences.	**Subjective and Socially Constructed**: Knowledge is socially co-constructed by individuals based on their experiences and interactions.	**Value-Laden**: Values influence knowledge construction. Subjectivity is embraced and reflexivity is encouraged.

Note. Reprinted from Pretorius (2024, p. 2701)

The table below summarizes a selection of commonly used methodologies and methods for post-positivism and constructivism.

Table 2.8: The Comparison of Post-Positivism and Constructivism: Research Designs and Methods

Paradigm	Common Research Designs	Common Methods
Positivism and Post-Positivism	Experimental or Quasi-Experimental Designs, Grounded Theory, Survey Research	Predominantly Quantitative Approaches - Quantitative Questionnaires, Scales, Experimental Methods - Descriptive Statistics, Inferential Statistics
Constructivism	Action Research, Autoethnography, Case Study, Document or Policy Analysis, Ethnography, Grounded Theory, Narrative Inquiry, Survey Research	Predominantly Qualitative Approaches - Diaries, Focus Groups, Interviews, Observations, Photos/Videos/ Artefacts, Narratives, Qualitative/ Open-Ended Surveys, Reflections - Content Analysis, Discourse Analysis, Thematic Analysis

Note. Reprinted from Pretorius (2024, p. 2709)

2.5.3. Situated Learning Theory and VR

The previous two approaches were in relation to the research design of the study; however, Situated Learning Theory (SLT) is introduced here to explain its relevance to VR. SLT was introduced by Lave and Wenger (1991) and they conceptualized learning as arising from learners' active engagement in socially constructed communities of practice embedded in authentic activities, contexts and cultural settings. This process involves interaction and collaboration in genuine contexts pertaining to those such as environmental, social or cultural dimensions (Brown et al.,1989; Stein 1998). It stresses learners interacting with materials in contextually appropriate and relevant settings for effective knowledge construction so that it enhances their learning and understanding of the meaning and practical values of what is learned (Herrington and Oliver, 2000; Huang and Liaw, 2011; Stein, 1998; Yan et al., 2023). Thus, research indicates that SLT supports language learning as it provides authentic communicative environments where language skills and subskills can be practiced and developed in realistic contexts (Chen and Hwang, 2020; Dhimolea et al., 2022; Epp, 2014; Lu, 2020; Zheng, 2020; Zuengler and Miller, 2006).

It has been observed that VR activities create meaningful conditions for situated learning by enabling authentic engagement and social interaction (Dawley and Dede, 2014; Enkin, 2022; Falconer, 2013; Huang and Liaw, 2011; Wu and Hung, 2022; Zheng, 2020). As learning is context-dependent, VR can be regarded as a viable technological tool for facilitating speaking skill development within a situated learning framework, where learners engage in role-play activities and interact with simulated environments (Yan et al., 2023; Yan et al., 2024).

2.6. RESEARCH GAP AND RATIONALE

2.6.1. Summary of Key Findings from Literature

Empirical research on VR in language learning indicates that it enhances learners' affective experiences such as boosting motivation and greater engagement; however, evidence of its effectiveness in creating measurable language gains remains mixed, which is particularly in relation to speaking. Studies examining speaking development report encouraging outcomes but these claims are frequently weakened by small samples, unclear assessment frameworks, insufficient rater calibration, and the absence of control groups. Furthermore, it is highlighted that learning gains depend less on immersion level and more on instructional design as HiVR does not

consistently outperform LiVR. Research specifically in relation to ImmerseMe also reflects similar patterns. To sum up, these findings indicate that VR is most reliable for improving affective-cognitive dimension but remains underexamined for speaking skills.

2.6.2. Identified Gaps in Research

A synthesis of recent studies has revealed gaps in VR based language learning research. Although scholars have repeatedly advocated for more quantitative, qualitative and mixed methods designs to strengthen the findings in terms of soundness and interpretability for VR in language learning (İnalöz and Yılmaz, 2025; Raman et al., 2024), especially systematic reviews consistently report that despite the rapid expansion of the literature, high-quality experimental and quasi-experimental research remains scarce as they lack control groups, transparent reporting, or reliable outcome measures (Butorova et al., 2024; Mallek et al., 2024). While positive effects of iVR have been reported, many studies depend largely on learners' self-reported perceptions or omit direct assessment of learning outcomes, which limit inferences about speaking development. Moreover, VR-based studies targeting speaking skills have been conducted without a control group, which further restricts the ability to attribute observed improvements in speaking performance specifically to VR itself. Furthermore, the scarcity of experimental or quasi-experimental evidence restricts the extent to which causal relationship between VR use and language improvement can be established (Ironsi, 2023). Hence, there is clear need to also investigate the cognitive processes such as perceptual processing, memory functions and attentional mechanisms in VR environments to understand these processes is essential for the developments of effective VR-based instructional design (Butorova et al., 2024), especially to investigate the role of VR-based instruction in enhancing students' speaking skills (Raman et al., 2024). This area necessitates further investigation and additional studies are required to continue examining the claims surrounding the use of VR (Soto-Lucena, 2024). Thus, these gaps point to a clear need for more meticulous research that integrates objective speaking performance data with learners' experiential accounts in relation to VR use.

2.6.3. Relevance of the Current Study

This present study is designed to address these gaps by conceptualizing speaking development as a measurable performance construct, integrating performance data with learners' accounts to explain how and why VR may support or fail to support speaking development, and situating the use of VR within a situated learning framework that emphasizes authentic and contextualized practice.

2.7. CONCLUSION

This chapter reviewed the centrality of speaking in EFL settings and the persistent barriers that prevent learners from gaining sufficient opportunities for authentic oral practice. It outlined key principles and challenges in speaking assessment, which emphasized that valid interpretation of speaking development depends on clear constructs, appropriate tasks, and reliable scoring. Building on this foundation, the chapter examined VR as a technology that has the capability of supporting situated speaking practice and improving affective conditions for communication, while also recognizing that findings across studies are mixed and often constrained by methodological and practical limitations. Thus, this chapter identified a need for research designs that combine sound outcome measures with learner-experience as evidence to clarify the conditions under which VR can meaningfully enhance speaking performance. These arguments provide the conceptual and empirical basis for the present study's design and research questions.

3. METHOD

This chapter aims to present the research methodology of this study. Therefore, it comprises 10 sections in total to clarify all that has been conducted as part of this study. These segments of this chapter are composed of the research design, researcher positionality, setting and participants data collection instruments and procedures, and the data analysis procedures. Hence, this chapter also displays the trustworthiness to elucidate in particular the reliability and validity of this study along with ethical considerations and limitations to provide as proof to having conducted a sound study.

This present study sought to gather answers to the following research questions with the first question being the overarching research question and the others being subsequent research questions:

i. What is the effect of virtual reality on students' speaking performance?
ii. What are the perceptions of students toward using VR for learning and speaking English?
    ii.i Are there any correlated items in the students' perceptions toward using VR for learning and speaking?
    ii.ii How do students describe their perception toward using VR for learning and speaking?
iii. Is there a difference between students' perceptions of the effect of virtual reality on their speaking performance and their actual speaking performance?

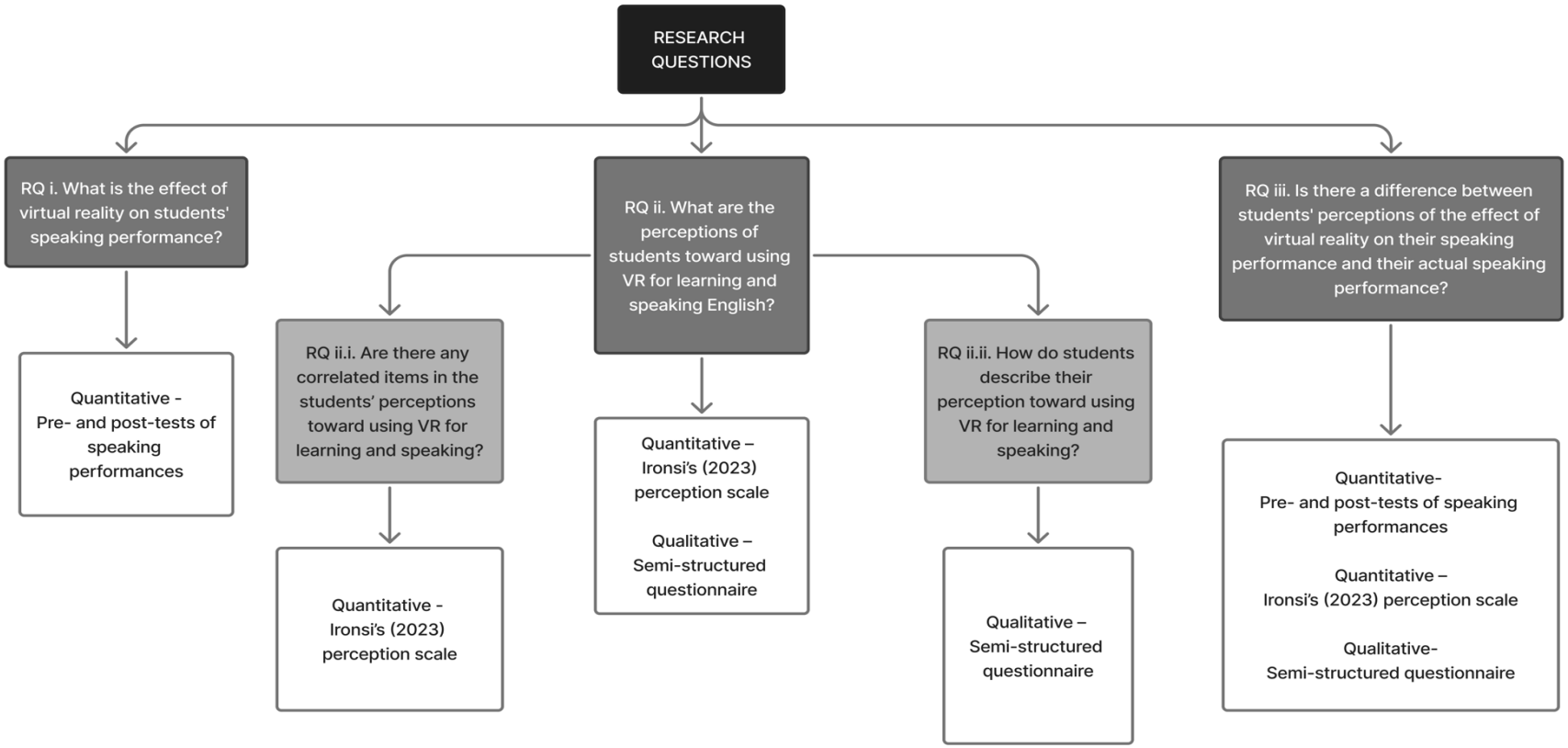


Figure 3.1: Mapping of Research Questions to Data Collection Instruments

3.1. RESEARCH DESIGN

This study adopts a mixed methods research design due to the subsequent questions requiring the reasons for the quantitative data obtained for the primary question. In other words, the main research question has entailed collecting qualitative data to gain insight about students' perceptions on the effect of ImmerseMe on their speaking performance. Thus, the sub-questions had to be developed to provide a holistic outlook to enable full comprehension on this issue. This is especially because without integration of the qualitative and quantitative components, a study is not classified as mixed methods (Creamer, 2018). Therefore, this study has specifically employed one of the core designs of mixed methods, which is the explanatory sequential design.

Explanatory sequential mixed methods design incorporates two phases. The first phase includes the collection of quantitative data and analyzing its results, this is then followed by the second phase of the collection of qualitative data and its analysis to explain the quantitative results that had been obtained (Creswell and Creswell, 2023; Creswell and Plano Clark, 2018). Thus, the aim of the explanatory sequential design is to gain further insight about the quantitative outcome with the qualitative data by connecting them so that the qualitative data can illuminate how the processes or causal connections of quantitative results operate (Creswell and Creswell, 2023). This is why the qualitative data is developed on the quantitative results (Creswell and Creswell, 2023; Creswell and Plano Clark, 2018). Below is a figure that represents this research design for this study.

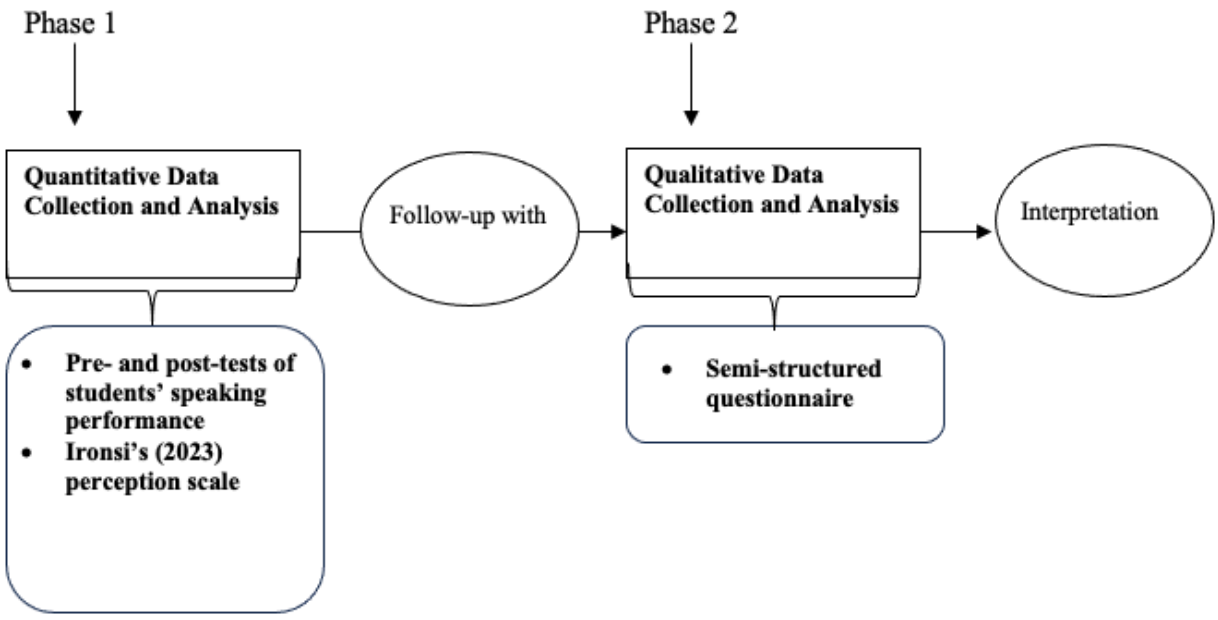


Figure 3.2: Phase 1 and Phase 2 Illustration of This Study

Note. Adapted from Creswell and Plano Clark (2018)

There are some challenges with this design even though it is suitable for beginners (Creswell and Creswell, 2023). The first is that expecting the type of qualitative data collection before the completion of the quantitative phase and the second is choosing the sample for the qualitative being proven to be another difficulty (Creswell and Creswell, 2023). As the qualitative phase is subsequent to the quantitative phase, further information is needed about the comparison of the students' pre- and post-test scores regardless of the outcome of the effect of ImmerseMe in order to better comprehend why or how they attained that score. This is why the semi-structured questionnaire has been utilized in this study. As the primary question is based on the effect of virtual reality, it necessitated having two groups, which were the experiment group and control group. Thus, the second issue that was previously presented is not a concern in this study as the sample for the qualitative data can only be based on the students in the experiment group since they were the ones having experienced learning and speaking English using VR. Therefore, their answers to the perception scale will present the quantitative aspects of learners' perceptions whereas the semi-structured questionnaire will delve deeper into the qualitative reasons for why their speaking performance may have been affected positively, negatively or remained the same.

The prototypical characteristic of this research design is that the paradigm foundation is different for each phase; the post-positivist approach is utilized in the first phase, whereas the constructivist approach is used in the second phase (Creswell and Plano Clark, 2018). This design philosophically commences quantitatively as it begins from aspects relating to post-positivism such as matters concerning creating instruments, measuring variables and assessing statistical outcomes (Creswell and Plano Clark, 2018). This later shifts to constructivism for phase 2 as qualitative data is prioritized so multiple perspectives are accepted and thick description is need (Creswell and Plano Clark, 2018). Hence, this research design portrays the need to shift between different philosophical lenses for each phase (Creswell and Plano Clark, 2018).

Post-positivist approach was employed to objectively measure students' speaking performance through pre- and post-tests in addition to measuring their perceptions for using VR to develop their speaking skills during phase 1 by using a scale developed by Ironsi (2023). The constructivist approach was utilized in phase 2 as the effects of ImmerseMe on students' speaking performance may have been positive, negative or neutral, which is dependent on each student in the experiment group. Thus, multiple meanings are acknowledged as understanding the reasons for the differences in the outcome of these individuals is essential. Especially, as this is based on the investigation of their experience of learning and practicing English in the virtual reality environment. In this respect, the constructivist approach of phase 2 is consistent with situated learning theory, as discussed in the literature review, which conceptualizes learning as emerging through participating in authentic and contextualized activities. In this study, this is provided through the use of ImmerseMe as it is an immersive virtual environment that allows learners to engage in role-play contexts with authentic speaking tasks that provide real-life simulation to support experiential and meaningful learning.

3.2. RESEARCHER POSITIONALITY

As the researcher, I had a direct relationship with the participants in this study as they were my own students from two different classes. Thus, I was the interlocutor during the students' pre- and post-speaking exams to reduce any anxiety that they might have had so that it didn't affect their speaking performance. Additionally, I was present throughout the whole study for the

students in both the experiment and the control group. I had a partner, who had no relationship with these students, to be the other rater and also individually score each student's speaking performance. Upon completion of giving our individual scores to each student's performance, we compared our scores and if there was a discrepancy between our score that was higher than 0.5, then we negotiated by replaying the audio of the student's performance and this time scoring the student's performance together. This assisted in enabling that the scoring was consistent and unbiased. Furthermore, as teachers with a few years of teaching experience, we were familiar with conducting speaking exams due to already having had training regarding this matter and having prior professional experiences. This ensured that we were equipped with the necessary skills and attitude to carry out the pre- and post-tests of the students' speaking performances. Hence, this established inter-rater reliability by having another rater score the students' performance alongside me as it validated the scoring process and ensured that no bias could impact the outcome of the study.

3.3. SETTING AND PARTICIPANTS

The sample of participants were students at the English Preparatory Program (BEPP) of Istanbul Bilgi University. As they had been placed into this level based on the scores that they received in the Bilgi English Language Test (BILET), this guaranteed that these students were indeed at the intermediate level of English proficiency. All of the students in this study were part of Level 4, which meant that they were in B1+ level at BEPP but as clarified in the literature review, to refer to the intermediate level in this study solely the term B1 will be used to avoid any confusion, especially due to Level 4 also incorporating much of B1 descriptors along with the B1+ (see Appendix 1). To provide evidence to this statement in Appendices 2 and 3, there are tables that offer the analysis of selected speaking activities; one unit from the start of the term, one from middle, and one from the later stages. As stated in the Bilgi Prep Program's student orientation documents, the level 4 speaking goals in alignment with CEFR (CoE, 2020) are as follows: Students will communicate with some confidence on familiar routines and non-routine matters related to interest areas by actively becoming involved in conversations to express opinions, or agree or disagree with basic arguments.

The grammatical structures taught in level 4 at BEPP are the following: Present Perfect, Present Perfect Continuous, the passive voice, comparative/superlative adjectives, modals of obligation/advice, defining/non- defining relative clauses, 2nd conditional, 3rd conditionals, and narrative tenses (past simple, past continuous and past perfect).

The two predominantly used coursebooks in level 4 are Empower B1+ intermediate level (Doff et al., 2022) and Prism 2 (Baker and Westbrook, 2022).

Empower B1+ intermediate level coursebook (Doff et al., 2022) has been used as the main course textbook in Level 4. The units covered include 2, 3A and C, 6B, 9, 10 A and B. To provide verification for the B1 and B1+ descriptors being seen and used consistently in the Empower B1+ intermediate coursebook (Doff et al., 2022), speaking tasks from one unit from early on in the term, one from mid-term, and one from the later stages of the term have been categorized. Thus, the speaking tasks in units 2A, 6B and 10B have been categorized in a table format in (Appendix 2).

Prism 2 (Baker and Westbrook, 2022) has its description being provided as intermediate level and the CEFR (CoE, 2020) level being B1. Even though Prism 2 is primarily a coursebook to improve reading skills, there are critical thinking questions at the end of each reading that require students to discuss together, which necessitates both higher-order thinking skills (create, evaluate, analyze) and lower-order thinking skills (apply, understand, remember). The units covered in level 4 classrooms include: U1R1, U1R2, U2R2, U3R1, U4R1, U5R1, U6R1. The critical thinking questions that are discussed in the classroom in these units include Bloom's taxonomy of understanding, apply, analyze, and evaluate. To provide evidence for the B1 and B1+ descriptors being seen and used consistently in the Prism 2 (Baker and Westbrook, 2022), one unit from the earlier days in the term, one unit from the middle of the term and one from the late part of the term have been categorized below. Thus, the critical thinking questions used for discussion in units 2, 4, and 6 have been categorized in the table in Appendix 3.

In brief, even though Level 4 is equivalent to intermediate level B1+, it still incorporates B1 descriptors as it can be seen in Appendix 1, which is why in this section descriptors relevant for B1 and B1+ have been included. The B1 level speaking descriptors provided by the CEFR (CoE, 2020) have been used as references or benchmarks for alignment to conceptualize the speaking proficiency of the participants and assess them in an appropriate manner within a well-

established framework. Therefore, in this study, as also highlighted above the term B1 collectively refers to B1 and B1+ levels.

The table in Appendix 1 presents the CEFR (CoE, 2020) aligned descriptors corresponding to the main speaking tasks in Empower B1+ (Doff et al., 2022) and Prism 2 (Baker and Westbrook, 2022) coursebooks. It illustrates the number of descriptors associated with each CEFR (CoE, 2020) subscale that are represented in both coursebooks. This table is included to further justify the inclusion of both B1 and B1+ descriptors in this study as both Empower B1+ (Doff et al., 2022) and Prism 2 (Baker and Westbrook, 2022) coursebooks incorporate speaking tasks aligned with descriptors from these two levels.

Furthermore, in this study, only the CEFR (CoE, 2020) subscales under Sustained Monologue were examined. These are describing experience, giving information and putting a case. This focus was determined by the nature of the assessment of speaking performance in this study, which included individual interviews with the students rather than interactive or collaborative tasks. The interview format mainly elicited an extended individual speech that demanded students to describe experiences, express opinions and provide explanations; all of which corresponds with the three subscales. Other CEFR (CoE, 2020) subscales (such as spoken interaction) were not included as they encompass descriptors such as reciprocal exchanges and turn taking that involve a partner. Thus, as they were beyond the scope of the assessment format in this study, they were not included. By limiting the analysis to these three subscales, the conceptual alignment between the assessment format, coursebook speaking task analysis and CEFR (CoE, 2020) communicative descriptors have been ensured to strengthen the validity and coherence of the study design. Thus, as CEFR (CoE, 2020) encompasses B1+ into B1, so does this study in accordance with it.

In short, the table in Appendix 1 shows a total of 94 descriptors provided across both books. Out of these, there are 59 B1 level descriptors and 35 B1+ descriptors. It needs to be taken into consideration that this difference is also due to B1 level having more descriptors in the CEFR (CoE, 2020) than B1+ level; naturally it is expected that there would be more B1 descriptors.

Describing experience is represented more heavily in Empower B1+ compared to Prism 2. This is because it has 30 tasks that pertain to this subscale whereas Prism 2 has 10 tasks. This suggests that Empower B1+ emphasizes personal narrative and experiential fluency more.

Giving information appears in both coursebooks; however, Prism 2 includes much more of B1+ descriptors in comparison to Empower B1+. This indicates that its speaking tasks are more in relation to academic or content-based speaking activities.

Putting a case displays the greatest difference in between the coursebooks. Empower B1+ includes both B1 and B1+ descriptors yet it is less compared to Prism 2. For these subscales, Prism 2 only incorporates B1+ descriptors with a total of 13. Thus, it can be said that Prism 2 entails a higher cognitive demand regarding critical thinking skills and forming arguments.

3.3.1. BILET Exam Procedures

As Istanbul Bilgi University uses English as a medium of instruction, students are obliged to take the BILET at the start of each academic year, which consists of two stages. BILET is devised to assess students' English language proficiency level to determine whether they have the competency and capability of following their studies in English in their corresponding faculty. Students who are incapable of receiving the benchmark language scores in BILET have to attend the English Language program until they attain the desired English level proficiency. The following information was provided by the members of the Assessment Unit at BEPP and the BEPP 20204-2025 Student Handbook.

BEPP has obtained accreditation from Commission on English Language Program Accreditation (CEA), which is the internationally renowned commission that provides certification for English language education. Having acquired this accreditation initially in 2015 to last five years, later extended by the CEA until 2031 upon meeting the entire 44 standards that were categorized under 11 headings that pertain to areas such as program aims, curriculum, assessment and evaluation, and faculty competencies. This has enabled BEPP to accentuate the superior quality of their language education. BEPP has also embodied the university's motto of 'Learning not for school, but for life', as it strongly implemented that the primary objective of BEPP is for students to not only achieve academic success but to also be prepared to communicate and interact in English in real-life situations by gaining the essential language skills for expression and exchange of information.

The first stage of the BILET level placement exam is an internationally acknowledged level placement exam because it is prepared by Cambridge University Press. This exam is conducted for two main reasons, which are to determine the students' English level and decide on the eligibility of which students can move onto the second stage of the BILET exam. This first exam incorporates multiple choice questions for 20 reading questions and 30 language use questions in addition to two written tasks. BILET 1 writing exam consists of two simple writing tasks; personal question and opinion question. For instance, for personal questions, it can be about the student's holiday with wh- questions provided to students as an aid. The aim of this writing task is to observe the students' comprehension and expression. The second writing task regarding an opinion piece is written so that it can help determine whether this student can take the BILET 2 exam. Students who receive a score that is at least 45 out of 70 are permitted to sit the BILET exam for stage 2. BILET 2 corresponds to the B2 level of CEFR (CoE, 2020). As this is the highest proficiency level that students can attain in BEPP, if students pass BILET 2, then they don't attend BEPP. This exam consists of reading, writing, and speaking. The reading exam consists of two reading texts with 10-14 multiple choice questions. These questions have a variety of question types such as skimming, scanning, making inferences and others, and they are asked in the order of the reading text. The writing exam includes an opinion writing piece that is one sided, which is suitable for B2 level. For instance, "what are the effects of...?" or an agree or disagree topic. Three prompts are given as main ideas and facts are also provided to assist students. For agree or disagree writing topics, three prompts and facts are given for both sides. The speaking exam involves two invigilators as raters to provide interrater reliability. One of the raters is the interlocutor who asks questions to the student. The other is the observer who takes notes at the back while the student is talking. The speaking exam rubric consists of four criteria: communication, fluency, vocabulary and accuracy and has a rating scale of 0-4 for each criterion. Communication is related to the comprehension of questions, appropriacy, relevancy and the elaboration and development of ideas. Fluency is about the pace of students' speech. Vocabulary is about the variety of vocabulary and expressions that the student can use. Lastly, accuracy is about the student's control of the level appropriate language structures and the intelligibility of their pronunciation. The speaking exam lasts approximately 5 minutes for each student. Upon completion, the raters negotiate the score the student will receive by focusing on each criterion. The whole speaking exam process is audio-recorded including the raters' negotiation. The BILET stage 2 exam provides the opportunity for

certain students to be completely exempt from attending BEPP. These students are those who attain a score of minimum 60 as the average score for all of the components of BILET 2, which are reading, writing and speaking. Any new registered students or current second year students have the permission to take this exam. Students that are successful in BILET 2 are granted the permission to embark on their relevant departmental studies. However, those who receive a score lower than 45 on the first exam or those who were unsuccessful in the second exam are assigned to a corresponding level in BEPP according to the outcome that they obtained from BILET 1. Below is a table that depicts how the results of BILET 1 is categorized for each level.

Table 3.1: BILET 1 Exam Placement Levels

Level Placement Exam Result	Assigned Level
0-15	Level 1 – A1
16-27	Level 2 – A2
28-36	Level 3 – B1
37-44	Level 4 – B1+
45-70	Level 5 – B2

Note. Reprinted from the Istanbul Bilgi University (2024, p.14)

The first week of the academic year provides the chance to determine that the results of the BILET exam for stage 1 are accurate and confirmed. This week offers students to complete common level determination tasks in each class that are completed within the classroom. This is aligned with the purpose to select any students who might have been incorrectly placed into a particular level. In the case that a student proves to be misplaced in a level, all the material completed by this student in the first week and the approved petition that has been signed by the student and their instructors is presented to the specific level coordinator. If approved by the respective level coordinator, the student is then allocated to the appropriate level at the start of the second week. The student cannot return to their initially assigned level once the level change has been completed. Students who do not take the BILET exam are automatically placed into Level 1.

Second year students, in other words those who couldn't pass BEPP within one year, are entitled to take the BILET 1 placement exam held at the start of the academic year as they are still enrolled in the program for an additional year. Only if the student has obtained a score to be placed into a higher level are their results taken into consideration. If this isn't the case, then the student

resumes learning at the same level that they finished their first year. It is important to mention that there were no second year students in this study.

Moreover, students can also be exempt from studying at BEPP provided that they can present internationally recognized test results, which have been obtained within the last two years of applying to the university.

The following table displays the internationally acknowledged exams and the minimum score for each one to be exempt from BEPP.

Table 3.2: Recognized Examinations and Minimum Exemption Scores for BEPP

Exam	BILET	TOEFL IBT	First Certificate in English (FCE)	Cambridge Certificate in Advanced English (CAE)	Cambridge Proficiency Exam (CPE)	KPDS/ ÜDS/ YDS/ e-YDS	PTE Academic
Result	60	75	B	C	C	65	60

Note. Reprinted from the Istanbul Bilgi University (2024, p.30)

Level change is only made during the first week of term 1 for regular students. However, late registered students or international students, and lateral transfer or vertical transfer students can be presented to sit a different placement test if they state that their proficiency level is higher than A1.

3.3.2. Description of Level 4

Level 4 is the first level that enables learners to become independent users of the language by converting passive knowledge into practice. This is the first level at BEPP where students learn about essay writing. First of all, the listening goals aim to improve listening for gist and listening for specific information in daily conversations, discussions, news and academic contexts. Moreover, the reading goals are based on developing the reading strategies of skimming and scanning along with making inferences. Furthermore, the writing goals consist of predominantly learning to write a 250-300 word long coherent essay about a specific topic and applying their micro-writing skills that have a stress on strengthening their main argument by presenting supporting techniques to develop their ideas. This incorporates level appropriate grammar and

vocabulary use as well. Lastly, the speaking goals focus on students gaining confidence while communicating and interacting in English. Similar to writing goals, students are expected to present sound arguments when expressing opinions or agreeing and disagreeing while trying to produce more complex sentences by using a variety of grammatical structures that they have learned. As for language subskills of grammar and vocabulary, students learn about a new level appropriate grammatical structure each week and new vocabulary is taught each week through theme-based material.

3.3.3. Research Setting

The setting of the research took place at Istanbul Bilgi University's Kustepe campus as BEPP is located there. The experiment group had to be taken to the computer lab to use the computers there to practice speaking using ImmerseMe. This computer lab had dividers and headsets for students to utilize. The control group was in a regular classroom. As also being the teacher, the researcher had two separate intermediate level classes to conduct this study. The researcher had to co-teach these classes with a teaching partner to ensure that these two separate classes have had the same teachers throughout the study. The total number of students in this study was 24; 12 in the experiment group and 12 in the control group. An informed consent form was given to each group before the study had started to sign and agree to the conditions of the research. One copy of the informed consent form was provided for the participant to keep and another copy of the same informed consent form that was filled out and signed by each participant was given to the researcher to be kept.

3.3.4. Sampling

The sampling method of this study was based on non-probability sampling as it incorporated a non-random selection of participants. Non-probability sampling does not rely on random selection and is also commonly used in research that include qualitative data (Guest et al., 2013). This is primarily because studies that employ qualitative data seldom employ inferential statistical analyses that depend on probability sampling (Guest et al., 2013). Moreover, this type

of research does not aim for statistical generalization as its typically small in sample sizes, which are not suitable for probability-based assumptions (Guest et al., 2013).

As the researcher had only two classes of intermediate level students to work with for this study, it was clear that those who volunteered to take part in this study would be the participants due to the convenience of the students being the easiest for the researcher to access. However, it is highly significant to stress that the researcher didn't select the students and the sampling was conducted by students volunteering. Thus, the sampling in this study is voluntary response sampling, which is a branch of non-random selection (Murairwa, 2015). Moreover, this study also employs purposive sampling as all the participants had to meet the requirement of intermediate level of proficiency in English. Since the naturalistic inquirer aims to capture the widest possible variety of information within the thick description, purposive sampling is typically adopted for that purpose, among other factors (Lincoln and Guba, 1985). This correlated especially to phase 2 of the study. If the objective is to explore a phenomenon rather than generalize to a population, the purposeful sampling is commonly used in mixed-methods studies (Onwuegbuzie and Collins, 2007). Thus, the total number of participants in this study is 24; 12 in the experiment group and 12 in the control group. Although, typically the sample size for the quantitative phase is larger than the qualitative phase (Creamer, 2018; Teddlie and Tashakkori, 2009), the sample size may include either a small number of cases or a large number of units of analysis depending on the research questions for mixed-methods sampling strategies (Teddlie and Tashakkori, 2009). That being said, purposive sampling usually necessitates usually less than 30 cases to explore personal experiences comprehensively (Teddlie and Tashakkori, 2009). Therefore, this study has a small sample size to control variables like teaching style, which might drastically influence the outcome of the students' speaking performance. Also, having a small sample size enables us to examine individual experiences of students in the experiment group in depth. As a result, each group included 12 students to ensure that all control variables were managed and that conditions remained consistent for both groups in order to obtain a deeper understanding.

3.3.5. Participant Profile

The experiment group was composed of 12 students as mentioned above. The students in this group were based on the volunteers chiefly from one intermediate level class; however, as

there weren't enough volunteers from one sole class, 3 students who volunteered from the other intermediate level class were gathered into the experiment group. These 3 students had been informed not to mention the intervention of VR to their own classmates so as to not have any influence on the results of the study.

The students in this group were all aged between 18-20, which consisted of 4 male and 8 female students. All of the students were of Turkish descent and their native language was Turkish. However, only one student also considered Farsi as their native language. These students' bachelor's departments included Architecture, Interior Design, Industrial Design, Digital Game Design, International Relationships, Law, Mechanical Engineering, New Media and Communication, Psychology, and Radio, Television and Cinema. 7 students reported that they had been learning English for more than 10 years, while 3 students expressed that they had been studying English for 5-10 years and only 2 students stated that it was less than 5 years since they had been learning English.

8 students mentioned that they had never used desktop-based VR or similar technologies previously, whereas 4 students noted that they had prior experience. On a Likert scale of 1-5, students expressed their comfort level for using a computer or technology in general, which ranged from 1 being very uncomfortable to 5 being very comfortable. 2 students expressed a neutral stance about their comfort level, whereas 8 students announced that they were comfortable, and only 2 students highlighted that they were very comfortable with using a computer or technological devices in general.

The control also had 12 students in total as expressed earlier. The students in this group were all recruited from volunteers of students being within the same intermediate level class and they remained oblivious to the VR intervention that was being implemented with the experiment group throughout the study. 10 students' ages in the control group ranged between 18-20, and 2 of these students' ages were between 21-23. In this group the gender distribution was more balanced compared to the experiment group as there were 6 male and 6 female students. All of the students were Turkish so their native language was also Turkish. Their bachelors' departments included Industrial Engineering, Radio, TV and Cinema, Advertising, Computer Engineering, New Media and Communication, Architecture, Business Administration, Interior Design, and International

Trade and Business. Thus, it is clear that the bachelors' degree programs in both the experiment and control groups portray a considerable overlap.

4 of the students conveyed that they had been studying English for more than 10 years, 4 shared that they had been learning English for between 5-10 years, while the remaining 4 revealed that they had been studying English for less than 5 years.

3.3.6. Variables

The independent variable is the variable that is assumed to influence or produce a particular outcome in an experimental study, whereas the dependent variable is the variable that is expected to change by the influence of the independent variable and represents the outcome that is measured (Field, 2009). As this study has employed an explanatory sequential mixed-methods design, the independent variable was ImmerseMe as the VR-based intervention. The dependent variable was students' speaking performance, which were measured through pre- and post-tests using the IELTS rubric to assess speaking skills.

Control variables are factors that researchers deliberately hold constant to prevent them from influencing the outcomes of a study (Webb, 2017). If the focus is on examining how an independent variable influences a dependent variable, then to ensure that this relationship is accurately assessed, other factors need to be controlled by minimizing, standardizing or eliminating their potential effects (Webb, 2017). The control variables of the participant profile and the intervention conditions included the session lengths as each group only spent 40 minutes a week in each session. Moreover, the English proficiency of these students was the same as they were all intermediate level students. Also, these learners were taught by the same two instructors to minimize concerns related to teaching style or other instructional differences. Additionally, the number of students in each group was the same and their ages were relatively the same. Furthermore, the time of the day for the designated sessions was the same for both groups; the sessions would start right after the morning shift classes had ended at BEPP. In addition, the topics to speak about during the sessions were kept the same. Last but not least, the speaking activities for both groups involved only role-play activities. It is important to mention that students in the experiment group were all well informed about ImmerseMe and how to navigate it, so that there

was no difference between any of the students' digital literacy in this group. The control variables for the assessment conditions involved being conducted by the same examiners in the pre- and post-tests. Also, learners were asked the same questions and topics in the pre- and post-tests. The same IELTS rubric was used for all of the students during the pre- and post-tests. As their teacher, I acted as the interlocutor during the pre- and post-tests to reduce potential speaking anxiety as the students were already familiar with me.

The extraneous variable is any factor other than the independent variable that may influence the dependent variable and interfere with the accurate interpretation of relationship between the independent variable and the dependent variable (Bowman, 2017). The extraneous variables could be students using language learning apps or having some other extra speaking practice during the length of this study; however, they were specifically asked not to do so that it doesn't affect the outcome of the study.

3.4. DATA COLLECTION INSTRUMENTS

The data disclosed in this study is attained from four different instruments. The first instrument was a demographic information survey to gather the background information of the learning in both groups. The second were the speaking questions and IELTS rubric to measure students' speaking performance in the pre- and post-tests. The third was the quantitative scale, which was Ironsi's (2023) scale about the students' perception on the use of VR. This scale was completed by the experiment group only after the post-test. The fourth was the semi-structured questionnaire, which was filled out by the experiment group to be used as the qualitative data to further understand the results of the quantitative results in relation to their perceptions on the use of VR to improve their speaking skills.

3.4.1. Demographic Information Survey

The demographic survey (see Appendix 4) for the experiment group contained three main headings; personal information, educational background, and technology use. The third heading was the only different one better the demographic surveys of the experiment and the control groups. Information about their prior knowledge, comfort and use of technology were necessary

to a certain extent but not as crucial since no HiVR was used in this study. Nonetheless, this information was essential to comprehend how well the students in the experiment group would feel navigating and practicing speaking skills using ImmerseMe.

The control group's demographic survey (see Appendix 5) encompassed two main headings; personal information and educational background. As previously stated, the heading in relation to the use of technology and VR was omitted so that they remain unaware of the intervention.

3.4.2. Pre- and Post-Tests

The speaking questions used for the pre- and post-tests (see Appendix 6) were similar to those that are used in the BILET stage 2 speaking exam and IELTS module 1 speaking exam. The exact questions of the BILET stage 2 speaking exam couldn't be used due to copyright. In other words, ethical and legal constraints. However, the speaking questions used in the pre- and post-tests extremely resemble those of the BILET stage 2 speaking exam. Moreover, in the first part of the IELTS speaking exam, the examiner poses general questions regarding the candidate and familiar topics about them such as home, family, occupation, academic life and personal interests, which the test-taker usually had 4-5 minutes to answer (Hashemi and Daneshfar, 2018). In short, this similarity also applies to the topics and the types of questions that were used to measure the proficiency of the students.

As stated earlier, the same speaking questions and topics were asked for each student to measure the difference in their speaking skills. The topics included home town, family and home, work, leisure, future plans, likes and dislikes, friendship, travelling and social media. Each topic consisted of 6 general questions to speak about that were appropriate for the intermediate level students.

The IELTS speaking scale (see Appendix 7) consisted of 9 speaking band descriptors and four main categories; fluency and coherence, lexical resource, grammatical range and accuracy, and pronunciation.

3.4.3. Quantitative Perception Scale

The quantitative perception scale (see Appendix 8 for original version and Appendix 9 for Turkish translation) about the experiment group students' (n=12) perceptions concerning the use of VR to improve their speaking skills was created by Ironsi (2023), which included a 5-point Likert scale with 13 items. 1 being strongly disagree and 5 being strongly agree. This scale measures one factor and includes some reverse-coded items.

The items in this scale had multiple conceptual dimensions related to the use of VR for learning and speaking English. These dimensions included effectiveness, pedagogical value, authenticity, engagement and enjoyment, usability, perceived learning gains, learning efficiency, practical and contextual constraints. Thus, this scale provides a comprehensive outlook of students' evaluations on the use of VR for learning and speaking English.

Moreover, it is significant to state that this scale had been translated into Turkish by 2 experts whose mother tongue is Turkish. One of the translators has a BA in Translation Studies from Istanbul University, the other translator has an MA in Translation Studies from Istanbul University. This was done so that the students wouldn't have any confusion about the items and scoring, especially since the researcher wouldn't have been present to clarify them as this scale was sent out to the participants after the post-test. Consent was given by the creator to use this scale in this study.

3.4.4. Semi-Structured Questionnaire

The semi-structured questionnaire (see Appendix 10 for the original version and Appendix 11 Turkish translation) in this study were used to delve deeper into the comprehension of the outcomes of the quantitative data, which were the pre- and post-tests and the quantitative scale of perception. The semi-structured questionnaire was in reference to experiment group students' (n=12) perceptions of the effect of VR on their speaking skills. This questionnaire could be categorized into 3 main sections; general experience with VR, perceived impact on speaking performance, and overall perception and suggestions. The general experience with VR had 11 questions, perceived impact on speaking performance had 17 and the overall perception and

suggestions had 4. That means this questionnaire had 32 questions in total with some being open-ended and others being closed-ended but had prompts to guide the students' while answering.

The questions related to students' general experience with VR for learning and speaking prompted learners to reflect on how they experienced VR overall for learning and speaking, which included aspects such as enjoyment, challenges and negative aspects of VR. Students were encouraged to comment on the perceived advantages and disadvantages of VR, its usability, and how it compared with traditional speaking activities. There were further questions on affective factors such as enjoyment, engagement, comfort, anxiety and confidence while using VR. These questions were worded in this manner to gain insight into the underlying reasons of students' experiences.

The questions in regard to the impact of VR on learners' speaking performance asked whether and how students believed VR influenced their speaking development. This included their overall speaking ability, fluency, vocabulary, grammar and pronunciation. Follow-up questions were asked to get students to explain their reasons behind their evaluations and give them the opportunity to elaborate on the specific features of VR that supported or hindered their speaking development. Several questions also asked perceived changes in confidence, awareness of language use, and the transfer of speaking skills that were practiced in VR to real-life communication and interaction.

The questions pertaining to the students' overall perceptions, evaluations and suggestions toward the use of VR for learning and speaking English encourages students to reflect back on their VR experience through a holistic perspective. This gave students the chance to express whether they would recommend the use of VR for future speaking lessons, propose ways to improve speaking activities in VR and share any other additional thoughts that might have not been covered in the questionnaire.

Overall, the semi-structed questionnaire was designed to provide rich qualitative data that the quantitative findings wouldn't be able to capture as it involved many different dimensions pertaining to the use of VR for learning and speaking English.

This semi-structured questionnaire was initially created in English by the researcher, who had it translated into Turkish by the same translators that translated Ironsi's (2023) perception scale for the same reasons as mentioned previously.

The figure on the following page illustrates how CEFR based descriptors for communicative language competence of B1was used for assessment of speaking performance in this study. Figure 3.3 presents the assessment framework for B1 level oral production in sustained monologue. It specifically displays the integration of sustained monologue subscales with selected linguistic and pragmatic competences that are relevant to the study's objectives.

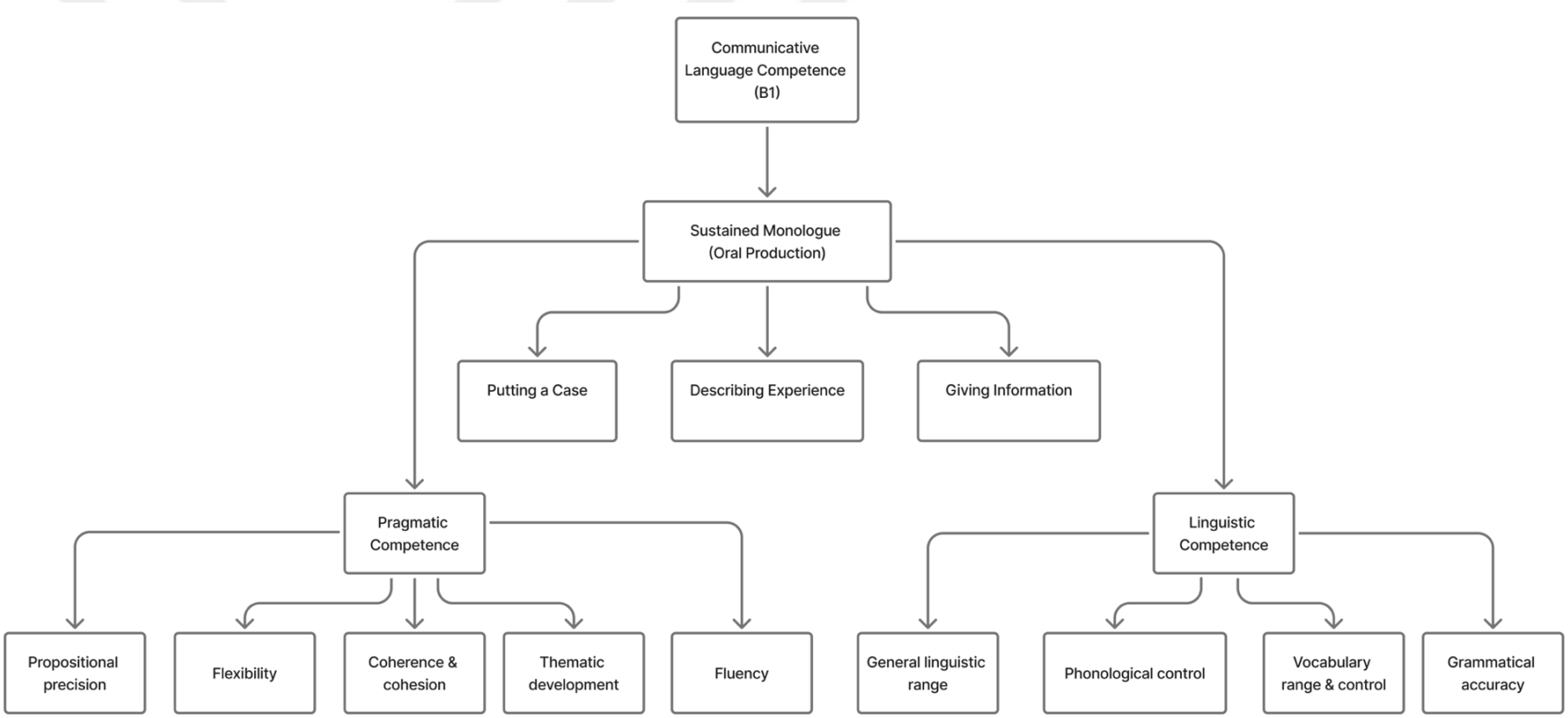


Figure 3.3: CEFR-Based Assessment Framework for B1 Level Speaking Performance

Note. Adapted from CEFR (CoE, 2020)

3.5. DATA COLLECTION PROCEDURE

Below is a figure that delineates the data collection procedure in the order that it was collected.

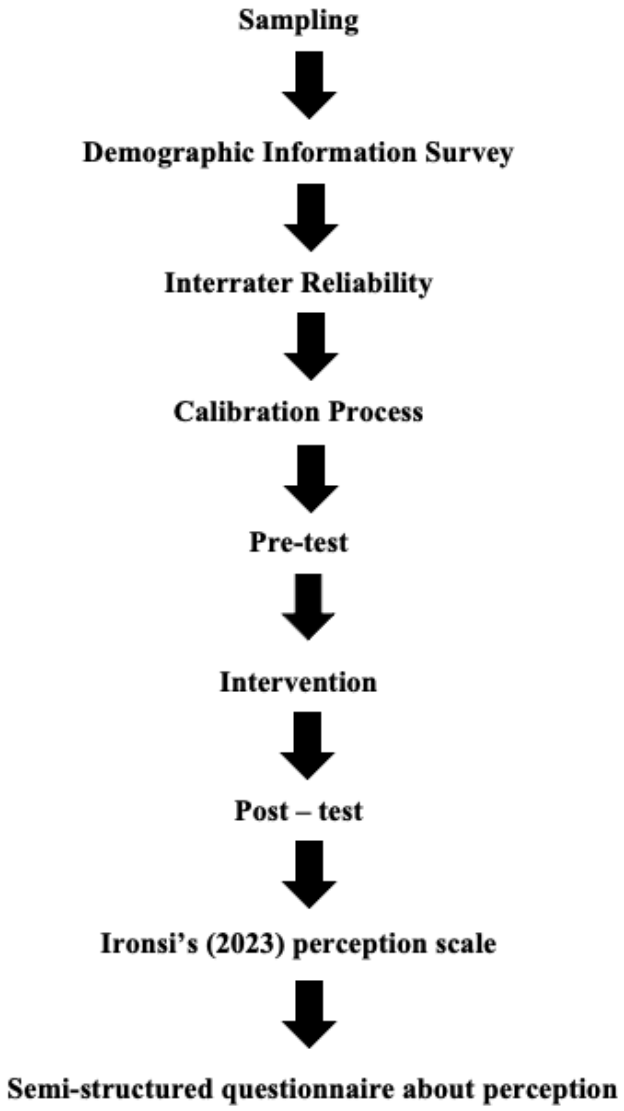


Figure 3.4: The Outline of Data Collection Procedure

Below is an overview of the data collection procedure that is categorized weekly.

Table 3.3: The Summary of The Data Collection Procedure

Week	Data Collection
Week 1	Interrater reliability – 2 raters individual scoring before pre-test
Week 2	Calibration process – with 5 other raters – individual and blind scoring before pre-test. Then, checked the 2 raters scores' against the 5 raters' scores so the total of 7 raters' scores were compared before the pre-tests to establish reliability and validity of the 2 raters' scoring.
Week 3	Pre-test and ImmerseMe accounts
Week 4	Intervention
Week 5	Intervention
Week 6	Intervention
Week 7	Intervention
Week 8	Intervention
Week 9	Last intervention week + Post-test
Week 10	Ironsi's (2023) perception scale and the quantitative data's statistical analysis conducted by a statistician who used SPSS.
Week 11	Semi-structured questionnaire about perception and thematic analysis of this questionnaire.

3.5.1. Raters

3.5.1.1. 2 Raters

The researcher as the rater is a female who was aged 26. The highest degree of educational level she attained was Bachelor's degree. She had 4 years of professionally teaching English to students in only English preparatory programs of universities. She had prior experience teaching all of the proficiency levels except A1 level of English. She also had 4 years of professionally assessing students' speaking performances. Thus, she has previous experience as an invigilator or a rater in speaking assessments. She had experience using standardized assessment rating scales and she was aware of the meaning of bands and descriptors of the IELTS speaking rubric. She also had substantial familiarity with rating speaking performance at a range of proficiency levels.

The partner rater who assisted with assessing the pre- and post-tests filled out a short questionnaire (see Appendix 12) to record their interpretations of the IELTS speaking rubric, and professional experience of teaching and assessing. She is a female and she was 25-years-old. Her highest degree of level of education was a Bachelor's degree. She had 5 years of experience teaching English professionally, and she taught in primary, middle, high school and university. She had 2 years of specifically teaching an English preparatory program. She reported teaching A1, A2, B1, B1+ and B2 levels. She expressed having 3 years of assessing students' speaking skills in a professional context, so she highlighted having previous experience acting as an invigilator or rater in speaking exams. She noted having previously used standardized assessment rubrics such as IELTS speaking rubric, especially as she also prepared students to take the IELTS speaking exams as a supplementary professional activity. She stated that she had high familiarity with rating speaking performances at different proficiency levels.

In short, the 2 raters possessed adequate qualifications and relevant experience to rate the pre- and post-test speaking performances. Despite being early in their professional careers, each had sufficient experience in teaching and assessing speaking across multiple proficiency levels in addition to having familiarity with standardized assessment rating scales. This alignment supports the methodological suitability of their involvement and the examination of interrater reliability.

3.5.1.2. 5 Raters

After volunteering to assist with the calibration process, the 5 raters completed a short questionnaire (see Appendix 13) that was designed to document their interpretations of the IELTS speaking rubric and their professional experiences of teaching and assessing. There were 4 females

and 1 male raters in this group. 4 of them stated that the highest degree they had was Master's Degree, while 1 mentioned Bachelor's Degree. 3 reported to be in the age range between 25-34, whereas the other 2 expressed their age to be in between 35-44.

In relation to their professional teaching experience of English, 2 mentioned between 11-15, 1 stated it to be between 7-10, 1 indicated it to be between 4-6 and the remaining 1 noted it to be between 1-3. The 5 raters reported teaching experiences across a range of educational contexts. All raters had experience teaching at university level, while most had also taught at high schools. In addition, several raters reported having experience teaching in primary and middle schools, private language schools and online teaching contexts. In relation to teaching experiences specifically at preparatory programs, 3 raters reported having between 4-6 years of experience, while the other 2 specified it as 1-3 years. Regarding the proficiency levels that they had taught, all raters reported experience teaching at A1 level. Most raters had also taught A2, B1, and B1+ levels, while 3 indicated teaching B2 level. 4 out of 5 raters had specifically taught B1+.

In terms of experience assessing students' language skills, 1 rater noted it to be between 11-15 years, 1 expressed it to be 7-10 years, 2 articulated it to be 4-6 years, and 1 reported it to be 1-3 years. All of them highlighted having prior experience in being an invigilator and/or rater in speaking tests. In addition, they had all emphasized having experience using standardized assessment rating scales such as IELTS speaking band descriptors. With respect to familiarity with rating speaking performance across different proficiency levels, the raters mentioned high levels of familiarity. 2 raters rated their familiarity at the highest level, while the remaining 3 indicated a high level of familiarity.

Overall, the 5 raters' qualifications and professional experience of teaching and assessing indicate that they were suitable to engage in the calibration process. While some studies indicate that brief and practical standardization trainings can reduce rater bias in the short term (Bijani et al, 2022; Kang et al., 2019), others reported that additional training often leads to minimal changes in scoring behavior among experienced raters, especially as they proceed to having severity of scoring (Bijani et al, 2022; Eckes, 2011). Within this context, the group of 5 raters reflect a balanced profile of having both experienced and less experienced raters so that this issue doesn't cause a problem in this study. As Fulcher (2003) emphasizes that reliance on trained raters alone cannot provide strong validity if their judgments are made primarily by their prior experience

rather than by the clarity of the rating scale. Thus, the calibration process was designed not to remove individual rater differences, but to support a common understanding of the IELTS speaking rubric's descriptors and to report the extent to which experienced and less experienced raters are consistent in their interpretations of assessing speaking performances based on this rubric. Thus, this approach provides a methodologically and theoretically sound basis to integrate this group of 5 raters in the calibration process, which is in fact a part of the interrater reliability process of establishing reliable and valid scores.

It is significant to clarify that no formal standardization training was conducted prior to the rating process in interrater reliability and calibration. Instead, all of the raters were provided with the IELTS speaking rubric and the IELTS and CEFR conversion chart, which were reviewed thoroughly together to ensure a shared understanding and interpretation of the IELTS speaking rubric and IELTS speaking bands. This approach is consistent with research that suggests that having comparative reference points such as examples or shared discussions support rater calibration, especially when descriptors are too vague (Davis, 2015), such as the descriptors in the IELTS speaking. From a psychological perspective, rater judgment is in its nature a comparative evaluation, as individuals evaluate performances by referencing similar examples either present in the environment or retrieved from memory (Pollitt, 2012; Stewart et al., 2006). Thus, this approach is justified by the previously presented research.

3.5.2. Week 1 - Inter-Rater Reliability

To ensure the validity and reliability of the assessment of students' speaking performance, interrater reliability needed to be conducted. The concerns about single rater bias and some shortcomings of the IELTS speaking exam had been identified by Karim and Haq (2014) regarding the potential bias of a single rater. Thus, two independent raters need to mark through individual scoring to minimize objective bias and establish reliability (Karim and Haq, 2014).

Before the pre-test, 30 individual recordings of students speaking performances in English at the intermediate level were used as samples in the interrater reliability process. These students were also from BEPP but they were different students than those who were the participants in the actual research of this study. In other words, these 30 students weren't part of the experiment nor

the control groups. These students were asked the same sets of speaking questions used in the pre- and post-tests.

The two raters conducted individual and blind scoring. Thus, each rater listed to the students' speaking performances' recording and gave their own score for the student without discussing the score to be given with one another. Also, they had no idea who the student in the audio recording was and had no connection to them. They only knew that this student was from an intermediate level class at BEPP.

3.5.3. Week 2 – Calibration Process

Rater calibration is the process of harmonizing raters' evaluation to promote consistency, accuracy and fairness in assessment outcomes (Monteiro and Spadoni, 2025). To ensure that the scores given by the two raters was in alignment with those of others, 5 raters volunteered to be of assistance for the calibration process that needed to be conducted. This was to guarantee that the two raters who will conduct the pre- and post-test could give accurate scores to each student's speaking performance. First, these five raters listened to the 30 sample recordings, which were the same ones used for the interrater reliability process, and they used the IELTS speaking rubric to also do individual and blind scoring. Before doing so, the IELTS speaking rubric and bands were explained to these five raters so that they had no confusion while grading the samples. The IELTS speaking bands and the CEFR (CoE, 2020) proficiency levels conversion chart was used for the 5 raters to fully grasp and interpret the IELTS speaking bands.

Then, all of the 7 raters' scores were compared against one another to check whether the two raters can give accurate scores for speaking performances. This was to show that the two raters who will grade the pre- and post-test speaking performances can give reliable and valid scores by using the IELTS speaking rubric accurately.

3.5.4. Weeks 3 and 4 – Pre-Test and ImmerseMe Accounts

The pre-test of the students' speaking performance included 2 raters and each student's speaking performance was audio-recorded. This was done for just in case purposes. The researcher as the rater was the interlocutor, and the other teacher was the observer, who took notes while

sitting at the back of the classroom and grading each student's performance based on each subskill. This procedure is the same as the BILET stage 2 speaking exam so the students are familiar with this procedure. In addition, this process is also similar to that of Cambridge English (2016) in assessing speaking performances as there are two raters with one being the interlocutor and the other being the observer. However, unlike Cambridge English (2016), both of the raters have used the same rubric to grade students' performance. All of this strengthens the criterion-related validity in this study.

The pre-test involved each student to enter individually and be assigned a random topic to talk about through an interview format. Each speaking topic had 6 questions. This manner of conducting the pre-test was similar to the way IELTS module 1 speaking exam is conducted. The speaking test in IELTS module 1 is conducted in a face-to-face manner through an interview that contains short questions, which aims to have the test-taker talk extensively about an ordinary topic in a guided discussion format (Hashemi and Daneshfar, 2018). Both the experiment and the control group's speaking performances were graded on the same day to also establish intrarater reliability. Upon completion of the speaking performance, the student just left the room with no knowledge about their speaking performance's grade.

After contacting the ImmerseMe team and informing the co-founder Scott Cardwell about this MA study being based on their VR platform, this research was granted a complimentary licensing by the company as it is academic research. In total 12 student accounts and one teacher account were provided by ImmerseMe. Each student received a username and a password for their individual account. These students were asked not to use ImmerseMe outside of these designated sessions so that each participant spoke the same amount of English during the study.

3.5.5. Weeks 4 to 8 – Intervention

In the computer lab at the Kustepe campus, students in the experiment group started their journey of improving their speaking skills while using ImmerseMe for 40 minutes in each session once a week, which was on Wednesdays for this group. The researcher was there with them throughout the entire session of the intervention duration. ImmerseMe provided a variety of speaking activities and topics for the students to practice their speaking skills with functional

speaking tasks and everyday speaking scenarios in the form of role-play. These role-play activities were created to resemble the authentic situations in daily life that enables students to have the opportunity to develop their speaking subskills such as fluency, pronunciation, grammar and vocabulary. All of the components that are found as criteria on the IELTS speaking rubric.

In these speaking activities students interacted with pre-recorded 360-degree videos of native speakers speaking in English. This is also why the responses students could give were limited as this platform didn't include conversational and generative AI within the VR environment; it only contained 360-degree videos. Thus, students practice speaking in a controlled and meaningful way through semi-scripted role-play activities.

To ensure consistency across all the participants in the experiment group, the students completed the order of the sequence of tasks presented to them in the platform. As they were intermediate level students, they only selected B1.1 level content, which was based on the CEFR's (CoE, 2020) intermediate level descriptors to guarantee that it was suitable for this group.

The content for practicing speaking on ImmerseMe included celebrations, travel and tourism, health and emergencies, food and complaints, holidays, customs and culture, art, technology and media, education and work, and problem-solving and conflicts. Only a browser was required to access this platform. In the computer lab, students had segments that would physically separate them from the others. They each had a computer and a pair of headphones with a microphone attached to it. The students covered the B1.1 content on ImmerseMe in the first 3 weeks of the study, when the accuracy accepted for their pronunciation was at 75%. Therefore, they had to repeat the B1.1 content while their accuracy of pronunciation was increased to 100% on the platform to make it slightly challenging this time.

The speaking activities for the control group (see Appendix 14 for a sample) were created by the researcher. The topics were attained from ImmerseMe to keep the content being learned as close to one another in both groups. The sessions with the control group lasted for 40 minutes as well, which were also once a week. For this group, the sessions were on Thursdays. These sessions incorporated a short warm up activity, a role-play scenario relating to the specific topic and had useful vocabulary and phrases that these students can use while speaking with their partner in pairs.

Thus, these role-play activities were written as Student A and Student B. Thus, the control group also used semi-scripted role-play activities to practice speaking just like the experiment group. Each week the students in this group spoke about three topics, which were gathered from the order displayed on ImmerseMe. Although the content of these speaking activities was identical to that of ImmerseMe, the activities themselves were not exactly the same but they were designed to achieve the highest possible similarity by being aligned with the topics and role-play descriptions and following the order of the topics provided on the platform. This was done to make the comparison between the two groups as fair and valid as possible.

The content for speaking activities of the control group can be found in the table below.

Table 3.4: Speaking Activity Content for the Control Group

Week	Content
Week 1	Comparing Traditional Holidays; Saying What Hurts; and Asking for Tourist Information and Posting a Parcel
Week 2	Describing Famous Artworks; Customs, Traditions, and Stereotypes; Virtual Reality and Social Media
Week 3	Makin g a Complaint about Food and Talking about Healthy Eating; Talking About Future Plans and Career Goals; Asking for Help in a Medical Emergency and Talking about Health
Week 4	Family Gathering - Describing Family Relationships; Hotel - Dealing with Check-In Issues; Reporting a Lost Phone
Week 5	Spending a Day at the Festival and Resolving a Conflict; Explaining School Rules and Giving an Opinion; Explaining New Technologies I Use
Week 6	Social Interactions – Asking for a Favor and Leaving a Phone Message; Travel – Renting a Car and Other Travel Scenarios; Workplace – Cultural and Professional Scenarios

3.5.6. Week 9 – Last Week of Intervention and Post-Test

The last week of the intervention and the post-test were conducted during the same week as it was the students' final week before their one-week break. As usual, the students had their sessions on their assigned days and a day later, which was the last day, the post-test was carried out.

The post-test was conducted in the same manner as the pre-test. All of the students were assessed on the same day to provide intra-rater reliability. Therefore, it was also audio-recorded for just in case purposes. As mentioned above, the students were asked the same set of questions in the post-test that they answered about in the pre-tests to be able to measure the difference in their speaking skills so that no other factor changes the outcome such as anxiety.

3.5.7. Week 10 – Ironsi's (2023) Perception Scale and Quantitative Data Analysis

The quantitative scale of Ironsi's (2023) perception scale was sent out to the experiment group by using google forms a short time after the post-test had been conducted. This was so that the students can recall their experiences easily to give honest answers. Hence, the students were given a few days to complete this quantitative scale. First, the quantitative data was analyzed starting with the statistical analysis of pre- and post-test students' speaking performance results. Then, as part of the quantitative data analysis the quantitative scale's results, which was created by Ironsi (2023), was analyzed.

3.5.8. Week 11 – Semi-Structured Questionnaire and Qualitative Data Analysis

Once the quantitative scale was completed, the semi-structured questionnaire was sent to the experiment group by google forms to obtain qualitative data about their perceptions of using VR for their learning experience and speaking skills. The students were given a week to complete this semi-structured questionnaire. The semi-structured questionnaire's answers were analyzed as part of the qualitative data analysis by conducting thematic analysis.

It should be noted that the analysis of both quantitative data and qualitative data is described in detail in section 3.7.

3.6. RELIABILITY AND VALIDITY

3.6.1. Reliability

"Reliability implies the degree to which it is possible to repeat or corroborate these measurements (i.e., Do the measures give results that are consistent, dependable, or stable?)" (Rosenthal and Rosnow, 1991, p.87). Reliability coefficients indicate the degree of correlation between two or more variables (for e.g., tests, items, or raters) assessing the same attribute (Drost, 2011).

Interrater reliability (IRR) is when behavior is assessed by raters or judges, the reliability of their evaluations and the combined internal consistency of their judgments are examined (Rosenthal and Rosnow, 1991). Interrater reliability was established in this study as two raters individually graded each student's speaking performance in the pre- and post-tests to ensure an unbiased and accurate assessment of students' performance. Also, as mentioned previously, the accuracy of these two raters scoring had been verified before the pre-test as well through the calibration process, which compared their scores against 5 other raters. "IRA is the degree to which two or more evaluators using the same scale assign the same rating to an identical observable situation (Bajpaj et al., 2015, p.20). The absolute value of ratings is essential for assessing IRA, whereas IRR emphasizes the consistency of those rating (Bajpai et al., 2015). To clarify, IRR is the broad concept which encompasses IRA, which is why it has been mentioned under IRR previously as well. However, as ICC with absolute agreement was applied to compare raters' scores directly in addition to the averaged rater score, this specifically assessed IRA. The result of the IRA is used to imply the reliability and consistency of scoring as part of establishing IRR.

Intra-rater reliability assesses how consistent a single rater is when measuring the same subjects over time (Stokes, 2011). The intra-rater reliability was ensured as the pre-tests of both of the groups were conducted on the same day, and likewise, the post-tests of both groups were conducted on the same day. Also, the intervention only lasted for 6 weeks so the pre- and post-test scores were not given months apart, which also establishes intra-rater reliability.

Lastly, as a reminder, the reliability and validity of the quantitative scale had already been established as mentioned in the literature review. Ironsi's (2023) perception scale of using VR had a Cronbach's alpha of .83, which showed it to be a reliable measure. Cronbach's alpha coefficient

is the most widely employed method for testing internal consistency in behavioral research (Drost, 2011).

Table 3.5: The Results of ICC for Interrater Reliability and Calibration Process

Measure	Raters (n)	Reliability Index	Value	Interpretation
Interrater reliability	2	ICC	.86	Good agreement
Interrater reliability	5	ICC	.98	Excellent agreement
Interrater reliability	7	ICC	.98	Excellent agreement

As it can be seen in the table above, the reliability analyses indicated high levels of agreement and consistency across all rater arrangements. As reliability increased with the number of raters, the highest coefficients were when ratings were averaged across 5 and 7 raters.

3.6.2. Validity

"...Validity implies the degree to which the measures are appropriate or meaningful in the way they claim to be (Rosenthal and Rosnow, 1991, p. 87). Construct validity indicates how well a measure assessed the intended concept without being affected by measurement error (O'Leary-Kelly and Vokurka, 1998), criterion-related validity is the degree to which a test or questionnaire correlates with relevant criteria (Rosenthal and Rosnow, 1991). Therefore, construct validity and criterion validity were established through the use of IELTS speaking rubric and the pre- and post-tests were administered in a similar manner to the IELTS module 1 speaking exam as they had an interview format and based the speaking questions on familiar topics. Moreover, having an interlocutor and an observer as the raters was similar to the procedures of Cambridge English (2016) and BEPP. This also made the administration process reliable and valid.

Moreover, convergent validity is established which is a measure that refers to the extent to which a measure shows meaningful associations with other instruments or tasks designed to assess the same underlying construct (Weiss and Adams, 2010). This suggests that measures intended to capture the same concept should be related (Weiss and Adams, 2010). Thus, it is categorized under construct validity and in this study incorporates examining the relationships within the quantitative

perception scale's items. To examine the internal coherence of the scale used in the study, inter-item relationships were explored by Pearson's correlation coefficient. Correlation analysis was conducted within this scale to assess whether items intended to measure the same construct demonstrated linear relationships. The results indicated that the majority of items within each scale were positively correlated, which suggests that the items functioned in a coherent manner and measured related aspects of the same underlying construct.

Content validity ensures that each item in a test or questionnaire appropriately represents the primary content areas it is intended to measure (Rosenthal and Rosnow, 1991). Content validity was examined through the speaking activities for both groups as they involved all aspects of IELTS speaking rubric; fluency and coherence, lexical resource, grammatical range and accuracy, and pronunciation.

Face validity implies that a test designed for practical applications should both be valid in terms of statistics or utility and appear relevant and appropriate for its intended purpose (Mosier, 1947). However, face validity is specifically concerned about whether the target audience perceives items as understandable, relevant and easy to respond to, rather than distressing, intrusive or judgmental; hence, participants should be the judge of it (Allen et al., 2023). Regardless of the debate of what face validity is or who should be the judge of it, IELTS speaking tests' formats can be said to have face validity due to its clear and well-established format (Karim and Haq, 2014). Thus, as this study followed a similar format and procedure, it can be said to provide face validity in the pre- and post-tests.

Internal validity refers to the extent to which a study accurately establishes a causal relationship between variables by ensuring that the results are not influenced by confounding factors, biases, or external circumstances (Drost, 2011). Internal validity was ensured through pre- and post-tests as the results demonstrated improvement in both groups; however, the experiment group had higher outcomes. This indicated that VR intervention likely contributed to the improvements especially since the other potential variables were under control and extraneous variables tried to be eliminated.

The external validity of a study reflects its ability to generalize findings across other participants, environments, and time points (Drost, 2011). Thus, external validity is questionable in this study because even though everything was kept pretty much the same for the experiment

and the control groups (except the VR intervention), the sample size in this study was too small to be able to generalize the results.

The validity of the interpretation of the quantitative findings was ensured through the use of appropriate statistical procedure that aligned with the research design and the characteristics of the datasets. The interpretation of the quantitative data was checked by the statistician who conducted the statistical analyses quantitative to provide an additional layer of external review.

3.6.3. Threats to Internal Validity

Campbell (1957) identified 7 categories of extraneous variables that, if not controlled, may pose a risk to the internal validity of the study. Threats to internal validity range from factors such as history and maturation to testing effects and instrument decay, and include statistical regression, selection and mortality (Flannelly et al., 2018). Possible threats to internal validity were carefully examined and mitigated throughout the study.

3.6.3.1. History

History refers to external events experienced by participants during the study that are not attributable to the intervention but may still affect the findings (Campbell, 1957). The history effects were controlled by instructing participants not to engage in additional speaking practice or use other language apps during the study. Also, the experiment group was specifically told not to use ImmerseMe outside of sessions.

3.6.3.2. Maturation

Maturation involves changes in participants over time such as fatigue, hunger, or other biological developments, which may impact the results (Campbell, 1957). The maturation was minimized by keeping the intervention period short (6 weeks) and ensuring both groups has equal exposure time (40 minutes each session; once a week).

3.6.3.3. Testing

Testing pertains to the effects that taking a pre-test or measurement may have on participants' performance in subsequent assessments (Campbell, 1957). The testing effects were reduced by administering pre- and post-tests under identical conditions, using the same questions in pre- and post-tests, raters using the same IELTS speaking rubric in pre- and post-tests, and the interlocutor remaining the same for all students in the pre- and post-tests.

3.6.3.4. Instrumentation

Instrumentation (formerly instrument decay) occurs when changes in measurement tools or procedures over time affect the consistency of data (Campbell and Stanley, 1966). The instrumentation threats were minimized by having expert raters and using the same standardized IELTS speaking rubric consistently for pre- and post-tests.

3.6.3.5. Statistical Regression

Statistical regression is the likelihood of extreme pre-test scores to move toward the mean on subsequent measurements (Campbell, 1957). The statistical regression was unlikely to jeopardize the internal validity of this study because participants were not selected on the bases of extreme pre-test speaking scores and both groups were formed from pre-existing classes that had the same assessment procedures in the study. Moreover, the comparison group allowed potential regression effects to be controlled as any regression to the mean would be expected to occur similarly across groups rather than systematically favoring the conditions of one of the groups.

3.6.3.6. Selection Bias

Selection bias arises when participants across groups are not comparable in important ways, which makes groups non-equivalent (Campbell, 1957). Selection bias and group differences were controlled by matching participants on English proficiency, age, group size, teaching exposure, and making sure that they had the same teachers to teach both groups.

3.6.3.7. Mortality

Mortality refers to unequal participant dropout across groups that may bias the study's findings (Campbell, 1957). There weren't any participants who dropped out so mortality isn't an issue in this study.

Though specifically not one of Campbell's (1957) threats to internal validity, the diffusion of treatment is when observed improvements cannot be attributed solely to the intervention as elements of the treatment were incorporated into both the experiment and control groups' conditions (Gundersen and Svartdal, 2010). The diffusion of treatment was prevented in this study by scheduling sessions separately and ensuring that only the experiment group had access to the VR platform.

3.6.4. Trustworthiness

Four criteria establish trustworthiness in naturalistic studies; “credibility, transferability, dependability, and confirmability” (Lincoln and Guba, 1985, p. 300). It has been expressed that the trustworthiness of naturalistic or qualitative studies is similar to that of quantitative studies. They draw parallels between terms in each type of research design; credibility and internal validity, transferability and external validity, dependability and reliability and confirmability and objectivity (Lincoln and Guba, 1985).

3.6.4.1. Credibility

Credibility related to how convincingly the qualitative findings reflect reality and whether the reported insights are accurately grounded in the data (Stahl and King, 2020). To establish credibility, prolonged engagement, triangulation, peer debriefing and referential adequacy were provided in this study.

For prolonged engagement, the researcher was with the experiment and the control groups throughout the 6 weeks for each session, which allowed the researcher to observe students’ engagement and performance across multiple sessions.

Moreover, the triangulation of data in this study was provided to strengthen the credibility. Hence, pre- and post-tests results, quantitative scale’s outcomes, and the semi-structured questionnaire were used to collect data that built upon one another in this study.

Peer debriefing was utilized in this study to enhance credibility as the qualitative data analysis was reviewed by an expert in English Language Teaching (ELT), who was not directly involved in the study. This expert provided general feedback on the integration and organization of the thematic analysis.

To establish referential adequacy, a qualitative data was set aside during the initial phase of analysis and was revisited later on to verify the stability and coherence of the emerging interpretations. Furthermore, all data, which includes audio recordings of speaking tasks, interrater reliability and calibration processes’ audio samples, and questionnaire responses, were securely stored to support this process. Thus, if required experts can verify that the conclusions in this study are grounded in the actual data of both quantitative and qualitative data analysis.

3.6.4.2. Transferability

Transferability incorporates the provision of thick description necessary for offering a detailed account that enables the evaluation of potential transfer (Lincoln and Guba, 1985). In other words, whether the findings form a study can apply to other resembling contexts or settings. To meet the transferability criterion, the researcher clarified the participant profile, setting, data collection instruments and processes together with the study's limitations (in section 3.9) to guide other studies in conducting valid research on this issue.

3.6.4.3. Dependability

Dependability describes consistent, stable and trustworthy findings in qualitative study, which reflects confidence in the research process and outcomes (Stahl and King, 2020). Dependability provides a basis for determining reliability (Lincoln and Guba, 1985). To ensure dependability, there was clear documentation of the research procedures, transparent reporting of data collection and analysis steps, and clear reporting of coding and theme development was provided.

3.6.4.4. Confirmability

Confirmability is to ensure no bias in the research (Lincoln and Guba, 1985). It indicates how strongly the findings are based on the data rather than the researcher's biases and can be confirmed by others (Stahl and King, 2020). To verify confirmability, transparency of the research process was provided, which could also be traced easily. The qualitative interpretations were based on the participants' own responses with representative excerpts to support analytic claims. In addition, all raw data and analytic decisions were systematically kept to allow for external review. Moreover, to further support the confirmability of the qualitative findings, an audit trail was maintained throughout the thematic analysis process. Anonymized examples of raw qualitative data excerpts, initial codes, and the finalized progression from revised codes to themes are provided in Appendix 15. This documentation allows readers to trace how the interpretations were derived directly from students' responses and aids in reassuring that the findings are driven by the data and not by the researcher's assumptions.

3.7. DATA ANALYSIS PROCEDURES

Below is a figure that presents the overall data analysis procedure of this study. It outlines the order of the interrater reliability analysis, quantitative data analyses of the pre- and post-tests and the perception scale, and thematic analysis of the qualitative data.

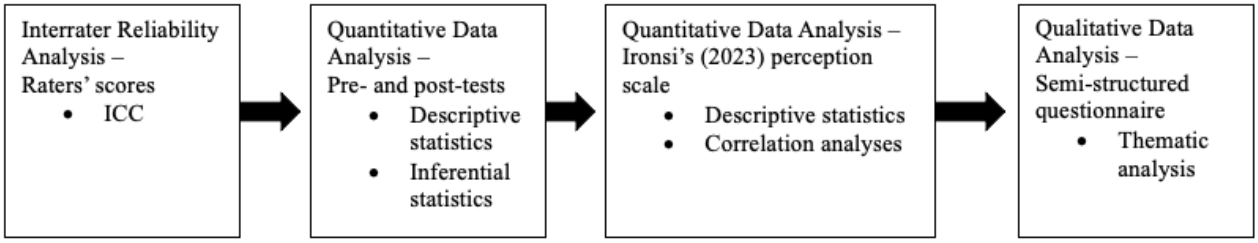


Figure 3.5: Overview of the Data Analysis Procedure

3.7.1. Interrater Reliability and Calibration Process

Interrater reliability and calibration process of the speaking scores was examined using the intraclass correlation coefficient (ICC) in Statistical Package for Social Sciences (SPSS). Specifically, ICC analyses were conducted on the speaking scores assigned by all raters to the same set of student performances during interrater reliability and calibration process; however, this was done successively as it was first 2 raters, then 5 raters, then all of the 7 raters.

A two-way mixed-effect model with absolute agreement and single measures was employed as the same fixed raters assessed all speaking performances and absolute agreement was of primary interest. ICC is a commonly used statistic for evaluating the agreement between numerical measurements that provide an overall estimate of concordance across two or more sources (Han, 2020). ICC is derived from analysis of variance by estimating variance components; however, its use should be limited to models that appropriately reflect the experimental design (Bartko, 1996; McGraw and Wong, 1996). This is especially due to different ICC forms producing varying results even when applied to the same dataset, and because ICC estimates are also influenced by the structure and variability of data (Koo and Li, 2016; McGraw and Wong, 1996). Nonetheless, its usefulness and relevance in social sciences have been well established (Bartko, 1966).

Before proceeding, it is important to clarify that fixed and mixed in this section refer to the same definition; SPSS names the model “mixed” and McGraw and Wong (1996) have called it “fixed”.

3.7.1.1. Selecting the Suitable Model

When the same fixed group of raters is used and the goal is to ignore differences between those raters, a two-way mixed-effects model is used (Bartko, 1966). In other words, if you are not trying to generalize the results to other raters and want to focus only on agreement, a two-way mixed-effects model is appropriate. Furthermore, absolute agreement refers to the extent to which different raters assign identical scores on the same subject (Koo and Li, 2016). Moreover, Shrout and Fleiss (1979) argue that a two-way mixed effects model is suitable for also assessing intra-rater reliability when multiple ratings are obtained from the same rater as the results are not intended to be generalized beyond that specific rater.

3.7.1.2. ICC Measurements

Single-measure ICC estimates the reliability of an individual rating provided by a single rater for a given subject, while average-measure ICC estimates the reliability of the mean score obtained by averaging ratings across multiple rater (McGraw and Wong, 1996). To put it simply, single measurements refer to one score given by one rater to a subject, whereas average measurements refer to the mean of scores given by multiple raters. Single-measure ICC was reported because interrater reliability reflects the reliability of an individual rater’s score, even though the averaged scores were later used in the main analyses.

3.7.1.3. Definition - Consistency or Agreement ICC

Consistency ICC so ICC(C) focuses on whether raters score participants in the same relative order ((McGraw and Wong, 1996). On the other hand, absolute agreement ICC so ICC(A) examines whether raters assign the same numerical scores. Thus, this study involved ICC(A) (McGraw and Wong, 1996). ICC(A,1) which was specifically used in this study was initially not defined by Shrout and Fleiss (1979) because it does not report true correlations in the strict statistical sense, so it is an addition made by McGraw and Wong (1996).

3.7.1.4. Column Variables as Fixed and Random Effects

A key factor in distinguishing between ICC types in a two-way model relates to whether the column variable is modelled as a random effect or as a fixed effect (McGraw and Wong, 1996). The distinction between random and fixed effects primarily influences how ICC results are

interpreted rather than how they are calculated; specifically, when the column factor is treated as a random effect, the findings can be generalized, whereas no such generalization is required when the factor is fixed (McGraw and Wong, 1996). That's why consideration of the size of the population is necessary to decide which model to apply (McGraw and Wong, 1996). In short, the computational formulas, CIs, and test statistics are identical across random effects and mixed-effects (McGraw and Wong, 1996).

3.7.1.5. Use Of ICC(A,1) in a Two-Way Mixed Effects Design

The ICC values of ICC(A,1) is of substantial practical importance for assessing the strength of the relationship because it represents a single-measure intraclass correlation based on absolute agreement, which unlike traditional correlation coefficients, it accounts for both the degree of association between scores and systematic mean differences that decrease when scores differ in absolute values even if they remain strongly correlated (McGraw and Wong, 1996). Simply, ICC (A,1) reported whether scores truly match and not just whether they move in the same direction as it check whether two scores are not only related, but also close to each other in actual values.

Below is a flowchart that illustrates the appropriate selection of ICC types.

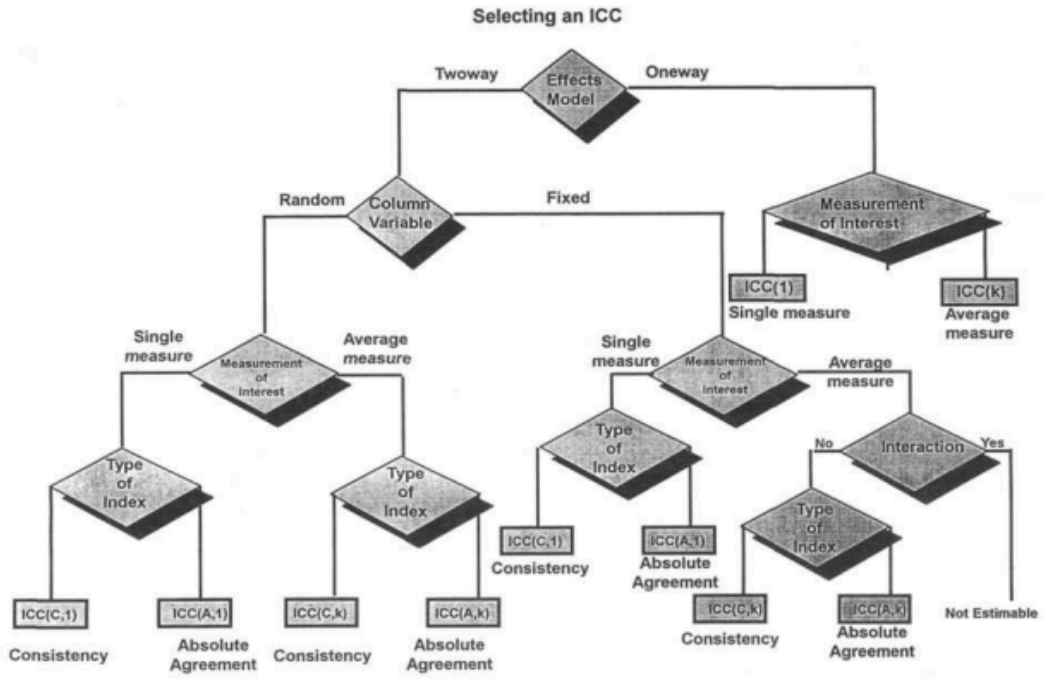


Figure 3.6: A Flowchart of Selecting an ICC

Note. Reprinted from McGraw and Wong (1996, p. 40).

3.7.1.6. Interpretation of ICC Values

It should be acknowledged that there is no universally agreed-upon threshold for determining adequate reliability when interpreting ICC values (Han, 2020; Koo and Li, 2016). A low ICC does not necessarily indicate poor rater agreement as it may also be due to restricted score variability, limited sample size, or a small number of raters (Han, 2020; Lee et al., 2012; Portney and Watkins, 2000). Thus, methodological guidelines commonly recommend at least 30 participants with heterogeneous performance levels and involving a minimum of 3 raters when conducting reliability analyses (Koo and Li, 2016). Thus, for the interrater reliability and calibration processes, there were 30 audio-recordings of students' speaking performances and 7 raters in total.

Under these conditions, ICC values are interpreted as in the table below (Koo and Li, 2016).

Table 3.6: Interpretation of ICC values

ICC Range	Interpretation of Reliability
<.50	Poor reliability
.50-.75	Moderate reliability
.75-.90	Good reliability
>.90	Excellent reliability

Note. Adapted from Koo and Li (2016)

Also, as part of the reliability analysis, in addition to the ICC, SPSS also reported the associated 95% confidence interval (CI) as part of the reliability analysis (Koo and Li, 2016). A 95% CI for ICC represents the range of values within which the true population ICC is expected to lie with 95% confidence that is based on the observed sample data (Koo and Li, 2016). In other words, the 95% CI shows that the reliability estimate is both high and precise, which means that even at its lowest plausible value, the level of agreement remains strong. The 95% CI is interpreted by using the established ICC interpretation bands, which means that the true ICC value falls within the commonly classified interpretation bands that are presented in the table (Koo and Li, 2016). It should be noted that the calculation of CI is more complex for ICC(A) because it includes the column variance in the denominator, unlike ICC(C), which excludes this source of variance (McGraw and Wong, 1996). In simple terms, once again ICC(A) accounts for differences between

raters whereas ICC(C) ignores these differences and focuses only on consistency (McGraw and Wong, 1996).

In summary, the choice of an ICC form for reliability analysis depends on the specific design of the reliability study and involves careful consideration of the appropriate model, type and definition for the analysis (Koo and Li, 2016). Thus, the ICC analysis was used as both a calibration-check and an indicator of interrater reliability before the pre- and post-tests.

The results of the interrater reliability for the two raters calculated by using the intraclass correlation coefficient (ICC) through a two-way mixed-effects model with absolute agreement was applied so it treated raters as fixed effects. Thus, it examined inter-rater agreement (IRA), whose analysis indicated good agreement between the two raters. The single-measures ICC was .75, while the average-measures ICC was .86, with a 95% confidence interval ranging from .70 to .93. The ICC value was statistically significant ($p<.001$), which displayed a high level of agreement between raters.

Upon completion of giving their score, the 5 raters were compared against one another to see whether they had internal agreement within their group so that their scores can be compared against the 2 raters, which would determine whether the 2 raters can accurately grade speaking performances. For the 5 raters, the interrater reliability was examined by assessing IRA by using ICC, which was also through a two-way mixed-effects model with absolute agreement that treated raters as fixed effects. The single measures ICC was .91, whereas the average-measures ICC reached .98. The ICC value was statistically significant ($p<.001$). This illustrated a nearly perfect agreement between the 5 raters.

The 7 raters' scores were calculated in the same manner as the previous interrater reliability calculations. Thus, interrater reliability was established by analyzing IRA among the 7 raters, which was evaluated through ICC involving a two-way mixed-effects model that used absolute agreement. The analysis demonstrated excellent agreement among raters. The single-measures ICC was .88, and the average-measures ICC was .98, with a 95% confidence interval ranging from .97 to .99. The ICC depicts statistical significance ($p<.001$), which highlights the almost perfect agreement across all raters. Following the establishment of high interrater agreement, the two raters were deemed sufficient for scoring pre- and post-tests.

The interpretations of the ICC results for establishing the reliability in this study can also be found summarized in Table 3.5 under section 3.6.1.

3.7.2. Pre- and Post- Test Results

To begin with, the pre- and post-tests were analyzed according to their normality. To check normality, as the sample size was small (12 students per group), Shapiro-Wilk was conducted on SPSS as it is a test to assess normality of the data's distribution (Field, 2009). For Shapiro-Wilk, if $p \geq .05$ then this indicated that the data was normally distributed (Field, 2009).

If the data was normally distributed so $p \geq .05$, then t-test was used. If the data was not normally distributed so $p < .05$, then Mann-Whitney U test was used. To decide whether the data was distributed normally, the p values were taken into consideration.

Independent samples t—test is used to compare the mean scores of exactly two groups (Field, 2009; Morgan et al., 2012; Stone, 2010). This is because it is a part of parametric statistics (Field, 2009; Morgan et al., 2012). It is typically applied to examine whether the population means differ and uses sample data to make inferences about the populations (Stone, 2010). Independent samples consist of different observations in each group (Field, 2009; Stone, 2010). Thus, the independent sample t-test compares two separate groups (Field, 2009; Stone, 2010), which is why it is also referred to as a two-sample t-test. The resulting t value helps indicate whether the difference between the two-sample means is statistically meaningful or likely to have occurred by chance (Field, 2009; Stone, 2010).

The Mann-Whitney U test is a nonparametric statistical procedure used to examine the differences between two independent groups (Field, 2009; Hart, 2001; Morgan et al., 2012). In the strictest statistical interpretation, it determines whether observations from one group are more likely to produce higher values than those from another group (Hart, 2001). When the sample size in each group is below 15 and the scores don't follow a normal distribution, the Mann-Whitney U test is used as the nonparametric counterpart to the independent samples t-test (Hart, 2001). In other words, the Mann-Whitney U test is the nonparametric version of the independent samples t-tests. It compares two separate groups and indicates which group tends to have higher scores without relying on strict statistical assumptions (Hart, 2001; Morgan et al., 2012). Therefore, the

Mann-Whitney U test has a statistical power that is only slightly lower than the t-test, which is why it makes it a good alternative to it (Morgan et al., 2012). The asymptotic significance is taken into consideration when there is a small sample size and if the value is $p<.05$, then the two groups are significantly different (Fisher, 1992).

To decide whether the meaning of the independent samples t-test or the Mann-Whitney U test, the p value was taken into account.

$P <.05$ value means that the independent samples t-test was meaningful (Fisher, 1992). In other words, it was statistically significant.

$P <.05$ is the critical value that means that the Mann-Whitney U test is meaningful (Fisher, 1992). In simple terms, it was statistically significant.

First, the experiment and control group pre-test scores were analyzed, then the post-tests of both groups were analyzed. Moreover, each group's own pre- and post-tests were compared within itself. In addition, each of IELTS speaking rubric's criterion was analyzed for the pre- and post-test results, so fluency and coherence, lexical resource, grammatical range and vocabulary, and pronunciation were also examined individually by each rater's scores regarding before and after the treatment. Finally, both of the groups' descriptive statistics were analyzed so the mean, and standard deviation (SD) of the pre- and post-tests were calculated.

Descriptive statistics refers to a group of statistical measures to summarize the main features and distributional properties of a dataset (Lee, 2009). In classical statistics, these include minimum and maximum, range, percentiles, mean, median, mode, SD, and variance (Lee, 2009). Descriptive statistics provide an overview of the data and form a basis for comparing patterns and differences across datasets (Lee, 2009).

Furthermore, to answer RQ iii, the individual mean scores of each student in the experiment group (n=12) was calculated for their pre- and post-tests to see the difference between their scores of the two stages. For this part of the study, the interpretation of the difference of the students' mean scores between the pre- and post-tests was relative to the IELTS speaking band movement as this was the speaking rubric used in this study to assess students' speaking performance. Thus, as the speaking scores derived from the IELTS speaking rubric, changes were interpreted in relation to IELTS band movement, in which differences of approximately 0.5 represent moderate improvement and differences of 1.0 or above indicate substantial improvement. The interpretation of the difference of the mean scores of learners for the pre- and post-test stages are listed in the table below.

Table 3.7: Pre- and Post-Test Individual Score Differences Interpretation

Difference range	Interpretation
0.00-0.49	Low improvement
0.50-0.99	Moderate improvement
≥ 1.00	High/ substantial improvement

3.7.3. Quantitative Perception Scale

First, the descriptive statistics (i.e., items' means range, mean, SD of the quantitative scales of Ironsi's (2023) was calculated as they summarized the experiment group's (n=12) responses to the 5-point Likert-scale items; 1 being strongly disagree and 5 being strongly agree. Negatively worded items were reverse-coded prior to analysis so that higher scores consistently reflected more positive perceptions toward VR use. Mean scores were calculated on Microsoft Excel for this scale by calculating the average of each item response and the overall average score of this scale. In addition to calculating the mean scores, SD was calculated on Microsoft Excel. For each item on the scale, the SD was calculated based on distribution of individual participants' responses around the item mean. Similarly, the SD for the overall scale score was calculated by using the participants'' averaged scale scores, which indicated the extent to which individual scores varied from the overall mean of the scale.

Also, individual mean scores were calculated and the SD for each student in the experiment group to answer the third research question. Individual mean scores of each student were calculated

by averaging their scores across the items of the scale. Afterwards, the SD of these individual mean scores were calculated to represent the degree of variability in each student's own scores of the scale. The following table presents how the interpretation of the item mean scores, the overall mean scores, and the individual mean scores of each student in the experiment group (n=12) of Ironsi's (2023) perception scale was conducted. The class intervals of the perception scale were calculated by following Alkharusi's (2022) approach for 5-point Likert-scale composite means.

Table 3.8: Interpretation of the Descriptive Statistics of Ironsi's (2023) Perception Scale Based on Alkharusi's (2022) Class Intervals

Scale range	Interval width	Mean score range	Interval midpoint	Interpretation
1-5	1.33	1.00-2.33	1.67	Low agreement
		2.34-3.67	3.01	Moderate agreement
		3.68-5.01	4.35	High agreement

Interpretation of the mean scores was conducted using equal class intervals by following Alkharusi's (2022) approach for average-based scores that includes fewer categories. Based on Alkharusi (2022) for 5 point Likert scale, first the scale range is calculated (5-1=4), then the interval width (4÷3=1.33), then the mean range and interpretation are based on interval width. This approach also includes the interval midpoint to compare the mean scores to.

As Ironsi's (2023) is a 5-point Likert scale (1-5), the interval width was calculated by dividing the range of the scale (4) by three. This produced and interval width of 1.33. Thus, mean scores between 1.00-2.33 meant low agreement, 2.34-3.67 indicated moderate agreement and 3.68-5.01 suggested high agreement. Interval midpoints represent the central value of each interpretive range. In other words, the interval midpoint is the average of the lower bounds and upper bounds of each class interval. They were calculated to provide a reference point for interpreting how closely the observed mean scores aligned with low, moderate or high levels of agreement. The same predefined intervals were applied to interpret both the overall mean score of the scale and the item means of it. The intervals were calculated once as the scale (1-5) never changes, and since all item means and the overall mean of the scale are on the same 5-point Likert scale, the same pre-defined intervals could be used to interpret them consistently.

Also, the SD of the overall score, the item-level analysis, and individual mean scores of each student in the experiment group of the perception scale was interpreted by using the table below as all were based on the same scale. The SD value of the overall perception scale displays how similar or different students' answers were from one another in their overall evaluation of VR experience. The SD values at item-level reflect variability across students for that item, whereas within group SD analysis was descriptively and comparatively reflected the consistency of the student's own responses. It is important to mention that there is no set rule for calculating SD. In a 5-point Likert scale, given that the one scale point represents a change between adjacent response categories, so the value of 0.60 was used as the cut-off point because it represents the upper boundary of less than half a Likert scale step on a 5-point scale, which allows for a small practical buffer. Thus, the interpretation of SD was calculated as the following in the table below.

Table 3.9: Interpretation of SD Ranges of Ironsi's (2023) Perception scale

SD range	Mathematical meaning	Interpretation
<0.60	<1/2 Likert step	Low variability
0.61-1.00	≈ 1 Likert step	Moderate variability
>1.00	> 1 Likert step	High variability

Following this, Pearson correlation analysis was conducted on SPSS to explore the relationship among items within this scale. Pearson's correlation coefficient I as a statistical measure used to describe the strength and direction of the relationship between two continuous variables (Field, 2009). The value of r ranges from -1 (perfect negative relationship) to 0 (no relationship) and to +1 (perfect positive relationship) (Field, 2009). Since this study aimed to understand learners' perceptions instead of improving the scale itself, Evan's (1996) framework was used to interpret the strength of the relationships as it is more flexible unlike those that include very strict measurements. Below is a table that presents the guidelines to interpret the strength of Pearson correlation coefficients that were provided by Evans (1996).

Table 3.10: Interpretation of Pearson Correlation Coefficient Strength

Pearson Correlation Coefficient	Relationship
.00-.19	Very weak
.20-.39	Weak
.40-.59	Moderate
.60-.79	Strong
.80-1.00	Very strong

Note. Correlation strength categories follow Evans (1996).

Even though the sample size was small for a correlation analysis to be conducted in a regular context, this study used correlation analysis as the perception scale employed in this study had not yet been widely adopted in empirical research of VR in EFL education for speaking skills. Hence, it was significant to examine not only overall perceptions scores but also the internal relationships among the scale items. That's why correlation analysis was conducted to show how conceptual dimensions in the scale were related to one another, which is why it doesn't aim to generalize the results but to gain insight into the internal coherence and interpretive structure of this scale.

Also, prior to the categorization of item-level correlation patterns, the items of Ironsi's (2023) perception scale were categorized into analytic groups based on their primary conceptual focus. In other words, this categorization was guided by the semantic content of each item. It is important to note that this is not thematic analysis as categorization was based on pre-defined dimensions of the items in the scale. This categorization was conducted to facilitate clearer interpretation of correlation patterns among items, which were then used for the top-down thematic analysis of the semi-structured questionnaire's data. The analytic categories of Ironsi's (2023) perception scale can be found below.

Table 3.11: Categorization of Items in Ironsi's (2023) Perception Scale

Item	Item Statement	Analytic Category
1	VR is not effective in foreign language acquisition classrooms*	Instructional effectiveness
2	VR enhances learning and speaking	Instructional effectiveness
3	VR does not improve speaking skills*	Instructional effectiveness
4	VR introduces virtual real-life situations helpful in speaking	Instructional value and authenticity
5	Teaching through VR improved classroom performance in speaking	Learning outcomes
6	The use of VR improved my speaking grades	Learning outcomes
7	VR makes speaking lessons boring*	Engagement and enjoyment
8	It is not difficult for me to handle and use VR during speaking lessons	Usability
9	VR gadgets are expensive to purchase	Practical constraints
10	VR is efficient in learning to speak	Learning efficiency
11	VR makes learning to speak fun and easy	Engagement and enjoyment
12	VR helps to achieve lesson objectives on speaking	Instructional effectiveness
13	I am concerned about privacy and security issues*	Privacy and security concerns

Note. The item statements were taken from Ironsi's (2023) perception scale (p.18).

The table below clarifies how items were conceptually grouped to support interpretation of the correlation patterns.

Table 3.12: Analytic Categorization of Items in Ironsi's (2023) Perception Scale

Analytic Category	Item(s)
Instructional effectiveness	1, 2, 3, 12
Instructional value and authenticity	4
Learning outcomes	5, 6
Engagement and enjoyment	7, 11
Usability	8
Practical constraints	9
Learning efficiency	10
Privacy and security concerns	13

The detailed Pearson correlation results within scale items for Ironsi's (2023) perception scale are presented in Appendix 16.

3.7.4. Semi-Structured Questionnaire

The semi-structured questionnaire was analyzed on MAXQDA by using thematic analysis. Thematic analysis is an analytic approach that is used to identify, examine, and present recurring patterns of themes within qualitative data by providing a structured yet flexible manner of organizing and describing the data while providing detailed and meaningful insights (Braun and Clarke, 2006).

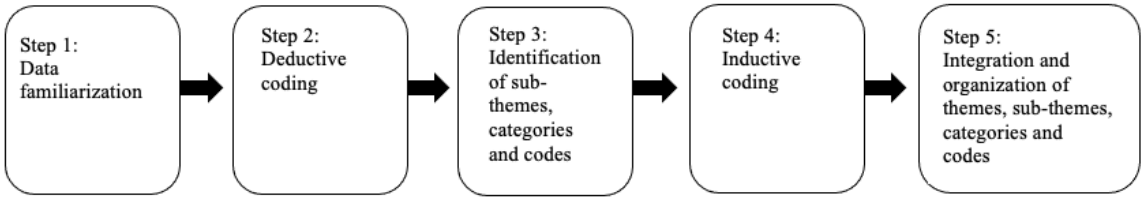


Figure 3.7: Qualitative Data Analysis Procedure

Above is a figure that portrays the step-by-step procedure of the qualitative data analysis procedure. It illustrates the stages from data familiarization to the integration and organization of the qualitative data.

Although the semi-structured questionnaire about perceptions included several closed-ended response options, these items were used to guide participants' reflections as they had prompts. Thus, these closed-ended response options were not used to generate independent quantitative results. Therefore, the qualitative data was analyzed holistically using thematic analysis with the open-ended responses and closed-ended responses considered together to support and provide context for the interpretation of themes, sub-themes and categories.

The analysis involved familiarization with the data, and then using the codes generated from the quantitative data's categories to find them in the qualitative data. Hence, first a top-down approach was utilized. The analytic categories of the quantitative scale eventually became the themes in this thematic analysis. Then, a further in-depth analysis was conducted as subthemes, categories, and codes were found within the data. Also, emergent themes appeared so open-coding was conducted for this part as these themes emerged from the descriptive data itself and wasn't observed in the quantitative scale's categories. Thus, this required open-coding. Finally, the pre-defined themes and the emergent themes were grouped along with their relevant subthemes and

categories. Deductive coding involves predefined categories informed by research questions and prior literature, whereas inductive coding refers to codes and themes emerging from the data itself without being pre-defined (Braun and Clarke, 2006). As deductive coding (top-down approach) and inductive coding (bottom-up approach) were both applied during the analysis of the qualitative data, a hybrid coding approach was used. Hybrid coding refers to the integration of deductive and inductive coding strategies, which allows researchers to apply pre-defined categories and remain open to themes that emerge from the data itself (Fereday and Muir-Cochrane, 2006).

It is important to mention that frequency counts (n) were also reported in this thematic analysis to indicate the prevalence of themes, subthemes and categories, but the analysis predominantly focused on identifying patterns of meaning in learners' responses rather than quantifying individual units of data. Thus, it does not include descriptive content analysis.

Below is a table describing the levels at which integration occurs in mixed-methods research that was proposed by Fetters et al. (2013).

Table 3.13: Integration Levels within Mixed-Methods Research

Integration Level	Approaches
Design	**3 basic designs**
	Exploratory sequential
	Explanatory sequential
	Convergent
	4 Advanced frameworks
	Multistage
	Intervention
	Case study
	Participatory—Community-based participatory research, and transformative
Methods	Connecting
	Building
	Merging
	Embedding
Interpretation and Reporting	Narrative—Weaving, contiguous and staged
	Data transformation
	Joint display

Note. Reprinted from (Fetters et al., 2013, p. 2136).

Based on the framework proposed by (Fetters et al., 2013), this study demonstrates integration at multiple levels. At the design level, an explanatory sequential mixed-methods design

was employed as quantitative data was collected and analyzed first, followed by the qualitative data to further explain and elaborate on the quantitative results. This design choice can be seen in the structure of the research questions, where students' speaking performances and perceptions were initially examined through quantitative measures, and qualitative data was subsequently used to provide explanatory depth and contextual understanding.

At the methods level, integration was achieved predominantly through connecting and merging quantitative and qualitative datasets. Connecting occurred through the quantitative and qualitative datasets being linked as they had the same participants since the same students contributed to both branches of data that enabled direct comparison of objectively measured outcomes and subjective accounts. Merging compasses both the quantitative and qualitative dataset integration in order to analyze them jointly and examine the similarities and differences between the findings, whereas connecting involves the same participants to contribute in both phases of the study (Fetters et al., 2013). In this study, merging was used at the analysis stage by bringing together the quantitative data and qualitative data's results for comparison after each type of data had been analyzed to determine where findings were similar, different or complementary to one another. This was specifically relevant for the second and third research questions. For the second research question, results from Ironsi's (2023) perception scale including its correlation between its items, and the responses from the semi-structured questionnaire was used. For the third research question, all of the results of the data collection instruments were used; pre- and post-test results of the speaking performances, findings of Ironsi's (2023) perception scale and the answers from the semi-structured questionnaire. These findings were compared to identify where the findings supported each other, complemented one another, or differed. This approach allowed the quantitative outcomes to be interpreted more meaningfully by incorporating students' perspectives and experiences. Therefore, it provided a more comprehensive understanding of the findings.

At the interpretation and reporting level, integration occurred through narrative strategies, which included a contiguous approach with elements of weaving during interpretation. Integration at the interpretation and reporting level was achieved through narrative integration as defined by Fetter et al. (2013), in which researchers present qualitative and quantitative findings together in a single report or across multiple reports. The contiguous approach refers to presenting the findings in a single report while reporting qualitative and quantitative results in separate sections while

maintaining a clear and consistent structure (Fetters, et al., 2013), which was seen in the organization of the results chapter in this study. In addition, elements of the weaving approach were employed at the interpretation stage, where qualitative and quantitative findings were discussed together by organizing them around shared themes or concepts (Fetters, et al., 2013). In this study, the weaving approach was applied during interpretation by discussing qualitative findings alongside quantitative results under shared themes to explain, complement or contrast the results. This combined narrative strategy allowed both the quantitative and qualitative data to clarify one another and supported an integrated interpretation of the findings.

The following is written to analyze each RQ in detail according to their integration level. The findings of RQ i presented the quantitative foundation of the explanatory sequential design, which established the extent of speaking development before any interpretive integration.

To address RQ ii, both quantitative and qualitative data were used. The integration for RQ ii occurred at the methods level through connecting, and through merging at the analysis stage as quantitative results of the perception scale were compared with the qualitative descriptions of the semi-structured questionnaire. Therefore, contiguous narrative integration was employed with quantitative and qualitative findings present in separate but sequential sections within the same report. At the interpretation level, integration was achieved through a weaving approach through which quantitative results were interpreted and contextualized using students' qualitative responses.

The following procedures were employed to integrate quantitative and qualitative data in addressing RQ iii. For the final research question, students' actual speaking performance was represented through their overall speaking scores from the pre- and post-tests. These scores were calculated by averaging the four speaking criteria across the two raters. To observe students' improvement over time, speaking gain scores were calculated as the difference between pre- and post-test scores. Furthermore, each student's perception of the use of VR for learning and speaking English were drawn from their overall mean score on Ironsi's (2023) perception scale. In addition, students' responses from the semi-structured questionnaire were examined to provide qualitative insights on how students' perceptions of this matter aligned with the reality of their speaking performance. In short, all of the outcomes of all of the data collected were used. Thus, RQ iii

represented the highest level of integration as it required the use of all sources. Integration occurred through connecting, and merging at the analysis stage, in which perception and performance were examined for alignment or differentiation. Through contiguous narrative reporting, quantitative and qualitative findings were first presented in separate sections and then interpreted by using elements of weaving as both types of data were considered together to identify patterns of similarities, differences of complementary results.

Below presents a figure that represents the information in regard to the integration levels of methods and interpretation and reporting of this study.

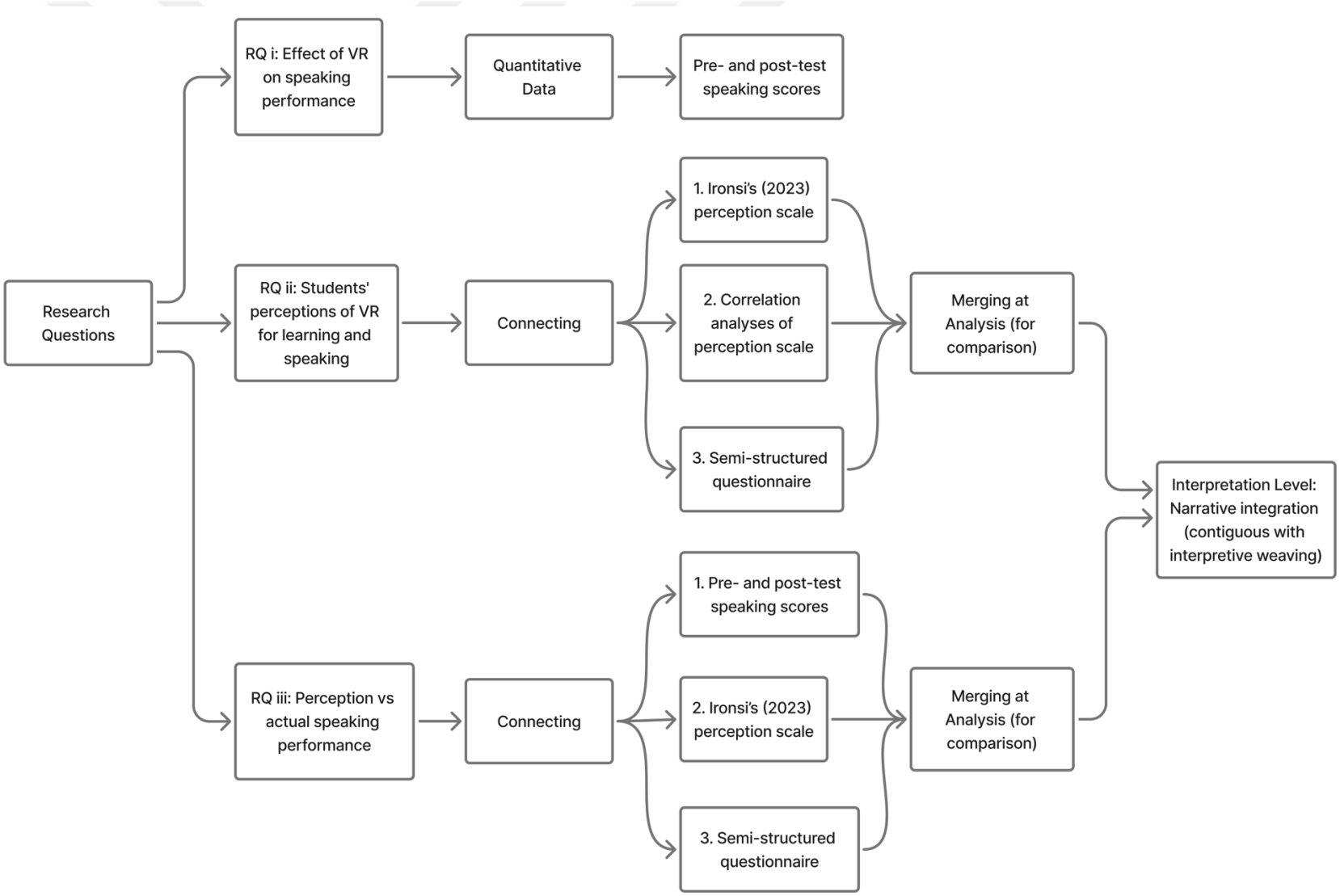


Figure 3.8: Integration Levels of Methods, and Interpretation and Reporting

Note. Adapted from Fetters et al. (2013) for this study

3.8. ETHICAL CONSIDERATIONS

The ethics committee of Istanbul Bilgi University had granted permission (see Ethics Committee Permission) for this research to be conducted along with receiving the official permission of BEPP (see Institutional Permissions). The topic of this study was approved by Istanbul University-Cerrahpaşa (see Appendix 17).

All of the participants in both the experiment group (see Appendix 18 for the English version and Appendix 19 for the Turkish version) and control group received an informed consent form (see Appendix 20 for the English version and Appendix 21 for the Turkish version) that contained all the necessary information about the study, which included details of the research, participation details, voluntary participation (being able to withdraw whenever they want), confidentiality, contact information about the researcher and the MA thesis advisor. The informed consent form for the control group was the same too, except without mention the intervention of VR to keep this group oblivious. These participants kept a copy of the informed consent form and the researcher kept a copy too. All participants were over the age of 18 and they had volunteered to be a part of this study. Their identities were kept confidential too as they were given code names. All participants had given permission for audio-recordings to be conducted in this study.

Moreover, the 30 students who volunteered to assist with the interrater reliability and calibration process had been informed through consent forms (see Appendix 22 for the English version and Appendix 23 for the Turkish version) as well, which included the purpose of the audio-recording of their speaking performance, their role, confidentiality, voluntary participation and contact information.

Additionally, the teachers who volunteered to assist with the interrater reliability and the calibration process had been given an informed consent form (for Rater 2's informed consent for see Appendix 24 and for the other 5 raters see Appendix 25) regarding the content of the study, their role along with the other headings mentioned for the other informed consent forms. The identities of these teachers were kept confidential as well.

Each informed consent form in this study was written in both English and Turkish to ensure that all was clear.

There were no risks associated with any humans in this study.

3.9. LIMITATIONS

The limitations in this study are related to the number of participants, length of study and using LiVR instead of HiVR as the intervention. Since the number of participants was small, the results of the study cannot be generalized. Furthermore, due to the scheduling program of BEPP, the study could only be conducted for 6 weeks. A better understanding of the effectiveness of VR could have been noted if it had been studied for longer to provide a more reliable outcome. Had this study been conducted through the use of HiVR, aspects pertaining to immersion and presence could have been studied to observe its implications on learners' motivation, engagement and how students process language as "being there". Moreover, the speaking performances were assessed exclusively through individual interviews that were aligned with CEFR (CoE, 2020) sustained monologue subscales, so interactive skills such as turn-taking and negotiation of meaning were beyond the scope of this study. Finally, learners' perceptions were measured through self-report instruments, which may be influenced by social desirability, recall bias or novelty effects, and the researcher's dual role as the instructor and researcher may have shaped participants' responses despite multiple measures to minimize bias. For these reasons, the findings should be interpreted within the specific educational context of the study and viewed as exploratory rather than broadly generalizable. In short, due to the pre-existing classes' having a small sample size, and the lack of random assignment of quantitative data, the findings of this study should be interpreted as exploratory rather than confirmatory. Therefore, the results of the study sole provide initial evidence of the potential effects of VR on students' speaking performance. Thus, future studies that address these limitations are more likely to produce better results to understand this issue at hand.

3.10. SUMMARY

This chapter outlined the methodological framework of this study by explaining the research design, participants, setting, data collection instruments and procedures, and data analysis methods that were employed to investigate the effect of VR on EFL learners' speaking performance and their perceptions. An explanatory sequential mixed-methods was adopted, which started with a quantitative phase to measure changes in students' speaking performances and their perceptions, followed by a qualitative phase to further explore learners' perceptions of using VR for practicing their speaking skills. The study was conducted with intermediate

level students at BEPP. Pre-existing classes were assigned to the groups of experiment and control. The data was collected through multiple instruments, which were pre- and post-tests of speaking performance using IELTS speaking rubric, a validated perception scale, and a semi-structured questionnaire. Meticulous procedures were implemented to ensure reliability and validity, which included interrater reliability, rater calibration, standardized assessment conditions, and expert review of both quantitative and qualitative analyses. The potential threats to internal validity were systematically addressed, and trustworthiness in the qualitative data was established through triangulation, peer debriefing, and transparent documentation. In short, this chapter presented that the methodological choices were theoretically grounded, contextually appropriate, and aligned with the research questions. By doing so, it has provided a solid foundation for interpreting the results and discussing their implications.

4. RESULTS

This chapter reports the findings of the current research study, which incorporated both quantitative and qualitative data to investigate the impact of VR on students' speaking performance and their perceptions toward this issue at hand. The quantitative results include analyses of pre- and post-test speaking scores of the experiment (n=12) and control (n=12) groups, along with responses of the experiment group for Ironsi's (2023) perception scales. The qualitative results are based on students' semi-structured questionnaire responses. The results are presented in relation to the research questions which have guided the study.

4.1. THE EFFECT OF VR ON STUDENTS' SPEAKING PERFORMANCE

In order to analyze the first research question regarding the effect of VR on students' speaking performance, an analysis of experiment group (n=12) and control group (n=12) students' pre-test and post-test scores was conducted. The speaking performance results of the pre- and posts of both groups are presented in a funnel manner; starting from the overall descriptive scores of both groups on pre- and post-tests, followed by Shapiro wilk normality results that determined the independent samples t-test or the Mann-Whitney U test for the overall scores and speaking subskills on the pre- and post-tests, and finally displaying the speaking subskills and overall mean score of each rater. Thus, first an overview of the speaking performance scores of the two groups and the two testing stages are provided, which is then followed by stating parametric or non-parametric inferential procedures were used. Finally, rater-specific speaking subskills and mean scores are presented for a more detailed view of how overall performance patterns appeared across different speaking components and raters.

4.1.1. Speaking Performance Results of Pre- and Post-Tests

4.1.1.1. Descriptive Statistics for Pre- and Post-Test Speaking Scores

Below Table 4.1 presents the descriptive statistics for the experiment and control groups' overall speaking performance scores in the pre-test and the post-test.

Table 4.1: Descriptive Statistics for Experiment and Control Groups on Pre- and Post-Test Speaking Scores

Measure	Group	N	M	SD
Pre-test Overall	Experiment	12	4.44	1.19
	Control	12	3.41	0.90
Post-test Overall	Experiment	12	5.45	0.53
	Control	12	4.56	0.78

Note. M = mean; SD = standard deviation

Table 4.1 presents that at the pre-test stage, the experiment group obtained a higher mean score (M=4.44, SD=1.19) than the control group (M=3.41, SD=0.90). After the intervention, the mean scores increased in both groups; however, the experiment group demonstrated a greater increase in overall speaking performance (M= 5.45, SD=0.53) than the control group (M= 4.56, SD= 0.78). The mean score difference between the experiment and control groups was 1.03 at the pre-test stage and 0.89 at the post-test stage. The mean score difference at the pre-test and post-test stages approximately corresponds to one IELTS speaking band, which especially indicates a substantial baseline difference.

The experiment group showed a mean increase of 1.01 points from pre-test to post-test, whereas the control group displayed a mean increase of 1.15 points. The standard deviation of the experiment group's mean scores decreased by 0.66, while the control group showed a decrease of 0.12.

4.1.2. Inferential Analysis of Speaking Performance in Pre- and Post-Tests

4.1.2.1. Shapiro-Wilk Normality Test Results for Pre-Test Scores

Prior to the main analyses of the pre- and post-test scores, the normality of the speaking performance scores was examined using the Shapiro-Wilk test.

Table 4.2: Normality Assessment and Selection of Statistical Tests for Pre-Test Scores

Measure	Group	Shapiro-Wilk P	Normality	Selected Test
Pre (Overall)	Experiment	.641	Normal	t
	Control	.69	Normal	
R1 Pre Fluency & Coherence	Experiment	.201	Normal	U
	Control	.021	Not Normal	
R1 Pre Lexical Resource	Experiment	.276	Normal	U
	Control	.006	Not normal	
R1 Pre Grammatical Range and Accuracy	Experiment	.276	Normal	t
	Control	.134	Normal	
R1 Pre Pronunciation	Experiment	.015	Not normal	U
	Control	.118	Normal	
R1 Pre Overall Average	Experiment	.444	Normal	t
	Control	.218	Normal	
R2 Pre Fluency & Coherence	Experiment	.259	Normal	t
	Control	.187	Normal	
R2 Pre Lexical Resource	Experiment	.228	Normal	t
	Control	.146	Normal	
R2 Pre Grammatical Range and Accuracy	Experiment	.654	Normal	t
	Control	.172	Normal	
R2 Pre Pronunciation	Experiment	.292	Normal	t
	Control	.053	Normal	
R2 Pre Overall Average	Experiment	<.001	Not normal	U
	Control	.249	Normal	

Note. Normality was assessed using the Shapiro-Wilk test. When $p \geq .05$, the distribution was considered normal and independent samples t-test were conducted. When $p < .05$, the normality assumption was violated and Mann-Whitney U tests were applied.

As shown in Table 4.2, when the normality assumption was met, independent samples t-tests were conducted, whereas Mann-Whitney U tests were applied when this assumption was

violated ($p<.05$) for the pre-test scores. The Shapiro-Wilk test results indicated that normality was met for several pre-test measures including the overall pre-test speaking score. However, violations of the normality assumption were observed in some rater and subskill specific measures. Thus, both parametric and non-parametric tests were employed to ensure that the statistical analyses were conducted according to the distributional characteristics of the data.

4.1.2.2. Shapiro-Wilk Normality Test Results for Post-Test Scores

Below table 4.3 presents the Shapiro-Wilk results for the post-test scores.

Table 4.3: Normality Assessment and Selection of Statistical Tests for Post-Test Scores

Measure	Group	Shapiro-Wilk P	Normality	Selected Test
Post (Overall)	Experiment	.50	Normal	t
	Control	.996	Normal	
R1 Post Fluency & Coherence	Experiment	.004	Not normal	U
	Control	.078	Normal	
R1 Post Lexical Resource	Experiment	.030	Not normal	U
	Control	.187	Normal	
R1 Post Grammatical Range & Accuracy	Experiment	.005	Not normal	U
	Control	.029	Not normal	
R1 Post Pronunciation	Experiment	<.001	Not normal	U
	Control	.003	Not normal	
R1 Post Overall Average	Experiment	.276	Normal	t
	Control	.613	Normal	
R2 Post Fluency & Coherence	Experiment	.006	Not normal	U
	Control	.187	Normal	
R2 Post Lexical Resource	Experiment	.030	Not normal	U
	Control	.069	Normal	
R2 Post Grammatical Range & Accuracy	Experiment	.012	Not normal	U
	Control	.069	Normal	
R2 Post Pronunciation	Experiment	<.001	Not normal	U
	Control	.133	Normal	

Table 4.3 (continued): Normality Assessment and Selection of Statistical Tests for Post-Test Scores

R2 Post Overall Average	Experiment	.004	Not normal	U
	Control	.478	Normal	

Note. Normality was assessed using the Shapiro-Wilk test. When p≥ .05, the distribution was considered normal and independent samples t-test were conducted. When p <.05, the normality assumption was violated and Mann-Whitney U tests were applied.

As indicated in Table 4.3, which followed the same procedure used for the pre-test scores, independent samples t-test were employed for post-test scores that met the normality assumption (p≥.05). However, Mann-Whitney U tests were used when this assumption was violated (p <.05). The Shapiro-Wilk test results revealed mixed distributional patterns across post-test measures and groups. While overall post-test speaking scores were normally distributed in both groups, several subskill-level measures in the experiment group deviated from normality. In contrast, the control group's post-test scores were predominantly normally distributed. Thus, both parametric and non-parametric tests were used for the post-test analyses to ensure alignment with the distributional characteristics of the data.

4.1.2.3. Inferential Analysis Group Comparisons of Overall Speaking Performance of Pre- and Post-Tests

Independent samples t-test results that compare the experiment and control groups' overall speaking performance scores are presented below in Table 4.4.

Table 4.4: Independent Samples t-Test Results of the Overall Scores for Pre- and Post-Tests

Measure	t	df	P	Mean Difference	SE Difference	95% CI for Mean Difference
Pre-test Overall	2.39	22	.026	1.03	0.43	[0.14, 1.92]
Post-test Overall	3.27	22	.004	0.89	0.27	[0.32, 1.45]

Note. The results are reported for the equal variances assumed condition. t = t statistics; df = degrees of freedom; SE = standard error; CI = confidence interval. All p values are two-tailed.

As it can be observed from Table 4.4, a statistically significant difference was found between the two groups at the pre-test stage (t (22) = 2.39, p=.026) as p <.05. At the post-test stage, the difference between the groups was also statistically significant (t (22) = 3.27, p=.004), so the experiment group achieved higher overall speaking performance scores than the comparison group.

4.1.3. Speaking Subskill and Overall Individual Rater Mean Results of Pre- and Post-Tests

4.1.3.1. Pre-Test Results of Speaking Subskills and Rater Mean Scores

Further analyses were conducted to examine group differences across specific speaking subskills in the IELTS speaking rubric, which were fluency and coherence, lexical resource, grammatical range and accuracy, and pronunciation as rated by each rater. This also included each rater's own overall average score.

Table 4.5 provides the independent samples t-test scores for the pre-test scores of speaking subskills given by each rater.

Table 4.5: Independent Samples t-Test for Pre-Test Speaking Measures

Phase	Measure	Rater	t	df	p
Pre	Grammatical Range & Accuracy	R1	1.62	22	.119
Pre	Overall Average	R1	2.07	22	.050
Pre	Fluency & Coherence	R2	2.08	22	.050
Pre	Lexical Resource	R2	2.40	22	.025
Pre	Grammatical Range & Accuracy	R2	1.76	22	.093
Pre	Pronunciation	R2	1.61	22	.122

Note. Independent samples t-tests were conducted for measures where normality assumption was met for both groups. Results are reported for the equal variances assumed condition. All p values are two-tailed.

As shown in the table above, no statistically significant differences were found between the experiment and control groups in grammatical range and accuracy as rated by Rater 1(t (22) = 1.62, p=.119) or Rater 2 (t (22) =1.76, p=.093), as all p values were greater than .05. These results indicate that the experiment and control groups were considered comparable in terms of grammatical range and accuracy at this stage. Similarly, no statistically significant difference was observed for pronunciation as rated by Rater 2 (t (22) =1.61, p=.122). This suggests that the two groups demonstrated similar levels of performance of pronunciation for this rater. Also, the differences observed for the overall average speaking performance as rated by Rater 1 (t (22) = 2.07, p=.050) and for fluency and coherence as rated by Rater 2 (t (22) = 2.08, p=.050) were not statistically significant as the p values were equal to .05 and they did not meet the $p < .05$ criterion. Thus, experiment and control groups were considered comparable for these measures depending on the rater.

In contrast, a statistically significant difference was identified for lexical resource as rated by Rater 2 (t (22) = 2.40, p=.025). Since p < .05, the independent samples t-test was meaningful (Fisher, 1992), with the experiment group obtaining higher lexical resource scores than the control group.

Table 4.6 presents the Mann-Whitney U test scores for the pre-test scores of speaking subskills given by each rater.

Table 4.6: Mann-Whitney U Test Results for Pre-Test Speaking Measures

Phase	Measure	Rater	U	z	p
Pre	Fluency & Coherence	R1	43.50	-1.73	.084
Pre	Lexical Resource	R1	40.00	-1.92	.055
Pre	Pronunciation	R1	35.50	-2.22	.027
Pre	Overall Average	R2	38.00	-1.97	.049

Mann-Whitney U tests were conducted for pre-test speaking measures that did not meet the normality assumption. As shown in Table 4.6, no statistically significant differences were found between the experiment and comparison groups for fluency and coherence (U=43.50, z=-1.73, p=.084) or lexical resource (U=40.00, z=-1.92, p=.055) as rated by Rater 1 as p <.05. Therefore, the experiment and control groups were considered comparable for fluency and coherence, and lexical resource for Rater 1.

However, a statistically significant difference was observed for pronunciation as rated by Rater 1 (U=35.50, z=-2.22, p=.027). Also, a statistically significant difference was found for the overall average scores as rated by Rater 2 (U=38.00, z=-1.97, p=.049). Since p < .05, the Mann-Whitney U test was meaningful (Fisher, 1992). This indicated a significant difference between the experiment and the control group for these measures.

4.1.3.2. Post-Test Results of Speaking Subskills and Rater Mean Scores

Additional analyses were carried out to identify post-test differences of each rater's scores between groups with respect to individual speaking subskills defined in the IELTS speaking rubric and the overall mean post-test scores assigned by each rater.

Table 4.7 provides the independent samples t-test scores for the post-test scores of the overall average score of the speaking performances given by Rater 1.

Table 4.7: Independent Samples t-Test for Post-Test Speaking Measures

Phase	Measure	Rater	t	df	p
Post	Overall Average	R1	2.81	22	.010

As depicted in Table 4.7, a statistically significant difference was found between the experiment and the control group in overall speaking performance at post-test stage as rated by Rater 1 (t (22) =2.81, p= .010). Since p < .05, the independent samples t-test was meaningful (Fisher, 1992). This indicates that the experiment group obtained higher overall speaking scores than the control group on the post-test for Rater 1.

Below Table 4.8 provides the Mann-Whitney U test results for the post-test scores.

Table 4.8: Mann-Whitney U Test Results for Post-Test Speaking Measures

Phase	Measure	Rater	U	z	p
Post	Fluency & Coherence	R1	39.00	-2.00	.045
Post	Lexical Resource	R1	46.00	-1.60	.110
Post	Grammatical Range & Accuracy	R1	36.00	-2.25	.024
Post	Pronunciation	R1	27.50	-2.87	.004
Post	Fluency & Coherence	R2	24.00	-2.92	.004
Post	Lexical Resource	R2	36.00	-2.17	.030
Post	Grammatical Range & Accuracy	R2	42.50	-1.79	.073
Post	Pronunciation	R2	28.50	-2.66	.008
Post	Overall Average	R2	30.50	-2.45	.014

For Rater 1, statistically significant differences were found in fluency and coherence (U=39.00, z=-2.00, p=.045), grammatical range and accuracy (U=36.00, z=-2.25, p=.024), and pronunciation (U=27.50, z=-2.87, p=.004) with the experiment group obtaining higher scores for these speaking subskills for Rater 1. Since p <.05 in these measures, the Mann-Whitney U tests were meaningful (Fisher, 1992). In contrast, lexical resource as rated by Rater 1 did not show a statistically significant difference (U=46.00, z=-1.60, p=.110) as the p value was greater than .05. Hence, in regard to lexical resource, the two groups were seen as comparable for this rater.

For Rater 2, statistically significant differences were observed in fluency and coherence (U=24.00, z=-2.92, p=.004), lexical resource (U=36.00, z=-2.17, p=.030), pronunciation (U=28.50, z=-2.66, p=.008), and overall average of speaking scores (U=30.50, z=-2.45, p=.014), which show that the experiment group obtained higher scores across these speaking measures for this rater. As all corresponding p values were smaller than .05, these results were statistically significant (Fisher, 1992). However, grammatical range and accuracy as rated by Rater 2 did not reach statistical significance (U=42.50, z=-1.79, p=.073). Therefore, the two groups were seen as comparable for grammatical range and accuracy for Rater 2.

In brief, taking into account across speaking subskills and overall speaking scores, the experiment and control groups were largely comparable at the pre-test stage as no statistically significant differences were observed for most measures of the speaking performance. However, there was a small number of rater-specific differences. At the post-test stage, statistically significant differences were observed across several speaking subskills and the overall speaking performance that were in favor of the experiment group, which showed that the experiment group especially obtained higher scores in fluency and coherence, pronunciation and overall average of speaking performance for both raters. Despite not all speaking subskills reaching a statistical significance for both raters, the overall pattern of results indicate that the experiment group had a better speaking performance in comparison to the control group on multiple speaking measures on the post-test.

4.2. STUDENTS' PERCEPTIONS OF USING VR FOR LEARNING AND SPEAKING ENGLISH

As the second research question is about the perception of using VR, only data from the experiment group (n=12) was gathered. Also, both quantitative and qualitative data were analyzed to address the second research question, which examines students' perception of using VR for learning and speaking English. The quantitative data for this question was obtained from Ironsi's (2023) perception scale, whereas the qualitative data was from the semi-structured questionnaire about students' perceptions toward using VR.

4.2.1. Quantitative Results of Students' Perceptions Toward Using VR

This subsection provides the descriptive statistics of students' perceptions by presenting the overall descriptive statistics and item-level descriptive statistical results of Ironsi's (2023) perception scale in addition to the Pearson correlation results of the items in Ironsi's (2023) perception scale.

4.2.1.1. Descriptive Statistics of Students' Perceptions

The overall descriptive statistics for students' scores on Ironsi's (2023) perception scale is presented in Table 4.9.

Table 4.9: The Overall Descriptive Statistics of Ironsi's (2023) Perception Scale

Scale	*N*	Items' means range (Min-Max)	Range	M	SD
Ironsi's (2023) perception scale	12	2.5-4.42	1.92	4.04	0.53

Note. Range refers to the difference between the minimum and maximum item mean values.

As reported in Table 4.9, the mean score was 4.04 (SD=0.53) with item means ranging from 2.50 to 4.42 for Ironsi's (2023) perception scale. The observed range of item means was 1.92, which indicates some variability across items for this scale. Based on Alkharusi's (2022) class intervals, the overall scale mean (M=4.04) was located within the high agreement interval (3.68-5.01), which supported the general positive perceptions toward using VR for learning and speaking. When compared to the interval midpoint (4.35) it falls slightly below it, which indicates overall high agreement but not at the strongest level observed. The SD of the overall mean of this scale indicates low overall variability in students' answers as it is <.60.

4.2.1.2. Item-Level Descriptive Statistical Results of Ironsi's (2023) Perception Scale

Table 4.10 presents the item-level means and standard deviations for Ironsi's (2023) perception scale.

Table 4.10: Mean Scores and Standard Deviation of Ironsi's (2023) Perception Scale

N	Item	Statement	M	SD
12	1	VR is not effective in foreign language acquisition classrooms*	3.83	1.47
12	2	VR enhances learning and speaking	4.25	0.45
12	3	VR does not improve speaking skills*	4.33	0.65
12	4	VR introduces virtual real-life situations helpful in speaking	4.25	0.87

Table 4.10 (continued): Mean Scores and Standard Deviation of Ironsi's (2023) Perception Scale

12	5	Teaching through VR improved classroom performance in speaking	4.42	0.51
12	6	The use of VR improved my speaking grades	4.17	0.58
12	7	VR makes speaking lessons boring*	4.33	0.65
12	8	It is not difficult for me to handle and use VR during speaking lessons	4.17	0.72
12	9	VR gadgets are expensive to purchase*	2.5	1
12	10	VR is efficient in learning to speak	4.25	0.62
12	11	VR makes learning to speak fun and easy	4.33	0.49
12	12	VR helps to achieve lesson objectives on speaking	4.25	0.45
12	13	I am concerned about privacy and security issues*	3.5	1.24

Based on Alkharusi's (2022) class intervals, the item-level mean scores of Ironsi's (2023) perception scale largely indicated a high level of agreement among students. Most items produced mean scores within the high agreement range (3.68-5.01), which suggested positive perceptions of VR use for learning and speaking English. Especially students reported high agreement regarding the effectiveness of VR in enhancing learning and speaking (Item 2, M= 4.25), improving speaking skills (Item 3, M=4.33), increasing classroom performance (Item 5, M=4.42), and making speaking lessons more engaging and enjoyable (Items 7 and 11, M=4.33; Item 10, M=4.25). Similarly, items related to instructional affordances such as providing realistic speaking situations (Item 4, M=4.25), supporting lesson objectives (Item 12, M=4.25), improving speaking grades (Item 6, M=4.17) and ease of use (Item 8, M=4.17) also demonstrated high agreement.

Items 3, 4, 7, 8 and 10 show moderate variability, which implies some variation in students' responses. While items 2, 5, 6, 11, and 12 showed low variability, which illustrated strong agreement among learners about perceptions related to instructional effectiveness, learning outcomes, and engagement and enjoyment.

However, even though Item 1 (M=3.83) also represented the effectiveness of VR for foreign language acquisition, it was classified within the high agreement range but it was comparatively lower than most other item means within this interval. Thus, this item demonstrated comparatively weaker agreement than most other items in this range. This item also displayed a high variability, which suggests divided views on the overall effectiveness of VR for language learning.

In contrast, comparatively lower mean scores were observed for items addressing potential concerns. Students expressed moderate agreement regarding the cost of VR equipment (Item 9, M=2.50), and privacy and security issues (Item 13, M=3.50). Item 9 indicates borderline high variability, which suggests notable differences in the students' perceptions of cost related constraints of VR, whereas Item 13 shows a high variability, which indicates mixed views of learners on this issue.

When interpreted using interval midpoints, most items mean clustered around the midpoint of the high agreement interval (4.35), which indicates strong agreement of VR-related statements. Items with means closest to this midpoint reflected the strongest levels of agreement, whereas Item 1, was farther from the midpoint when compared to the other item means so it as mentioned previously it showed weaker agreement than the others. Items falling in the moderate agreement range displayed variation when compared to the moderate interval midpoint (3.01). Item 13 indicated stronger moderate agreement and Item 9 reflected weaker moderate agreement.

4.2.1.3. Pearson Correlation Results of Ironsi's (2023) Perception Scale

Pearson correlation analysis was conducted to examine the relationship between students' perceptions toward VR to answer research question ii.i. This analysis was based on Ironsi's Perception scale as a correlation within the items of this scale was analyzed.

Table 4.11 offers the Pearson correlation results for within Ironsi's (2023) perception scale.

Table 4.11: Summary of Pearson Correlation Results for Ironsi's (2023) Perception Scale

Correlation Category	Description of Item Relationships
Positive Correlations	The positive correlation analysis revealed a highly consistent pattern of associations across analytic categories related to instructional effectiveness, instructional value and authenticity, learning outcomes, learning efficiency, usability, and engagement and enjoyment. Perceptions of VR as effective for speaking instruction were positively associated with perceptions of authenticity, improved speaking outcomes, efficient learning, ease of use, enjoyment and achievement of lesson objectives. In addition, lower perceived practical constraints and lower privacy and security concerns were positively associated with several instructional, learning and affective categories, which indicated that favorable learning experiences tended to co-occurred with reduced concern about cost and data-related issues.
No Linear Correlations	No linear correlations were observed between Item 7 (VR makes speaking lessons boring) and Item 9 (VR gadgets are expensive to purchase), and Item 8 (ease of use of VR during speaking lessons) and Item 13 (privacy and security concerns). These findings suggest that students' affective engagement in VR-based speaking lessons functioned independently of cost-related perceptions and that ease-of-use judgments were not systematically related to concerns about privacy and security.
Negative Correlations	The negative correlation analysis demonstrated consistent inverse relationships between variables after negatively worded items were reverse-coded to ensure aligned scoring directions. Higher perceptions of instructional effectiveness, authenticity, learning outcomes, learning efficiency, usability, and engagement were associated with lower agreement of ineffectiveness beliefs, reduced boredom, fewer cost-related constraints and fewer privacy and security concerns. These inverse relationships indicate that as students' positive evaluations of VR-based instructions increased, opposing negative beliefs and concerns tended to decrease.

Note. N=12. Pearson correlation coefficients were calculated using pairwise deletion. Item numbers correspond to statements in Ironsi's (2023) perception scale. Negatively worded items were reverse-coded prior to analysis so that higher scores consistently reflected more positive perceptions. See appendix 16 for the full item-level correlation matrix.

Pearson correlation coefficients were calculated to examine the linear relationships among the items of this scale. The magnitude of the Pearson correlation was interpreted following Evans' (1996) classification of correlation strength; values between .00-.19 were considered very weak, values between .20-.39 were weak, values between .40-.59 were moderate, values between .60-.79 were strong, and values that fall between .80-1.00 were very strong. This was used to interpret both the positive and negative correlations.

The results indicated that most Item pairs demonstrated positive linear correlations, which suggested that there is internal coherence among the scale items. Some item pairs demonstrated strong and very strong positive correlations. Several item pairs showed weak to moderate positive relationships, and some displayed very weak relationships. In addition, some item pairs portrayed negative correlations. Negative correlations within this scale were also observed across item pairs through a range of strengths, which were from very weak to very strong as well. As negatively worded items were reverse-coded prior to analysis, these negative correlations indicate inverse linear associations between aligned positive constructs rather than the co-occurrence of negative perceptions. A small number of item pairs displayed no linear

correlation. In short, both positive and negative correlations of varying strengths were observed within the scale.

Figure 4.1 on the following page presents the Pearson correlation matrix that illustrates the item relationships among the items of Ironsi's (2023) perception scale items.

Correlations

		G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12	G13
G1	Pearson Correlation	1	-,069	-,349	,322	-,261	-,072	-,063	,288	,310	-,050	,084	-,069	-,050
	Sig. (2-tailed)		,832	,266	,307	,413	,825	,845	,364	,327	,878	,795	,832	,878
	N	12	12	12	12	12	12	12	12	12	12	12	12	12
G2	Pearson Correlation	-,069	1	-,617*	,290	,683*	,522	-,309	,420	,101	,728**	,408	,556	-,081
	Sig. (2-tailed)	,832		,033	,360	,014	,082	,329	,174	,756	,007	,188	,061	,803
	N	12	12	12	12	12	12	12	12	12	12	12	12	12
G3	Pearson Correlation	-,349	-,617*	1	-,645*	-,090	-,322	,357	-,259	-,419	-,674*	-,472	-,309	,112
	Sig. (2-tailed)	,266	,033		,024	,780	,307	,254	,416	,175	,016	,121	,329	,728
	N	12	12	12	12	12	12	12	12	12	12	12	12	12
G4	Pearson Correlation	,322	,290	-,645*	1	-,051	,273	-,161	,073	,472	,380	,213	,290	-,211
	Sig. (2-tailed)	,307	,360	,024		,875	,391	,617	,821	,121	,223	,506	,360	,510
	N	12	12	12	12	12	12	12	12	12	12	12	12	12
G5	Pearson Correlation	-,261	,683*	-,090	-,051	1	,663*	-,361	,533	-,088	,497	,478	,683*	-,071
	Sig. (2-tailed)	,413	,014	,780	,875		,019	,248	,074	,785	,100	,116	,014	,826
	N	12	12	12	12	12	12	12	12	12	12	12	12	12
G6	Pearson Correlation	-,072	,522	-,322	,273	,663*	1	-,322	,366	-,157	,633*	,746**	,870***	-,507
	Sig. (2-tailed)	,825	,082	,307	,391	,019		,307	,242	,625	,027	,005	<,001	,093
	N	12	12	12	12	12	12	12	12	12	12	12	12	12
G7	Pearson Correlation	-,063	-,309	,357	-,161	-,361	-,322	1	-,065	,000	-,225	-,472	-,617*	,561
	Sig. (2-tailed)	,845	,329	,254	,617	,248	,307		,841	1,000	,483	,121	,033	,058
	N	12	12	12	12	12	12	12	12	12	12	12	12	12
G8	Pearson Correlation	,288	,420	-,259	,073	,533	,366	-,065	1	,380	,102	,600*	,420	,000
	Sig. (2-tailed)	,364	,174	,416	,821	,074	,242	,841		,223	,753	,039	,174	1,000
	N	12	12	12	12	12	12	12	12	12	12	12	12	12
G9	Pearson Correlation	,310	,101	-,419	,472	-,088	-,157	,000	,380	1	-,073	,185	-,101	,219
	Sig. (2-tailed)	,327	,756	,175	,121	,785	,625	1,000	,223		,821	,566	,756	,493
	N	12	12	12	12	12	12	12	12	12	12	12	12	12
G10	Pearson Correlation	-,050	,728**	-,674*	,380	,497	,633*	-,225	,102	-,073	1	,297	,404	,059
	Sig. (2-tailed)	,878	,007	,016	,223	,100	,027	,483	,753	,821		,348	,192	,856
	N	12	12	12	12	12	12	12	12	12	12	12	12	12
G11	Pearson Correlation	,084	,408	-,472	,213	,478	,746**	-,472	,600*	,185	,297	1	,816**	-,594*
	Sig. (2-tailed)	,795	,188	,121	,506	,116	,005	,121	,039	,566	,348		,001	,042
	N	12	12	12	12	12	12	12	12	12	12	12	12	12
G12	Pearson Correlation	-,069	,556	-,309	,290	,683*	,870***	-,617*	,420	-,101	,404	,816**	1	-,728**
	Sig. (2-tailed)	,832	,061	,329	,360	,014	<,001	,033	,174	,756	,192	,001		,007
	N	12	12	12	12	12	12	12	12	12	12	12	12	12
G13	Pearson Correlation	-,050	-,081	,112	-,211	-,071	-,507	,561	,000	,219	,059	-,594*	-,728**	1
	Sig. (2-tailed)	,878	,803	,728	,510	,826	,093	,058	1,000	,493	,856	,042	,007	
	N	12	12	12	12	12	12	12	12	12	12	12	12	12

*. Correlation is significant at the 0.05 level (2-tailed).

**. Correlation is significant at the 0.01 level (2-tailed).

***. Correlation at 0.001(2-tailed)

Figure 4.1: Pearson Correlation Matrix of Ironsi's (2023) Perception Scale Items (G1-G13)

4.2.1.4. Analytically Derived Categories Based on Item-Level Correlation Patterns

Below Table 4.12 that presents the categories identified based on the item-level positive correlation patterns in Ironsi's (2023) perception scale. The positive relationships among the analytic categories show that students' positive views of one aspect of VR-based speaking instruction were often accompanied by positive views of other aspects of the experience.

Table 4.12: Positive Correlation Patterns Across Analytic Categories in Ironsi's (2023) Perception Scale

Analytic category	Categories positively correlated with	Description of positive correlation
Instructional effectiveness	Instructional value and authenticity Learning outcomes Learning efficiency Usability Engagement and enjoyment Practical constraints Privacy and security concerns	Higher perceived instructional effectiveness was associated with stronger perceptions of authenticity, improved learning outcomes, greater efficiency, better usability, higher engagement, fewer practical constraints and fewer privacy concerns. This indicates that when students felt VR worked well for learning, they also tended to think that it was realistic, useful, easy to use, enjoyable and not very problematic in terms of cost or privacy.
Instructional value and authenticity	Instructional effectiveness Learning outcomes Learning efficiency Usability Engagement and enjoyment Practical constraints	Perceiving VR as authentic and instructionally valuable was positively related to effectiveness, learning gains, efficiency, ease of use, engagement and reduced cost-related constraints. This means that when students saw VR as realistic and useful for learning, they were more likely to feel that it worked well, helped them learn, was easy and efficient to use, enjoyable and not very problematic in terms of cost
Learning outcomes	Instructional effectiveness Instructional value and authenticity Learning efficiency Usability Engagement and enjoyment	Stronger perceived learning outcomes in speaking were associated with higher instructional effectiveness, authenticity, efficiency, usability, and engagement. This suggests that when students felt they improved in speaking, they also tended to see VR as effective, realistic, easy and efficient to use and enjoyable
Learning efficiency	Instructional effectiveness Learning outcomes Usability Engagement and enjoyment Privacy and security concerns Instructional value and authenticity	Viewing VR as an efficient learning tool was positively associated with effectiveness, authenticity, learning gains, usability, engagement and lower levels of privacy and security concerns. This indicates that when students felt VR helped them learn efficiently, they also tended to see it as effective, realistic, easy to use, enjoyable and less worrying in terms of privacy and security.

Table 4.12 (continued): Positive Correlation Patterns Across Analytic Categories in Ironsi's (2023) Perception Scale

Usability	Instructional effectiveness Instructional value and authenticity Learning outcomes Learning efficiency Engagement and enjoyment Practical constraints	Greater perceived ease of use was associated with stronger instructional effectiveness, authenticity, learning outcomes, efficiency, engagement, and fewer cost-related constraints. This means that when students found VR easy to use, they were more likely to see it as effective, realistic, helpful for learning, efficient, enjoyable and not concerned with the cost.
Engagement and enjoyment	Instructional effectiveness Instructional value and authenticity Learning outcomes Learning efficiency Usability Practical constraints Privacy and security concerns	Higher engagement and enjoyment in VR-based speaking activities were associated with stronger perceptions of instructional effectiveness, authenticity, learning outcomes, learning efficiency, and usability as well as lower perceptions of practical constraints, and privacy and security concerns. This indicates that when students enjoyed and felt engaged in VR speaking activities, they tended to see VR as more effective, realistic, easy and efficient to use, and less problematic in terms of cost, and privacy and security.
Practical constraints	Instructional effectiveness Instructional value and authenticity Learning efficiency Usability Engagement and enjoyment Privacy and security concerns	Lower perceived cost-related constraints were associated with more positive instructional authenticity, learning, usability, and engagement perceptions as well as fewer privacy and security concerns. This means that when students did not see cost as a big problem, they were more likely to view VR as realistic, useful for learning, easy to use, enjoyable, and less of a concern in terms of privacy and security.
Privacy and security concerns	Instructional effectiveness Learning efficiency Engagement and enjoyment Practical constraints	Lower privacy and security concerns were positively associated with higher instructional effectiveness, learning efficiency, engagement and fewer perceived practical constraints. This means that when students were less worried about privacy and security, they were more likely to see VR as effective and efficient, enjoy using it more, and feel that practical issues like cost were less of a problem.

Note. See Appendix 16 for detailed item-level correlation results.

Overall, the pattern of positive correlations presented in Table 4.12 suggests that students' perceptions of VR-based speaking instruction were closely interconnected rather than isolated. When students evaluated one aspect of the VR experience positively, they tended to report similarly positive perceptions across other instructional, learning and affective dimensions, while practical and privacy-related concerns were lower. This consistent grouping of positive

perceptions indicates that learners experienced VR-based speaking instruction as a coherent and integrated learning environment rather than as a set of unrelated features.

Below Table 4.13 displays the categories derived from item-level no linear correlation patterns in Ironsi's (2023) perception scale. These findings indicate instances where students' perceptions across different dimensions of VR-based speaking experience did not show a consistent directional association.

Table 4.13: No Linear Correlation Patterns Across Analytic Categories in Ironsi's (2023) Perception Scale

Analytic category	Categories with no linear relationship	Description of no linear correlation
Engagement and enjoyment	Practical constraints	No linear relationship was observed between students' engagement or boredom in VR-based speaking lessons and their perceptions of the cost of VR gadgets. This indicates that affective engagement operated independently of cost-related concerns.
Usability	Privacy and security concerns	No linear relationship was found between perceptions of ease of use of VR and concerns about privacy and security. This suggests that usability judgments were not systematically related to data-related or security concerns.

Note. See Appendix 16 for detailed item-level correlation results.

In brief, the absence of linear relationships in these category pairings suggests that certain dimensions function independently rather than together. Especially students' engagement and enjoyment were not systematically linked to cost-related concerns and perceptions of usability were not directly related to privacy and security concerns. These findings highlight that positive or negative evaluations in one area of the VR experience did not necessarily predict perceptions in these other domains.

Below Table 4.14 presents the categories derived from item-level negative correlation patterns in Ironsi's (2023) perception scale. As all negatively worded items were reverse-coded prior to analysis, the negative correlations reported in this table represent inverse relationships

between aligned positive constructs rather than the co-occurrence of negative perceptions. In other words, negative correlations below are real and not caused by wording or scoring errors that might have reflected fake or misleading information.

Table 4.14: Negative Correlation Patterns Across Analytic Categories in Ironsi's (2023) Perception Scale

Analytic category	Categories negatively correlated with	Description of negative correlation
Instructional effectiveness	Instructional effectiveness Learning outcomes Learning efficiency Engagement and enjoyment Privacy and security concerns Instructional value and authenticity Usability Practical constraints	Higher perceived instructional effectiveness was associated with lower agreement of ineffectiveness beliefs, stronger learning outcomes, greater efficiency, higher engagement, better usability, stronger authenticity perceptions, and fewer practical constraints, and privacy concerns. This means that when students felt VR worked well for learning, they also tended to think that it helped their learning, was easy and efficient to use, enjoyable, realistic and caused fewer problems in terms of cost and privacy.
Instructional value and authenticity	Instructional effectiveness Learning outcomes Engagement and enjoyment Privacy and security concerns	Stronger perceptions of authenticity were associated with lower instructional ineffectiveness beliefs, stronger learning outcomes, higher engagement, and fewer privacy and security concerns. This suggests that when students felt VR was realistic and authentic, they were less likely to think it was ineffective, more likely to feel they learned better, enjoyed it more and had fewer worries about privacy and security. On a side note, the only positively worded items that displayed a negative correlation between instructional value and authenticity, and learning outcomes. This indicates that higher learning gains were not consistently accompanied by higher authenticity ratings of VR. However, as there was also a positive correlation between these two items, this suggests that the findings reflected individual differences in how VR was experienced and perceived.
Learning outcomes	Instructional effectiveness Engagement and enjoyment Privacy and security concerns Practical constraints Instructional value and authenticity	Stronger perceived learning outcomes were associated with lower instructional ineffectiveness, reduced boredom, fewer privacy and security concerns, fewer cost-related constraints, and stronger authenticity perceptions. This indicates that when students felt their speaking really improved, they were less likely to see VR as ineffective or boring, had fewer worries about cost and privacy, and were more likely to see it as realistic and authentic.

Table 4.14 (continued): Negative Correlation Patterns Across Analytic Categories in Ironsi's (2023) Perception Scale

Learning efficiency	Instructional effectiveness Engagement and enjoyment Practical constraints	Higher perceived learning efficiency was associated with lower instructional ineffectiveness beliefs, reduced boredom, and fewer cost-related constraints. This means that when students felt VR helped them learn efficiently, they were less likely to think it was ineffective or boring and were less concerned about cost issues.
Usability	Engagement and enjoyment Instructional effectiveness	Greater ease of use was associated with lower boredom and weaker agreement of instructional ineffectiveness beliefs. This suggests that when students found VR easy to use, they were less likely to feel bored and less likely to think that VR was ineffective for learning.
Engagement and enjoyment	Usability Engagement and enjoyment Learning efficiency Instructional effectiveness Learning outcomes Instructional value and authenticity	Higher engagement and enjoyment were associated with lower boredom, stronger instructional effectiveness, greater efficiency, better usability, stronger learning outcomes, and higher authenticity perceptions. This indicates that when students enjoyed and felt engaged in VR activities, they were less bored and more likely to see VR as effective, efficient, realistic and easy to use and helpful for learning.
Practical constraints	Learning efficiency Instructional effectiveness Learning outcomes	Lower perceived cost-related constraints were associated with higher instructional effectiveness, greater learning efficiency, and stronger learning outcomes. This means that when students did not see cost as a big problem, they were more likely to feel that VR worked well, helped them learn efficiently, and improved their learning outcomes.
Privacy and security concerns	Instructional effectiveness Engagement and enjoyment Learning outcomes Instructional value and authenticity	Fewer privacy and security concerns were associated with higher instructional effectiveness, higher engagement, stronger learning outcomes and greater authenticity perceptions. This suggests that when students were less worried about privacy and security, they were more likely to see VR as effective, enjoy using it, feel that they learned more, and view it as realistic and authentic.

Note. See Appendix 16 for detailed item-level correlation results.

In short, the negative correlation patterns indicate that stronger positive perceptions in one dimension of the VR-based speaking experience were consistently associated with weaker agreement of opposing beliefs or concerns in other dimensions. In particular, higher perceptions

of instructional effectiveness, learning outcomes, efficiency, usability, engagement, and authenticity tended to link with lower levels of boredom, perceived ineffectiveness, cost-related constraints, and privacy and security concerns. These inverse relationships further highlight the internal coherence of students' perceptions, which suggests that favorable evaluations of VR-based speaking instruction were accompanied by a reduction in perceived limitations and concerns.

4.2.2. Qualitative Results of Students' Perceptions on the Use of VR for Learning and Speaking English

Below Table 4.15 presents the results of the thematic analysis that was conducted on the responses in the semi-structured questionnaire. As mentioned in the methodology chapter, the analytic categories present in the items of Ironsi's (2023) perception scale were used to conduct deductive coding of the semi-structured questionnaire through a top-down approach. Thus, these categories became the themes in the thematic data analysis due to them being vague and incorporating sub-themes and categories. This table summarizes the hierarchical coding structure as it outlines the theme, sub-theme, category and associated codes. The codes specifically emerged from the students' reflection on the VR-enhanced speaking experiences. Frequency counts (n) are provided to indicate the relative prevalence of each category across the participant group; however, these frequencies are reported descriptively and do not imply statistical generalization, which is due to the number of participants in this study. To sum up, the themes illustrate students' instructional, affective, experiential and practical perspectives on the use of VR for learning and speaking as it provides a comprehensive overview of the qualitative patterns that were identified in the data.

Table 4.15: Thematic Analysis of Semi-Structured Questionnaire

Theme	Sub-theme	Category	Codes
Instructional effectiveness (n=31)	General reflections on VR use (n=6)	Overall perceptions and future expectations of VR (n=6)	• Optimistic future expectations • Overall satisfaction with VR • Perceived future relevance of VR
	Perceived usefulness and endorsement of VR (n=14)	Instructional usability and practicality (n=3)	• Practical and comfortable use • Smoother lesson flow • Efficient learning experience
		Pedagogical value and learning support (n=6)	• Perceived contribution to overall language development • An alternative to traditional instruction

Table 4.15 (continued): Thematic Analysis of Semi-Structured Questionnaire

			• Provides corrective feedback • Effective learning strategy
		Engagement-oriented instruction (n=3)	• Enjoyable learning experience • Engaging and educational learning experience
		Affective support within instruction (n=2)	• Supportive and low stress speaking environment • Enhancement of speaking confidence • Learning through trial-and-error in a non-judgmental environment
	Comparative instructional effectiveness (n=11)	Perceived instructional advantages of VR (n=11)	• Engagement-oriented instructional design • Instructional flexibility and access • Instructional configuration and learning support • Affective support in instruction • Form-focused instruction and feedback • Learner control and strategic learning
Instructional value and authenticity (n=20)	Authenticity of VR-based speaking practice (n=15)	Authentic speech and pronunciation (n=2)	• Exposure to natural production models • Perceived realism of spoken output • Authentic pronunciation practice in context • Natural speech patterns in VR dialogues
		Authentic interaction (n=7)	• Realistic conversational exchanges • Contextually appropriate spoken interaction • Simulation of real-life communicative situations
		Immersion and sense of presence (n=6)	• Heightened sense of presence due to real-life simulation
	Experiential and intercultural learning (n=5)	Exposure to culturally situated experiences (n=5)	• Gaining cultural experience through virtual scenarios • Learning about unfamiliar cultures through VR • Perceived value of interaction with native speakers
Learning outcomes (n=172)	Learning outcomes of VR-enhanced speaking practice (n=145)	Perceived global improvement in speaking ability (n=3)	• Improved speaking performance • Feeling good after speaking development • Development of performance accuracy over time

Table 4.15 (continued): Thematic Analysis of Semi-Structured Questionnaire

		Perceived development of speaking skills (n=142)	• Pronunciation development • Fluency development • Vocabulary development • Grammar and accuracy development • Increased naturalness of speech
	Limited or uncertain learning outcomes (n=27)	Limited improvement in speaking subskills (n=13)	• Limited improvement in grammar • Limited improvement in vocabulary • Limited improvement in fluency
		Uncertain improvement in speaking subskills (n=14)	• Uncertain improvement in grammar • Uncertain improvement in vocabulary • Uncertain improvement in fluency
Learning efficiency (n=8)	Low-effort practice (n=8)	Reduced cognitive load and time-efficient learning (n=6)	• Effortless repetition without fatigue • Efficiency and enjoyment • Faster learning through VR-enhanced practice
		Easy access to speaking practice and repetition (n=2)	• Easy access to speaking practice and repeatability
Usability (n=14)	Technical challenges (n=14)	Issues related to hardware (n=2)	• Audio or microphone problems during speaking activities • Technical malfunctions during VR use
		Issues related to speech recognition (n=4)	• Failure of the system to recognize correct responses • Disruption of task flow due to recognition errors
		Absence of VR use challenges (n=8)	• No technical difficulties experienced
Engagement and enjoyment (n=42)	Interest and focus (n=15)	Enhanced interest and focus in VR-based instruction (n=15)	• Facilitated interest and focus through enjoyment • Facilitated interest and focus through impactful real-life simulation experience • Enhanced focus through engaging practice • Increased focus through individual and comfortable learning conditions • Increased interest and focus through novelty of experience

Table 4.15 (continued): Thematic Analysis of Semi-Structured Questionnaire

	Motivation and enjoyment (n=27)	Perceived motivational and enjoyment related effectiveness of VR (n=27)	• Increased motivation to practice speaking • Enjoyment through social interaction • Increased efficiency through enjoyment • Increased enjoyment through non-traditional lesson format • Variety of topics
Practical constraints (n=29)	Resource-based constraints (n=3)	Cost-related constraints (n=3)	• Cost-related barriers to accessing VR-enhanced instruction
	Instructional constraints (n=26)	Design and system level constraints (n=13)	• Interactive and collaborative instruction • Gamification and engagement-oriented design • Authenticity and realism of speaking tasks • Adaptive instructional feedback • Pronunciation-gated progression
		Pedagogical constraints (n=11)	• Instructional scaffolding and feedback support • Language input diversity • Task and content diversity • Technological and resource constraints • Institutional and implementation constraints
		No suggestions for improvement (n=2)	• Unable to identify areas for improvement
Affective factors in speaking (n=93)	Speaking confidence (n=64)	Post VR confidence level (n=27)	• Increased self-confidence after VR use • Feeling more comfortable speaking English
		Sources of increased confidence (n=28)	• Confidence gained through repeated practice • Confidence gained through individualized interaction • Confidence gained through real-life simulation • Confidence gained due to practicing in a non-judgmental environment • Confidence gained after improvement of speaking skills • Confidence gained in real-life simulation enabling transfer of skills to real world

Table 4.15 (continued): Thematic Analysis of Semi-Structured Questionnaire

		Neutral or unchanged confidence perceptions (n=9)	• No noticeable change in speaking confidence • Already confident before VR use
	Speaking anxiety (n=29)	Reduced anxiety in a non-judgmental environment (n=20)	• Reduced anxiety due to low stress virtual environment • Reduced fear of negative evaluation • Reduced negative preconceptions toward speaking • Reduced fear of making mistakes • No anxiety experienced in VR-based speaking practice • No pre-existing anxiety
		Anxiety regulation mechanisms (n=9)	• Individualized practice • Perceived gradual improvement • Ability to repeat the speaking activity • Perceived positive affective outcomes

All in all, the qualitative findings summarized in Table 4.15 indicate that students' experiences of using VR for learning and speaking were predominantly positive and multifaceted. In relation to Ironsi's (2023) perception scale, the thematic analysis presents all of the analytic categories except privacy and security concerns. The emergent theme that appeared in this thematic analysis which wasn't initially in Ironsi's (2023) perception scale was affective factors in speaking that encompassed speaking confidence and speaking anxiety. Each theme was reflected across multiple categories, with categories representing distinct dimensions of students' perception within the same overarching theme. This is in line with Can (2020), who employed a qualitative study through thematic analysis to explore pre-service teachers' experiences with machinima production in virtual environments, the results of which presented that themes were reflected across multiple categories rather than being strictly represented in a single analytic dimension.

The theme of instructional effectiveness (n=31) consisted of three sub-themes, which were general reflections on VR use (n=6), perceived usefulness and endorsement of VR (n=14), and comparative instructional effectiveness (n=11). The sub-theme of general reflections of VR use described learners' perceptions in general use of VR and their future expectations of it, which can be summarized as positive views toward using VR as a learning tool. Students stated "I believe that it will develop much further and reach a much better level in future" (S08), "it was enjoyable, it was a good experience" (S09), and "I feel fortunate to have had the opportunity to experience this application in advance as it is likely to be used more widely in the future" (S07).

The sub-theme of perceived usefulness and endorsement of VR included four categories; Instructional usability and practicality, pedagogical value and learning support, engagement-oriented instruction, and affective support within instruction. These categories emphasized the use of VR as a practical and comfortable instructional tool that assisted in developing their overall language skills through an enjoyable and engaging learning experience due to a low-stress environment that heightened learners' confidence in speaking. Learners expressed "VR makes speaking English both enjoyable and educational" (S03), "It is a platform where everyone can express themselves comfortably" (S06), "I think it is beneficial as the lessons have a smoother flow" (S09), and "VR will help strengthen students' confidence while speaking" (S07).

The sub-theme of comparative instructional effectiveness includes students' perceptions of the advantage of VR in comparison to other traditional learning methods. The codes in this category express a range of VR's strengths as they mention its instructional value on engagement, flexibility, different modalities of learning, learning support, affective support, providing form-meaning connection and feedback, learners' having control of their learning and the VR platform incorporating strategic learning. The students match the general consensus of VR being superior articulated "The technological features of the VR environment made it better than conventional learning" (S07), "There is not always enough time to speak in the classroom, but VR focuses only on you and gives you plenty of time to speak" (S03), and "In the classroom some students feel anxious about their teachers' or their classmates' reactions, but in VR, since there is no real person in front of them, they can feel more comfortable and be more open to improvement" (S02). The one student who disagreed with VR being superior to traditional learning voiced that "In the

classroom, the facial expressions are perceived as real, which I think makes learning more effective" (S04).

The theme of instructional value and authenticity (n=20) was categorized into two sub-themes; authenticity of VR-based speaking practice (n=15) and experiential and intercultural learning (n=5). Authenticity of VR-based speaking practice included students' responses related to authentic speech and pronunciation, authentic interaction, and immersion and sense of presence. These categories contained perceptions regarding natural production models, realistic spoken output, contextually appropriate interaction, and feeling as if the interaction was real due to real-life simulation. Student mentioned "Speaking in a chat-like format was very helpful" (S11), "Talking about different topics in different places was what I enjoyed the most because it created a realistic and sincere effect" (S03), "The environment and the feeling were as if you were really talking to someone" (S04) and "It provides exposure to more natural speech and accent" (S12).

Experiential and intercultural learning was in relation to students having experiences of different cultures based on specific situations through VR. Thus, they learned and gained cultural experience through the virtual scenarios. Learners expressed "I virtually saw places I had never been to" (S03) and "I learned many things about cultures I was not familiar with" (S05).

The most frequently represented theme was learning outcomes (n=172). It had two sub-themes; Learning outcomes of VR-enhanced speaking practice (n=145) and Limited or uncertain learning outcomes (n=27). The sub-theme of learning outcomes of VR-enhanced speaking practice encompassed two categories, which were perceived global improvement in speaking ability and perceived development of speaking skills. Perceived global improvement in speaking ability was in regards to students' general views of the overall speaking performance as learners stated "It improved my speaking skills" (S10) and "I found it developing" (S05). Whereas, perceived development of speaking skills reflected students' strong emphasis on perceived development in speaking skills in terms of fluency and naturalness of speech, vocabulary use, grammatical accuracy, and pronunciation. Learners' remarks included "At first, I made frequent mistakes, but later I noticed that I reduced these errors and almost reached an error-free level" (S07), "While

repeating, some words naturally stay in mind. Even though I did not focus specifically on vocabulary, it was still beneficial" (S09) ""With each correct and fluent speech, it is difficult to even describe the feeling of achievement" (S01), "I spoke more naturally and more quickly" (S02 and S08) and "I started making fewer grammar mistakes" (S07).

In contrast, the sub-theme of limited or uncertain learning outcomes expressed quite a few sets of responses that indicated students' perceptions of limited or uncertain learning gains. This displays a variation in how students experienced the instructional impact of VR. Students commented "I am not sure whether VR improved my fluency" (S10), "I am not sure whether there was a significant difference in the pace or naturalness of my speaking" (S04), "I am not sure whether VR improved my grammar" (S09), "VR helped improve my grammar but not very much" (S11) and "VR improved my vocabulary very slightly" (S01).

The theme of learning efficiency (n=8) was the least stated by the students. However, it highlighted learners' perception of VR as easy to use, providing repetitive practice and being time-efficient, so it was still beneficial. It includes learners' perception of VR as easy to use, providing repetitive practice and being time-efficient through the sole sub-theme of low-effort practice (n=8) as the codes displayed learners' effortless practice of speaking and being able to repeat the speaking activities without getting tired, which enabled easier and quicker learning. Students stated "It allowed me to repeat without feeling tired" (S09), "I noticed that we learned more quickly" (S08), "Rather than learning by speaking in front of others, learning in this way was much easier and not stressful" (S10), and "It was a process where I learned very efficiently while having fun" (S12).

The theme of usability (n=14) contained one sub-theme; technical challenges (n=14). This is the theme in which students highlighted issues related to hardware and speech recognition, which were related to audio or microphone problems, technical malfunctions, or the system failing to recognize speech. In addition to these challenges, there have also been mentions of an absence of experiencing difficulties while using VR, which shows that not all students struggled with these issues during the VR intervention. Students' remarks contained "There were minor issues with the

microphone and technical equipment" (S05 and S07), "The system sometimes did not detect properly" (S04), and "I did not experience any difficulties" (S12).

Moreover, the theme of engagement and enjoyment (n=42) was a significant theme that had two sub-themes; interest and focus (n=15), and motivation and enjoyment (n=27). As students frequently highlighted their interest, motivation and enjoyment during VR-enhanced speaking practice activities. The sub-theme of interest and focus highlights students' explanations of increased concentration during the VR use, which were also connected to their enjoyment, individualized and comfortable learning conditions, and the novelty experience of using VR. Learners mentioned their reasons as "You work individually and feel more comfortable" (S10), "When I first heard about it, I was very curious" (S11), "I liked the interaction I had with the person in those environments" and "VR was more helpful in terms of focus" (S09).

The sub-theme of motivation and enjoyment emphasized learners' accentuated motivation and enjoyment from practicing speaking, social interaction, efficiency, its non-traditional lesson format and inclusion of a variety of speaking topics. Learners commented that "I had fun with my classmates during the lesson" (S04), "VR kept my motivation high" (S09), "VR encouraged me to do more pronunciation and speaking practice" (S06), "It did not feel like a lesson; it felt like an experience, so we did not get bored when using VR for English" (S10), and "Communicating with others and the application's content were very colourful" (S12).

The theme of practical constraints (n=29) encompassed two sub-themes; resource-based constraints (n=3) and instructional constraints (n=26). Resource-based constraints were linked to cost-related barriers that affected the access of VR. Students highlighted "It may be costly. It is not easily accessible to everyone" (S03 and S10).

Whereas, instructional constraints comprise both design and system-level constraints and pedagogical constraints. These included a range of needs ranging from interactive and collaborative instruction, enhancement of gamification, greater authenticity to having a diversity of language input, and diversity of task and content. Students articulated "I would like it to feel as if you are talking to a real person wherever you want. There could be someone who corrects you

when you form incorrect sentences, but otherwise allows you to speak freely" (S10), "I would prepare speaking tests based on everyone's interests" (S07), "I would include more realistic backgrounds and facial expressions, and I would also include both male and female interlocutors, as the application we used only had male characters" (S04), "There could be daily goals that learners can realistically achieve and they could receive rewards or gains in return. This would be more motivating" (S06), and "Learners could speak with people from different professions, which could both help students with career choices and be more engaging" (S03). However, a small number of students mentioned no suggestions for improvement, which means that they weren't able to find areas for change in this VR platform as one learner expressed "I cannot respond because I do not keep up with VR applications" (S09).

The emergent theme of affective factors in speaking (n=93) had two sub-themes: speaking confidence (n=64) and speaking anxiety (n=29). Speaking confidence was in regard to students frequently mentioning their increased confidence in speaking after the VR intervention especially due to feeling more comfortable. In addition to this, they expressed where their sources for increased confidence came from such as repeated practice, individualized interaction, engagement in real-life simulations, practicing in a non-judgmental environment, being able to transfer their cultivated speaking skills to real-life situations thanks to VR, and also believing that their speaking skills had developed. Students mentioned "Before, I lack confidence when speaking, but VR made it easier" (S06), "I noticed it not during the sessions but later in my daily life. I realized that my confidence had increased and that I was pronouncing words more accurately" (S09), "It helped me a little to speak English more fluently and confidently" (S11), "After practicing in VR, I feel very confident about using newly learned words in real life" (S02), and "I felt more confident because in the classroom, people naturally feel tense, but VR does not create that kind of environment" (S03). Neutral or unchanged perceptions of their confidence levels were only stated by a small number of students. These students have expressed "I was already quite confident; I am not shy about these kinds of practices" (S04), and "I think there was no change in the confidence I have when speaking English" (S01).

Speaking anxiety was about how VR reduced students' anxiety due to it being a virtual environment with in its nature presents a low-stress learning environment as there is no negative

evaluation and students aren't afraid of making mistakes. Learners said "I felt more comfortable because I had the chance to make mistakes" (S07), "Because it increased my confidence, so my anxiety decreased" (S02), "I felt more comfortable because the person we were speaking to could not see us in real time" (S08), "It made me feel more comfortable because there was no real person in front of me" (S02) and "I felt comfortable because the person in front of me was in a virtual world and could not interfere with my speech or judge me" (S03). This sub-theme also contained anxiety regulation mechanisms as students describe how VR supported their anxiety management through individualized practice, observation of their gradual speaking development, being able to repeat the activity again and it generally reinforcing a positive affective outcome. Learners mentioned "It made me feel more comfortable because it was a one-to-one dialogue (S06 and S11), "It was helpful because I was able to say words that I could not say before, so my anxiety decreased" (S08), "I felt comfortable because I felt that I was gradually improving" (S12), and "I felt more comfortable because I always had the chance to try again; normally it is not possible to change what you have already said" (S09).

To conclude, the Table 4.15 presents a holistic view that covers a wide-range of themes that includes essential dimensions such as pedagogical, affective, experiential and practical in detail to comprehend the learners' experiences of using VR for learning and speaking.

4.3. COMPARISON OF STUDENTS' PERCEIVED AND ACTUAL SPEAKING PERFORMANCES

To address the third research question, students' actual speaking performance was represented by their overall speaking scores obtained from the pre- and post-test, while students' perceptions were measured by using the overall mean scores from Ironsi's (2023) perception scale. In addition, qualitative data from the semi-structured questionnaire were used to provide insight into the relationship between perceived and actual speaking improvement.

4.3.1. Quantitative Results of Pre- and Post-Test Scores and Ironsi's (2023) Perception Scale

4.3.1.1. Experiment Group Students' Individual Pre- and Post-Test of Speaking Performance

This subsection provides the individual pre- and post-test speaking performance scores of the students in the experiment group (n=12). For each student, the pre- and post-test speaking performance scores represent the mean of the scores given by the two raters as interrater reliability had been established.

Table 4.16: Pre- and Post-Test Speaking Scores and Score Differences for the Experiment Group

Student	Pre-Test	Post-Test	Difference
S01	6.38	6.50	0.13
S02	5.00	5.38	0.38
S03	2.50	5.25	2.75
S04	5.50	5.00	0.50
S05	4.88	5.38	0.50
S06	4.63	5.75	1.13
S07	4.75	5.50	0.75
S08	5.25	5.38	0.13
S09	3.63	5.25	1.63
S10	3.50	5.38	1.88
S11	5.25	5.38	0.13
S12	3.38	4.25	0.88

As shown in the Table 4.16, students in the experiment group demonstrated varying degrees of change between the pre-test and post-test speaking scores. While several students displayed noticeable increases in their post-test scores, others showed smaller gains, and a limited number of students experienced slight gains. The size of the scores differed from one student to another, which implies that improvement in speaking was not the same for everyone.

4.3.1.2. Experiment Group Students' Individual Descriptive Statistics of Ironsi's (2023) Perception Scale

This subsection presents the individual descriptive statistics of the experiment group students' (n=12) scores on Ironsi's (2023) perception scale. Mean scores and standard deviations are reported for each student to illustrate individual variation in perceptions of the effect of VR on their speaking performance.

Table 4.17: Individual Students' Mean Score and Standard Deviation of Ironsi's (2023) Perception Scale

Student	M	SD
S01	3.46	1.13
S02	4.38	1.33
S03	4	1.08
S04	3.92	0.49
S05	3.77	0.60
S06	3.92	0.28
S07	4.08	0.76
S08	4.54	0.88
S09	4.08	1.44
S10	3.77	0.73
S11	3.77	0.60
S12	4.85	0.38

Table 4.17 presents each student's mean score and standard deviation of Ironsi's (2023) perception scale. For each student, the mean score reflects the overall level of perceived impact of VR on speaking performance, while the standard deviation indicates the variability in their responses across the scale items. As shown in Table 4.17, students' mean perception scores varied across individuals and differences were also observed in the degree of the consistency of their responses. These results illustrate individual variation in students' perceptions of the effect of VR on their speaking performance within the experiment group.

4.3.1.3. Comparison Of Students' Perceptions and Speaking Performance Outcomes

Below Table 4.18 presents the relationship between students' perceptions of the impact of VR on their speaking performance and their actual speaking performance outcomes. The students were grouped according to their level of speaking performance change and corresponding perception levels.

Table 4.18: Alignment Between Students' Speaking Performance Change and Perception Levels of VR-Based Speaking Instruction

Students	Number	Speaking Performance change	Perception level
S01	1	Low improvement	Moderate
S02; S08; S11	3	Low improvement	High
S05; S07; S12	3	Moderate improvement	High
S03; S06; S09; S10	4	High improvement	High
S04	1	Moderate decrease	High

The meaning of the mean of pre- and post-test scores, differences were interpreted in relation to IELTS speaking band movement, in which differences of levels are based on 0.5 differences. For the perception scale, Alkharusi's (2022) class intervals were used. The distribution of students across speaking performance change and perception levels show high perceptions of VR were reported across all levels except one. Three students demonstrated low improvement with high perceptions, while one student in this group reported a moderate perception. Three students displayed moderate improvement and all of them reported high perceptions. In addition, four students demonstrated high improvement in speaking performance and also reported high perceptions. However, a significant finding to include here is that one student experienced a moderate decrease in speaking performance while reporting high perceptions of VR. S04 received mean scores of 5.25 and 5.75 in the pre-test from Rater 1 and Rater 2 (respectively). In the post-test, these mean scores changed from 5.50 for Rater 1 and 4.50 for Rater 2. Even though the overall score of Rater 1 showed a 0.25 increase, there were lower ratings on certain speaking criteria in the post-test for both raters when compared to the pre-test. This implies that the decline in S04's speaking performance is unlikely to be due to scoring inconsistency. Overall, Table 4.18 indicates that although most students expressed high perceptions of VR for learning and speaking English, their speaking performance outcomes varied, which ranged from a decrease in performance to high improvement in performance.

4.3.2. Qualitative Results from Semi-Structured Questionnaires: Students' Descriptions of Their Speaking Performance

While both sections 4.2.2 and 4.2.3 draw on the same thematic coding of the qualitative data, the analyses are different in terms of their focus. Section 4.2.2 examines students' perceptions of VR as a learning tool through a holistic perspective, whereas section 4.3.2 focuses specifically on students' perceptions of VR on their speaking performance.

The qualitative findings related to students' speaking performance indicate that learners predominantly perceived VR as supportive of the development of their speaking skills. As summarized in Table 4.15, the most frequently reported theme was learning outcomes especially as it had the highest category reported, which was perceived development of speaking skills (n=142). The students stated their perceived improvements across multiple speaking subskills, which included pronunciation, fluency, grammatical accuracy, lexical use, and naturalness of speech. Additionally, a small number of students articulated an overall sense of improved speaking performance and increased accuracy that gradually occurred over time. Nevertheless, some students voiced limited or uncertain learning gains especially in relation to grammar, vocabulary and fluency. This reflects a variation of learners' perceptions on the use of VR for learning and speaking.

Aside from describing learning outcomes, students also conveyed affective factors in speaking that incorporated their reduced anxiety while speaking during the use of VR and increased confidence in speaking both during and after the use of VR. These affective changes were frequently linked to opportunities for repeated practice, individualized interaction, and engagement in a low-stress and non-judgmental virtual environment. All in all, these findings indicate that students perceived the impact of VR on their speaking performance as positive since it was reported by them to involve both linguistic development and affective factors; however, their perceptions of learning gains did indeed highlight individual variation.

4.4. SUMMARY OF RESULTS

In conclusion, the findings of this chapter present both the measured effect of the use of VR for learning and speaking and how students described their experiences of learning and speaking during this intervention. The quantitative results of the pre- and post-test scores illustrated that the experiment group had a better speaking performance than the control group as statistical differences were present especially for the speaking subskills of fluency and coherence, and pronunciation. Ironsi's (2023) perception scale results revealed a generally high level of agreement of students for the use of VR for learning and speaking, especially in terms of effectiveness, usability, authenticity, engagement and efficiency. The correlation analysis' results of this scale also suggested positive perceptions across educational, learning, and affective dimensions as they were observed to be closely interconnected. The qualitative findings from the semi-structured questionnaire complemented these quantitative results as students highlight chiefly their positive experiences during the VR intervention due to factors such as perceived development of speaking skills, increased confidence in speaking, reduced speaking anxiety, and a higher engagement in the learning provided by VR. Nevertheless, a small number of reports included differing viewpoints pertaining to learning gains and practical constraints.

5. DISCUSSION

This study examines the effect of VR on EFL students' speaking performance along with their perceptions toward using VR for learning and speaking. It specifically explores whether LiVR speaking practice leads to improvement in overall speaking performance, which was assessed through pre- and post-test speaking assessments. The study also investigates students' views on the effectiveness, usability, learning outcomes and emotional aspects of their VR experience. Furthermore, it inspects learners' personal experiences of using VR for speaking tasks, which was especially in relation to engagement, confidence, anxiety and perceived authenticity of the VR learning environment.

To develop an in-depth and evidence-based understanding of these issues, this study adopts an explanatory sequential mixed-methods design that integrates quantitative and qualitative data sources. The quantitative component consists of students' speaking performance scores obtained from the pre- and post-tests, as well as their responses to Ironsi's (2023) perception scale. The qualitative component draws on data collected through a semi-structured questionnaire, which provides richer insight into learners' experiences by including their perceived advantages and challenges associated with VR use. This chapter addresses the research questions by presenting the relevant quantitative and/or qualitative findings and offers an interpretive discussion of these results by examining the ways in which VR use shaped students' speaking development and overall learning experiences.

5.1. DISCUSSION OF FINDINGS IN RELATION TO RQ I: WHAT IS THE EFFECT OF VIRTUAL REALITY ON STUDENTS' SPEAKING PERFORMANCE?

5.1.1. Pre-test Comparisons Between Groups

5.1.1.1. Discussion of the Descriptive Statistics for the Pre-Test

The descriptive statistics provide an important overview of how students' overall speaking performance changed over time in both groups. At the pre-test stage, the experiment group demonstrated higher speaking performance than the control group, which indicates a clear difference in initial oral proficiency, especially as the difference between the two groups was approximately one IELTS speaking band. This indicates a meaningful baseline advantage for the experiment group in terms of overall communicative ability. Thus, the post-test group differences need to be interpreted with reference to this baseline difference.

5.1.1.2. Discussion of the Normality of the Pre-Test Scores

Before interpreting the effect of the VR intervention on students' speaking performance, it is important to consider how the pre-test scores were distributed across the participants in both groups since they provide insight into the comparability and variability of the two groups. The Shapiro-Wilk test indicated that the majority of pre-test speaking performance measures were normally distributed across both groups and rater criteria. This overall pattern suggests that the students in the experiment and control groups demonstrated relatively stable and comparable distributional characteristics for certain speaking subskills prior to the intervention. Nonetheless, a small number of pre-test measures deviated from normality, especially for isolated subskills of rater-specific scores. These deviations indicate that there is naturally a variability in certain aspects of speaking performance such as pronunciation and lexical resource. Since the students were at the intermediate level, it is normal to expect such differences among them. However, these non-normal distributions were limited and did not show a systematic difference between the two groups. Therefore, the pre-test normality results suggest that any differences seen in the post-test are unlikely to be due to the irregularities in the pre-test data.

5.1.1.3. Discussion of the Overall of the Pre-Test

The independent sample t-test results indicate that the experiment group obtained higher overall speaking scores than the control group at the pre-test stage. The fact that there is a statistically meaningful difference of the overall scores at the pre-test stage suggests that the experiment group entered the study with a higher level of oral proficiency. This is an important contextual factor when interpreting the post-test outcomes as also expressed above.

5.1.1.4. Discussion of Pre-Test Results of Speaking Subskills as Rated by Two Raters

The combined results of the independent samples t-tests and Mann-Whitney U tests indicate that the experiment and control groups were broadly comparable across speaking subskills at the pre-test stage, while also revealing specific baseline differences that depended on the rater and analytic approach. Both parametric and non-parametric analyses showed no statistically significant differences between the groups in grammatical range and accuracy and to a large extent fluency and coherence. This suggests that the groups entered the study with similar abilities in these core areas of speaking. However, in particular, the Mann-Whitney U test findings highlighted differences in pronunciation as rated by Rater 1, and in overall speaking performance as rated by Rater 2. Whereas the t-test results identified a significant difference in lexical resource as rated by Rater 2. When considered together, these findings suggest that prior to the VR intervention, the experiment group may have possessed selective linguistic advantages especially in relation to vocabulary use and pronunciation instead of having a superiority across all speaking subskills. The fact that these differences emerged based on rater and test related manner, this implies the multidimensional nature of speaking assessment and the sensitivity of different raters to particular aspects of oral production. Thus, while the overall pattern supports baseline comparability between the groups, the presence of specific pre-test differences should be taken into consideration when interpreting the post-test subskill development and evaluating the contribution of the use of VR for learning and speaking English.

5.1.2. Post-Test Comparison Between Groups

5.1.2.1. Discussion of Descriptive Statistics of the Post-Test Speaking Scores

Following the intervention period, both groups displayed improvement in their speaking scores, which suggests that progress occurred in both groups over the instructional period. In fact, the mean difference between the two groups slightly decreased, but was still close to one IELTS

speaking band. This suggests that the gap did not widen between the two groups following the intervention. Instead, the slightly smaller post-test difference may indicate that the control group may have partially closed the oral proficiency gap, despite the experiment group's continued advantage. Therefore, as established in the pre-test discussion, post-test results should be interpreted in relation to the initial oral proficiency differences rather than post-test means alone.

An important distinction emerged in terms of performance consistency when within group change across the pre- and post-test stages was examined. The experiment group showed a noticeable reduction in score variability between the pre-test and post-test. Whereas, the control group's improvement was accompanied by only a slight reduction in score variability between the pre-test and post-test. This pattern may indicate that despite both groups showing similar levels of development (approximately one IELTS speaking band development in each group), the experiment group's development may have been more stable and evenly distributed among students. At the same time, these post-test outcomes need to be interpreted in relation to their starting oral proficiency levels as the experiment group demonstrated higher speaking performance at the pre-test stage. Thus, the descriptive findings suggest that changes in speaking performance were not only evident in terms of improvement but also in terms of how consistently students' speaking skills developed.

5.1.2.2. Discussion of the Normality of the Post-Test Scores

When the post-test Shapiro-Wilk results were examined, it was found that there was noticeable change in how the speaking performance scores were distributed, especially within the experiment group. In contrast to the largely normal distributions that were observed in the pre-test, a substantial number of post-test speaking measures were found to deviate from normality in the experiment group across both raters. This pattern suggests that after the VR-enhanced speaking instruction, students in the experiment group may have reflected greater variability and differentiation in their speaking performance. The findings further indicate that learners' progress was not uniform as their scores varied across speaking subskills, which produced a less symmetrical distribution of scores. This variability is pedagogically meaningful because it could indicate individualized learning trajectories, increased willingness to experiment with language use, and differing levels of engagement with VR-based speaking tasks. In contrast, the control

group's post-test scores largely remained as normal distributions across most speaking measures, which indicated a more stable and homogeneous pattern of performance development. This difference between the two groups further supports the interpretation that the VR intervention could have shaped how students developed their speaking skills and not just how much their overall mean scores had increased. It is essential to mention that despite the non-normal distributions in several post-test measures, the overall averages of the post-test for both raters in the control group as well as the overall average of the post-test of the experiment group for Rater 1 followed a normal distribution. Even though some speaking subskills showed non-normal distribution after the intervention, students' overall speaking performance remained stable and evenly distributed when scores were averaged, especially for the control group, and the experiment group as rated by Rater 1. Thus, the use of non-parametric Mann-Whitney U test when normality was violated strengthened the validity of the group comparisons.

Overall, the normality findings of the post-test complement the pre-test results by revealing that the experiment and control groups were broadly comparable prior to the intervention. Therefore, the differences observed after the intervention in the post-test normality findings highlight that changes in the distribution of speaking performance scores may have occurred during the VR-enhanced speaking practice period. However, these distributional changes are interpreted descriptively rather than causally and should be understood as indicative of variability in development. In other words, they should not be interpreted as direct evidence of instructional impact. In brief, these differences in the distribution of scores may suggest that VR-enhanced instruction could have influenced not only the learners' speaking outcomes but also how their speaking skills developed.

5.1.2.3. Discussion of the Overall of the Post-Test

The independent sample t-test results indicate that the results of the overall scores of the post-test were consistent with the baseline difference; the experiment group continued to outperform the control group at the post-test stage. In other words, there was a statistically significant difference at the post-test stage just like the pre-test stage. Nonetheless, this pattern does not imply that the observed post-test difference was solely the result of the VR intervention, which was also mentioned above. Thus, the higher post-test scores should be interpreted in the

light of the experiment group's stronger starting point. For this reason, the t-tests depict differences between the groups at both time points, but they should be considered together with their initial oral proficiency levels and how their scores changed over time within the group, which was reported above in subsection 5.1.2.1. As a result, these results show the importance of looking at improvement over time and using qualitative findings when evaluating the role of VR-enhanced speaking practice in students' speaking development.

5.1.2.4. Discussion of Post-Test Results of Speaking Subskills as Rated by Two Raters

The combined post-test results from the independent samples t-test and the Mann-Whitney U test indicate that clear group differences emerged after the VR intervention period across several dimensions of speaking performance, which generally favored the experiment group. The t-test results revealed a statistically significant difference in overall speaking performance as rated by Rater 1, which suggests that through a holistic lens, students who experienced learning and speaking through the use of VR achieved higher post scores than the control group for this rater. This finding points to a meaningful distinction in overall speaking outcomes between the groups at the end of the intervention period.

The Mann-Whitney U test results further extend this pattern by showing statistically significant post-test differences across multiple speaking measures, especially in fluency and coherence, pronunciation and the overall average of speaking scores for both raters, grammatical range and accuracy for Rater 1 and lexical resource for Rater 2. For both raters, fluency and coherence, and pronunciation consistently showed significant differences, which suggests that these subskills may have been especially affected by the VR-enhanced speaking practice. These aspects of speaking are closely related to sustained oral production, repeated practice and increased opportunities for interaction, which are core features of the VR environment used in this study. Nonetheless, not all subskills demonstrated a uniform improvement across raters. Lexical resource rated by Rater 1 and grammatical range and accuracy rated by Rater 2 did not reach statistical significance, which indicates that post-intervention development was selective rather than evenly distributed across all components of speaking.

In short, the similarities between the parametric and non-parametric findings suggest that speaking practice may have supported differential development in specific speaking subskills, while also contributing to higher overall speaking performance. However, as particularly

highlighted in the discussion of this section, these post-test differences should be interpreted in relation to baseline proficiency and rater-specific sensitivity instead of being the direct evidence of the impact of VR. Thus, the results highlight the importance of examining both the overall performance and qualitative findings when evaluating the contribution of VR on learning and speaking English.

In short, to answer RQ i, the findings suggest that VR-enhanced instruction assisted in shaping students' speaking performance, especially in terms of fluency and coherence, and pronunciation and how their speaking skills developed over time. Nevertheless, the impact of VR needs to be understood within the broader learning context as it works together with students' starting levels and regular instruction rather than being the only reason for improvement.

These findings align with those in Raman et al.'s (2024) study as they too reported observed notable gains in learners' fluency and accuracy after eight weeks of using ImmerseMe; however, they also reported substantial individual variability, which suggests VR-enhanced speaking practice is shaped by learner characteristics and tasks demands rather than guaranteeing a development for all and across all speaking subskills. These results are also similar to Yan et al.'s (2023) which found significant gains in speaking ability with the most pronounced improvements occurring in fluency, vocabulary and grammar. However, the findings in this study presented that improvement was seen more in fluency and coherence, and pronunciation, since the raters did not consistently identify vocabulary and grammar to show improvements.

It is important to note that these findings are different than Akay and Kessler's (2024) who used ImmerseMe as LiVR for over a six-week period to measure their oral and pragmatic abilities found that although the VR group had a higher average than the control group, the difference was not statistically meaningful. However, in this study, the difference between the control group and experiment group was not just higher in average, it was also statistically significant for the overall scores of the pre- and post-tests. It is significant to mention that the baseline differences of certain speaking subskills of the two groups in this study was significantly different, whereas in their study, they might have started with balanced groups. Thus, this study highlights how VR might

function differently for students who already possess baseline differences in certain speaking subskills.

5.2. DISCUSSION OF FINDINGS IN RELATION TO RQ II: WHAT ARE THE PERCEPTIONS OF STUDENTS TOWARD USING VR FOR LEARNING AND SPEAKING ENGLISH?

5.2.1. Discussion of Descriptive Statistics from Ironsi's (2023) Perception Scale

5.2.1.1. Discussion of the Overall Mean of Ironsi's (2023) Perception Scale

The descriptive results of Ironsi's (2023) perception scale indicate that students generally had positive views toward the use of VR for learning and speaking. The overall mean score falling within the high agreement range suggests that participants perceived VR as a beneficial and supportive tool for learning English and developing their speaking skills. The fact that the overall mean of the scale was slightly below the interval midpoint further implies that students' perceptions were positive but also had criticism. In other words, learners also recognized the limitations of VR. As the SD of the overall of this scale had low variability, it meant that students' responses were largely consistent. In other words, they generally perceived VR as effective for learning and speaking English.

5.2.1.2. Discussion of Item-Level Descriptive Results of Ironsi's (2023) Perception Scale

The item-level analysis of students' perceptions provides a more detailed outcome of how VR-enhanced speaking practice was experienced by the students. Overall, learners expressed strongly positive perceptions toward the use of VR for learning and speaking as most items fell within the high agreement range. Especially items related to instructional effectiveness, speaking improvement, classroom performance, and engagement received consistently high scores, which suggested that students perceived VR as a valuable instructional tool that supported both their speaking development and their overall learning experience. High agreement on items that addressed realistic speaking situations, lesson objectives and ease of use were observed, which further indicated that students viewed VR as pedagogically useful and manageable within the speaking lessons. Additionally, the item-level results revealed a more nuanced pattern rather than a consensus. Although item 1, which addressed the effectiveness of VR for foreign language acquisition and fell within the high agreement range, it was comparatively lower in terms of mean score compared to the other high agreement range items. This suggests that students were slightly more cautious in evaluating VR's overall effectiveness when compared to its role in speaking

specific outcomes. While students showed moderate concern about privacy and security, they expressed a greater concern for the cost of VR equipment, which suggests that financial accessibility may be a more important limitation according to the descriptive statistics. Overall, these findings suggest that students largely viewed VR as an effective and engaging tool for speaking; however, they also acknowledged its practical constraints and potential concerns. This balanced perception aligns with the performance findings of the study, which highlights that VR-enhanced instruction was experienced as beneficial even though it has limitations, so it should be evaluated along with both learning outcomes and contextual factors.

The findings reported by Ironsi (2023) align closely with the results of the present study; it was mentioned that students perceived VR-enhanced speaking practice as helpful and enjoyable in terms of engagement and affective support. However, Ironsi (2023) also reported that students expressed some uncertainty about the overall effectiveness of VR, which is also the same in this study. Also, the participants in Ironsi's (2023) study voiced their concerns about the cost of VR, which was acknowledged by a few students but not stressed upon by the participants in this study. In Ironsi's (2023) study, it was mentioned that participants found VR difficult to use as he used HiVR in his study, so this outcome was not the same as this study employed LiVR.

The item-level SD patterns provide insight into how consistently students evaluated different aspects of VR for learning and speaking English. Overall, the low SD values for items related to instructional effectiveness, enjoyment, and perceived learning gains, and achievement of speaking objectives indicate a high degree of consensus among students, which implies that these core pedagogical benefits of VR were widely shared and experienced in a similar manner. In contrast, higher SD values were observed for items indicating broader effectiveness in language-learning, cost-related constraints and privacy and security issues, which suggested that students had differing views among one another. This variability implies that while students largely agreed on VR's value for speaking practice within the educational context of the study, their evaluations became less uniform when considering external or contextual factors that were not part of the immediate use of VR. The presence of moderate to high variability in certain items highlight the role of individual learner perspectives, their prior experience, and considerations of situational

factors that may influence their perceptions of VR. This reinforces the notion that positive instructional experiences exist together with concerns about practical, ethical or institutional constraints.

Overall, the item-level mean scores and the variability patterns indicate that positive evaluations of VR were mostly stable and shared when tied directly to speaking-related experiences, whereas perceptions became more differentiated when broader implementation-related considerations were taken into account.

5.2.2. Discussion of Findings in Relation to RQ ii.i: Are There Any Correlated Items in the Students' Perceptions Toward Using VR for Learning and Speaking?

5.2.2.1. Overall Interpretation of Item-Level Correlation Patterns

The pattern of the Pearson correlation suggests that the scale functions as a largely coherent tool as most items show positive linear relationships with one another. The presence of strong and very strong positive correlations among several item pairs indicates that related aspects of students' perceptions tend to move together, which supports the internal consistency of the scale. At the same time, the existence of weak, very weak and moderate correlations reflect that items do not duplicate one another but instead they capture related yet distinct dimensions of the construct. The occurrence of negative correlations across item pairs further highlights the multidimensional structure of this scale. As negatively worded items were reverse-coded prior to the analysis, these inverse relationships depict meaningful contrast between aligned positive constructs rather than as evidence of conflicting or negative perceptions. Moreover, the range of correlation strengths present in this scale displays a balanced structure in which items are sufficiently related to reflect a common underlying construct while still measuring different aspects of it. Finally, the small number of item pairs portray a no linear relationship, which may indicate certain aspects of learners' perceptions to function more independently. This further supports the argument that the scale represents students' experiences in a detailed but non-repetitive way.

5.2.2.2. Interpretation of Positive Correlation Patterns across Analytic Categories

Instructional effectiveness. The positive relationships surrounding instructional effectiveness suggest that this category acts as a central reference point in students' evaluations of VR-based speaking practice. When VR was perceived as effective, students also tended to see it

as authentic, supportive of learning outcomes, efficient, easy to use, engaging and enjoyable, while practical and privacy-related concerns became less noticeable. Thus, effectiveness appeared to shape their overall impressions by making other aspects of the experience feel more meaningful and coherent. Through this perspective, effectiveness seems to carve students' overall judgments of the VR learning environment by helping students make sense of VR as a well-functioning and trustworthy learning environment rather than a problematic or risky tool.

Instructional value and authenticity. The patterns of association that involve instructional value and authenticity indicate that realism plays an important role in how students interpret their VR experience. When VR felt authentic, students were more likely to view it as educationally valuable rather than it being merely a novel instructional tool. Especially when students perceived it as authentic, they were more likely to view it as effective, helpful for learning, efficient, usable, and engaging, while cost-related concerns were less of a concern. This suggests that perceived realism assisted in legitimizing VR as a serious educational tool for speaking practice.

Learning outcomes. The observed relationships that involve learning outcomes imply that students' perceptions of improvement are closely linked to the quality of the learning environment. When students felt they learned more, they also tended to evaluate VR as effective, authentic, efficient, easy to use and engaging. Therefore, instead of evaluating their progress independently, students appeared to connect their sense of speaking development to how effective, engaging and supportive the VR experience felt in a general sense, which shows how supportive and well-designed the VR environment felt overall.

Learning efficiency. The findings related to learning efficiency suggest that students valued VR not only for its learning benefits but also how efficiently those benefits were achieved. When VR was perceived as being efficient, students were also more likely to perceive it as effective, authentic, supportive of learning outcomes, easy to use and engaging, while privacy and

security concerns were less of a worry. This implies that efficiency helped students feel that VR was worth their time and effort by reducing uncertainty or hesitation.

Usability. The positive associations that involve usability highlight the importance of ease of use that contours the students' experiences. When students found VR easy to use, they were more likely to see it as authentic, effective, efficient, engaging and helpful for learning, while cost-related concerns were less influential. Thus, when technical interaction was smooth, students appeared better able to concentrate on speaking tasks themselves, which enabled instructional and learning benefits to be more fully experienced. This suggests that usability enabled students to focus on speaking practice itself rather than on managing technical difficulties.

Engagement and enjoyment. The strong connections that include engagement and enjoyment suggest that affective involvement played a key role in students' perceptions. When students enjoyed and felt more engaged in VR-enhanced speaking activities, they were more likely to perceive VR as authentic, efficient, usable and beneficial for learning, while practical and privacy concerns became less significant. Thus, engagement and enjoyment appeared to support participation and learning, which reinforces the perceived effectiveness and value of VR-enhanced speaking activities.

Practical constraints. The patterns related to practical constraints indicate that when barriers such as cost were perceived as low, students were more able to focus on instructional and learning benefits. Especially when students perceived cost as not a major barrier, students evaluated VR more favorably across instruction, learning, usability, and engagement dimensions, while privacy concerns were also lower. This suggests that reduced practical concerns allowed the positive aspects of VR to become more prominent in students' evaluations.

Privacy and security concerns. The relationships that encompass privacy and security concerns suggest that feeling safe while using VR supported students' willingness to engage with

the technology. Especially when these concerns were low, students were more likely to see VR as effective, efficient, engaging and less constrained by cost. In other words, when these concerns were minimal, students appeared to be more comfortable while participating in speaking activities, which likely contributed to more positive overall learning experiences.

5.2.2.3. Interpretation of No Linear Correlation Patterns across Analytic Categories

Engagement and enjoyment in relation to practical constraints. The lack of relationship between engagement and enjoyment, and practical constraints suggests that students' emotional involvement in VR-based speaking activities developed independently of cost-related constraints. In other words, feeling interested, motivated or bored during VR speaking practice does not appear to depend on whether students perceive VR as expensive or financially demanding. This indicates that affective engagement is more related to the learning experience itself rather than by external concerns.

Usability in relation to privacy and security concerns. The absence of a relationship between usability and privacy or security concerns indicate that students viewed these dimensions separately. Finding VR easy or difficult to use did not seem to influence how safe or secure students felt when using it. This implies that technical ease and ethical or data-related concerns were processed as distinct aspects of the VR experience rather than as interconnected judgments.

5.2.2.4. Interpretation of Negative Correlation Patterns across Analytic Categories

Instructional effectiveness. The negative relationships linked to instructional effectiveness suggest that this category represents a clear contrast between positive and opposing beliefs. When students experienced VR as instructionally effective, beliefs related to ineffectiveness, boredom, inefficiency, and concerns appeared to fade into the background. This indicates that instructional effectiveness plays a protective role by reducing doubts and concerns once students are convinced that VR genuinely supports learning. Moreover, perceptions of learning outcomes, efficiency, usability, engagement and authenticity strengthened. This indicates

that instructional effectiveness also functions as a stabilizing perception that also enhances positive evaluations.

Instructional value and authenticity. The inverse relationships that involve authenticity suggest that realism assists in the reduction of skepticism. When VR felt authentic, students were less likely to question its instructional effectiveness and value or feel uncertain about its privacy and security concerns. This implies that authenticity may function as a reassurance mechanism by helping students accept VR as a credible and trustworthy learning environment. Furthermore, it increases engagement that assists in constructing VR as a meaningful learning environment by strengthening learning-related perceptions and easing underlying concerns. It is important to mention that both positive and negative correlations were seen between instructional effectiveness and learning outcomes as positively worded items. This pattern suggests that students seemed to judge VR either by results or by experience but not always by both. Thus, it can be interpreted as students having prioritized different criteria when evaluating VR as some students associated authentic and instructionally valuable VR experiences through stronger learning outcomes, others appeared to assess learning outcomes not in relation to the authenticity of the VR experience.

Learning outcomes. The negative correlation patterns associated with learning outcomes suggest that feeling genuine improvement in speaking reshaped students' overall perceptions of VR. When students believed their speaking had improved, they were less likely to see VR as ineffective or boring, less concerned about cost and privacy issues, and more tended to perceive the instruction as authentic. This indicates that when students feel they are learning more, it helps offset both emotional disengagement and practical concerns.

Learning efficiency. The inverse relationships that involve learning efficiency suggest that efficiency reduces resistance toward VR-based instruction. When students felt VR supported efficient learning, beliefs related to ineffectiveness and boredom weakened, and cost-related concerns became less of a concern. This implies that when VR feels efficient, students are more willing to accept the effort and resources it requires, so limitations seem less important.

Usability. The negative correlations linked to usability suggest that ease of use plays a key role in preventing disengagement and doubt. When VR was perceived as easy to use, boredom was less likely to occur and beliefs questioning instructional effectiveness weakened. This indicates that usability supports sustained engagement by reducing frustration and allowing students to focus on learning rather than technical challenges.

Engagement and enjoyment. The inverse pattern that surrounds engagement and enjoyment suggests that affective involvement helps reduce negative views about instruction and learning. When students felt engaged and enjoyed VR-enhanced speaking activities, boredom declined while perceptions of effectiveness, learning outcomes, efficiency, usability and authenticity strengthened. This indicates that engagement and enjoyment not only enhance positive experiences but also suppress opposing beliefs and disengagement.

Practical constraints. The negative relationships associated with practical constraints suggest that perceived affordability shapes learners' confidence in instructional value. When cost-related concerns were lower, students were less likely to question instructional effectiveness or learning outcomes and more likely to perceive learning as efficient. This implies that reduced practical barriers allow instructional and learning benefits to become more significant in students' evaluations.

Privacy and security concerns. The inverse correlation patterns linked to privacy and security concerns suggest that psychological safety supports positive engagement with VR. When students were less worried about privacy and security, they were more likely to experience VR as effective, authentic, engaging, and supportive of learning outcomes. This indicates that reduced security concerns assist in creating conditions in which instructional and learning benefits can be more fully realized.

Overall, the correlation analysis demonstrates that students' perceptions toward using VR for learning and speaking were meaningfully interrelated rather than isolated. Across the analytic categories, many perceptions tended to move together, which indicates that how students evaluated one aspect of the VR experience was closely connected to how they evaluated others. Instructional, learning, usability, and affective dimensions formed a coherent connection in which positive evaluations reinforced one another, while opposing beliefs, boredom and practical or privacy-related concerns tended to weaken as students' confidence in VR increased. At the same time, the presence of negative correlations indicated inverse relationships between aligned positive constructs instead of conflicting perceptions. In addition, no linear correlation suggests that students did not evaluate VR in a simplistic or uniform way because they distinguished experiential, technical, and ethical dimensions when forming their judgments. In short, these patterns indicate that students' perceptions of VR-enhanced speaking instruction were multidimensional yet internally coherent as they reflect an integrated learning experience, in which instructional quality, engagement, learning benefits, usability, and perceived constraints were intertwined.

A key limitation of the correlation patterns lies in the small number of participants included in the analysis. Because the Pearson correlation coefficients were calculated based on data from only 12 students, the strength and stability of the observed relationships should be interpreted with caution. With such a limited sample size, correlation estimates are more sensitive to individual response patterns, which means that a small number of extreme or unusual responses may have had a disproportionate influence on the results. In addition, the limited sample reduces statistical power by making it difficult to draw firm conclusions about the generalizability of the correlation patterns beyond the present group. Consequently, while the correlations provide useful insights into how students' perceptions may be related within this specific context, they should be viewed as exploratory rather than definitive.

5.2.3. Interpretation of Thematic Analysis in Relation to RQ ii.ii: How Do Students Describe Their Perception toward Using VR for Learning and Speaking?

5.2.3.1. Instructional Effectiveness

Students mostly viewed VR as a valuable educational tool rather than it being just a novel technology. In general, the students' responses to the semi-structured questionnaire revealed an overall sense of satisfaction with the VR experience for learning and speaking English. It also included their optimistic expectations of this educational tool in relation to its future relevance, which implied that VR was seen as a sustainable and valuable component of language education.

Furthermore, students' endorsement of VR was in relation to its perceived usefulness in facilitating smoother lesson flow, providing a practical and comfortable learning environment and supporting efficient learning processes. Moreover, learners stressed VR's contribution to overall language development by highlighting its role to be an alternative to traditional learning methods that allowed for corrective feedback and strategic learning methods. The engagement-oriented aspect of VR such as providing enjoyable and interactive experience, further strengthened learners' positive perceptions. Thus, naturally affective support emerged as a complementary benefit of this type of instruction. The low-stress, non-judgmental nature of VR environments appeared to encourage trial-and-error learning and enhance speaking confidence.

In addition, the comparison of the traditional and the VR instruction method revealed that students perceived distinct advantages of VR, particularly in terms of flexibility, learners having control, and the integration of form-focused feedback within communicative speaking activities. However, the presence of a single differing view shows that instructional effectiveness was not perceived as the same in terms of preference of traditional and the VR instruction method as one student stated the preference for traditional classroom interaction, which suggest that perceptions of effectiveness in determining traditional or VR based learning may be influenced by perceived authenticity of interpersonal cues, such as facial expressions and real-time human interaction. Therefore, while nearly all of the students, this exception highlights individual differences in learning preferences, expectations and how they each respond to the use of technology in education. This aligns with the findings in Soto-Lucena (2024) who reported a variation in learners' views pertaining to the use of ImmerseMe, which implied that this difference could have been caused by learners' individual characteristics.

When seen through a holistic lens, instructional effectiveness of VR is in regard to its capability of integrating pedagogical, affective and strategic dimensions of learning, which positions it as a medium of supportive instruction that develops speaking skills when it is thoughtfully embedded within appropriately contextualized language learning contexts. Nonetheless, its perceived value may still depend on how closely the learning environment aligns with learners' notions of authentic communication.

5.2.3.2. Instructional Value and Authenticity

Students viewed the use of VR for learning and speaking to be a highly contextualized and realistic learning environment that supported the use of meaningful language. Even though it was form-focused, unlike the traditional decontextualized speaking drills, this environment was seen as valuable since it facilitated exposure to natural speech models and authentic pronunciation patterns in a contextualized manner in real-life-like settings. This enabled students to engage with spoken language in ways that resembled communication in the real world. The learners' reports of realistic conversational exchanges and contextually appropriate interactions further demonstrated that VR simulations can assist in bridging the gap between classroom speaking practice and real-life communicative situations. The heightened sense of immersion and presence stated by the learners appears to have reinforced this authenticity as it allowed learners to experience speaking as a situated action instead of it being merely an isolated language production. Moreover, this study partially aligns with Loughlin and Mentzer's (2025) as they also mentioned Immerse for fostering a sense of presence and confidence in pronunciation. However, they used HiVR in this study. Thus, the reporting of presence was labelled as strong, whereas in this study as LiVR was used, learners didn't state the intensity of this presence as being as high as HiVR. This was especially unexpected to have observed as LiVR was used in this study instead of HiVR, yet learners still mentioned immersion and presence by being felt as if they were there. This suggests that LiVR can still provide immersion and presence even though the intensity of it can be debatable when compared to HiVR. At a broader perspective, these findings align with the systematic review conducted by Mallek et al. (2024) who found that learners reported positive perceptions of flexible and immersive nature of VR-enhanced speaking practice. However, they had also reported personalized practice through VR in their review, which was not reported in this study. Once again, this systematic review assists in the contextualization of these findings instead of direct comparison with the present results.

Moreover, experiential and intercultural learning emerged as an important dimension of this theme as learners expressed that they had the opportunity to encounter culturally situated scenarios and interact with native-like speaking in meaningful contexts. This implies that the perceived authenticity of VR for speaking practice may enhance learners' engagement and willingness to communicate by providing purposeful speaking activities that have a cultural basis for the communicative situation. In brief, these findings display that the instructional value of VR is related to technological immersion along with its capacity to create rich speaking environments that are authentic, experiential, and cultural to support the development of communicative competence.

5.2.3.3. Learning Outcomes

Learners viewed VR as especially beneficial for the development of their speaking skills since they most frequently stated improvements in pronunciation, fluency, vocabulary use, grammatical accuracy and the overall naturalness of their speech. These perceptions suggest that VR may support the holistic development of speaking skills as it enables repeated and contextualized practice that allows learners to refine form and meaning in time. The reports of improved speaking performance and perceived gains in accuracy further imply that learners experienced a sense of progression themselves during the VR intervention. This may have contributed to positive outcomes of self-evaluation and increased satisfaction with their speaking ability. These findings align with Akay and Kessler (2024) as they found that students generally valued the VR experience as the majority of the students perceived ImmerseMe to be beneficial for their oral and pragmatic abilities despite facing difficulties.

Nonetheless, there have been accounts of a small number of students that have mentioned limited and uncertain learning gains on speaking skills through the use of VR. This highlights that not all the students viewed VR's effect to be beneficial when compared with the rest. In particular, uncertainty or limited improvement in fluency, grammar and fluency indicates that some learners may require more explicit scaffolding or longer exposure to VR-enhanced practice in order to perceive tangible outcomes in relation to these speaking subskills. As pronunciation was exempt from the category of uncertain or limited outcomes, this shows that all students agreed that VR was most beneficial in supporting pronunciation. Nevertheless, this variation in students' perceptions in relation to learning outcomes related to the effect of VR suggests that there is a need

of interpreting the perceived learning outcomes as developmental and dependent on the learners' judgments instead of as immediate or guaranteed effect of the use of VR.

In short, as this theme had the highest number of reports (most of which was positive), it shows that VR has strong potential to support speaking development; however, its effectiveness is up to the evaluations of each individual learner and the extent to which its tasks are pedagogically aligned with targeting speaking subskills.

5.2.3.4. Learning Efficiency

Students mentioned that the use of VR for learning and speaking required low-effort and was time-efficient. Thus, it facilitated sustained engagement without cognitive overload or emotional strain. Learners have mentioned that being able to repeat speaking tasks without fatigue assisted having more frequent practice. Moreover, the reduced perceived difficulty of the VR environment may have been helpful for learners to pay attention to their spoken production instead of trying to manage the task or deal with external pressures such as judgments, which might have enhanced learning efficiency. In addition, the combination of efficiency and enjoyment appears to have strengthened learners' willingness to engage in repeated speaking practice. These findings align with Yan et al. (2023) who reported that VR assisted learners with comprehending speaking topics more easily, recalling vocabulary and sentences more effectively while remaining focused and engaged in the immersive environment, and developing stronger motivation to improve their speaking ability. This indicates that affective and cognitive dimensions of learning efficiency were intertwined. Having easy access to speaking and being able to repeat whichever activities learners wanted highlighted VR's potential to remove logistic barriers that are commonly associated with speaking practice such as limited classroom time or peer availability. These findings are in alignment with Soto-Lucena (2024) who expressed that many participants perceived ImmerseMe as a supportive tool that was useful to compensate for limited real-world speaking opportunities. When considered altogether, these findings imply that VR may be an efficient tool for supplementary support that allows learners to increase their opportunities for speaking practice without feeling tired.

5.2.3.5. Usability

The learners have expressed clearly that using VR was generally manageable and functional. This was important to note as technical factors are significant in shaping learners' instructional experiences since if learners experience many technical issues, this disrupts the flow

of their learning and negatively influences engagement. Thus, it was not a major barrier in learners' participation in speaking activities. A number of students reported technical challenges pertaining to issues related to hardware and speech recognition such as microphone problems or the system's failure to accurately detect speech, which prevented smooth interaction during the speaking activities. Since these technical issues affected the continuous practice while learning, it also potentially caused learners to divert their attention away from communicative goals to problem-solving these technical issues. These findings align with Soto-Lucena's (2024) findings which state that participants also reported frustrations with automatic speech recognition. In addition, these findings partially align with Loughlin and Mentzer's (2025) who used HiVR and have mentioned technical challenges and issues related to having limited independent actions, which were also difficulties some students voiced in this research.

However, there was a notable number of students who mentioned that they didn't experience any technical difficulties while using VR. This variation implies that technical challenges may depend on contextual factors such as the stability of the system, quality of the devices used, and user familiarity. This pattern denotes that the effectiveness of VR depends partly on technical reliability and not on flaws that can be found in the VR approach itself. Nonetheless, to a certain extent, these findings do highlight usability concerns that may influence learners' perceptions of task authenticity and immersion, which may be even due to minor usability issues. That's why being able to ensure technical reliability and providing sufficient support are essential to maximize the benefits of VR by preventing technical issues from interfering with learners' engagement and learning outcomes.

5.2.3.6. Engagement and Enjoyment

The findings pertaining to this theme suggested that speaking practice through the use of VR was significant in the sense that it sustained learners' interest, focus, and motivation throughout the intervention period. Students' accounts indicate that VR environments augmented their attention and engagement, which were attributed to not only its novelty but also the immersive and realistic nature of the simulated speaking contexts. The presence of scenarios that simulated real-life appears to have made speaking practice more meaningful. Thus, this allowed learners to maintain their concentration by VR reducing their feelings of boredom, which is often linked to repetitive speaking practice. This finding partially aligns with Gracia (2024) in terms of affective benefits as it was reported that learners found the VR experience enjoyable and perceived it as

realistic, which were outcomes also mentioned in this study. However, it is important to mention that HiVR was used in Gracia's (2024) study.

Moreover, as VR was able to provide individualized and comfortable learning conditions, learners were able to engage with speaking tasks at their own pace without the pressure of external factors. This sense of autonomy, combined with opportunities for social interaction with peers and within the VR environment during the intervention appears to support both cognitive engagement and emotional involvement in speaking activities.

In regard to motivation, learners expressed that their willingness to practice speaking had boosted during the VR intervention period. This signals that enjoyment through the use of VR encourages sustained engagement of students than them just being impressed by this novel technology. The enjoyment of the students by using VR was further supported through experiencing social interaction, having exposure to a variety of topics, and learning through a non-conventional lesson format. The finding of learners seeing VR not as a traditional lesson aligns with Yan et al. (2023) who stated that participants in their study reported the VR experience as feeling more like play than participation in a formal lesson, which is the same in the qualitative findings of this study. This indicates that the immersive and enjoyable nature of VR sometimes blurred the lines between learning and entertainment. The findings also illustrate that engagement and enjoyment in VR-based instruction are closely intertwined with heightened motivational levels of learners and positive affective outcomes that supports sustained speaking practice.

This underlines the salience of VR in developing speaking skills as it provides insight into not only what learners practice but also how their learning experience in a meaningful virtual learning environment can influence their willingness to engage in speaking over time. These findings align with Restall et al. (2023) who stated students' enjoyment of the use of VR. Also, Yan et al's (2023) observations are aligned closely with the results of the present study, in which learners reported increased understanding, motivation, and engagement. From a review perspective, these findings also align with the systematic review conducted by Mallek et al. (2024) who found that students highlighted increased engagement, enjoyment, and perceived authenticity by using VR. Thus, this systematic review contextualizes these findings rather than being directly parallel to this study's results.

5.2.3.7. Practical Constraints

Students observed the practical constraints of VR to be in relation to its instructional effectiveness as it was shaped by limitations pertaining to context and implementation. Cost-related barriers were related to constraints based on resources. This implies financial accessibility to be a concern especially in institutional contexts with limited technological resources. From a broader perspective, these findings align with those in the systematic review conducted by Mallek et al. (2024) also mentioned challenges regarding instructional design, accessibility, high cost in addition to other difficulties. This aids in contextualizing the findings in this study.

Furthermore, instructional contexts highlighted learners' sensitivity to design and pedagogy related elements of VR environments. Students' reflections show a need for interactive and collaborative opportunities, greater use of gamification, enhanced authenticity of speaking tasks and more adaptive feedback mechanisms. In regard to the need for more enhanced authenticity of speaking tasks could be the argument which was previously brought up in the theme of instructional value and authenticity as it could be in reference to the difference of LiVR and HiVR in presenting authentic speaking tasks. This depicts that learners evaluated the use of VR not only as a technological tool but through a pedagogical perspective as they compared what it could do and what they needed to improve their learning.

The pedagogical constraints further reinforced the salience of scaffolding, having a diversity of language input and variation in task and content design. These findings align with Soto-Lucena (2024) who expressed that the participants in her study expressed mixed feelings about grammar-focused and repetitive activities. This implies that insufficient pedagogical support may limit learners' ability to fully benefit from using VR for learning and speaking. However, there were also reports of learners who stated no suggestions for improvement, which itself suggests that not all the students perceived constraints the same way and it may in fact be dependent on individual expectations and prior learning experiences. Overall, these findings may reflect that VR is effective in speaking instruction when it is grounded in sound pedagogy and institutional support.

5.2.3.8. Affective Factors in Speaking

Students voiced this theme to be one with a substantial impact on them during the VR intervention as it carved their emotional experiences that pertained to speaking confidence and anxiety regulation. The sub-themes of speaking confidence and speaking anxiety were closely

interrelated and functioned in a complementary manner such that increases in speaking confidence were accompanied by corresponding reductions in speaking anxiety.

A large number of students mentioned increased self-confidence in speaking along with greater comfort. This indicates that VR environments may create conditions that support positive affective outcomes that also support the development of speaking skills. Moreover, learners frequently attributed the increase of the confidence in speaking to repeated practice opportunities, individualized interaction, transfer of what they have learned to real-life and engagement in realistic simulations. This appears to have reduced the psychological pressure that learners face in the classrooms while speaking in front of others. Thus, the non-judgmental and low-stress nature of VR was also stated by the learners in reducing their fear of making mistakes and experiencing negative evaluations of other people. This seems to have lowered their speaking anxiety for many of them. As many have articulated that speaking with virtual interlocutors is what reduced their anxiety and so in turn increased their confidence, it can be attributed to the anonymity that VR provides as the students' identities remain hidden and so they are shielded from the negative evaluations of others regardless of who they may be. By removing the pressure of being visibly evaluated by others in an immediate environment, VR seems to offer a level of safety in terms of psychological protection that is difficult to achieve in face-to-face contexts. There is a fact that pertains to real-life interactions in regard to not being able to avoid being influenced by others reactions, which makes real-life environments to be seen as more threatening to the sense of self. In contrast, as this isn't an issue in virtual environments, learners are able to take risks, experiment with language and engage more persistently in speaking activities. Thus, virtual environments are especially beneficial for those learners who struggle with speaking anxiety or have reluctance to speak in the classic classroom setting. This affordance that is provided by VR makes it invaluable in terms of offering a supportive and safe learning environment especially in comparison to the conventional classroom environment.

However, there were students that articulated neutral or unchanged confidence in speaking. This was mentioned especially among those students who already felt confident prior to the use of VR, which suggests that affective benefits of VR may be more beneficial for learners who initially experience higher levels of anxiety or lower confidence.

Furthermore, anxiety regulation mechanisms were identified in the students' responses as pertaining to gradual improvement, individualized pacing and interaction, being able to repeat the tasks whenever they wanted, and perceived positive emotional outcomes in general towards the use of VR. This further reinforces the use of VR for supporting speaking development as it may facilitate affective self-regulation since they could practice speaking in a safe environment and repeat the activities. In short, these findings suggest that VR's contribution to speaking development goes beyond improving performance outcomes as it has the potential to function as a learning environment that supports learners affectively by fostering confidence and reducing anxiety, which are critical conditions for sustained engagement in speaking practice.

These affective benefits are consistent with Soto-Lucena's (2024) and Yudintseva's (2024) findings. Soto-Lucena's (2024) stated that learners reported heightened comfort, reduced anxiety, and increased willingness to communicate in the ImmerseMe environment due to realism, low-stress environment and support for pronunciation and lexical development. In addition, Yudintseva (2024) had found that both LiVR and HiVR had led to reduced speaking anxiety and increased speaking confidence, which is also present in the results identified in this study by using LiVR. Furthermore, these findings align with those in Restall et al. (2023) again as they also mentioned that improved confidence and perceived competence were connected to students' enjoyment. This was also reported by the students in this study. These findings also align with Yan et al's (2023) as they reported learners' greater confidence in applying language in realistic contexts through VR.

The absence of privacy and security concerns in the students' responses in the semi-structured questionnaire indicates that this matter was not a prominent issue for students. This means that learners were either mainly unconcerned about privacy-related issues or they did not view them as being directly relevant to their speaking experience in the VR environment.

5.2.4. Integrated Interpretation of Quantitative and Qualitative Findings in Relation to RQ ii.

This subsection integrates the quantitative item-level results and correlation findings with the qualitative thematic analysis results to provide a comprehensive interpretation of students' perceptions toward the use of VR for learning and speaking English.

The comparison of the quantitative and qualitative findings shows a strong alignment and complementarity in the interpretations of the manner students perceive VR for learning and speaking English. Through the quantitative perspective, the correlation analysis depicted that students' perceptions across instructional effectiveness, learning outcomes, authenticity, engagement and enjoyment, usability and learning efficiency were strongly interrelated. In addition, item-level descriptive results indicated that most statements on Ironsi's (2023) perception scale fell within the high agreement range, especially those related to VR's effectiveness in enhancing learning, improving speaking skills and classroom performance, providing realistic speaking settings and making speaking lessons enjoyable and engaging. These patterns suggest that students evaluated VR as a coherent learning experience as their positive judgments in one dimension also tended to strengthen positive evaluations in others. In regard to the negative beliefs such as those pertaining to boredom, inefficiency or concern were weakened as their confidence in VR increased. Through a qualitative perspective, this interconnected structure was also seen in students' responses to the semi-structured questionnaire as they described the interwoven manner of instructional, affective and experiential dimensions of VR as being holistic instead of describing them as separate features. In other words, learners did not discuss effectiveness, engagement, learning or confidence as being independent since they framed VR as an integrated environment that simultaneously supported the development of speaking skills, motivation and feeling emotionally comfortable.

When examining the themes specifically, the quantitative findings highlighted instructional effectiveness and instructional value and authenticity as being central perceptions as students articulated that when VR was seen as effective or realistic, they were more likely to evaluate it positively across the different dimensions of learning outcomes, usability, learning efficiency and engagement and enjoyment. This pattern is supported by high item-level agreements on statements

that indicate that VR enhances learning, improves speaking performance and grades and introduces realistic speaking situations. Moreover, at item level, although the statement on VR's effectiveness for foreign language acquisition displayed high agreement, it was slightly lower in mean when compared to the other high agreement means, which indicates that students were cautious about VR's overall effectiveness than about its impact on speaking skills. This alignment closely with the qualitative findings, in which students repeatedly emphasized VR's value as a realistic tool for learning and speaking English. Learners' views of VR enabling a smoother lesson flow, providing meaningful feedback, learners having the control, being immersive, and including realistic interaction assists in the explanation of students seeing effectiveness and authenticity as stabilizing and legitimizing perceptions in the quantitative data.

Furthermore, the quantitative correlations indicated that learning outcomes, engagement and enjoyment, and learning efficiency were interconnected. This implies that learners' sense of improvement was closely tied to how engaging, efficient and supportive the VR experience felt for them in a general sense. At the item level, strong agreement with statements that indicated that VR makes speaking lessons fun and easy, supports lesson objectives, and enables efficient learning reinforces this relational pattern. Especially as there was a strong agreement with statements that indicated that VR makes speaking lessons fun and easy, supports lesson objectives and enables efficient learning. The qualitative findings deepen this interpretation by displaying the construction of these perceptions as students connected perceived learning gains in pronunciation, fluency, vocabulary, grammar and naturalness in speech to repeated practice, reduced boredom, low cognitive load and tasks that were enjoyable and immersive. Thus, as the quantitative results present relational strength among these constructs, the qualitative findings provide insight into how VR supported the development of speaking skills such as through repetition, comfort and motivation.

In terms of constraints, both types of data presented mixed results. In the quantitative findings, practical constraints and privacy and security concerns tended to weaken as positive perceptions became stronger. This shows that these concerns became less salient when students trusted the instructional effectiveness and value of VR. This pattern is also reflected in comparatively lower item-level mean scores for practical constraints and privacy and security concerns, which fell within the moderate agreement range. Item-level results further indicated that

cost-related concerns were of more importance than privacy and security issues, which suggests that financial accessibility of VR may represent a more meaningful limitation for students. This pattern of answers is only limitedly mentioned in qualitative findings in relation to practical constraints and privacy and security issues were not mentioned at all, even though these concerns were included in the perception scale. This difference indicates that while learners acknowledge these concerns in the structured measurement of the perception scale, they did not play a major role in their actual experiences of using VR for learning and speaking English. In contrast, the qualitative findings emphasized practical constraints in relation to instructional and pedagogical aspects such as task design, feedback and collaboration, which did not appear in the correlations but aided in the identification of concrete areas for the development of VR.

All in all, this comparison portrays that the quantitative findings establish the coherence of the structure of students' perceptions while the qualitative findings shed light on the experiences that are underlying those structures. When taken into consideration together, they indicate that students viewed VR as an effective, engaging, authentic and affectively supportive environment for learning and speaking English. However, they also recognized its impact being dependent on pedagogical design, technical reliability and instructional context. Overall, the alignment of findings strengthens the validity of the interpretations and confirms that students' perceptions of the use of VR for learning and speaking English were multidimensional, coherent, and based on their real learning experiences.

5.3. DISCUSSION OF FINDINGS IN RELATION TO RQ III: IS THERE A DIFFERENCE BETWEEN STUDENTS' PERCEPTIONS OF THE EFFECT OF VIRTUAL REALITY ON THEIR SPEAKING PERFORMANCE AND THEIR ACTUAL SPEAKING PERFORMANCE?

5.3.1. Interpretation of Students' Individual Perceptions of VR on Their Speaking Performance and Their Actual Speaking Performance Outcomes

By comparing the individual pre- and post-test speaking scores of the experiment group and the mean and standard deviation of each students' perceptions on Ironsi's (2023) perception scale certain conclusions were drawn for each student to answer RQ iii.

S01. This student demonstrated low improvement in speaking performance between the pre- and post-test stages, while reporting a moderately positive perception of VR with a high response variability. Also, this student obtained the lowest mean score on this scale in this group. As S01's mean score is above the moderate agreement midpoint, this indicates a moderately positive perception of VR that is closer to the upper end of the moderate range. This suggests that although S01 viewed VR somewhat positively, this perception was not strongly reflected in measurable speaking improvement, which indicates a possible emphasis on experiential or affective benefits rather than performance gains. S01 reported in the semi-structured questionnaire that VR was enjoyable and increased their speaking confidence while using it. These affective experiences are consistent with the qualitative themes of experiencing enjoyment and having an increase in speaking confidence while using VR, which indicates that the perceived benefits of VR for this student may have been due to affective factors rather than learning gains. In the semi-structured questionnaire, this student has also added that VR didn't assist in developing all of the subskills, such as mentioning that VR had a limited impact on vocabulary and highlighted peculiar aspects of the practical constraints of VR in relation to pedagogical limitations such as the lack of comprehensive content. These mixed remarks potentially aid in explaining why S01 had a high variability of their perception score since this student mentioned both the positive and negative aspects.

S02. This student showed low improvement in speaking performance and reported a high mean perception score with a high variability. As S02's mean score is slightly above the high agreement midpoint, this indicates a strong and clearly positive perception of the effect of VR on speaking performance. This alignment suggests that S02's positive perceptions of VR were generally consistent with their observed speaking development even though the improvement was small. S02 reported perceived improvements in fluency and vocabulary in the semi-structured questionnaire due to VR having scenarios to practice speaking in. This experience is consistent with the qualitative themes of learning outcomes especially in relation to fluency development and vocabulary growth through scenario-based practice, which implies that the student experienced VR as a highly beneficial learning environment despite having a low improvement in speaking performance. This means that the students' perceptions suggest that meaningful learning occurred

beyond what the scores could capture. The only potentially negative statement in this student's answers of the semi-structured questionnaire was that they had difficulty in focusing at the start of the VR intervention; however, this on its own isn't enough to clearly explain the high variability of the perception score. Nonetheless, it can imply that the student may have temporarily had reduced engagement that may have resulted in them giving less consistent answers across the perception scale.

S03. This student exhibited a substantial increase in speaking performance along with a high mean perception score with a relatively high variability. As S03's mean score falls within the high agreement range but below the midpoint of that interval, this suggests a positive perception that is strong but not at the highest intensity level. This close alignment indicates that S03's strong perceived impact of VR was also seen in meaningful speaking gains. In the semi-structured questionnaire, S03 mentioned enjoyment that came from the immersive nature of the VR environment along with perceived improvement through having exposure to natural English models and providing realistic speaking practice. The experience is consistent with qualitative themes related to enjoyment that is derived from immersive and situated VR environments and perceived improvement through exposure to natural English models and authentic communicative practice. This suggests that VR may have functioned as both an effective and motivating learning tool for this student. However, S03 had also commented on the practical constraints in regard to VR's cost, and added remarks such as limited or uncertain views on certain speaking subskills such as fluency and vocabulary use. These comments potentially aid in understanding the relatively high variability of their perception score as the student's answers seem to have differing views depending on the specific question at hand.

S04. This student demonstrated a moderate decrease in post-test speaking performance despite reporting a high mean perception with low variability. As S04's mean score is above the high agreement threshold but below the midpoint, this indicates a consistently positive perception of VR with room for stronger agreement across all items. This mismatch suggests that positive perceptions did not translate into measurable speaking improvement for this student, which highlights a divergence between subjective experience and objective performance. S04 expressed

in the semi-structured questionnaire that while the VR experience was enjoyable, contextual and that constraints related to implementation limited its contribution to learning. This response is consistent with qualitative themes that indicate that due to contextual or implementation-related constraints, the VR experience was perceived more as enjoyable rather than as fully realized learning opportunity. It is noteworthy to mention that S04 explicitly stated a preference for face-to-face interaction especially due to seeing real facial expressions that makes learning more effective. This comment suggests that despite positive perceptions of VR, the learner may have experienced perceived limitations in interpersonal authenticity, which could have influenced the depth of engagement and learning transfer.

Also, it is important to mention that the pre-test speaking scores of S04 were characterized by relatively stable performance across the criteria, including one specifically high score that elevated the overall mean. In contrast, the post-test results showed the absence of such high performance and the emergence of lower scores in some criteria, especially Rater 2's ratings. This resulted in a moderate decrease in the overall speaking score. As this pattern was observed across both raters, the decline is unlikely to reflect scoring inconsistency. As VR increased confidence and willingness to speak, the learner may have attempted to use more complex grammar and new vocabulary so this increased risk-taking may have negatively affected criterion-based rating in the short term, particularly for accuracy-oriented dimensions. While students are credited for trying to use a range of grammar and vocabulary, if this student constantly made errors while trying to use new grammar and vocabulary this could have resulted in a lower performance score for the post-test. Also, as this decline might be due to the learner-related factors such as attention, fatigue, test-day anxiety, that might have had an influence on the student that resulted in a decline in their speaking performance score. However, in this case, the moderate decline may reflect situational performance variability rather than a stable reduction in speaking ability.

S05. This student displayed a moderate improvement in speaking performance and reported a high mean perception score with low variability. However, as S05's mean score is just above the lower boundary of the high agreement range, this indicates a moderately strong positive perception as it is positioned near the transition from moderate to high agreement. This pattern suggests a relatively consistent relationship between perceived benefits of VR and actual speaking

development. In the semi-structured questionnaire, S05 reported that VR supported the gradual improvement through repeated, and low-effort speaking practice. This response is consistent with qualitative themes related to learning outcomes and learning efficiency, where VR was perceived as a tool that supported gradual improvement through repeated, low-effort speaking practice rather than immediate performance gains. The low variability of their perception score hints that the student stayed consistent in giving answers to the different dimensions in relation to using VR for learning and speaking English. To put simply, their responses didn't display much fluctuation.

S06. This student depicted a high improvement in speaking performance and reported a high perception mean score with low variability. As S06's mean score falls within the high agreement range but below the midpoint, this suggests a clearly positive and stable perception of VR's contribution to speaking skills and learning but not at the highest intensity level. This alignment indicates that S06's perceptions were both stable and closely reflected in their speaking gains, which suggests a strong correspondence between perceived and actual impact. S06 mentioned that repeated speaking practice in real-life simulations through the use of VR supported learning gains and contributed to holistic speaking development. These experiences align with qualitative themes of instructional effectiveness and learning outcomes, where students described VR learning gains through repeated practice in real-life simulations and that also enabled holistic speaking development, which may explain the substantial gains observed for this student. The low variability of the perception score suggests that the students answered consistently across different dimensions of using VR for learning and speaking English. In other words, their answers didn't fluctuate much.

S07. This student portrayed a moderate improvement in speaking performance along with a high perception mean score with moderate variability. As S07's mean score is within the high agreement range but slightly below the high agreement midpoint, this indicates a strong positive perception but not at the highest intensity level. This suggests that S07 perceived VR as beneficial and this perception was mostly consistent with their observed improvement. In the semi-structured questionnaire, S07 reported that VR supported the gradual development of speaking skills through being comfortable, having an increased motivation due to an increase in confidence while

speaking. This experience is consistent with qualitative themes on affective factors in speaking and engagement and enjoyment, where VR was seen as supporting developing speaking skills in a low-stress environment that enabled the better production of speaking. This student's remarks also included technical challenges in relation to usability due to problems caused by the hardware, and also added having uncertain views of VR for improving certain subskills such as vocabulary and grammar. This might explain the moderate variability of the perception score, as certain challenges might have led to fluctuations in the student's overall evaluation of VR experience.

S08. This student showed a low improvement along with a high perception mean score with moderate variability. As S08's mean score is well above the high agreement midpoint, this indicates a very strong positive perception of VR's effect on speaking skills and learning. This discrepancy suggests that S08's positive evaluation of VR may have been caused more by affective and experiential factors rather than by substantial performance gains. S08 articulated in the semi-structured questionnaire experiencing enjoyment, reduced performance pressure and increased comfort when speaking while using VR. This experience is consistent with the qualitative themes related to enjoyment, reduced performance pressure, and increased speaking comfort in virtual interaction that indicates that emotional ease and confidence may have played a central role in shaping this students' positive perceptions despite limited observable improvement. However, S08 mentioned concerns in relation to practical constraints such as cost-related aspects and pedagogical constraints due to the inadequacy of task and content diversity. These comments potentially help explain the moderate variability of their perception score as their answer could have differed according to the specific dimension in relation to using VR for learning and speaking English.

S09. This student exhibited a high improvement in speaking performance and reported a high perception mean score with high variability. As S09's mean score is slightly below the high agreement midpoint, this indicates a strong positive perception of VR but not at the highest intensity level. This suggests that while S09 experienced meaningful speaking development, their perceptions of VR were less consistent, which possibly reflected mixed experiences across different aspects of VR use. In the semi-structured questionnaire, S09 mentioned uncertainty regarding the extent to which VR supports the development of all speaking subskills and raised

concerns about its broader use. This uncertainty may help explain the high variation of the perception score despite having a strong performance gain. This experience aligns with the qualitative themes that reflect uncertainty or limitations about the development of VR and its widespread use with learners expressing mixed views on whether it supports all subskills of speaking.

S10. This student demonstrated a substantial improvement in speaking performance along with a high perception mean score with moderate variability. As S10's mean score is just above the high agreement threshold, this indicates a positive perception that is closer to the lower boundary of high agreement rather than the midpoint. This pattern suggests that S10 made a meaningful improvement in speaking and perceived VR as beneficial, but also indicates mixed evaluations across different aspects of the VR experience. S10 reported reduced speaking anxiety and the ease of learning in a low-stress VR environment along with perceived gains in pronunciation and vocabulary in the semi-structured questionnaire. Also, S10 noted practical constraints in relation to the cost of VR, and task-related constraints such as reliance on guided prompts that limited opportunities for spontaneous speech production. Despite having a substantial improvement in speaking performance, these instructional constraints may help explain the moderate variability of the perception score despite substantial performance gains since the student experienced clear linguistic support but with restricted communicative freedom. This experience is consistent with qualitative themes related to reduced speaking anxiety, ease of learning in a low-stress environment and perceived gains in pronunciation and vocabulary, and with the qualitative findings that indicated task-related constraints such as guided prompts.

S11. This student depicted a low improvement in speaking performance and reported a high perception mean score with low variability. As S11's mean score falls slightly above the high-agreement threshold, this indicates a moderately strong positive perception of VR. This suggests that while S11 viewed VR as beneficial, this perception was not accompanied by noticeable speaking improvement. In the semi-structured questionnaire, S11 reported increased enjoyment and greater ease of speaking in the VR environment but also identified instructional constraints in relation to pedagogy that might have limited the extent to which speaking performance developed.

This experience is consistent with qualitative themes related to increased enjoyment and perceived ease of speaking in VR as well as instructional constraints in relation to pedagogy. These findings suggest that while this experience was enjoyable for this student, instructional constraints may have limited the extent to which their speaking improved. The low variability of the perception score suggests that the students' responses were consistent with one another across various dimensions of using VR for learning and speaking English, which indicate minimal fluctuation in their evaluations.

S12. This student exhibited a moderate improvement in speaking performance and reported the highest perception mean score with low variability. As S12's mean score is well above the high-agreement midpoint, this represents the strongest positive perception of VR among all students. This suggests that the student experienced noticeable improvement in speaking performance and consistently perceived VR as beneficial, which indicates a strong and coherent perception of VR's contribution to their speaking development. In the semi-structured questionnaire, S12 strongly agreed with VR being effective and enjoyable but also showed awareness of its practical constraints that were related to its implementation. This experience is consistent with the qualitative themes that reflect strong endorsement of VR as both effective and enjoyable but along with constructive awareness of practical constraints. Thus, even though this student has positive perceptions, their moderate improvement in speaking performance could be due to the constraints in relation to the implementation conditions. The low variability of the perception score implies that S12 provided stable and consistent responses across the various dimensions of the use of VR for learning and speaking English.

To conclude, the findings show that students' perceptions of VR's impact on their speaking performance did not always align with their measured learning gains in their speaking skills. This revealed both similarities and differences across individual students. While most students reported high perceptions of VR's effectiveness, the degree of consistency in these perceptions also varied across students and their actual speaking performance scores also differed. Four students demonstrated low improvement despite holding moderate to high perceptions with varying levels of perceptual variability. Three students showed moderate improvement along with high

perceptions that also showed differences in perceptual variability, and four students exhibited both high improvement and high perceptions also with differences in perceptual variability. Nonetheless, four students indicated a closer alignment between perception and performance for these learners. It is highly important to state that one student experienced a moderate decrease in speaking performance while still reporting high perceptions of VR which was accompanied with low variability. This highlights that consistent perceptions of VR may reflect learners' subjective learning experiences rather than short-term measurable improvement in speaking performance. Nonetheless, the students' perception scores displayed varying levels of variability that reflected the consistency of their evaluations of using VR for learning and speaking English. While low variability across several students indicated stable and coherent perceptions of this issue across multiple dimensions, the moderate to high variability, which was observed among students' whose qualitative remarks displayed mixed experience as there were positive and negative evaluations or criticism in regard to the use of VR for learning and speaking English such as those pertaining to instructional effectiveness, pedagogical limitations or practical constraints. Hence, the variability of the students sheds light on the degree to which students' experiences with VR were internally consistent across different dimensions of this experience.

At a broader level, these findings are consistent with Butorova et al. (2024) as they conducted a systematic review that highlighted that while learners consistently reported positive emotional outcomes and perceived learning gains, several studies showed no objective improvement in actual language skills, which was the same for some students in this study. As this is a systematic review, it aids in contextualizing the results in this rather than mirroring its findings. Also, these findings partially align with Akay and Kessler (2024) as they mentioned that students perceived VR as beneficial for their oral and pragmatic abilities, but the mean increase of the experiment group didn't result in a statistical difference, which shows that students' perceptions and noticeable learning gains didn't match in their study as well, which was also the same for some students in this study.

This implies when VR was experienced as effective, authentic and supported learning through repeated practice, these perceptions were reflected in their speaking skills for some. However, strongly positive perceptions were not accompanied by comparable performance

improvement for others, which suggests that VR was often experienced as emotionally supportive instead of immediately being the cause for performance enhancement. In these cases, the qualitative findings highlight increased enjoyment, reduced anxiety, emotional comfort, and ease of participation as key factors that shaped learners' positive perceptions even when speaking development was limited, all of which were connected to anonymity, engaging content, being easy to use are only some among many other reasons. Moreover, practical constraints in relation to pedagogical constraints and design and system level constraints such as task design, limited opportunities for spontaneous production and implementation conditions, emerged as factors that influenced how such constraints affected the learners' learning outcomes. In short, these results indicate that although students generally viewed VR as a valuable and engaging tool for learning and speaking English, the relationship between perception and actual speaking performance was variable and learner dependent as it was shaped by affective support, pedagogical design and opportunities for sustained and autonomous speaking practice.

5.4. SUMMARY OF THE DISCUSSION

This chapter discussed the findings of the study by integrating quantitative and qualitative results to address the three research questions about the effect of VR on EFL students' speaking performance, their perceptions toward using VR for learning and speaking English, and the relationship between perceived and actual speaking development.

In relation to RQ i, the discussion demonstrated that the use of VR for learning and speaking contributed to shaping students' speaking performance, especially in relation to fluency and how students' speaking skills developed over time. While both the experiment and the control groups showed improvement from the pre-test stage to the post-test stage, there were differences in score distributions, variability patterns and selective subskill developments. This implies that VR influenced learners' speaking trajectories as reflected in variability patterns and selective subskill development rather than being the sole factor in determining the overall performance gains. Moreover, these effects were interpreted based on the fact that there were baseline proficiency differences in specific speaking measures, rater perspectives and contextual factors.

Thus, it indicates that VR functioned as a supportive educational medium along with regular classroom instruction instead of replacing it.

With regard to RQ ii, the discussion revealed that students generally held positive perceptions toward the use of VR for learning and speaking English. Quantitative results from the perception scale and correlation analysis indicated that instructional effectiveness, authenticity, learning outcomes, usability, engagement, and learning efficiency were meaningfully interconnected, which formed a coherent perceptual structure. These findings were reinforced by the qualitative thematic analysis, which showed that students experienced VR as an engaging, authentic, effective and affectively supportive learning environment. Furthermore, learners demonstrated a balanced stance by acknowledging practical and pedagogical constraints, which were in relation to cost, task design, and instructional implementation. In brief, these results suggest that students' perceptions of VR were multidimensional, internally consistent and grounded in their actual learning experiences.

In addressing RQ iii, the integrated discussion highlighted that students' perceptions of VR's impact on their speaking performance did not always align with their measured speaking gains. While some learners displayed a close correspondence between positive perceptions and observable improvements, others reported strong and consistent perceptions of VR despite limited or no measurable performance gains, and specifically in one case, there was a moderate decrease in post-test speaking performance of a learner. This variability underscores the role of affective support, confidence, anxiety reduction, and experiential engagement in determining learners' evaluations of VR, even if the short-term performance outcomes did not exactly reflect these perceptions. Thus, VR was often experienced as emotionally and motivationally beneficial with its instructional impact being mediated by individual learner characteristics, pedagogical design and implementation conditions. However, the students' perception scores did display different levels of variability, which reflected their differences in the consistency of their evaluations of using VR for learning and practicing English. Low variability indicated stable perceptions, whereas moderate to high variability depicted fluctuations in students' evaluation due to having both positive and negative evaluation toward using VR for learning and speaking English, which

were related to effectiveness, instructional limitations or practical constraints. These findings indicate that variability of their perceptions scores reflect the internal consistency of their experiences of using VR for learning and speaking English.

In conclusion, the discussion findings position the use of VR for learning and speaking English as a valuable and supportive educational tool that can foster engagement, confidence, and selective speaking development, while also revealing that its effects aren't the same on each learner and that learners' characteristics play an important role in being influenced by this type of instruction. The chapter demonstrates that VR's contribution to speaking development should be comprehended as contextual, developmental, and learner-dependent, rather than being seen as an immediate or guaranteed source that provides performance gains. Moreover, as the sample size of this study is small; 24 students in total, and 12 students each in the experiment and the control group, these discussions and interpretations should be taken into account cautiously. These insights provide a foundation for the concluding chapter, which will outline the pedagogical implications, limitations of the study and directions for future research.

6. CONCLUSION AND RECOMMENDATIONS

This study aimed to investigate the effect of VR on students' speaking performance and to dive deeper into this issue by exploring students' perceptions toward using VR for learning and speaking English. By using an explanatory sequential mixed-methods design, this study integrated the quantitative date of the speaking performance with perceptual evidence of a scale and qualitative findings to provide a comprehensive understanding of how the use of VR for learning and speaking English functions for EFL learners within an English preparatory program.

6.1. SUMMARY AND OVERALL INTERPRETATION OF KEY FINDINGS

The findings in regard to RQ i, which examined the effect of VR on students' speaking performance, revealed that both the experiment and the control groups displayed improved from the pre-test stage to the post-test stage. However, the experiment group consistently obtained higher speaking scores across both tests. Thus, a meaningful baseline difference was identified at the pre-test stage with the experiment group obtaining approximately one IELTS speaking advantage over the control group. Hence, this baseline difference necessitated a cautious interpretation of post-test results and prevented attributing the post-test group differences solely to the VR intervention.

Despite this limitation, post-test analyses revealed clearer group differences across several speaking subskills particularly fluency and coherence, and pronunciation for both raters. These findings suggest that the use of VR for speaking practice may have contributed to more stable and selective development in specific dimensions of speaking rather than it providing a uniform development across all speaking subskills. The reduction in score variability observed within the experiment group further implies that VR may have supported a more consistent speaking

development among learners even though similar overall gains were observed in both groups as each group improved approximately one IELTS speaking band from pre-test to post-test.

In relation to RQ ii, which explored students' perception of using VR for learning and speaking English, the results of Ironsi's (2023) perception scale indicated that students generally had positive opinions toward the use of VR for learning and speaking English. Students perceived VR as effective, engaging, authentic and supportive of learning outcomes, while also acknowledging practical constraints such as cost and, to a lesser extent, privacy and security concerns. The item-level correlation analysis revealed that students' perceptions were multidimension but at the same time they were also internally coherent as categories such as instructional effectiveness, engagement and enjoyment, usability, and perceived learning outcomes were closely interconnected with one another.

The qualitative findings provided a further enriched understanding of these results by illustrating how students experienced VR as a low-stress, immersive and supportive environment that encouraged participation, reduced speaking anxiety and enhanced confidence. Learners frequently emphasized the value of realistic speaking scenarios, opportunities for repeated practice and the ability to use language without fear of negative evaluation. However, individual differences emerged as one student expressed a preference for traditional learning, which especially highlighted the impact of learner characteristics and expectations in defining the VR experience.

With respect to RQ iii, the findings indicate that students' perceptions of VR on their speaking performance did not consistently align with their actual speaking performance score as measured by the pre- and post-tests. While some learners demonstrated a close correspondence between positive perceptions and objective improvements of observable speaking performance, others reported strong perceptions of VR based on its effectiveness for learning and speaking English despite limited, uneven or even negative changes in the objective measurement of their speaking performance. This difference implies that learners' evaluations of VR were strongly

influenced by affective and experiential factors such as reduced anxiety, increased confidence, and experiencing real-life simulation as authentic, instead of focusing on the short-term performance gains alone. The analysis of the variability of the students' perception scores revealed that they had differences in their answers due to fluctuations that occurred since some mentioned both positive and negative aspects of the use of VR for learning and speaking English. While the positive judgments were predominantly made on its instructional effectiveness, engagement and enjoyment, and increasing confidence, the negative criticism usually involved pedagogical and practical constraints. This resulted in learners having moderate to high variability since their answers were not consistent or stable, whereas the others who had low variability of their perception score showed more of a union in the consistency of their answers. Hence, these findings enabled to delve deeper into understanding the degree of which and how learners made sense of the use of VR for learning and speaking English. In short, the relationship between perceived and actual speaking development appears to be complex, individualistic and informed by learner characteristics and instructional context rather than being in a linear pattern.

Briefly, the findings imply that the use of VR for learning and speaking English appeared as a supportive educational tool that shaped how speaking developed over time especially in relation to fluency and coherence, and pronunciation components of speaking. This alignment between performance results and students' largely positive perceptions indicates that VR may enhance the quality and consistency of speaking practice even when the overall learning gains are influenced by baseline proficiency and ongoing of the conventional classroom instruction. The study crucially demonstrated that learners' subjective experiences of VR for learning and speaking English do not always align perfectly with their objectively measured performance gains. While most students viewed VR to be beneficial and motivating while learning and practicing speaking, the degree of observable improvement varied across individuals and speaking subskills. This highlights the salience of the interpretation of VR related outcomes to be conducted through both quantitative and qualitative perspectives. In addition, caution needs to be added against the equivalence of positive perceptions of learners immediately or uniformly resulting in performance gains.

6.2. PEDAGOGICAL IMPLICATIONS

The findings of this study suggest that VR can be a valuable pedagogical tool in EFL speaking educational settings when it is implemented with clear instructional purposes and integrated into a broader curriculum. It shouldn't be presented as a sole solution for improving speaking performance. Thus, VR should be viewed as a supportive learning environment that enhances speaking practice by influencing how learners engage within their own experience of using this novel device and its reflections on their oral communication abilities.

One key pedagogical implication is that VR is particularly suitable for supporting the development of fluency and coherent as well as pronunciation through repeated and contextualized oral production activities. VR environments allow learners to engage in meaningful speaking tasks that resemble real-world communicative situations by encouraging sustained oral production and real-life language use. Therefore, teachers can use VR environments to enable students to repeat similar scenarios multiple times by gradually increasing the communicative demands as VR platforms such as ImmerseMe tend to have level appropriate learning material that can easily be selected to be changed. This repetition within authentic contexts can assist learners to develop more spontaneous speech through greater automaticity and confidence in speaking skills.

Another important implication concerns the affective dimension of speaking. VR offers a low-stress environment in which learners can interact with virtual interlocutors without negative evaluations. This anonymity that is provided by VR allows students to hide their real identities from being impacted negatively by external judgements and reduces the social pressure that is commonly connected with speaking in front of others regardless of who they may be. This is one of the most crucial affordances of VR as it is impossible to hide one's own self or not be influenced by others' negative evaluations in the real world. In other words, real-life is much more of a threat to self-identity and not as safe as a virtual environment in terms of speaking in a foreign language. As a result, learners are more willing to take risks, experiment with language, and persist in speaking practice. This makes VR especially beneficial for students who experience speaking anxiety or are hesitant in speaking in the conventional classroom settings.

The effectiveness of VR for learning and speaking English is dependent on its pedagogical design rather than pursuing its use as a novel instrument of learning. VR activities should be aligned with speaking objectives and interwoven with a structured educational order. The effective

implementation of VR involves preparation beforehand, such as introducing target language, communicative goals and success criteria so that the guided VR practice can aid in focusing learners on particular speaking subskills. Especially as the students were part of an English preparatory program, they were already part of a well-established institution with a strict and intense curriculum for learning English. Thus, these students in the experiment group were ready to tackle the level specific speaking activities in respect to VR being presented as a supportive tool. Otherwise, they might have been overwhelmed if many grammatical structures, lexical resource, task types and other factors were extremely unfamiliar to them. Also, post-VR follow-up activities that include receiving feedback, reflection or transfer of skills into real life. Without such scaffolding, VR risks being seen as engaging but not producing the desired learning outcomes, which would make it pedagogically superficial. As seen in this study, the students' own experiences and accounts of learning and speaking English through VR assisted in the comprehension of why their scores improved and to what extent. Thus, according to the responses provided by learners, actions can be taken so that the VR experience provides better learning outcomes, whether this means changing to another VR platform or getting in touch with the VR company to offer recommendations on how to develop it.

The findings also highlight the significance of learner characteristics and flexibility for using VR for learning and speaking English. Learners vary in how they benefit from VR due to personal factors. Thus, even positive learning experiences don't always result in immediate performance gains for all students. Hence, teachers should take into consideration of providing varied task types, additional support and opportunities for reflections of learners in which they express what is going well, what is lacking, and what they are having difficulties with to tailor to the needs of individual learners. By allowing learners to voice their opinions about such matters can further enhance learner autonomy and engagement as learners will experience having the chance to be heard through follow-up actions that portray the necessary changes had been adjusted according to their requirements or desires.

Furthermore, VR that is used as a supplementary tool of speaking instruction can be combined with classroom discussion, peer interaction and teacher-led feedback as this can help bridge the gap between immersive experience and speaking outcomes. It can especially be effective to prepare learners for in-class speaking tasks or to reinforce language that had been

taught within the traditional classroom setting. In addition, VR can be used to provide feedback to learners in regard to their speaking skills especially for pronunciation since it can provide immediate and focused feedback. Many VR platforms offer built-in features such as speech recognition and model utterances, which allow learners to notice the difference in their own production and the norms of the target language. This form of feedback can raise learners' metalinguistic awareness especially in relation to grammar, phonology and vocabulary by drawing attention to aspects such as stress, intonation and accuracy immediately or gradually over time. This is also valuable for learners to receive feedback within a VR environment can reduce the affective pressure that many consider to be quite a nightmare in relation to corrective feedback in face-to-face classroom contexts. As feedback through VR is delivered through the system rather than directly by the teacher or peers, learners are more likely to perceive it as less threatening, which encourages repeated practice and self-correction. As a result, the use of VR in EFL contexts can support the development of more accurate and intelligible linguistic elements while sustaining the confidence of the learners and their willingness to engage in speaking practice.

Practical considerations are also central to effective use of VR in learning and speaking English. Teachers and institutions must be able to ensure technical reliability, provide clear guidelines on using the platform, and establish routines for troubleshooting. In contexts where HiVR is not feasible, educational institutions can opt for LiVR as it can still offer many pedagogical benefits that were discussed in this study. Moreover, although privacy and security concerns were not very salient for most learners in this study, teachers should inform their learners on how data are handled and ensure that VR use aligns with institutional guidelines.

Overall, the pedagogical implications of this study suggest that VR has strong potential to enrich learning and speaking English in EFL contexts by fostering engagement, reducing anxiety, and supporting meaningful speaking practice. Through its thoughtful and pedagogical integration into a curriculum, VR can be of assistance as a powerful complementary tool that enhances students' speaking experiences and aids in the development of core speaking skills within a learner-centred educational framework.

To sum up, the findings of this study offer several pedagogical implications for EFL contexts. To begin with, VR can be effectively integrated as a supplementary tool for speaking practice that provides learners with increased exposure to meaningful, contextualized and

opportunities for speaking in a low-stress environment. Next, VR appears especially well-suited to develop fluency and pronunciation through repeated interaction and sustained oral production. Finally, the successful implementation of VR needs to have a careful instructional design, technical accessibility, and alignment with learners' proficiency levels and expectations. Without thinking about such considerations, the potential benefits of VR may not be fully actualized.

6.3. LIMITATIONS

Despite its contributions, this study has several limitations that should be considered. The relatively small sample size (n=24) as 12 students were in the experiment group and 12 students were in the control group and the use of pre-existing classes without random assignment limits the generalizability of the results, which is especially in relation to the correlation analysis of the perception scale as the number of students was 12. In addition, the baseline differences of the mean scores of the pre- and post-tests and certain speaking subskills between the experiment and control groups limited the interpretations of the effect of VR as directly having an impact on learners' speaking performance. Moreover, the short duration of the intervention and LiVR being used further limited the extent to which the long-term development and immersive effects could be examined. Finally, the reliance on self-reported perception findings and the researcher's dual role as the instructor may have influenced participants' responses despite efforts to minimize this bias. Based on these limitations, the findings should be seen as exploratory and specific for this context.

6.4. SUGGESTIONS FOR FUTURE RESEARCH

Future research may benefit from larger and more diverse groups of participants that involve a longer intervention period. This is especially for the correlation analysis as they couldn't be used to generalize the results, so larger and more diverse samples are needed to confirm whether correlation structures emerge across different learner populations and educational contexts. Furthermore, future studies could examine within-group development more closely and study the long-term effects of VR on speaking skills and investigate how learner-related factors and learning preferences interact with using VR for learning along with its impact on other aspects of speaking such as those pertaining to interaction. Last but not least, the effects of HiVR on speaking skills in

the long-term should be looked into to gain insight into how true immersion and presence and not just pretending affects learners over a period of time, which could also be compared to LiVR.

6.5. CONCLUDING REMARKS

To conclude, this study adds to the developing body of research on VR in language education by providing detail and being based on evidence in relation to its role in the development of speaking skills for EFL learners. While there is a clear acknowledgement of the value of VR for this issue, it is not viewed as a transformative solution. It is instead presented as a tool that is valuable in supporting learning and speaking skills through different contexts and enhancing engagement when it is carefully integrated into existing educational practices. The learners' voices in this study were of immense value as they highlighted the salience of evaluating educational technologies not just in terms of measurable outcomes but also in regard to how learners themselves experience and make sense of this novel learning environment, which once again emphasizes the importance of hearing and listening to the perspective of humans and their own experiences of stories as each narrates unique aspects that are best interpreted together as part of a whole.

REFERENCES

Akay, M. and Kessler, M., 2024, The impact of virtual reality (VR) pedagogy on L2 English learners' oral communication and pragmatic competence, *Language, learning and technology,* 28, 1-11. Doi: 10.64152/10125/73568.

Al-Tamimi, A. and Shuib, M., 2009, Motivation and attitudes towards learning english: a study of petroleum engineering undergraduates at hadhramout university of sciences and technology, *GEMA online journal of language studies,* 9(2), 29–53.

Alderson, J. C. and Wall, D., 1993, Does washback exist? *Applied linguistics,* 14, 115-129. http://dx.doi.org/10.1093/applin/14.2.115

Alderson, J. C., 1991, Bands and scores. In: Alderson J. C. and North B. (eds.) *Language testing in the 1990s.* London: Modern English Publications and the British Council, 71-86.

Alexander, N. and LeBaron, M., 2009, Death of the role-play. In: Honeyman C., Coben J. and De Palo G. (Eds.), *Rethinking negotiation teaching: Innovations for context and culture,* St. Paul, MN: DRI Press, 179–197.

Alkharusi, H., 2022, A descriptive analysis and interpretation of data from Likert scales in educational and psychological research, *Indian journal of psychology and education,* 12(2), 13–16.

Allen, M., Robson, D. and Iliescu, D., 2023, Face validity, *European journal of psychological assessment*, 39, 153-156. 10.1027/1015-5759/a000777.

Almaktary, H. M. A. and Al-Kadi, A. M. T., 2017, CALL in post-method era, *Indonesian journal of EFL and linguistics,* 2(2), 133-146.

Almutairi, A. and Alfarwan, S., 2024, Employing drama techniques in the classroom to ımprove the speaking skills of EFL university students, *Arab world english journal,* 15, 311-327. 10.24093/awej/vol15no4.20.

Altun, M., 2015, Using role-play activities to develop speaking skills: A case study in the language classroom, *International Journal of social sciences and educational studies,* 1(4), 27–33.

Alwehebi, K. A. I., 2025, The effectiveness of using role-play in developing EFL speaking skills among saudi preparatory students, *Theory and practice in language studies,* 15(8), 2750-2758. https://doi.org/10.17507/tpls.1508.33

Alzubi, A., Nazim, M. and Ahmad, J., 2025, Identifying English as a foreign language students' attitude to improving speaking skills through collaboration, *Journal of education and learning,* 19, 180-190. Doi:10.11591/edulearn.v19i1.21333.

Anthony, E. M., 1963, Approach, method and technique, *English language teaching,* 17, 63-67. https://doi.org/10.1093/elt/XVII.2.63

Ay Kaya, E., and Yangın Ersanlı, C., 2025, The effect of flip discussions on EFL students' speaking skills and attitudes, *Participatory educational research,* 12(2), 57–74. https://doi.org/10.17275/per.25.19.12.2

Ayuningtias, D. O., Wulandari, and Yana, Y., 2019, The use of role play to improve students'speaking skill, *Project (professional journal of english education),* 2(3), 416-420.

Bachman, L. F., 1990, *Fundamental considerations in language testing.* Oxford: Oxford University Press.

Bachman, L. F., and Palmer, A. S., 1996, *Language testing in practice.* Oxford: Oxford University Press.

Bailey, K. M., 2003, *Speaking*. In: Nunan D. (ed.), Practical english language teaching, International Edition, McGraw-Hill, 47-66.

Bajpai, S., Bajpai, R and Chaturvedi, H., 2015, Evaluation of inter-rater agreement and inter-rater reliability for observational data: an overview of concepts and methods, *Journal of the indian academy of applied psychology,* 41, 20-27.

Baker, B. and Turner, C., 2022, *Rating scales and rubrics in language assessment.* In: Chapelle C. A. (ed.), The encyclopedia of applied linguistics, 1-7. 10.1002/9781405198431.wbeal1045.pub2.

Baker, C., 1992, *Attitudes and language.* Clevedon: Multilingual Matters Ltd.

Baker, L., and Westbrook, C., 2022, *Prism 2: reading student's book.* Cambridge University Press.

Bao, Y. Y., and Liu, S. Z., 2021, The influence of affective factors in second language acquisition on foreign language teaching, *Open journal of social sciences,* 9, 463-470. https://doi.org/10.4236/jss.2021.93030

Barcelos, A. M. F., and Kalaja, P., 2003, *Conclusion: exploring possibilities for future research on beliefs about SLA*. In: Kalaja P. and Barcelos A. M. F. (eds.), Beliefs about SLA: New research approaches, Kluwer Academic Publishers, 231–238.

Bartko, J. J., 1966, The intraclass correlation coefficient as a measure of reliability, *Psychological reports,* 19(1), 3–11.

Bhandari, B. L. and Bhandari, L.P., 2025, Exploring teachers' perceptions of post-method pedagogy in ELT, *International journal of language and literary studies,* 7(2), 203-216. https://doi.org/10.36892/ijlls.v7i2.2030

Bijani, H., Hashempour, B., Ibrahim, K., Orabah, S., and Heydarnejad, T., 2022, Investigating the effect of classroom-based feedback on speaking assessment: a multifaceted Rasch analysis, *Language testing in asia,* 12, 1-28. https://doi.org/10.1186/s40468-022-00176-3.

Bowman, N., 2017, *Extraneous variables, control of.* In: Allen M. (ed.), The SAGE encyclopedia of communication research methods, SAGE Publications, 484–486. https://doi.org/10.4135/9781483381411.n177

Braun, V., and Clarke, V., 2006, Using thematic analysis in psychology, *Qualitative research in psychology*, 3(2), 77–101.

British Council, 1983, *VOTE: Oral testing*. British Council English Languages Services Department and the Design, Production and Publishing Department.

Brookhart, S. M. and Chen, F., 2015, The quality and effectiveness of descriptive rubrics, *Educational review,* 67(3), 343-368, DOI: 10.1080/00131911.2014.929565

Brookhart, S. M., 2018, Appropriate criteria: Key to effective rubrics, *Frontiers in education,* 3(22), 1-12, doi: 10.3389/feduc.2018.00022

Brown, A. V., 2009, Students' and teachers' perceptions of effective foreign language teaching: A comparison of ideals, *Modern language journal,* 93, 46–60.

Brown, H. D., 1997, English language teaching in the "post-method" era: Toward better diagnosis, treatment, and assessment, *PASAA,* 27(1), 1-11. DOI: 10.58837/CHULA.PASAA.27.1.1

Brown, H. D., 2004, *Language assessment principles and classroom practices*, White Plains, NY: Pearson Education.

Brown, H. D., and Abeywickrama, P., 2019, *Language assessment: Principles and classroom practices,* 3rd ed., Pearson Education.

Brown, J. S., Collins, A., and Duguid, P., 1989, Situated cognition and the culture of learning. *Educational researcher,* 18(1), 32–42.

Buck, C., 1989, *The ACTFL OPI: Tester training manual.* ACTFL.

Buck, G., 1989, Written tests of pronunciation: Do they work? *ELT journal,* 43(1), 50–56.

Butorova, A. S., Tarasov, D. A., Kosachenko, A. I., and Sergeev, A. P., 2024, Learning foreign languages by adults using immersive virtual reality systems: Review of recent studies (2014–2024), *8th Scientific School Dynamics of Complex Networks and Their Applications (DCNA)*, IEEE, 43–46. https://doi.org/10.1109/DCNA63495.2024.10718463

Caesar, Y. C., 2023, Critical thinking activities to improve senior high school students' speaking skills, *Al-ishlah: jurnal pendidikan,* 15(3), 3804–3817. Doi: https://doi.org/10.35445/alishlah.v15i3.3612

Camacho-Fidalgo, G., Zúñiga-Argüello, R., Cubillo-Araya, M., Parkins, K. C., and Rojas-Muñoz, E., 2025, So yu waan tu lern Limon Kryol? Considerations of virtual reality for endangered language preservation and learning. *2025 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops,* IEEE, 162–164. https://doi.org/10.1109/VRW66409.2025.00040

Cambridge English Language Assessment, 2016, *Principles of Good Practice: Research and innovation in language learning and assessment*.

Campbell, D. T., 1957, Factors relevant to the validity of experiments in social settings, *Psychological bulletin,* 54(4), 297–312. doi:10.1037/h0040950

Campbell, D. T., and Stanley, J. C., 1966, *Experimental and quasi-experimental designs for research*. Chicago, IL: Rand McNally and Company.

Campbell, R. and Wales, R., 1970, *The study of language acquisition*. In: Lyons J. (ed.), New horizons in linguistics. Harmondsworth: Penguin Books.

Can, T., 2020, Training pre-service English language teachers with 3-D machinima, *Turkish online journal of educational technology,* 19, 53–65.

Cardwell, S., 2025, *What is the pedagogy and academic research behind ImmerseMe?* https://immerseme.zendesk.com/hc/en-gb/articles/900000403223-What-is-the-pedagogy-and-academic-research-behind-ImmerseMe [Retrieved: December 2 2025]

Carey, M. D., Mannell, R. H., and Dunn, P. K., 2011, Does a rater's familiarity with a candidate's pronunciation affect the rating in oral proficiency interviews? *Language testing*, 28(2), 201-219. https://doi.org/10.1177/0265532210393704

Celce-Murcia, M., 2007, *Rethinking the role of communicative competence in language teaching.* In: Soler E. A. and Jordà M. S. (eds.), Intercultural language use and language learning, Springer, 41-57. https://doi.org/10.1007/978-1-4020-5639-0_3

Chen, C., and Yuan, Y., 2025, Effectiveness of virtual reality on chinese as a second language vocabulary learning: perceptions from international students, *Computer assisted language learning,* 38(3), 397–425. https://doi.org/10.1080/09588221.2023.2192770

Chen, M.-R. A., and Hwang, G.-J., 2020, Effects of experiencing authentic contexts on English speaking performance, anxiety, and motivation of EFL students with different cognitive styles, *Interactive learning environments,* 30(9), 1619–1639. https://doi.org/10.1080/10494820.2020.1734626

Cho, B., 2015, Improving learners' oral skills through two types of role-play, *Journal of performative teaching, learning, research,* 4(1), 35-52. https://doi.org/10.33178/scenario.9.1.2

Chomsky, N., 1965, *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Cobern, W. W., 1993, Constructivism, *Journal of educational and psychological consultation,* 4(1), 105-112.

Cohen, A. D., 2001, *Second language assessment.* In: Celce-Murcia M. (ed.), Teaching English as a second or foreign language, 3rd ed., Heinle and Heinle, 515-534.

Conti, G., 2025, *No, you don't need sentence builders to teach the EPI way.* https://gianfrancoconti.com/2025/11/12/no-you-dont-need-sentence-builders-to-teach-the-epi-way/ [Retrieved: November 12 2025]

Cooper, C. R., 1977, *Holistic evaluation of writing.* In: Cooper C. R. and Odell L. (eds) Evaluating writing: describing, measuring, judging, NCTE, 3-32.

Council of Europe (CoE), 2001, *Common European Framework of Reference for Languages: Learning, Teaching, Assessment.* Cambridge University Press.

Council of Europe (CoE), 2020, *Common European Framework of Reference for Languages: Learning, Teaching, Assessment—Companion Volume*. Council of Europe Publishing. https://www.coe.int/en/web/common-european-framework-reference-languages

Cowie, N., and Alizadeh, M., 2025, Virtual reality for language learning, *ELT journal*. https://doi.org/10.1093/elt/ccaf025

Creamer, E. G., 2018, *An introduction to fully integrated mixed methods research.* Sage. https://doi.org/10.4135/9781071802823

Creswell, J. W., and Plano Clark, V. L., 2018, *Designing and conducting mixed methods research*, 3rd ed., Thousand Oaks, CA: SAGE.

Creswell, J.W. and Creswell, J.D., 2023, *Research design: Qualitative, quantitative and mixed methods approaches.* Sage Publications Ltd.

Crystal, D., 2003, *English as a global language*, 2nd ed., Cambridge University Press.

Çeltek, E., 2020, *Progress and development of virtual reality and augmented reality technologies in tourism: a review of publications from 2000 to 2018.* In: Çeltek E. (ed.), Handbook of Research on Smart Technology Applications in the Tourism Industry, 1–23.

Çetin Köroğlu, Z., 2025, A study on pre-service English teachers' perceptions of virtual reality to develop speaking skills, *Technology, Pedagogy and Education*, 1-18. DOI: 10.1080/1475939X.2025.2495869

Daskalovska, N., Hadzi-Nikolova, A., and Todorova, M., 2025, The role of methods in the postmethod era. *12th International New York academic research congress on social, humanities, administrative, and educational sciences: Congress proceedings book*, BZT Turan Publishing House, 973-981. https://doi.org/10.30546/19023.978-9952-8545-4-1.20980

Davis, L., 2015, The influence of training and experience on rater performance in scoring spoken language, *Language testing*, 33(1), 117-135. https://doi.org/10.1177/0265532215582282

Dawley, L., and Dede, C., 2014, *Situated learning in virtual worlds and immersive simulations.* In: Spector J. M., Merrill M. D., Elen J. and Bishop M. J. (eds.), Handbook of research on educational communications and technology, Springer, 723–734. https://doi.org/10.1007/978-1-4614-3185-5_58

Dennis, K. P., 2024, Using AI-powered speech recognition technology to ımprove english pronunciation and speaking skills, *IAFOR journal of education: technology in education,* 12(2), 107-126.

Derwing, T. M., and Munro, M. J., 2009, Putting accent in its place: Rethinking obstacles to communication, *Language teaching,* 42, 1–15.

Despagne, C., 2010, The difficulties of learning english: perceptions and attitudes in mexico, *Canadian and ınternational education,* 39(2), 55–74.

Dhimolea, T. K., Kaplan-Rakowski, R., and Lin, L., 2022, A systematic review of research on high-immersion virtual reality for language learning, *TechTrends,* 66, 810–824. https://doi.org/10.1007/s11528-022-00717-w

Dinçer, M. N. and Bal Gezegin, B., 2024, Teachers' perspectives on assessing english speaking skills: a post-new exam model ınvestigation, *Journal of language research,* 8(2), 30-43.DOI: https://doi.org/10.51726/jlr.1566625

Doff, A., Thaine, C., Puchta, H., Stranks, J., and Lewis-Jones, P., 2022, *Empower intermediate/B1+ student's book,* 2nd ed., Cambridge University Press.

Doğan, C. D., and Uluman, M., 2017, A comparison of rubrics and graded category rating scales with various methods regarding raters' reliability, *Educational sciences: theory and practice,* 17, 631–651. http://dx.doi.org/10.12738/ estp.2017.2.0321

Drost, E.A., 2011, Validity and reliability in social science research, *Education research and perspectives,* 38, 105- 123.

Druckman, D. and Ebner, N., 2008, Onstage or behind the scenes? Relative learning benefits of simulation role-play and design, *Simulation and gaming,* 39(4), 465-497.

Ducasse, A. and Brown, A., 2009, Assessing paired orals: Raters' orientation to interaction, *Language testing,* 26, 423-443. 10.1177/0265532209104669.

Eckes, T., 2011, *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater- mediated assessments.* Frankfurt: Peter Lang.

Elder, C., Knoch, U., Barkhuizen, G., and von Randow, J., 2005, Individual feedback to enhance rater training: Does it work? *Language assessment quarterly,* 2(3), 175–196. https://doi.org/10.1207/s15434311laq0203_1

Ellis, R., 2005, Principles of instructed language learning, *System,* 33(2), 209–224. https://doi.org/10.1016/j.system.2004.12.006

Ellis, R., 2008, *Individual differences in second language learning*. In: Davies A. and Elder C. (eds.), The handbook of applied linguistics, Blackwell, 525–551 https://doi.org/10.1002/9780470757000.ch21

Enkin, E., 2022, Comparing two worlds: Spanish learners' face-to-face and immersive social VR speaking experiences, *Computer assisted language learning,* 23(1), 22–42. https://callej.org/index.php/journal/article/view/369

Epp, C. D., 2014, Mobile adaptive communication support for vocabulary acquisition, *Journal of learning analytics,* 1(3), 173–175. https://doi.org/10.18608/jla.2014.13.15

Erciyes, E., 2020, Paradigms of inquiry in the qualitative research, *European scientific journal,* 16(7), 181-200. https://doi.org/10.19044/esj.2020.v16n7p181

Ericsson, E., 2023, *Experiences of speaking with conversational AI in language education* [Unpublished doctoral dissertation], Department of Applied Information Technology, University of Gothenburg.

Ericsson, E., and Sofkova Hashemi, S., 2024, (Dis)engagement and usefulness: Students' experiences of practicing speaking German in a virtual 360° environment*, Nordic journal of language teaching and learning,* 12(1), 44-69. DOI 10.46364/njltl.v12i1.1229

Evans, J. D., 1996, Straightforward statistics for the behavioral sciences, Thomson Brooks/Cole Publishing.

Falconer, L., 2013, Situated learning in virtual simulations: Researching the authentic dimension in virtual worlds, *Journal of Interactive learning research,* 24(3), 285–300.

Fathi, J., Rahimi, M., and Derakhshan, A., 2024, Improving EFL learners' speaking skills and willingness to communicate via artificial intelligence-mediated interactions, *System,* 121, 1-17. https://doi.org/10.1016/j.system.2024.103254.

Fereday, J., and Muir-Cochrane, E., 2006, Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development. *International journal of qualitative methods*, 5(1), 80-92.

Fetters, M. D., Curry, L. A., and Creswell, J. W., 2013, Achieving integration in mixed methods designs—Principles and practices, *Health services research*, 48, 2134–2156. https://doi.org/10.1111/1475-6773.12117

Field, A., 2009, *Discovering statistics using SPSS,* 3rd ed., SAGE Publications.

Fisher, R. A., 1992, *Statistical methods for research workers*. In: Kotz S. and Johnson N. L. (Eds.), Breakthroughs in statistics, Springer, 66–70. https://doi.org/10.1007/978-1-4612-4380-9_6

Flannelly, K. J., Flannelly, L. T. and Jankowski, K. R. B., 2018, Threats to the internal validity of experimental and quasi-experimental research in healthcare, *Journal of health care chaplaincy*, 1-24. DOI: 10.1080/08854726.2017.1421019

Fox, N. J., 2008, Post-positivism, *The SAGE encyclopedia of qualitative research methods*, 2(1), 659-664.

Fulcher, G., 2003, *Testing Second Language Speaking,* Peason Education.

Galante, A., 2014, English language teaching in the post-method era, *Contact magazine,* 40(3), 57–62.

Grab, M. Ö., 2025, Integrated AI chatbot practice: A pathway to improved ESL speaking skills*, Social sciences and humanities open,* 12, 1-9. https://doi.org/10.1016/j.ssaho.2025.101933

Gracia, E., 2024, *Immersive virtual reality for L2 pragmatics instruction in an intensive English program at a US university* [Unpublished doctoral dissertation], Arizona State University.

Gu, L., 2025, Beyond the classroom: Exploring the impact of virtual reality exposure on foreign language anxiety with the mediating role of ESL Chinese learners' communicative confidence and fluency, *Humanities and social sciences communications,* 12, 1-13. https://doi.org/10.1057/s41599-025-05030-4

Guba, E. G., and Lincoln, Y. S.,1994, *Competing paradigms in qualitative research.* In Denzin N. K. and Lincoln Y. S. (Eds.), Handbook of qualitative research, Sage Publications, 105–117.

Guest, G., Namey, E. E., and Mitchell, M. L., 2013, *Collecting qualitative data: A field manual for applied research,* SAGE Publications.

Gundersen, K., and Svartdal, F., 2010, Diffusion of treatment interventions: exploration of 'secondary' treatment diffusion, *Psychology, crime and law,* 16(3), 233–249. https://doi.org/10.1080/10683160802612924

Hale Feinstein, A., Mann, S., and Corsun, D. L., 2002, Charting the experiential territory: Clarifying definitions and uses of computer simulation, games, and role play, *Journal of management development*, 21(10), 732-744.

Han, X., 2020, On Statistical Measures for Data Quality Evaluation, *Journal of geographic information system,* 12, 178-187. https://doi.org/10.4236/jgis.2020.123011

Harding, L., 2012, *Pronunciation assessment.* In: Chapelle C. A. (ed.), The encyclopedia of applied linguistics*,* Wiley-Blackwell, 4708–4713. DOI: 10.1002/9781405198431

Harmer, J., 1989, *Teaching and Learning Grammar.* Longman.

Harmer, J., 2007, *The practice of English language teaching.* (4th Edn.) London: Longman.

Harper-Whalen, S. and Morris, S., 2005, Using a role-play activity in training. *Training Solutions,* 9, 1-4.

Hart, A., 2001, Mann–Whitney test is not just a test of medians: Differences in spread can be important, *BMJ*, 391-393. DOI: 10.1136/bmj.323.7309.391

Hashemi, A. and Daneshfar, S., 2018, A review of the IELTS test: Focus on validity, reliability, and washback. *Indonesian Journal of English Language Teaching and Applied Linguistics, 3(1),* 39–52.

Hasumi, T. and Chiu, M., 2024, Technology-enhanced language learning in English language education: Performance analysis, core publications, and emerging trends, *Cogent education,* 11(1), 1-22, DOI: 10.1080/2331186X.2024.2346044

He, A. W. and Young, R., 1998, *Language proficiency interviews: A discourse approach.* In: Young R. and He A. W. (eds.), Talking and testing: Discourse approaches to the assessment of oral proficiency, John Benjamins, 1–24.

Herrington, J. and Oliver, R., 2000, An instructional design framework for authentic learning environments, *Educational technology research and development,* 48(3), 23–48. https://doi.org/10.1007/BF02319856

Ho-Minh, T. and Suppasetseree, S., 2025, Enhancing Vietnamese EFL students' speaking skills through augmented reality: A mixed-method study in english for tourism and hospitality, *Educational process: International journal,* 15, 1-23. https://doi.org/10.22521/edupij.2025.15.139

Horwitz, E. K., Horwitz, M. B. and Cope, J., 1986, Foreign language classroom anxiety, *The modern language journal*, 70(2), 125–132. https://doi.org/10.2307/327317

https://doi.org/10.1057/9780230514577

Huang, H. and Liaw, S., 2011, Applying situated learning in a virtual reality system to enhance learning motivation, *International journal of ınformation and education technology,* 1(4), 298–302. https://doi.org/10.7763/IJIET.2011.V1.48

Huang, H.-T. D., 2018, Modeling the relationships between anxieties and performance in second/foreign language speaking assessment, *Learning and individual differences,* 63, 44–56. https://doi.org/10.1016/j.lindif.2018.03.002

Hughes, A., 2003, *Testing for Language Teachers,* 2nd ed., Cambridge University Press. https://doi.org/10.1017/CBO9780511732980

Hymes, D., 1972, *On communicative competence*. In: Pride J. and Holmes J. (eds.), Sociolinguistics, Harmondsworth: Penguin Books, 269-285.

IELTS, 2025, *Academic institutions.* https://ielts.org/organisations/academic-institutions [Retrieved: August 20 2025]

ImmerseMe, 2025, *ImmerseMe: Using virtual reality to learn languages.* https://immerseme.co/ [Retrieved: December 2 2025]

Inalöz, A. B., and Yılmaz, K., 2025, The effect of augmented and virtual reality applications in teaching Turkish as a foreign language in University, *e-Kafkas journal of educational research,* 12, 329-344. doi: 10.30900/kafkasegt.1477439

Ironsi, C.S., 2023, Investigating the use of virtual reality to improve speaking skills: insights from students and teachers, S*mart learning environments,* 10(53)*,* 1-21. https://doi.org/10.1186/s40561-023-00272-8

Isaacs, T. and Harding, L., 2017, Pronunciation assessment, *Language teaching,* 50(3), 347–366. doi:10.1017/S0261444817000118

Isaacs, T. and Thomson. R. I., 2013, Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions, *Language assessment quarterly,* 10(2), 135–159.

Isaacs, T. and Trofimovich, P., 2012, Deconstructing comprehensibility, *Studies in Second Language Acquisition,* 34*,* 475-505. doi:10.1017/S0272263112000150

Isaacs, T., 2016, *Assessing speaking.* In: Tsagari D. and Banerjee J. (eds.), Handbook of second language assessment, DeGruyter Mouton, 131–146

Islam, P. and Islam, T., 2012, Effectiveness of role play in enhancing the speaking skills of the learners in a large classroom: An investigation of tertiary level students, *Stamford journal of English,* 7*,* 218-233.

Istanbul Bilgi University, 2024, *English Preparatory Program student handbook (2024–2025 academic year)*, İstanbul Bilgi University.

Johnson, K. and Johnson, H., 1998, *Encyclopedic dictionary of applied linguistics: A handbook for language teaching,* Blackwell Publishers, Oxford.

Kang, O., Rubin, D. and Kermad, A., 2019, The effect of training and rater differences on oral proficiency assessment, *Language testing,* 36(4), 481-504. https://doi.org/10.1177/0265532219849522

Kaplan-Rakowski, R. and Gruber, A., 2019, Low-immersion versus high-immersion virtual reality: Definitions, classification, and examples with a foreign language focus. *Proceedings of the ınnovation in language learning ınternational* conference *2019*. Pixel.

Kaplan-Rakowski, R., Papin, K. and Hartwick, P., 2023, Language teachers' perceptions and use of extended reality. *Calico journal,* 40, 1-23. https://doi.org/10.1558/cj.2275

Karagöl, E., Yıldırım Bilgen, D. and Korkmaz, C. B., 2025, The impact of AI applications on pre-service teachers' public speaking anxiety and academic speaking skill in the context of oral presentations: A mixed-methods study, *Teaching and teacher education,* 166, 1-18. https://doi.org/10.1016/j.tate.2025.105180

Karim, S. and Haq, N., 2014, An assessment of IELTS speaking test, *International journal of evaluation and research in education,* 3(3), 152-157.

Kim, J. and Leathem, T., 2018, Virtual reality as a standard in the construction management curriculum, *Proceedings of the ınternational conference on construction futures 2018*, University of Wolverhampton.

Knoch, U., 2011, Investigating the effectiveness of individualized feedback to rating behavior – a longitudinal study, *Language testing,* 28*,* 179–200.

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

Kormos, J. and Préfontaine, Y., 2016, Affective factors influencing fluent performance: French learners' appraisals of second language speech tasks, *Language teaching research*, 21(6), 699-716. https://doi.org/10.1177/1362168816683562

Krashen, S.D., 1981, *Second language acquisition and second language learning,* Pergamon Press Inc., Oxford.

Krebt, D. M., 2017, The effectiveness of role play techniques in teaching speaking for EFL college students, *Journal of language teaching and research,* 8(5), 863-870.

Kuiken, F. and Vedder, I., 2021, *Scoring approaches: scales/rubrics.* In: Winke P. and Brunfaut T. (eds.), The routledge handbook of second language acquisition and language testing*,* Routledge, 125-134. https://doi.org/10.4324/9781351034784-14

Kumaravadivelu, B., 2001, Toward a Postmethod pedagogy, *TESOL quarterly,* 35*,* 537-557. http://dx.doi. org/10.2307/3588427

Lackman, K., 2010, *Teaching speaking sub-skills: Activities for improving speaking.* Toronto, Canada: Lackman and Associates.

Lado, R., 1961, *Language testing: The construction and use of foreign language tests.* London: Longman.

Larsen-Freeman, D., 2000, *Techniques and principles in language teaching*. New York: Oxford University Press.

Lather, P., 1991, *Getting smart: Feminist research and pedagogy with/in the postmodern.* New York: Routledge.

Lave, J. and Wenger, E., 1991, *Situated learning: Legitimate peripheral participation.* Cambridge University Press.

Lazaraton, A., 1996, Interlocutor support in oral proficiency interviews: the case of CASE, *Language testing,* 13*,* 151-172. https://doi.org/10.1177/026553229601300202

Lee, A., 2025, Assessing speaking skills in virtual reality: Impacts and implications. *English teaching,* 80(2), 57–81. DOI: https://doi.org/10.15858/engtea.80.2.202506.57

Lee, J., 2009, *Statistics, descriptive*. In: Kitchin R. and Thrift N. (eds.), International encyclopedia of human geography*,* Elsevier, 422–428. https://doi.org/10.1016/B978-008044910-4.00534-4

Lee, K. M., Lee, J., Chung, C. Y., Ahn, S., Sung, K. H., Kim, T. W., Lee, H. J. and Park, M. S., 2012, Pitfalls and important issues in testing reliability using intraclass

correlation coefficients in orthopaedic research. *Clinics in orthopedic surgery,* 4(2), 149–155. Doi:10.4055/cios.2012.4.2.149

Lee, S.-M., Wang, X., Park, I. and Lestiono, R., 2024, It feels so real! Situated authentic language learning in immersive virtual reality, *Education and information technologies,* 29, 24023–24045. https://doi.org/10.1007/s10639-024-12807-8

Lestiono, R. and Lee, S.-M., 2024, Virtual reality exposure therapy for reducing students' public speaking anxiety: A systematic review, *The JALT CALL journal,* 20(2), 1–24. https://doi.org/10.29140/jaltcall.v20n2.1515

Lestiono, R. and Setyaningrum, R., 2025, Developing immersive virtual reality space for public speaking training: English debate room in Spatial platform. *19th International conference on ubiquitous information management and communication,* IEEE, 1-5. https://doi.org/10.1109/IMCOM64595.2025.10857513

Levis, J. M., 2006, *Pronunciation and the assessment of spoken language*. In: Hughes R. (ed.), Spoken English, TESOL and applied linguistics: Challenges for theory and practice, Palgrave Macmillan, 245–270.

Lincoln, Y. S. and Guba, E. G., 1985, *Naturalistic Inquiry,* Beverly Hills, CA: Sage Publications, Inc.

Liskin-Gasparro, J. E., 1998, Linguistic development in an immersion context: How advanced learners of Spanish perceive SLA, *Modern language journal,* 82*,* 159–175.

Littlewood, W., 2014, Communication-oriented language teaching: Where are we now? Where do we go from here? *Language teaching,* 47, 349-362. https://doi.org/10.1017/S0261444812000134

Littlewood, W.,1981, *Communicative language teaching,* Cambridge: Cambridge University Press.

Lou Y., 2025, The impact of virtual reality environments on English language acquisition: Innovative immersive learning technologies for communication skills development, *Journal of computational methods in sciences and engineering*, 25(5), 4616-4632. doi:10.1177/14727978251337950

Loughlin, E., and Mentzer, K., 2025, From headsets to language fluency: Investigating engagement and learning with VR in middle school Spanish, *Proceedings of the ISCAP Conference*, ISCAP, 1–11.

Lu, K., 2020, Experimental research of English teaching management model based on personalized multimedia technology, *Open access library journal,* 7(4), 1–13. https://doi.org/10.4236/oalib.1106251

Lu, S., 2025, Interactive virtual reality environments for English language learning: Immersive technologies to enhance language acquisition and communication skills, *Journal of computational methods in sciences and engineering,* 25(4), 3093–3106. https://doi.org/10.1177/14727978251318803

Lundeberg, O K., 1929, Recent developments in audition-speech tests, *The modern language journal,* 14(3), 193–202.

Luoma, S., 2004, *Assessing speaking*. Cambridge University Press.

Lutfi, N., Sunardi, A. I., Wahyu Utami, A. N. E., and Nafiah, R., 2025, Integrating virtual reality into English language instruction: assessing efficacy and student motivation, J*ournal on education,* 7(2), 10975-10989. https://doi.org/10.31004/joe.v7i2.8154

Ma, C., Haire, G., Smith, S., and Newbutt, N., 2025, Enhancing English as a foreign language education with virtual reality and augmented reality in K–12 classrooms: A systematic review from 2019 to 2023, *Technology in language teaching and learning,* 7(3), 1-30. https://doi.org/10.29140/tltl.v7n3.102535

Magoon, A. J., 1977, Constructivist approaches in educational research, *Review of educational research,* 47*,* 651-693.

Mahboob-Eshghdoost, S., and Mall-Amiri, B., 2019, The comparative effect of scripted and unscripted role-play on efl learners' speaking complexity, accuracy and fluency, *International journal of English and education,* 8(2), 84-96.

Maksimovic, J., and Evtimov, J., 2023, Positivism and post-positivism as the basis of quantitative research in pedagogy, *Research in pedagogy,* 13(1), 208-218.

Mallek, F., Mazhar, T., Abbas Shah, S. F., Ghadi, Y. Y., and Hamam, H., 2024, A review on cultivating effective learning: Synthesizing educational theories and virtual reality for enhanced educational experiences, *PeerJ computer science,* 10. https://doi.org/10.7717/peerj-cs.2000

McGraw, K. O., and Wong, S. P., 1996, Forming inferences about some intraclass correlation coefficients, *Psychological methods,* 1(1), 30–46. https://doi.org/10.1037/1082-989X.1.1.30

McNamara, T, and Roever, C., 2006, Psychometric approaches to fairness: Bias and DIF, *Language learning,* 56(2), 81–128. https://doi.org/10.1111/j.1467-9922.2006.00381.x

McNamara, T. and Ryan, K., 2011, Fairness versus justice in language testing: The place of English literacy in the Australian citizenship Test, *Language assessment quarterly,* 8(2), 161–178. https://doi.org/10.1080/15434303.2011.565438

McNamara, T., 2001, Language assessment as social practice: Challenges for research, *Language testing,* 18(4), 333–349.

Milgram, P. and Kishino, F., 1994, A taxonomy of mixed reality visual displays. *IEICE Trans. Information Systems*, 77(12), 1321-1329.

Mills, N., Pajares, F. and Herron, C., 2007, Self-efficacy of college intermediate French students: Relation to achievement and motivation, *Language learning,* 57*,* 417–442.

Monteiro, S. and Spadoni, G., 2025, *OSCE rater calibration.* MERIT and Department of Medicine, 1-6.

Morgan, G. A., Leech, N. L., Gloeckner, G. W. And Barrett, K. C., 2012, *IBM SPSS for introductory statistics: Use and interpretation,* 5th ed., Routledge.

Mosier, C. I., 1947, A critical examination of the concepts of face validity, *Educational and psychological measurement*, 7(2),191–205. Doi: https://doi.org/10.1177/001316444700700201

Mrabet-Abid, S., 2024, Voices in ELT: Exploring Tunisian EFL teachers' attitudes towards English language skills development, *International journal for multidisciplinary research,* 6(5), 1-13.

Munro, M. J. And Derwing, T. M., 1999, Foreign accent, comprehensibility, and intelligibility in the speech of second language learners, *Language learning,* 49, 285–310.

Murairwa, S., 2015, Voluntary sampling design, *International journal of advanced research in management and social sciences,* 4(2), 185–200.

Musa, M., Rahman, P. and Buhalis, D., 2022, *Virtual reality (VR) types.* In: Buhalis D. (ed.), Encyclopedia of Tourism Management and Marketing, Edward Elgar, 679–683. https://doi.org/10.4337/9781800377486.virtual.reality

Nakatsuhara, F., Inoue, C. and Khabbazbashi, N., 2020, *Measuring L2 speaking.* In: Winke P. and Brunfaut T. (eds.), The Routledge handbook of second language acquisition and language testing, Routledge, 285–295.

Nakatsuhara, F., Khabbazbashi, N. and Inoue, C., 2021, *Assessing speaking*. In: Fulcher G and Harding L. (eds.) the Routledge handbook of language testing, Routledge, 209-222.

Neupane, B., 2019, Effectiveness of role play in improving speaking skill, *Journal of NELTA Gandaki,* 1, 11-18.

Ni'mah, U., Nasihah, M. and Munfaati, F., 2025, Exploring self-regulated learning strategies to enhance English speaking skills among EFL students, *Indonesian journal of applied linguistics,* 14(3), 472-483. https://doi.org/10.17509/ijal.v14i3.72000

Nunan, D., 1989, *Understanding language classroom*. Cambridge: Prentice Hall.

Nunan, D., 2003, *Methodology*. In: Nunan D. (ed.), Practical english language teaching. Mc Graw Hill

O'Leary-Kelly, S.W. and Vokurka, R.J., 1998, The empirical assessment of construct validity, *Journal of operations management,* 16, 387-405. https://doi.org/10.1016/S0272-6963(98)00020-5

O'Loughlin, K. J., 2001, *The equivalence of direct and semi-direct speaking tests.* Cambridge University Press.

O'Sullivan, B. and Rignall, M., 2007, *Assessing the value of bias analysis feedback to raters for the IELTS writing module*. In: Taylor L. and Falvey P. (eds.), IELTS Collected Papers: Research in speaking and writing assessment, Cambridge: Cambridge University Press, 446–478.

Oktaviani, A. L., Rapiudin, T. U., Mahyati, M., Supriyanto, E. and Suryadi, A., 2025, Enhancing EFL students' speaking skills through the debate technique: A classroom action research study. *Journal of english teaching and applied linguistics,* 6(2), 177–187. https://doi.org/10.36655/jetal.v6i2.1808

Oller, J., 1979, *Language tests at school: A pragmatic approach,* Longman, London.

Onuf, N., 2015, *Constructivism: A user's manual*. In: Kusálková V., Onuf N. G. and Kowert P. (eds.), International relations in a constructed worlds, Routledge, 58–78.

Onwuegbuzie, A. J. and Collins, K. M. T., 2007, A typology of mixed methods sampling designs in social science research, *The qualitative report,* 12(2), 281-316.

Oradee, T., 2012, Developing speaking skills using three communicative activities (discussion, problem-solving, and role-playing), *International journal of social science and humanity,* 2(6), 533-535. DOI: 10.7763/IJSSH.2012.V2.164

Panhwar, A. H., Ansari, S. and Shah, A. A., 2017, Post-positivism: An effective paradigm for social and educational research*, International research journal of arts and humanities,* 45(45), 253-259.

Pani, P. and Ray, A., 2024, Immersive tech solutions to overcome speaking anxiety. *2nd International conference on signal processing, communication, power and embedded system*, IEEE, 1–5. https://doi.org/10.1109/SCOPES64467.2024.10991164

Patterson, A., Temple, C., Anderson, N., Rogalski, C. and Mentzer, K., 2025, The virtual stage: virtual reality ıntegration in effective speaking courses. *Information systems education journal,* 23(4), 57-68. DOI: https://doi.org/10.62273/

Peixoto, B., Gonçalves, G., Bessa, M., Pereira Bessa, L. C. and Melo, M., 2024, Learning a foreign language: Traditional listening vs interactive immersive virtual reality. *International Conference on Graphics and Interaction,* IEEE, 1–8. https://doi.org/10.1109/ICGI64003.2024.10923758

Phillips, D. C., and Burbules, N. C., 2000, *Postpositivism and educational research,* Bloomsbury Publishing PLC.

Phillipson, R., 1992, *Linguistic Imperialism,* Oxford: Oxford University Press.

Phisutthangkoon, K., 2024, Thai EFL university students' beliefs in English-speaking fluency, *rEFLections,* 31(2), 478-500.

Pituxcoosuvarn, M., Tanimura, M., Murakami, Y. and White, J.S., 2025, Enhancing EFL speaking skills with AI-powered word guessing: A comparison of human and AI partners. *Informatio,* 16*,* 1-31. https://doi.org/10.3390/ info16060427

Piumsomboon, T., Day, A., Ens, B., Lee, Y., Lee, G. and Billinghurst, M., 2017, Exploring enhancements for remote mixed reality collaboration, *Proceedings of the SA '17 symposium on mobile graphics and ınteractive applications*, ACM, 1–5. https://doi.org/10.1145/3132787.3139200

Pollitt, A., 2012, Comparative judgement for assessment, *International journal of technology and design education,* 22, 157–170. https://doi.org/10.1007/s10798-011-9189-x

Popper, K., 1959, *The logic of scientific discovery*, Routledge. https://doi.org/10.1063/1.3060577

Portney, L. G., and Watkins, M. P., 2000, *foundations of clinical research: applications to practice,* Prentice Hall.

Pouliot, V., 2004, The essence of constructivism. *Journal of ınternational relations and development,* 7(3), 319-336.

Prabhu, N., 1990, There is no method? *TESOL quarterly,* 24(2), 161-176.

Pretorius, L., 2024, Demystifying research paradigms: Navigating ontology, epistemology, and axiology in research, *The qualitative report,* 29(10), 2698-2715. https://doi.org/10.46743/2160-3715/20

Raman, K., Hashim, H. and Ismail, H. H., 2024, Exploring the impact of VR integration on ESL learners' English verbal communication skills: A case study in a Malaysian high school, *Arab world English journal special ıssue on CALL,* 10, 134–159. https://dx.doi.org/10.24093/awej/call10.10

Rambe, F. and Daulay, E., 2025, EFL undergraduate students' perspectives on the use of SmallTalk2Me to practice speaking skills. *Frasa: English education and literature journal,* 6(2), 260–269. https://doi.org/10.47701/frasa.v6i2. 5047

Restall, G., Yao, Y. and Niu, X., 2023, Exploring the experience of Year 10 South Korean students' English language learning in immersive virtual reality. *TESOL in context,* 31(2), 21-67. 10.21153/tesol2023vol31no2art1731

Richards, J. C. and Rodgers, T. S., 2001, *Approaches and methods in language teaching*, 2nd ed., Cambridge University Press.

Ritchie, S. M. and Rigano, D. L., 2001, Researcher–participant positioning in classroom research, *International journal of qualitative studies in education,* 14(6), 741–756. https://doi.org/10.1080/09518390110078413

Roberts, J. T., 2004, The communicative approach to language teaching: The king is dead! Long live the king! *International journal of English studies,* 4(1), 1-37.

Rodpradit, S. and Sinwongsuwat, K., 2012, Using scripted and non-scripted roleplays to improve English speaking performance of hotel engineering staff. *APHEIT journal*, 18(1), 17-25.

Rohima, A. and Pitaloka, N. L., 2025, Personality, motivation, and EFL speaking performance: Evidence from an English education program, *International journal of education research and development,* 5(2), 195–205. https://doi.org/10.52760/ijerd.v5i2.123

Romadhona, G., Munir, A. and Rizqiyah, A., 2023, Students' experiences toward scripted and unscripted role playing in junior high school, *Inspiring*, 133-150.

Rosenthal, R. and Rosnow, R. L., 1991, *Essentials of behavioral research: Methods and data analysis*, 2nd ed., New York: McGraw Hill.

Ruscica, P., Daxberger, H., Resch, G., Hadzovic, A., Dalili, S. and Arhonditsis, G. B., 2025, Transforming education and research with extended reality technologies: How virtual reality can shape the future of data interactions in earth and environmental sciences? *Ecological informatics*, 1-28. https://doi.org/10.1016/j.ecoinf.2025.103535

Ryan, A. B., 2006, *Post-positivist approaches to research*. In Researching and writing your thesis: A guide for postgraduate students, Maynooth Adult and Community Education, 12–26.

Saito, K. and Plonsky, L., 2019, Effects of second language pronunciation teaching revisited: A proposed measurement framework and meta-analysis. *Language learning,* 69(3), 652-708.

Santos, J. C. and Luque-Agulló, G, 2025, Beliefs and practices of Ecuadorian EFL preservice teachers about teaching speaking skills. *Profile: ıssues in teachers' professional development,* 27(2), 137–153. https://doi.org/10.15446/profile.v27n2.116833

Sasirekha, P., Devendraiah, K. M., Reddy, C. R. and Mohammad, R., 2024, Boosting EFL learners' speaking skills and motivation through dual graph neural networks, *International conference on distributed systems, computer networks and cybersecurity*, IEEE, 1-5.

Schratz, M. and Walker, R., 1995, *Research as social change: New opportunities for qualitative research*, Routledge. https://doi.org/10.4324/9780203014004

Schroeder, R., 1993. Virtual reality in the real world: History, applications and projections, *Futures*, 25, 963–973. https://doi.org/10.1016/0016-3287(93)90062-X

Schütze, U., 2024, *Virtual reality, artificial intelligence, and language learning: The need for attention*, John Benjamins. https://doi.org/10.1075/bpa.19. xvi+164.

Schwaiger, M., Krajnčan, M., Vukovič, M., Jenko, M. and Doz, D, 2024, Educators' opinions about VR/AR/XR: An exploratory study, *Education and information technologies,* 29*,* 24861–24880. https://doi.org/10.1007/s10639-024-12808-7

Schwandt, T. A., 2001, *Dictionary of qualitative inquiry*, 2nd ed., Sage Publications.

Seedhouse, P. and Nakatsuhara, F., 2018, *The discourse of the IELTS speaking test: Interactional design and practice,* Cambridge University Press.

Segar, T. and Asmawi, A., 2024, Exploring ESL students' use of multimedia in an English-speaking classroom, *Malaysian online journal of educational sciences,* 12(3), 61-75. https://doi.org/10.22452/aldad.vol12no3.5

Segvan, C. and Saeed, K. M., 2025, Perspectives and policy recommendations for EFL speaking skills in Rojava's secondary education, *TESOL today,* 2(1), 21–37. https://doi.org/10.53898/tesol2025212

Shrout, P. E. and Fleiss, J. L., 1979, Intraclass correlations: Uses in assessing rater reliability, *Psychological bulletin,* 86(2), 420–428.

Skehan, P., 2001, *Tasks and Language Performance Assessment*. In: Bygate M., Skehan P. and Swain M. (eds.), Researching Pedagogic Tasks Second Language Learning, Teaching and Testing, Pearson Education, 167-185.

Smith, M. and Loewen, S., 2018, Situational language teaching., *The TESOL encyclopedia of english language teaching,* 1-6. Doi:10.1002/9781118784235.eelt0174

Somé. K. J., 2025, Teachers' agency and teaching challenges in multilingual spaces: An exploratory study of how EFL teachers address listening and speaking skills needs in centralized systems, *Teaching English as a Second Language Electronic Journal,* 28(4). https://doi.org/10.55593/ej.28112a7

Soto-Lucena, I., 2024, *The use of immersive virtual environments to enhance language learning, improve learners' experience, and foster willingness to communicate: A study with ImmerseMe* [Unpublished doctoral dissertation]. University of Pittsburgh.

Stahl, N.A. and King, J.R., 2020, Expanding approaches for research: understanding and using trustworthiness in qualitative research, *Journal of developmental education,* 44, 26-28. http://www.jstor.org/stable/45381095

Stein, D., 1998, Situated learning in adult education, *ERIC Digests,* 195, 1-7

Stern, H. H., 1992, *Issues and options in language teaching,* Oxford: Oxford University Press.

Stewart, N., Chater, N. and Brown, G. D. A., 2006, Decision by sampling, *Cognitive psychology,* 53(1), 1–26. https://doi.org/10.1016/j.cogpsych.2005.10.003

Stokes, E. K., 2011, *Reliability: error in measurement.* In: Stokes E K. (ed.), Rehabilitation Outcome Measures, Churchill Livingstone, 27-34.

Stone, E, 2010, *t test, independent samples*. In: Salkind N. J. (ed.), Encyclopedia of research design, SAGE Publications, 1552–1556. https://doi.org/10.4135/9781412961288.n475

Sukmara, E., Friatin, L. Y. and Tarwana, W, 2025, Boosting EFL speaking skills through role play: Evidence from seventh grade students, *Journal of applied linguistics and literacy*, 9(2), 230–240.

Suratin, N. and Sribayak, V., 2025, Factors contributing to speaking anxiety and anxiety reduction techniques in Thai adult EFL learners, *LEARN journal: Language Education and Acquisition Research Network,* 18(1), 294-319. https://so04.tci-thaijo.org/index.php/LEARN/index

Sutarsyah, C., 2017, An analysis of student's speaking anxiety and its effect on speaking performance, *Indonesian journal of english language teaching and applied linguistics* 1(2), 143-152. https://doi.org/10.21093/ijeltal.v1i2.14

Swain, M., 1985, *Communicative competence: Some roles of comprehensible input and comprehensible output in its development*. In: Gass S. and Madden C. (eds.), Input in second language acquisition, Rowley, MA: Newbury House, 235-253.

Tafazoli, D., 2024, From virtual reality to cultural reality: Integration of virtual reality into teaching culture in foreign language education, *Journal for multicultural education,* 18(12), 6–24. https://doi.org/10.1108/JME-12-2023-0135

Taman, P., 2019, The validity and reliability of IELTS speaking rubric for inlingua international teachers, *Lexeme: Journal of linguistics and applied linguistics,* 1(1), 50-56. doi:10.32493/ljlal.v1i1.2483.

Tao, Y., Mofreh, S. A. M. and Salem, S., 2025, Examining the impact of identity on English speaking proficiency among Chinese undergraduate EFL learners: a structural equation modeling approach, *Eurasian journal of applied linguistics,* 11(1), 20-32. Doi: https:// doi.org/10.32601/ejal.11102

Taş, T. and Khan, Ö., 2020, On the models of communicative competence. *Proceedings of GLOBETS online: International conference on education, technology and science*, June 5–6 2020, ISBN: 978-605-82371-3-1.

Teddlie, C. B. and Tashakkori A., 2009, *Foundations of mixed methods research: Integrating quantitative and qualitative approaches in the social and behavioral sciences*, Sage.

Tran, M. T. N., Thach, L. S. and Nguyen, P. N. T., 2025, Effects of the two formats of authentic role-play on EFL young learners' speaking performance, European *journal of english language studies,* 5(1), 49-63. https://doi.org/10.12973/ejels.5.1.49

Tran, M., 2021, *The use of virtual reality technology in foreign language education: from teachers' perspectives* [Unpublished master's thesis], University of Oxford.

Vallarino, M., Cersosimo, R., Torre, I. and Vercelli, G., 2025, Effects of using depth and motion in virtual reality to enhance language learning, *Virtual reality,* 29, 1-22. https://doi.org/10.1007/s10055-025-01203-3

Van Lier, L. (1996). *Interaction in the language curriculum: Awareness, autonomy, and authenticity.* London: Longman

Vo, T.-V., Nguyen, M. and Le, H., 2018, Augmented reality on mobile platform: A new way to instantly view and display foreign currency exchange rate. *2018 International conference on image and vision computing New Zealand*, IEEE, 1–5. https://doi.org/10.1109/IVCNZ.2018.8634773

Waluyo, B. and Bakoko, R., 2022, Effects of affective variables and willingness to communicate on students' English-speaking performance in Thailand, S*tudies in english language and education, 9(1),* 45-61. https://doi.org/10.24815/siele.v9i1.21090.

Wang, L., 2024, A study on the application of virtual reality technology in English language education relying on immersive learning environments and interactive experience design, *Journal of electrical systems, 20*(6), 169–176. 10.52783/jes.2626

Wang, Q. and Li, Y., 2024, How virtual reality, augmented reality and mixed reality facilitate teacher education: A systematic review, *Journal of computer assisted learning,* 40, 1276-1294. 10.1111/jcal.12949.

Wang, W., Rezaei, Y. and Izadpanah, S., 2024, Speaking accuracy and fluency among EFL learners: the role of creative thinking, emotional ıntelligence, and academic enthusiasm, *Heliyon,* 10*,* 1-14. doi:10.1016/j.heliyon.2024.e37620.

Webb, N., 2017, *Variables, control*. In: Allen M. (ed.), The SAGE encyclopedia of communication research methods*,* SAGE Publications, 1842–1842. https://doi.org/10.4135/9781483381411.n658

Weigle, S. C., 1998, Using FACETS to model rater training effects, *Language testing*, 15(2), 263-287. https://doi.org/10.1177/026553229801500205

Weir, C. J., 2005, *Language testing and validation,* Palgrave Macmillan.

Weir, C. J.,1990, *Communicative language testing,* London: Prentice Hall International

Weiss, A. and Adams, M. J., 2010, *Personality, temperament, and behavioral syndromes*. In: Koob G. F., Le Moal M. and Thompson R. F. (eds.), Encyclopedia of behavioral neuroscience*,* Academic Press, 47–53. https://doi.org/10.1016/B978-0-08-045396-5.00117-2

Wesely, P.M., 2012, Learner attitudes, perceptions, and beliefs in language learning, *Foreign language annals,* 45*,* 98-117. http://dx.doi.org/10.1111/j.1944-9720.2012.01181.x

Wilkins, D. A., 1972, *Linguistics in language teaching*, Cambridge: MFT Press.

Wilkinson, A., 1968, *The testing of oracy.* In: Davies A. (ed.) Language Testing Symposium*,* Oxford: Oxford University Press, 117-32.

Williams, M. and Burden, R. L., 1999, Students developing conceptions of themselves as language learners, *Modern language journal,* 83, 193–201.

Willis, D. and Willis, J., 2001, *Task-based language learning*. In: Carter R. and Nunan D. (eds.), The Cambridge guide to teaching English to speakers of other languages, Cambridge University Press, 173–179.

Willis, D., 1990, *The lexical syllabus,* London, United Kingdom, HarperCollins.

Willis, J., 1996, *A Framework for task-based learning.* Harlow: Longman.

Winke, P., Gass, S. and Myford, C., 2013, Raters' L2 background as a potential source of bias in rating oral performance, *Language testing,* 30(2), 231–252.

Wolcott, H. F., 1990, *Writing up qualitative research*. Sage Publications.

Wu, J. G., Zhang, D., Lee, S.-M. and Xian, J., 2025, Game on! Digital gaming and augmented reality/virtual reality in language learning, *International journal of computer-assisted language learning and teaching,* 15(1), 1-15. https://doi.org/10.4018/IJCALLT.367601

Wu, Y.-H. S. and Hung, S.-T. A., 2022, The effects of virtual reality–infused instruction on elementary school students' English-speaking performance, willingness to communicate, and learning autonomy, *Journal of educational computing research,* 60(6), 1558–1587. https://doi.org/10.1177/07356331211068207

Yan, W., Lowell, V. and Yang, L., 2023, Exploring the impact of virtual reality on developing EFL learners' speaking skills in situated learning, *Research square,* 1–22. https://doi.org/10.21203/rs.3.rs-3711927/v1

Yan, W., Lowell, V. L. and Yang, L., 2024, Developing English language learners' speaking skills through a situated learning approach in VR-enhanced learning experiences, *Virtual reality,* 28, 1-23. https://doi.org/10.1007/s10055-024-01061-5

Yudintseva, A., 2024, An exploration of low- and high-immersive virtual reality modalities for willingness to communicate in English as a second language, *Computers and education: X reality 5*, 1-19. https://doi.org/10.1016/j.cexr.2024.100076

Zhang, C., Schießl, J., Plößl, L., Hofmann, F. and Gläser-Zikuda, M., 2023, Acceptance of artificial intelligence among pre-service teachers: a multigroup analysis, *International journal of educational technology in higher education,* 20(4), 1-22. https://doi.org/10.1186/s41239-023-00420-7

Zhang, D., 2025, Construction of immersive English listening and speaking training platform in virtual reality environment and its effect evaluation system, *2025 International conference on electronics, energy systems and power engineering* IEEE, 1162–1166. https://doi.org/10.1109/EESPE63401.2025.10987535

Zhang, Q., 2025, Immersive AI-powered language learning experience in virtual reality: A gamified environment for Japanese learning, *2025 Conference on virtual reality and 3D user interfaces abstracts and workshops,* IEEE, 1620–1621. https://doi.org/10.1109/VRW66409.2025.00455

Zheng, C., 2020, Situated learning as legitimate peripheral participation, *Philos study* 10(10), 649–653. https://doi.org/10.17265/215 9-5313/2020.10.006

Zuengler., J. and Miller, E. R., 2006, Cognitive and sociocultural perspectives: two parallel SLA worlds? *TESOL quarterly,* 40(1), 35–58. https://w ww.jstor.org/stable/40264510

APPENDICES

APPENDIX-1: DISTRIBUTION OF B1 AND B1+ LEVEL DESCRIPTORS IN LEVEL 4 COURSEBOOKS

Subscale	Level	Descriptors	Empower B1+ Total	Prism 2 Total	Sum
SUSTAINED MONOLOGUE: Describing Experience	**B1+**	Can clearly express feelings about something experienced and give reasons to explain those feelings.	7	3	10
SUSTAINED MONOLOGUE: Describing Experience	**B1**	Can give straightforward descriptions on a variety of familiar subjects within their field of interest.	4	4	8
SUSTAINED MONOLOGUE: Describing Experience	**B1**	Can reasonably fluently relate a straightforward narrative or description as a linear sequence of points. Can give detailed accounts of experiences, describing feelings and reactions.	7	1	8
SUSTAINED MONOLOGUE: Describing Experience	**B1**	Can describe dreams, hopes and ambitions.	1	2	3
SUSTAINED MONOLOGUE: Describing Experience	**B1**	Can describe events, real or imagined.	5	0	5
SUSTAINED MONOLOGUE: Describing Experience	**B1**	Can narrate a story.	6	0	6
SUSTAINED MONOLOGUE: Giving information	**B1+**	Can explain the main points in an idea or problem with reasonable precision.	1	6	7

SUSTAINED MONOLOGUE: Giving information	**B1**	Can report straightforward factual information on a familiar topic, for example to indicate the nature of a problem or to give detailed directions, provided they can prepare beforehand.	6	6	12
SUSTAINED MONOLOGUE: Putting a Case	**B1+**	Can develop an argument well enough to be followed without difficulty most of the time.	0	6	6
SUSTAINED MONOLOGUE: Putting a Case	**B1+**	Can give simple reasons to justify a viewpoint on a familiar topic.	5	7	12
SUSTAINED MONOLOGUE: Putting a Case	**B1**	Can briefly give reasons and explanations for opinions, plans and actions.	9	0	9
SUSTAINED MONOLOGUE: Putting a Case	**B1**	Can express opinions on subjects relating to everyday life, using simple expressions.	7	0	7
SUSTAINED MONOLOGUE: Putting a Case	**B1**	Can say whether or not they approve of what someone has done and give reasons to justify their opinion.	1	0	1

APPENDIX-2: SAMPLE CEFR-BASED ANALYSIS OF SPEAKING ACTIVITIES IN EMPOWER B1+ COURSEBOOK

<table>
<tr><td colspan="4">Empower B1+ - Unit 2A</td></tr>
<tr><td colspan="4">Objectives relating to speaking skills:
Use a lexical set of work-related words correctly
Use the present perfect and past simple to talk about experiences or past events
Talk about work and training experiences</td></tr>
<tr><td>Textbook question (p.22)</td><td>Related CEFR descriptor</td><td>CEFR level</td><td>CEFR subscale/ category</td></tr>
<tr><td rowspan="3">“You are going to talk about your experiences. Tick three things you have done.
• Got qualifications (which?)
• Learned practical skills (what?)
• Studied worked in a foreign country (where?)
• Chosen a career
• Worked for no money
• Visited an interesting office or factory
• Written a CV
• Given a talk or presentation
• Done some online learning
• Studied or worked as part of a team
• Been in charge of a project
Then make notes about your experiences. Take turns to talk about your work and training experiences. Ask question to find out more information.” (p.22)</td><td>“Can give detailed accounts of experiences, describing feelings and reactions” (p.62)
“Can describe events, real or imagined” (p.62)
“Can narrate a story” (p.62)</td><td>B1</td><td rowspan="2">Sustained monologue: describing experience.</td></tr>
<tr><td>“Can clearly express feelings about something experienced and give reasons to explain those feelings” (p.62)
“Can give straightforward descriptions on a variety of familiar subjects within their field of interest” (p.62)</td><td>B1+</td></tr>
<tr><td>“Can express opinions on subjects relating to everyday life, using simple expressions.” (p.64)
“Can briefly give reasons and explanations for opinions, plans and actions.” (p.64)</td><td>B1</td><td>Sustained monologue: putting a case.</td></tr>
</table>

	"Can report straightforward factual information on a familiar topic, for example to indicate the nature of a problem or to give detailed directions, provided they can prepare beforehand." (p.63)	B1	Sustained monologue: giving information.

Empower B1+ - Unit 6B			
Objectives relating to speaking skills: Use a lexical set to describe food and talk about preparing food Use comparative and superlative adjectives correctly Talk about a special meal that they have made or eaten			
Textbook question (p.73)	Related CEFR descriptor	CEFR level	CEFR subscale/ category
You are going to talk about a special meal. Make notes about one of these meals. Use the ideas in the box to help you plan what to say. • The most special meal you've ever made • The most delicious meal you have ever eaten • A meal you'll never forget Where? When? Who with? Ingredients? How was the food	"Can give detailed accounts of experiences, describing feelings and reactions" (p.62) "Can describe events, real or imagined" (p.62) "Can narrate a story" (p.62)	B1	Sustained monologue: describing experience.
	"Can clearly express feelings about something experienced and give reasons to explain those feelings" (p.62)	B1+	
	"Can express opinions on subjects relating to everyday life, using simple expressions." (p.64) "Can briefly give reasons and explanations for opinions, plans and actions." (p.64)	B1	Sustained monologue: putting a case.
	"Can give simple reasons to justify a viewpoint on a familiar topic." (p.64)	B1+	

cooked? Taste, smell, colour? Take turns to describe your meals. Then talk about which of the meals sounds the most delicious.	"Has enough language to get by, with sufficient vocabulary to express themselves with some hesitation and circumlocutions on topics such as family, hobbies and interests, work, travel and current events, but lexical limitations cause repetition and even difficulty with formulation at times." (p.130)	B1	Linguistic competence: General linguistic range

Empower B1+ - Unit 10B			
Objectives relating to speaking skills: Talk about a time when something good happened to them Use a lexical set of expressions with do, make and take Use the third conditional appropriately Talk about a past event that made their lives better			
Textbook question (p.121)	Related CEFR descriptor	CEFR level	CEFR subscale/ category
You're going to tell the story of a past even that made your life better. Make notes about one of the topics below. • A good friend, and how you met them.	"Can give detailed accounts of experiences, describing feelings and reactions" (p.62) "Can describe events, real or imagined" (p.62) "Can narrate a story" (p.62)	B1	Sustained monologue: describing experience.
	"Can clearly express feelings about something experienced and give reasons to explain those feelings" (p.62)	B1+	

• A sport or hobby and how you started doing it. • An accident and how it happened. • A job and how you got it. • A school and why you went there. • A big decision and how you made it. Think about the important events in your story. What were the consequences of what happened? How would your life have been different if you had done something differently? Take turns to tell your stories. Ask each other questions to find out more about what might have been happened if things had been different.	"Can express opinions on subjects relating to everyday life, using simple expressions." (p.64) "Can briefly give reasons and explanations for opinions, plans and actions." (p.64)	B1	Sustained monologue: putting a case.
	"Can give simple reasons to justify a viewpoint on a familiar topic." (p.64)	B1+	

APPENDIX-3: SAMPLE CEFR-BASED ANALYSIS OF SPEAKING ACTIVITIES IN PRISM 2 COURSEBOOK

Unit 2 Reading 2				
Critical thinking skill	Textbook question (p.45)	Related CEFR descriptor	CEFR level	CEFR subscale/ category
Understand	"As the world's climate changes, which places will have too much water? Which places will become drier?"	"Can report straightforward factual information on a familiar topic" (p.63).	B1	Sustained monologue: giving information
		"Can express opinions on subjects relating to everyday life, using simple expressions." (p.64) "Can briefly give reasons and explanations for opinions, plans and actions." (p.64)	B1	Sustained monologue: putting a case.
		"Can give straightforward descriptions on a variety of familiar subjects within their field of interest." (p.62)	B1	Sustained monologue: describing experience.
Evaluate	"How do both the melting of glaciers and deforestation contribute to the extinction of species?"	"Can explain the main points in an idea or problem with reasonable precision." (p.63)	B1+	Sustained monologue: giving information
		"Can develop an argument well enough to be followed without difficulty most of the time." (p.64) "Can give simple reasons to justify a viewpoint on a familiar topic." (p.64)	B1+	Sustained monologue: putting a case.

Unit 4 Reading 1				
Critical thinking skill	Textbook question (p.75)	Related CEFR descriptor	CEFR level	CEFR subscale/ category
Understand	"Which customs are similar in Brazil, Japan and India? Which are different?"	Can report straightforward factual information on a familiar topic" (p.63)	B1	Sustained monologue: giving information.
Apply	"Which of these countries would you most like to visit? Why?"	"Can give simple reasons to justify a viewpoint on a familiar topic." (p.64)	B1	Sustained monologue: putting a case.
Evaluate	"Have you ever made a cultural mistake in a foreign country? Describe it."	"Can give detailed accounts of experiences, describing feelings and reactions" (p.62) "Can describe events, real or imagined" (p.62) "Can narrate a story" (p.62) "Can give straightforward descriptions on a variety of familiar subjects within their field of interest." (p.62) "Can describe dreams, hopes and ambitions." (p.62)	B1	Sustained monologue: describing experience.
		"Can clearly express feelings about something experienced and give	B1+	

		reasons to explain those feelings" (p.62)		

Unit 6 Reading 1				
Critical thinking skill	Textbook question (p.111)	Related CEFR descriptor	CEFR level	CEFR subscale/ category
Apply	"What are the advantages of copying from nature?"	"Can report straightforward factual information on a familiar topic" (p.63).	B1	Sustained monologue: giving information
		"Can explain the main points in an idea or problem with reasonable precision." (p.63)	B1+	
		"Can give simple reasons to justify a viewpoint on a familiar topic." (p.64)	B1	Sustained monologue: putting a case.
Analyze	"Do you think biomimicry will be more common in the future? Why or why not?"	"Can explain the main points in an idea or problem with reasonable precision." (p.63)	B1+	Sustained monologue: giving information
		"Can develop an argument well enough to be followed without difficulty most of the time." (p.64)	B1+	Sustained monologue: putting a case.

APPENDIX-4: DEMOGRAPHIC INFORMATION SURVEY OF EXPERIMENT GROUP

1. What is your age?

□18-20
□21-23
□24-26
□27+

2. What is your gender?

□Female
□Male
□Prefer not to say
Other: _____

3. What is your nationality?

_____

4. What is your native language?

_____

5. What is your department that you will start to study next year?

_____

6. How many years have you been studying English?

□Less than 5 years
□5-10 years
□More than 10 years

7. Have you ever used desktop virtual reality (VR) or similar technologies before this study?

□Yes
□No

8. How comfortable are you with using technology in general?

	1	2	3	4	5	
Not comfortable at all	○	○	○	○	○	Very comfortable

APPENDIX-5: DEMOGRAPHIC INFORMATION SURVEY OF CONTROL GROUP

1. What is your age?
□18-20
□21-23
□24-26
□27+

2. What is your gender?
□Female
□Male
□Prefer not to say
Other: _____

3. What is your nationality?
_____

4. What is your native language?
_____

5. What is your department that you will start to study next year?
_____

6. How many years have you been studying English?
□Less than 5 years
□5-10 years
□More than 10 years

APPENDIX-6: PRE- AND POST-TEST SPEAKING QUESTIONS

Home Town:

1. Can you describe your hometown? What do you like most about it?
2. How has your hometown changed since you were a child?
3. Are there any famous landmarks in your hometown?
4. What activities or events are popular in your hometown?
5. How would you recommend someone to spend a day visiting in your hometown?
6. What local food or dish would you suggest someone to try when they are visiting your hometown?

Family and Home:

1. Tell me about your family. Do you have any siblings? What are they like?
2. What is your favourite family tradition or celebration?
3. Describe your home. What do you like most about it?
4. How do you and your family spend time together on weekends?
5. Do you have any specific responsibilities at home?
6. What changes would you like to make to your home in the future?

Work:

1. What are you currently studying? Can you describe what a typical day looks like for you?
2. Why did you choose your current field of study?
3. How do you balance your studies with your personal life and hobbies?
4. What skills do you think are important for success in your field of study?
5. Describe a challenging situation you've faced during your studies. How did you deal with it?
6. Do you have any specific goals related to your academic journey or future career?

Leisure:

1. What hobbies or activities do you enjoy in your free time?
2. Can you share a special moment from a recent leisure activity?
3. How do you usually spend your weekends or days off?
4. Do you prefer indoor or outdoor activities for leisure? Why?
5. Are there any specific places you like to visit for relaxation or entertainment?
6. If you could try a new hobby or activity, what would it be and why?

Future Plans:

1. Where do you see yourself in five years? What are your future plans?
2. Are there any specific career or educational goals you want to achieve?
3. How do you plan to prepare to achieve your future goals?
4. What factors will influence your decisions about your future?
5. Do you have any dreams that you hope to achieve in the long term?

6. How would you like your life to be different in the future compared to now?

Likes and Dislikes

1. What are your favorite hobbies or activities in your free time? Why do you enjoy them?
2. Can you share a book, movie, or TV show that you really like? What do you find interesting about it?
3. When it comes to music, do you have a favorite genre or specific artists you enjoy listening to?
4. Describe a type of food that you love. What makes it your favorite?
5. Is there a sport or physical activity that you find exciting or enjoyable? Why?
6. Discuss something you dislike, whether it's a food, activity, or any other thing. What don't you like about it?

Friendship

1. Can you describe your best friend? What qualities make them a good friend?
2. How do you usually make new friends? Are you outgoing or more reserved in social situations?
3. Share a special moment you had with a friend. What made it special?
4. What do you think is the most important quality in a friendship? Why?
5. Discuss a challenge you faced in a friendship. How did you solve the problem?
6. If you could choose one famous person to be your friend, who would it be and why?

Travelling

1. Have you been to any interesting places recently? Tell me about your last trip.
2. What type of travel do you prefer – by car, train, or plane? Why?
3. Can you share a special moment from a holiday you took in the past?
4. If you could visit any country in the world, where would you go and why?
5. How do you usually plan your trips? Do you like to have a detailed plan or go with the flow?
6. Describe a travel adventure or activity you would like to try in the future.

Social Media

1. How often do you use social media, and which platforms do you prefer?
2. Can you describe a positive experience you've had on social media?
3. Did you face any problems while using social media?
4. Do you think social media is a good way to stay connected with friends and family? Why or why not?
5. What are your thoughts on the impact of social media on daily life and well-being?
6. If you had to take a break from social media for a week, how do you think it would affect you?

APPENDIX-7: IELTS SPEAKING RUBRIC

IELTS Speaking Band Descriptors (public version)

Band	Fluency and coherence	Lexical resource	Grammatical range and accuracy	Pronunciation
9	• speaks fluently with only rare repetition or self-correction; any hesitation is content-related rather than to find words or grammar • speaks coherently with fully appropriate cohesive features • develops topics fully and appropriately	• uses vocabulary with full flexibility and precision in all topics • uses idiomatic language naturally and accurately	• uses a full range of structures naturally and appropriately • produces consistently accurate structures apart from 'slips' characteristic of native speaker speech	• uses a full range of pronunciation features with precision and subtlety • sustains flexible use of features throughout • is effortless to understand
8	• speaks fluently with only occasional repetition or self-correction; hesitation is usually content-related and only rarely to search for language • develops topics coherently and appropriately	• uses a wide vocabulary resource readily and flexibly to convey precise meaning • uses less common and idiomatic vocabulary skilfully, with occasional inaccuracies • uses paraphrase effectively as required	• uses a wide range of structures flexibly • produces a majority of error-free sentences with only very occasional inappropriacies or basic/non-systematic errors	• uses a wide range of pronunciation features • sustains flexible use of features, with only occasional lapses • is easy to understand throughout; L1 accent has minimal effect on intelligibility
7	• speaks at length without noticeable effort or loss of coherence • may demonstrate language-related hesitation at times, or some repetition and/or self-correction • uses a range of connectives and discourse markers with some flexibility	• uses vocabulary resource flexibly to discuss a variety of topics • uses some less common and idiomatic vocabulary and shows some awareness of style and collocation, with some inappropriate choices • uses paraphrase effectively	• uses a range of complex structures with some flexibility • frequently produces error-free sentences, though some grammatical mistakes persist	• shows all the positive features of Band 6 and some, but not all, of the positive features of Band 8
6	• is willing to speak at length, though may lose coherence at times due to occasional repetition, self-correction or hesitation • uses a range of connectives and discourse markers but not always appropriately	• has a wide enough vocabulary to discuss topics at length and make meaning clear in spite of inappropriacies • generally paraphrases successfully	• uses a mix of simple and complex structures, but with limited flexibility • may make frequent mistakes with complex structures, though these rarely cause comprehension problems	• uses a range of pronunciation features with mixed control • shows some effective use of features but this is not sustained • can generally be understood throughout, though mispronunciation of individual words or sounds reduces clarity at times
5	• usually maintains flow of speech but uses repetition, self-correction and/or slow speech to keep going • may over-use certain connectives and discourse markers • produces simple speech fluently, but more complex communication causes fluency problems	• manages to talk about familiar and unfamiliar topics but uses vocabulary with limited flexibility • attempts to use paraphrase but with mixed success	• produces basic sentence forms with reasonable accuracy • uses a limited range of more complex structures, but these usually contain errors and may cause some comprehension problems	• shows all the positive features of Band 4 and some, but not all, of the positive features of Band 6
4	• cannot respond without noticeable pauses and may speak slowly, with frequent repetition and self-correction • links basic sentences but with repetitious use of simple connectives and some breakdowns in coherence	• is able to talk about familiar topics but can only convey basic meaning on unfamiliar topics and makes frequent errors in word choice • rarely attempts paraphrase	• produces basic sentence forms and some correct simple sentences but subordinate structures are rare • errors are frequent and may lead to misunderstanding	• uses a limited range of pronunciation features • attempts to control features but lapses are frequent • mispronunciations are frequent and cause some difficulty for the listener
3	• speaks with long pauses • has limited ability to link simple sentences • gives only simple responses and is frequently unable to convey basic message	• uses simple vocabulary to convey personal information • has insufficient vocabulary for less familiar topics	• attempts basic sentence forms but with limited success, or relies on apparently memorised utterances • makes numerous errors except in memorised expressions	• shows some of the features of Band 2 and some, but not all, of the positive features of Band 4
2	• pauses lengthily before most words • little communication possible	• only produces isolated words or memorised utterances	• cannot produce basic sentence forms	• speech is often unintelligible
1	• no communication possible • no rateable language			
0	• does not attend			

APPENDIX-8: IRONSI'S (2023) PERCEPTION SCALE (P.18)

Questionnaire

Instructions: Please read the instructions carefully before answering the questions. Use the rating scale provided to assess the level of agreement with the statements on using Virtual Reality to improve speaking skills. Use the boxes to indicate your responses and tick the boxes that best apply to you.

Strongly Agree (5)	Agree (4)	Neutral (3)	Disagree (2)	Strongly Disagree (1)

Items	**5**	**4**	**3**	**2**	**1**
VR is not effective in foreign language acquisition classrooms					
VR enhances learning and speaking					
VR does not improve speaking skills					
VR introduces virtual real-life situations helpful in speaking					
Teaching through VR improved classroom performance in speaking					
The use of VR improved my speaking grades					
VR makes speaking lessons boring					
It is not difficult for me to handle and use VR during speaking lessons					
VR gadgets are expensive to purchase					
VR is efficient in learning to speak					
VR makes learning to speak easy and fun					
VR helps to achieve lesson objectives on speaking					
I am concerned about privacy and security issues					

APPENDIX-9: TURKISH TRANSLATION OF IRONSI'S (2023) PERCEPTION SCALE

Not. Çoktan seçmeli maddelerde öğrencilerin yalnızca bir seçeneği işaretlemesine izin verilmiştir.

1. VR, yabancı dil derslerinde sınıf içi kullanımda etkili değildir.
☐ Kesinlikle Katılıyorum
☐ Katılıyorum
☐ Kararsızım
☐ Katılmıyorum
☐ Kesinlikle Katılmıyorum

2. VR, öğrenmeyi ve konuşmayı geliştirir.
☐ Kesinlikle Katılıyorum
☐ Katılıyorum
☐ Kararsızım
☐ Katılmıyorum
☐ Kesinlikle Katılmıyorum

3. VR, konuşma becerilerini geliştirmez.
☐ Kesinlikle Katılıyorum
☐ Katılıyorum
☐ Kararsızım
☐ Katılmıyorum
☐ Kesinlikle Katılmıyorum

4. VR, gerçek hayatta karşılaşılabilecek durumları sanal olarak sunarak konuşma becerisinin geliştirilmesine yardımcı olur.
☐ Kesinlikle Katılıyorum
☐ Katılıyorum
☐ Kararsızım
☐ Katılmıyorum
☐ Kesinlikle Katılmıyorum

5. Derslerde VR kullanmak, öğrencilerin sınıf içi konuşma performansını geliştirdi.
☐ Kesinlikle Katılıyorum
☐ Katılıyorum
☐ Kararsızım
☐ Katılmıyorum
☐ Kesinlikle Katılmıyorum

6. VR kullanımı, konuşma becerisiyle ilgili notlarımı yükseltti.
☐ Kesinlikle Katılıyorum
☐ Katılıyorum
☐ Kararsızım
☐ Katılmıyorum
☐ Kesinlikle Katılmıyorum

7. VR, konuşma derslerini sıkıcı hale getirir.
□ Kesinlikle Katılıyorum
□ Katılıyorum
□ Kararsızım
□ Katılmıyorum
□ Kesinlikle Katılmıyorum

8. Konuşma derslerinde VR'ı kullanmak ve kontrol etmek benim için zor değil.
□ Kesinlikle Katılıyorum
□ Katılıyorum
□ Kararsızım
□ Katılmıyorum
□ Kesinlikle Katılmıyorum

9. VR cihazlarını satın almak maliyetlidir.
□ Kesinlikle Katılıyorum
□ Katılıyorum
□ Kararsızım
□ Katılmıyorum
□ Kesinlikle Katılmıyorum

10. VR, konuşmayı öğrenmede verimlidir.
□ Kesinlikle Katılıyorum
□ Katılıyorum
□ Kararsızım
□ Katılmıyorum
□ Kesinlikle Katılmıyorum

11. VR, konuşmayı öğrenmeyi kolay ve eğlenceli hale getirir.
□ Kesinlikle Katılıyorum
□ Katılıyorum
□ Kararsızım
□ Katılmıyorum
□ Kesinlikle Katılmıyorum

12. VR, konuşma derslerindeki hedeflere ulaşmaya yardımcı olur.
□ Kesinlikle Katılıyorum
□ Katılıyorum
□ Kararsızım
□ Katılmıyorum
□ Kesinlikle Katılmıyorum

13. Gizlilik ve güvenlik konularında endişeliyim.
□ Kesinlikle Katılıyorum
□ Katılıyorum
□ Kararsızım
□ Katılmıyorum
□ Kesinlikle Katılmıyorum

APPENDIX-10: ORIGINAL VERSION OF THE SEMI-STRUCTURED QUESTIONNAIRE

Note. Students were only allowed to select one option for the multiple-choice items.

1. How would you describe your overall experience with virtual reality (VR) in your English speaking lessons? Why?

2. What did you enjoy most about using VR in your speaking lessons? Why?

3. Were there any aspects of the VR experience that you found challenging, uncomfortable or less effective? If so, what were they?

4. What do you think are the biggest advantages of using VR for English speaking practice? Why?

5. What do you think are the biggest challenges of using VR for English speaking practice? Why?

6. Do you think using VR improved your English-speaking skills?
   Yes → How do you think it improved your ability to speak English
   No → Why do you think it didn't have an impact?

7. To what extent do you think VR has affected your English speaking skills overall?
   □ Strongly agree: VR has significantly improved my ability to speak English more fluently and confidently
   □ Agree: VR has helped me speak English more fluently and confidently to some extent.
   □ Neutral: I'm not sure if VR has had a clear impact on my English speaking skills.
   □ Disagree: VR has had little impact on my English speaking skills.
   □ Strongly disagree: VR has made no difference in my English speaking skills.

8. Can you explain how VR has specifically helped or not helped improve your English speaking skills overall?

9. To what extent do you think VR has improved your fluency in English?
   □ Strongly agree: VR has improved my fluency to a great extent.
   □ Agree: VR has improved my fluency to some extent.
   □ Neutral: I'm not sure if my fluency has improved with VR.
   □ Disagree: VR has only minimally improved my fluency.
   □ Strongly disagree: VR has not improved my fluency at all.

10. Can you describe specific ways in which VR helped or didn't help improve your fluency in English?

11. To what extent do you think VR has expanded your vocabulary in English?
    - ☐ Strongly agree: VR has greatly expanded my vocabulary in English.
    - ☐ Agree: VR has somewhat expanded my vocabulary in English.
    - ☐ Neutral: I'm not sure if my vocabulary has improved with VR.
    - ☐ Disagree: VR has only slightly expanded my vocabulary.
    - ☐ Strongly disagree: VR has not expanded my vocabulary in English at all.

12. Can you describe specific ways in which VR helped or didn't help improve your vocabulary in English?

13. To what extent do you think VR has helped you improve your grammar in English?
    - ☐ Strongly agree: VR helped me improve my grammar to a great extent.
    - ☐ Agree: VR helped me improve my grammar, but not significantly.
    - ☐ Neutral: I'm not sure if my grammar has improved with VR.
    - ☐ Disagree: VR had very little effect on improving my grammar.
    - ☐ Strongly disagree: VR did not help me improve my grammar at all.

14. Can you describe specific ways in which VR helped or didn't help improve your grammar in English?

15. To what extent do you think VR has enhanced your pronunciation in English?
    - ☐ Strongly agree: VR significantly enhanced my pronunciation in English.
    - ☐ Agree: VR somewhat improved my pronunciation.
    - ☐ Neutral: I'm not sure if my pronunciation improved with VR
    - ☐ Disagree: VR had little effect on my pronunciation.
    - ☐ Strongly disagree: VR did not improve my pronunciation at all.

16. Can you describe specific ways in which VR helped or didn't help improve your pronunciation in English?

17. Can you describe a specific moment during a VR session when you felt your speaking performance improved? What made that moment stand out?
    If you didn't have such a moment, what was the reason?

18. Do you think VR helped you practice speaking English in ways that traditional classroom methods did not?
    Yes → How did it differ?
    No → Why do you feel there was no difference?

19. What do you think could be improved in the VR experience to make it more effective for enhancing English speaking skills? Why?

20. Would you recommend using VR in future English speaking classes? Why or why not?

21. If you had the chance to design a VR activity for English speaking practice, what features would you include to make it more engaging and effective? Why

22. Would you like to share any additional thoughts on your experience using VR to practice and improve your English-speaking skills?

23. Did VR make speaking English more enjoyable and engaging? Why or why not?

24. In your opinion, did VR make it easier or harder to stay engaged and interested in practicing English speaking? Why?

25. Did VR make you feel more or less comfortable when speaking in English? Why

26. Did VR help you overcome any fears or anxieties about speaking English? Why or why not?

27. How confident were you in your speaking abilities during the VR sessions compared to other classroom activities? Please explain.

28. How confident do you feel speaking English after participating in the virtual reality sessions?
    - □ Very confident, I feel much more capable.
    - □ Somewhat confident, I feel a little more capable than before
    - □ Neutral, I feel the same as before the sessions.
    - □ Not very confident, I feel slightly unsure when speaking English.
    - □ Not confident at all, I still feel very unsure when speaking English

29. Did the VR activities help you speak more naturally or more quickly in English?
    - □ Strongly Agree: Yes, I spoke more naturally and quickly
    - □ Agree: Yes, I spoke more naturally, but not faster.
    - □ Neutral: I'm not sure if there was a significant difference in speed or naturalness of my speech.
    - □ Disagree: No, I spoke more slowly and less naturally
    - □ Strongly Disagree: No, I did not notice any improvement in speed or naturalness.

30. How confident do you feel about using new vocabulary in real-life situations after practicing in VR?
    - □ Strongly agree: I feel very confident about using new vocabulary in real-life situations after practicing in VR.
    - □ Agree: I feel somewhat confident about using new vocabulary in real-life situations

after practicing in VR.

□ Neutral: I am neither confident nor unconfident about using new vocabulary in real-life situations after practicing in VR.

□ Disagree: I don't feel very confident about using new vocabulary in real-life situations after practicing in VR.

□ Strongly disagree: I don't feel confident at all about using new vocabulary in real-life situations after practicing in VR

31. During the VR activities, did you become more aware of your grammar mistakes?
    □ Strongly agree: I was very aware of my mistakes and corrected them.
    □ Agree: I was aware of my mistakes, but I didn't focus on correcting them.
    □ Neutral: I neither noticed nor ignored my grammar mistakes.
    □ Disagree: I was somewhat unaware of my mistakes, but I noticed a few.
    □ Strongly disagree: I did not notice any grammar mistakes at all.

32. Did practicing pronunciation in VR help you feel more confident speaking English in real-life situations?
    □ Strongly agree: Yes, I feel much more confident.
    □ Agree: Yes, I feel somewhat more confident.
    □ Neutral: I'm unsure about how it affected my confidence.
    □ Disagree: No, I don't feel more confident, but it didn't affect my confidence negatively.
    □ Strongly disagree: No, I feel less confident or more unsure than before.

APPENDIX-11: TURKISH TRANSLATION OF THE SEMI-STRUCTURED QUESTIONNAIRE

Not. Çoktan seçmeli maddelerde öğrencilerin yalnızca bir seçeneği işaretlemesine izin verilmiştir.

1. İngilizce konuşma derslerinizde sanal gerçeklik (VR) kullanımınızla ilgili deneyiminizi genel olarak nasıl tanımlardınız? Neden?

2. Konuşma derslerinizde VR kullanmayla ilgili en çok keyif aldığınız şey ne oldu? Neden?

3. VR deneyiminin zorlayıcı, rahatsız edici veya etkisiz bulduğunuz herhangi bir yönü oldu mu? Olduysa, bunlar nelerdi?

4. İngilizce konuşma pratiği yapmak için VR kullanımının en büyük avantajları sizce nelerdir? Neden?

5. İngilizce konuşma pratiği yapmak için VR kullanımının en büyük zorlukları sizce nelerdir? Neden?

6. Sizce VR kullanmak İngilizce konuşma becerilerinizi geliştirdi mi?
   Evet → İngilizce konuşma yeteneğinizi nasıl geliştirdiğini düşünüyorsunuz?
   Hayır → Bunun neden etkisiz olduğunu düşünüyorsunuz?

7. Sizce VR, genel olarak İngilizce konuşma becerilerinizi ne kadar etkiledi?
   □ Kesinlikle katılıyorum: VR, İngilizceyi daha akıcı ve özgüvenli konuşma yeteneğimi önemli ölçüde geliştirdi.
   □ Katılıyorum: VR, İngilizceyi daha akıcı ve özgüvenli konuşmamda biraz yardımcı oldu.
   □ Kararsızım: VR'nin İngilizce konuşma becerilerim üzerinde net bir etkisi olup olmadığından emin değilim.
   □ Katılmıyorum: VR'nın İngilizce konuşma becerilerim üzerinde çok az etkisi oldu.
   □ Kesinlikle katılmıyorum: VR İngilizce konuşma becerilerimde hiçbir fark yaratmadı.

8. VR'nin genel olarak İngilizce konuşma becerilerinizi geliştirmenize nasıl yardımcı olduğunu veya olmadığını açıklayabilir misiniz?

9. Sizce VR, İngilizce konuşmadaki akıcılığınızı ne ölçüde geliştirdi?
   □ Kesinlikle katılıyorum: VR, akıcılığımı büyük ölçüde geliştirdi.
   □ Katılıyorum: VR, akıcılığımı biraz geliştirdi.
   □ Kararsızım: VR'nin akıcılığımı geliştirip geliştirmediğinden emin değilim.
   □ Katılmıyorum: VR, akıcılığımı çok az geliştirdi.
   □ Kesinlikle katılmıyorum: VR, akıcılığımı hiç geliştirmedi.
10. VR'nin İngilizce konuşurkenki akıcılığınızı özellikle hangi açılardan geliştirdiğini veya geliştirmediğini açıklayabilir misiniz?

11. Sizce VR, İngilizce kelime dağarcığınızı ne ölçüde geliştirdi?
    □ Kesinlikle katılıyorum: VR, İngilizce kelime dağarcığımı büyük ölçüde geliştirdi.

- □ Katılıyorum: VR, İngilizce kelime dağarcığımı biraz geliştirdi.
- □ Kararsızım: VR'nin kelime dağarcığımı geliştirip geliştirmediğinden emin değilim.
- □ Katılmıyorum: VR, kelime dağarcığımı çok az geliştirdi.
- □ Kesinlikle katılmıyorum: VR, İngilizce kelime dağarcığımı hiç geliştirmedi.

12. VR'nin İngilizce kelime dağarcığınızı özellikle hangi açılardan geliştirdiğini veya geliştirmediğini açıklayabilir misiniz?

13. Sizce VR, İngilizce dilbilginizi ne ölçüde geliştirdi?
    - □ Kesinlikle katılıyorum: VR, dilbilgimi büyük ölçüde geliştirdi.
    - □ Katılıyorum: VR, dilbilgimi geliştirmeme yardımcı oldu, ancak çok fazla değil.
    - □ Kararsızım: VR'nin dilbilgimi geliştirip geliştirmediğinden emin değilim.
    - □ Katılmıyorum: VR, dilbilgimi çok az geliştirdi.
    - □ Kesinlikle katılmıyorum: VR, dilbilgimi hiç geliştirmedi.

14. VR'nin İngilizce dilbilginizi özellikle hangi açılardan geliştirdiğini veya geliştirmediğini açıklayabilir misiniz?

15. Sizce VR, İngilizce telaffuzunuzu ne ölçüde geliştirdi?
    - □ Kesinlikle katılıyorum: VR, İngilizce telaffuzumu önemli ölçüde geliştirdi.
    - □ Katılıyorum: VR, telaffuzumu biraz geliştirdi.
    - □ Kararsızım: VR'nin telaffuzumun geliştirip geliştirmediğinden emin değilim.
    - □ Katılmıyorum: VR, telaffuzuma çok az etki etti.
    - □ Kesinlikle katılmıyorum: VR, telaffuzumu hiç geliştirmedi.

16. VR'nin İngilizce telaffuzunuzu özellikle hangi açılardan geliştirdiğini veya geliştirmediğini açıklayabilir misiniz?

17. VR oturumları sırasında konuşma performansınızın geliştiğini hissettiğiniz belirli bir anı tarif edebilir misiniz? O anı özel kılan neydi? Eğer böyle bir an yaşamadıysanız, nedeni neydi?

18. VR'nin, geleneksel sınıf yöntemlerinin yapamadığı şekilde İngilizce konuşma pratiği yapmanıza yardımcı olduğunu düşünüyor musunuz?
    Evet → Ne bakımdan farklılık gösterdi?
    Hayır → Neden bir fark olmadığını düşünüyorsunuz?

19. Sizce, İngilizce konuşma becerilerini geliştirmek amacıyla VR deneyiminin daha etkili hale getirilmesi için neler geliştirilebilir? Neden?

20. Gelecekteki İngilizce konuşma derslerinde VR kullanımını önerir misiniz? Neden?
21. Eğer İngilizce konuşma pratiği için bir VR etkinliği tasarlama şansınız olsaydı, bu aktiviteyi daha ilgi çekici ve etkili hale getirmek için hangi özellikleri eklerdiniz? Neden?

22. İngilizce konuşma becerilerinizi geliştirmek ve pratik yapmak için VR kullanma deneyiminizle ilgili paylaşmak istediğiniz başka düşünceleriniz var mı?

23. VR, İngilizce konuşmayı daha eğlenceli ve ilgi çekici hale getirdi mi? Neden?

24. Sizce, VR İngilizce konuşma pratiği yapmaya ilgi duymanızı ve odaklanmanızı kolaylaştırdı mı yoksa zorlaştırdı mı? Neden?

25. VR, sizi İngilizce konuşurken daha rahat mı yoksa daha rahatsız mı hissettirdi? Neden?

26. VR, İngilizce konuşmayla ilgili herhangi bir korku veya endişenizi aşmanıza yardımcı oldu mu? Neden?

27. VR oturumları sırasında diğer sınıf etkinliklerine kıyasla konuşma becerilerinize ne kadar güveniyordunuz? Neden?

28. VR oturumlarına katıldıktan sonra İngilizce konuşurken kendinize ne kadar güveniyorsunuz?
    □ Kendime çok güveniyorum: İngilizce konuşurken çok daha fazla şey başarabileceğimi hissediyorum.
    □ Kendime biraz güveniyorum: İngilizce konuşurken eskiye göre biraz daha fazla şey başarabileceğimi hissediyorum.
    □ Nötr: İngilizce konuşurken kendime duyduğum güvende bir değişiklik olmadığını düşünüyorum.
    □ Kendime çok güvenmiyorum: İngilizce konuşurken biraz tereddütlü hissediyorum.
    □ Kendime hiç güvenmiyorum: İngilizce konuşurken hala çok tereddütlü hissediyorum.

29. VR aktiviteleri, İngilizceyi daha doğal veya daha akıcı konuşmanıza yardımcı oldu mu?
    □ Kesinlikle Katılıyorum: Daha doğal ve daha hızlı konuştum.
    □ Katılıyorum: Daha doğal konuştum ama daha hızlı konuşmadım.
    □ Kararsızım: Konuşmamdaki hızda ya da doğallıkta önemli bir fark olup olmadığından emin değilim.
    □ Katılmıyorum: Daha yavaş ve daha az doğal konuştum.
    □ Kesinlikle Katılmıyorum: Hızda ya da doğallıkta herhangi bir gelişme fark etmedim.

30. VR'da pratik yaptıktan sonra, yeni öğrendiğiniz kelimeleri gerçek hayatta kullanma konusunda kendinize ne kadar güveniyorsunuz?
    □ Kesinlikle Katılıyorum: VR'de pratik yaptıktan sonra yeni öğrendiğim kelimeleri gerçek hayatta kullanma konusunda kendime çok güveniyorum.
    □ Katılıyorum: VR'de pratik yaptıktan sonra yeni öğrendiğim kelimeleri gerçek hayatta kullanma konusunda kendime biraz güveniyorum.
    □ Kararsızım: VR'de pratik yaptıktan sonra yeni öğrendiğim kelimeleri gerçek hayatta kullanma konusunda kendime ne güveniyorum ne de güvensizim.
    □ Katılmıyorum: VR'de pratik yaptıktan sonra yeni öğrendiğim kelimeleri gerçek hayatta kullanma konusunda kendime çok güvenmiyorum.
    □ Kesinlikle Katılmıyorum: VR'de pratik yaptıktan sonra yeni öğrendiğim kelimeleri gerçek hayatta kullanma konusunda kendime hiç güvenmiyorum.

31. VR aktiviteleri sırasında dil bilgisi hatalarınızın daha fazla farkında oldunuz mu?
    □ Kesinlikle Katılıyorum: Hatalarımın çok farkındaydım ve onları düzelttim.
    □ Katılıyorum: Hatalarımı fark ettim ama düzeltmeye odaklanmadım.
    □ Kararsızım: Dil bilgisi hatalarımı ne fark ettim ne de görmezden geldim.
    □ Katılmıyorum: Hatalarımı biraz fark ettim ama çoğunu gözden kaçırdım. Kesinlikle
    □ Katılmıyorum: Hiçbir dil bilgisi hatası fark etmedim.

32. VR'da telaffuz pratiği yapmak, gerçek hayatta İngilizce konuşurken kendine daha fazla güven duymana yardımcı oldu mu?
    □ Kesinlikle Katılıyorum: Kendime çok daha fazla güveniyorum.
    □ Katılıyorum: Kendime biraz daha güveniyorum.
    □ Kararsızım: Kendime güvenimi nasıl etkilediğinden emin değilim.
    □ Katılmıyorum: Kendime daha fazla güvenmiyorum ama güvenimi olumsuz etkilemedi.
    □ Kesinlikle Katılmıyorum: Eskiye göre kendime daha az güveniyorum ya da daha belirsiz hissediyorum.

APPENDIX-12: RATER 2'S TEACHING EXPERINCE QUESTIONNAIRE

1. Gender:
   - □ Male
   - □ Female
   - □ Prefer not to say
2. Age: _____
3. Highest level of education: _____
4. How many years of experience do you have teaching English? ____
5. In which contexts have you taught English? (Please check all that apply)
   - □ Primary School
   - □ Middle School
   - □ High School
   - □ University
   - □ Private Language School
   - □ Online Teaching
   - □ Other:_____
6. How many years of experience do you have teaching specifically at a preparatory school?
7. Which levels have you taught? (Please check all that apply.)
   - □ A1
   - □ A2
   - □ B1
   - □ B1+
   - □ B2
   - □ C1
8. How many years of experience do you have in assessing students' language skills?
9. Have you previously participated as an invigilator or rater in speaking tests?
   - □ Yes
   - □ No
10. Have you used any standardized speaking assessment tools (e.g., CEFR scales, IELTS speaking band descriptors)?
    - □ Yes
    - □ No
11. How familiar are you with rating speaking performance at different levels of proficiency (e.g., A1, B1, C1)?

Mark only one oval.

	1	2	3	4	5	
Not	◯	◯	◯	◯	◯	Extremely familiar

12. Any additional information you'd like to share about your teaching or assessment experience:_____

APPENDIX-13: 5 RATERS' TEACHING EXPERIENCE QUESTIONNAIRE

1. Gender:
   - □ Male
   - □ Female
   - □ Prefer not to say

2. Age:
   - □ Under 25
   - □ 25-34
   - □ 35-44
   - □ 45-54
   - □ 55+

3. Highest level of education:
   - □ Bachelor's Degree
   - □ Master's Degree
   - □ Doctorate
   - □ Postdoctoral

4. How many years of experience do you have teaching English?
   - □ Less than 1 year 1-3
   - □ 4-6
   - □ 7-10
   - □ 11-15
   - □ More than 15 years

5. In which contexts have you taught English? (Please check all that apply)
   - □ Primary School
   - □ Middle School
   - □ High School
   - □ University
   - □ Private Language School
   - □ Online Teaching
   - □ Other:_____

6. How many years of experience do you have teaching specifically at a preparatory school?
   - □ Less than 1 year
   - □ 1-3
   - □ 4-6
   - □ 7-10
   - □ 11-15
   - □ More than 15 years

7. Which levels have you taught? (Please check all that apply.)
   - □ A1

- ☐ A2
- ☐ B1
- ☐ B1+
- ☐ B2
- ☐ C1

8. How many years of experience do you have in assessing students' language skills?
   - ☐ Less than 1 year
   - ☐ 1-3
   - ☐ 4-6
   - ☐ 7-10
   - ☐ 11-15
   - ☐ More than 15 years

9. Have you previously participated as an invigilator or rater in speaking tests?
   - ☐ Yes
   - ☐ No

10. Have you used any standardized speaking assessment tools (e.g., CEFR scales, IELTS speaking band descriptors)?
    - ☐ Yes
    - ☐ No

11. How familiar are you with rating speaking performance at different levels of proficiency (e.g., A1, B1, C1)?



12. Any additional information you'd like to share about your teaching or assessment experience:_____

APPENDIX-14: SAMPLE OF ROLE-PLAY ACTIVITY FOR THE CONTROL GROUP (WEEK 4)

Role-Play Activity 1: Family Gathering - Describing Family Relationships

Warm-Up Activity (2-3 minutes):
Family Relationship Discussion:
Ask students to think about their family and name one person in their family who they admire or get along with well. Briefly share examples of family relationships (e.g., siblings, parents, cousins) and discuss one interesting fact about them.

Scenario:
Student A and Student B are meeting for the first time at a gathering, such as a wedding or holiday celebration. You are introducing yourselves to each other and explaining your family relationships, quirks, and stories.

Role Instructions:

1. **Student A:**
   You are going to talk about your family to Student B. Describe your family to them, including details about a few family members who stand out. Share an interesting or funny story about a family member and explain any unique family traditions.

   Things to Mention:
   "My cousin [Name] is a bit of a troublemaker; once, they…"
   "Our family has a tradition where we always [tradition] every [holiday]."
   "I'm very close to my [sibling / cousin / aunt]; we always go [activity] together."

2. **Student B:**
   You are meeting Student A for the first time. Ask questions about their family to get to know them better. Talk about your own family and a family member who has an interesting story or habit. Compare your family traditions with theirs.

   Questions to Ask:
   "It sounds like your family has a lot of fun! Do you all live close by?"
   "Do you have any funny stories about growing up with your family?"
   "In my family, we have a tradition of [tradition]; what about you?"

Goals:
Use descriptive language to talk about family members and traditions.
Practice storytelling and asking follow-up questions to show interest.

Useful Vocabulary:

- **Family Members:** aunt, uncle, cousin, grandparent, sibling, in-law, step-sibling
- Relationships & Descriptions: close-knit, distant, supportive, funny, strict, adventurous, traditional
- **Family Events:** family reunion, gathering, celebration, tradition, memory, activity

Useful Phrases:

- "In my family, we have a tradition of [activity] every [holiday]."
- "I'm very close to my [sibling / cousin / aunt]; we always [activity] together."
- "One thing about my [family member] is that they are very [adjective] – they love to [hobby or interest]."
- "When we get together, we usually [family activity]."
- "My [relative] always tells the funniest stories about…"
- "Our family gatherings are usually [description], and we all enjoy [activity]."

Role-Play Activity 2: Hotel - Dealing with Check-In Issues

Warm-Up Activity (2-3 minutes):

In pairs, students have 2-3 minutes to brainstorm the top three features or services they think make a hotel stay ideal (e.g., friendly service, clean rooms, breakfast, Wi-Fi).
Then, students share what they value most in a hotel stay and why.

Scenario:
Student A is a guest who booked a hotel room in advance, but upon arriving, they encounter an issue with their reservation. Student B is the receptionist who must find a solution and ensure the guest has a pleasant stay.

Role Instructions:

1. Student A (Guest):

You arrive at the hotel after a long journey, but the receptionist informs you that there is a mix-up with your reservation. Explain that you had a reservation and need help with a room. You also want to ask about other hotel amenities and services once the problem is solved.

Questions to Ask and Points to Mention:
"I booked a room online under [Your Name] for three nights. Could you check again?"
"Is it possible to get a room with a view?"
"Could you tell me more about the hotel amenities, like breakfast and Wi-Fi?"

2. Student B (Receptionist):

You discover that there was an issue with Student A's reservation, but you have some options to solve it. Be polite and offer solutions, such as upgrading their room, providing a discount, or arranging nearby accommodations if necessary. Once resolved, answer their questions about the hotel's services.

Solutions and Information to Provide:
"I'm sorry about the mix-up. We do have a room available, and I can upgrade you at no additional cost."
"Breakfast is served from 7 to 10 a.m., and we also have a gym and spa available."
"If there's anything else you need, please feel free to ask!"

Goals:
Practice problem-solving language and formal expressions for customer service.
Role-play a realistic scenario involving customer care, politeness, and compromise.

Useful Vocabulary:

- **Hotel Facilities:** reservation, reception, upgrade, room with a view, amenities, concierge, booking confirmation, complimentary
- **Common Requests:** room service, extra towels, late check-out, Wi-Fi, breakfast, shuttle service, luggage storage

Role-Play Activity 3: Reporting a Lost Phone

Warm-Up Activity (2-3 minutes):

Ask students to think about a time when they lost something important (e.g., keys, wallet, phone, a favorite accessory).
In pairs, students briefly share what the item was, where they lost it, and if they found it again. Also, students talk about **how they felt after losing the item** and when they found it again.

Scenario:
Student A recently lost their phone while on public transportation and is now at the Lost and Found office, where Student B works. Student A needs to describe the phone, explain how they lost it, and find out if it was turned in. Student B will ask for more information and offer suggestions.

Role Instructions:

1. **Student A (Person who lost the phone):**
   You're worried about your lost phone and want to check if it has been turned in. Describe where and how you lost it, and provide details about the phone. If it hasn't been found, ask about the next steps and mention any important information on the phone.
   Explain how it happened:
   - "I lost it on the 5 p.m. train, and I think it fell out of my bag."

   Describe the phone in detail:
   - "It's a black smartphone with a cracked screen and a blue case."

   Ask for assistance:
   - "Do you need my contact details in case someone turns it in?"
   - "Is there anything else I can do to increase the chances of finding it?"

2. **Student B (Lost and Found Employee):**
   You work at the Lost and Found office. Listen to Student A's situation, then ask questions to understand it better. Explain what actions Student A can take if the phone hasn't been found and offer advice to help them secure their information.
   Ask clarifying questions:
   - "Do you remember the exact train line or time?"
   - "Were there any other distinguishing features on the phone?"

   Offer advice on next steps:
   - "If you haven't already, it's a good idea to call your provider and report it lost."
   - "If you leave your contact information, we'll notify you if it's turned in."

Provide reassurance:

- "We'll keep an eye out, and sometimes lost items do get returned."

Goals:

Practice using past tense for describing past events and lost items.
Use clear, specific language to describe objects and situations.

Useful Vocabulary:

- **Describing a Lost Item:** phone, smartphone, screen, case, charger, cracked, lost and found, item description, report, tracking
- **Relevant Actions:** turn in, file a report, contact, call provider, lost property

APPENDIX-15: SAMPLE CODING OF QUALITATIVE DATA FOR THEMATIC ANALYSIS

Raw data of excerpt (English translation)	Initial code	Revised code	Category	Sub-theme	Theme
"It helped reduce my negative preconceptions about speaking." (S06)	Reduced negative judgments	Reduced negative preconceptions toward speaking	Reduced anxiety in a non-judgmental environment	Speaking anxiety	Affective factors in speaking
"It made me feel more comfortable because I could focus individually." (S10)	Through individual focus	Increased focus through individual and comfortable learning conditions	Anxiety regulation mechanisms	Speaking anxiety	Affective factors in speaking
"I think it was very helpful, especially for pronunciation." (S02)	Improved pronunciation	Pronunciation development	Perceived development of speaking skills	Learning outcomes of VR-enhanced speaking practice	Learning outcome
"I think that the content could be more comprehensive with more themes included so that it would be suitable for everyone." (S01)	Expansion of topic variety for speaking practice	Task and content diversity	Pedagogical constraints	Instructional constraints	Practical constraints
"Sometimes the system didn't recognize my speech" (S04)	Speech recognition issues	Failure of the system to recognize correct responses	Issues related to speech recognition	Technical challenges	Usability
"Speaking in a conversational format was very helpful." (S11)	Authentic exchange	Realistic conversational exchanges	Authentic interaction	Authenticity of VR-based speaking practice	Instructional value and authenticity
"Because everyone could speak at the same time and we had more opportunities to speak." (S12)	Simultaneous participation	Instructional configuration and learning support	Perceived instructional advantages of VR	Comparative instructional effectiveness	Instructional effectiveness
"Talking about different topics in different places was the thing I enjoyed the most because it felt realistic and sincere." (S03)	Real-life simulation	Facilitated interest and focus through impactful real-life simulation experience	Enhanced interest and focus in VR-based instruction	Interest and focus	Engagement and enjoyment
"I realized that we learned faster." (S08)	Time-efficient speaking practice	Faster learning through VR-enhanced practice	Reduced cognitive load and time-efficient learning	Low-effort practice	Learning efficiency

APPENDIX-16: PEARSON CORRELATION RESULTS FOR IRONSI'S (2023) PERCEPTION SCALE

Positive Correlation

Item	Category	Correlated Item	Description of correlation
1. VR is not effective in foreign language acquisition classrooms* - Instructional effectiveness	Instructional value and authenticity	4. VR introduces virtual real-life situations helpful in speaking	Perceiving VR as effective was associated with viewing VR as providing useful real-life speaking situations
	Usability	8. It is not difficult for me to handle and use VR during speaking lessons	Perceiving VR as effective was associated with being easy to handle and use
	Practical constraints	9. VR gadgets are expensive to purchase*	Students who viewed VR as effective were less likely to perceive cost as a major limitation
	Engagement and enjoyment	11. VR makes learning to speak fun and easy	Higher perceived effectiveness was associated with greater enjoyment and ease of learning when using VR
2. VR enhances learning and speaking - Instructional effectiveness	Instructional value and authenticity	4. VR introduces virtual real-life situations helpful in speaking	Students who perceived VR as enhancing learning and speaking also tended to believe that VR provides useful virtual real-life speaking situations
	Instructional effectiveness	5. Teaching through VR improved classroom performance in speaking	Positive perceptions of VR's effectiveness were associated with the view that teaching through VR improved classroom speaking performance
	Learning outcomes	6. The use of VR improved my speaking grades	Students who believed VR enhanced learning were also more likely to report that their speaking grades improved through VR use
	Usability	8. It is not difficult for me to handle and use VR during speaking lessons	Agreement that VR enhances learning and speaking was positively related to perceptions that VR was easy to handle and use during speaking lessons
	Practical constraints	9. VR gadgets are expensive to purchase*	This positive correlation indicates that students who viewed VR as effective were less likely to perceive the cost of VR gadgets as a major limitation
	Learning efficiency	10. VR is efficient in learning to speak	Perceiving VR as effective was associated with viewing VR as an efficient tool for learning to speak English

	Engagement and enjoyment	11. VR makes learning to speak fun and easy	Students who perceived VR as enhancing learning also tended to report that VR made learning to speak fun and easy
	Instructional effectiveness	12. VR helps to achieve lesson objectives on speaking	Positive views of VR's effectiveness were linked to the belief that VR helps achieve lesson objectives in speaking
3. VR does not improve speaking skills* -Instructional effectiveness	Engagement and enjoyment	7. VR makes speaking lessons boring*	Students who perceived VR as more effective for improving speaking skills also tended to report higher engagement and enjoyment
	Privacy and security concerns	13. I am concerned about privacy and security issues*	Students who viewed VR as more effective in developing speaking skills were also less likely to express concerns about privacy
4. VR introduces virtual real-life situations helpful in speaking - Instructional value and authenticity	Learning outcomes	6. The use of VR improved my speaking grades	Students who perceived VR as providing authentic real-life speaking situations also tended to report improved speaking grades
	Usability	8. It is not difficult for me to handle and use VR during speaking lessons	Viewing VR as authentic and realistic was positively associated with perceiving VR as easy to use
	Practical constraints	9. VR gadgets are expensive to purchase*	Students who valued VR's authenticity were less likely to perceive cost as a major limitation
	Learning efficiency	10. VR is efficient in learning to speak	Positive perceptions of authentic VR speaking contexts were associated with viewing VR as an efficient tool for developing speaking skills
	Engagement and enjoyment	11. VR makes learning to speak fun and easy	Students who valued VR's real-life speaking situations also tended to report higher enjoyment and engagement
	Instructional effectiveness	12. VR helps to achieve lesson objectives on speaking	Perceiving VR as authentic and realistic was positively related to believing that VR supports the achievement of speaking lesson objectives
5. Teaching through VR improved classroom performance in speaking - Learning outcomes	Learning outcomes	6. The use of VR improved my speaking grades	Students who perceived improvements in classroom speaking performance also tended to report higher speaking grades
	Usability	8. It is not difficult for me to handle and use VR during speaking lessons	Perceived improvement in speaking performance was positively associated with viewing VR as easy to use
	Learning efficiency	10. VR is efficient in learning to speak	Students who believed VR improved classroom speaking performance also tended to

			view VR as an efficient tool for learning to speak
	Engagement and enjoyment	11. VR makes learning to speak fun and easy	Positive perceptions of speaking performance gains were associated with enjoyment and engagement in VR-based speaking activities
	Instructional effectiveness	12. VR helps to achieve lesson objectives on speaking	Students who perceived VR as improving classroom speaking performance also tended to believe that VR supports the achievement of speaking lesson objectives
6. The use of VR improved my speaking grades - Learning outcomes	Usability	8. It is not difficult for me to handle and use VR during speaking lessons	Students who reported improved speaking grade through VR also tended to perceive VR as easy to handle and use
	Learning efficiency	10. VR is efficient in learning to speak	Perceived improvements in speaking grades were positively associated with viewing VR as an efficient method for learning to speak
	Engagement and enjoyment	11. VR makes learning to speak fun and easy	Students who experienced improved speaking grades also tended to report greater enjoyment and engagement in VR-based speaking activities
	Instructional effectiveness	12. VR helps to achieve lesson objectives on speaking	Students who reported higher speaking grades also believed that VR supports the achievement of speaking lesson objectives
7. VR makes speaking lessons boring* - Engagement and enjoyment	Privacy and security concerns	13. I am concerned about privacy and security issues*	Students who reported higher engagement and enjoyment in VR-based speaking practice also tended to report fewer concerns about privacy and security issues
8. It is not difficult for me to handle and use VR during speaking lessons - Usability	Practical constraints	9. VR gadgets are expensive to purchase*	Students who found VR easy to handle and use were less likely to perceive the cost of VR gadgets as a major limitation
	Learning efficiency	10. VR is efficient in learning to speak	Perceiving VR as easy to use was positively associated with viewing VR as an efficient tool for learning to speak
	Engagement and enjoyment	11. VR makes learning to speak fun and easy	Students who reported higher usability also tended to report greater enjoyment and engagement in VR-based speaking activities
	Instructional effectiveness	12. VR helps to achieve lesson objectives on speaking	Ease of use was positively related to the belief that VR supports the achievement of speaking lesson objectives

9. VR gadgets are expensive to purchase* -Practical constraints	Engagement and enjoyment	11. VR makes learning to speak fun and easy	Students who were less likely to perceive VR gadgets as expensive also tended to report higher enjoyment and engagement
	Privacy and security concerns	13. I am concerned about privacy and security issues*	Students who perceived fewer financial constraints related to VR also tended to report lower privacy and security concerns
10. VR is efficient in learning to speak - Learning efficiency	Engagement and enjoyment	11. VR makes learning to speak fun and easy	Students who perceived VR as an efficient way to learn speaking also tended to report higher engagement and enjoyment
	Instructional effectiveness	12. VR helps to achieve lesson objectives on speaking	Students who viewed VR as efficient also tended to believe that it supports the achievement of speaking lesson objectives
	Privacy and security concerns	13. I am concerned about privacy and security issues*	Students who perceived VR as efficient were also less likely to report privacy and security concerns
11. VR makes learning to speak fun and easy - Engagement and enjoyment	Instructional effectiveness	12. VR helps to achieve lesson objectives on speaking	Students who experienced higher enjoyment and engagement in VR-based speaking activities also tended to believe that VR helped them achieve speaking lesson objectives

No linear correlation

Item	Category	Correlated item	Description of correlation
7. VR makes speaking lessons boring* - Engagement and enjoyment	Practical constraints	9. VR gadgets are expensive to purchase*	This indicates that whether students found lessons boring or engaging was independent of whether they perceived VR gadgets as expensive
8. It is not difficult for me to handle and use VR during speaking lessons - Usability	Privacy and security concerns	13. I am concerned about privacy and security issues*	This suggests that perceiving VR as easy or difficult to use was not systematically related to concerns about privacy and security

Negative Correlation

Item	Category	Correlated Item	Description of correlation
1. VR is not effective in foreign language acquisition classrooms* - Instructional effectiveness	Instructional effectiveness	2. VR enhances learning and speaking	Students who perceived VR as more effective were less likely to agree that VR is ineffective
	Instructional effectiveness	3. VR does not improve speaking skills*	Agreement with VR's effectiveness was associated with disagreement that VR fails to improve speaking skills
	Learning outcomes	5. Teaching through VR improved classroom performance in speaking	Students who perceived improvements in classroom speaking performance through VR were less likely to view VR as ineffective
	Learning outcomes	6. The use of VR improved my speaking grades	Higher perceived speaking grade improvement was associated with lower perceptions of VR ineffectiveness
	Engagement and enjoyment	7. VR makes speaking lessons boring*	Students who found VR-based speaking practice more engaging were less likely to perceive VR as ineffective
	Learning efficiency	10. VR is efficient in learning to speak	Students who perceived VR as an efficient learning tool were less likely to consider it ineffective
	Instructional effectiveness	12. VR helps to achieve lesson objectives on speaking	Students who believed that VR helped achieve lesson objectives disagreed with the view that VR is ineffective
	Privacy and security concerns	13. I am concerned about privacy and security issues*	Students who reported fewer privacy and security concerns were less likely to perceive VR as ineffective
2. VR enhances learning and speaking - Instructional effectiveness	Instructional effectiveness	3. VR does not improve speaking skills*	Students who believed VR enhances learning and speaking were less likely to agree that VR does not improve speaking skills
	Engagement and enjoyment	7. VR makes speaking lessons boring*	Students who perceived VR as enhancing learning and speaking were less likely to find VR- based speaking lessons boring
	Privacy and security concerns	13. I am concerned about privacy and security issues*	Students who viewed VR as effective for learning and speaking tended to report fewer concerns about privacy and security
3. VR does not improve speaking skills* - Instructional effectiveness	Instructional value and authenticity	4. VR introduces virtual real-life situations helpful in speaking	Students who perceived VR as providing authentic real-life speaking situations were less likely to believe that VR does not improve speaking skills
	Learning outcomes	5. Teaching through VR improved classroom performance in speaking	Perceived improvement in classroom performance were associated with lower

			agreement that VR fails to improve speaking skills
	Learning outcomes	6. The use of VR improved my speaking grades	Students who reported improved speaking grades were less likely to perceive VR as ineffective for speaking development
	Usability	8. It is not difficult for me to handle and use VR during speaking lessons	Perceived greater ease of use was associated with lower agreement of the view that VR does not improve speaking skills
	Practical constraints	9. VR gadgets are expensive to purchase*	Students who perceived fewer cost-related constraints were less likely to view VR as ineffective for improving speaking skills
	Learning efficiency	10. VR is efficient in learning to speak	Perceiving VR as an efficient learning tool was associated with rejecting the view that VR does not improve speaking skills
	Engagement and enjoyment	11. VR makes learning to speak fun and easy	Higher enjoyment and engagement were associated with lower agreement that VR fails to improve speaking skills
	Instructional effectiveness	12. VR helps to achieve lesson objectives on speaking	Believing that VR helps achieve speaking lesson objectives was associated with disagreement that VR does not improve speaking skills
4. VR introduces virtual real-life situations helpful in speaking - Instructional value and authenticity	Learning outcomes	5. Teaching through VR improved classroom performance in speaking	Students who reported greater improvements in classroom speaking performance tended to report lower perceptions of VR's authenticity
	Engagement and enjoyment	7. VR makes speaking lessons boring*	Students who perceived VR speaking lessons as more engaging were more likely to view VR as providing authentic real-life speaking situations
	Privacy and security concerns	13. I am concerned about privacy and security issues*	Students who reported fewer privacy and security concerns were more likely to perceive VR as offering authentic and useful speaking contexts
5. Teaching through VR improved classroom performance in speaking - Learning outcomes	Engagement and enjoyment	7. VR makes speaking lessons boring*	Students who perceived improvements in speaking performance through VR were less likely to find VR-based speaking lessons boring
	Practical constraints	9. VR gadgets are expensive to purchase*	Students who perceived stronger speaking performance gains were less likely to view the cost of VR costs as a major limitation

	Privacy and security concerns	13. I am concerned about privacy and security issues*	Students who reported improved classroom performance through VR tended to express fewer privacy and security concerns
7. VR makes speaking lessons boring* - Engagement and enjoyment	Usability	8. It is not difficult for me to handle and use VR during speaking lessons	Students who perceived VR as easier to handle and use were less likely to find VR-based speaking lessons boring
	Learning efficiency	10. VR is efficient in learning to speak	Students who viewed VR as an efficient tool for learning to speak were less likely to perceive VR speaking lessons as boring
	Engagement and enjoyment	11. VR makes learning to speak fun and easy	Students who experienced greater enjoyment and ease in VR-based speaking activities were less likely to describe those lessons as boring
	Instructional effectiveness	12. VR helps to achieve lesson objectives on speaking	Students who believed VR helped achieve speaking lesson objectives were less likely to perceive VR speaking lessons as boring
9. VR gadgets are expensive to purchase* - Practical constraints	Learning efficiency	10. VR is efficient in learning to speak	Students who perceived VR as more efficient for learning to speak were less likely to view the cost of VR gadgets as a major limitation
	Instructional effectiveness	12. VR helps to achieve lesson objectives on speaking	Students who believed VR helped achieve speaking lesson objectives were less likely to perceive VR gadgets as a major constraint
11. VR makes learning to speak fun and easy - Engagement and enjoyment	Privacy and security concerns	13. I am concerned about privacy and security issues*	Students who reported higher enjoyment and engagement in VR-based speaking activities tended to express fewer privacy and security concerns
12. VR helps to achieve lesson objectives on speaking - Instructional effectiveness	Privacy and security concerns	13. I am concerned about privacy and security issues*	Students who perceived VR as more effective in helping achieve speaking lesson objectives tended to report fewer privacy and security concerns

APPENDIX-17: PERMISSION GRANTED BY ISTANBUL UNIVERSITY-CERRAHPASA

İÜC Tarih ve Sayı: 16.05.2024-988111

T.C.
İSTANBUL ÜNİVERSİTESİ-CERRAHPAŞA
REKTÖRLÜĞÜ
Lisansüstü Eğitim Enstitüsü Müdürlüğü

Sayı : E-13022260-199-988111 16.05.2024
Konu : Sümeyra Mürvet AYDIN

DAĞITIM YERLERİNE

Lisansüstü Eğitim Enstitüsü Yönetim Kurulu'nun 15.05.2024 tarih ve 19 sayılı oturumunda alınan 60 numaralı karar aşağıda belirtilmiştir.

Gereğini rica ederim.

Karar 60. Yabancı Diller Eğitimi Anabilim Dalı, İngiliz Dili Eğitimi Tezli Yüksek Lisans Programı'na 4801220311 numara ile kayıtlı Sümeyra Mürvet AYDIN'ın, Doç. Dr. Tuncer CAN danışmanlığında "Yabancı Dil Olarak İngilizce Sınıflarında Sanal Gerçekliğin Öğrencilerin Konuşma Performansı Üzerindeki Etkisi" [The Effect of Virtual Reality on Students' Speaking Performance in EFL Classes] olarak teklif edilen tez adı ve tez konusu öneri formu hakkında Anabilim Dalı Başkanlığı'ndan gelen 10.05.2024 tarih ve 984657 sayılı yazı görüşüldü; oybirliğiyle uygun görüldü.

Prof. Dr. Serdar AYDIN
Enstitü Müdürü

DAĞITIM:
Gereği:
Yabancı Diller Eğitimi Anabilim Dalı Başkanlığına

Bilgi:
Sayın Doç. Dr. Tuncer CAN

Bu belge, güvenli elektronik imza ile imzalanmıştır.

Belge Doğrulama Kodu :BS4SRUTP93 Belge Takip Adresi : https://www.turkiye.gov.tr/istanbul-universitesi-cerrahpasa-ebys

Adres:İstanbul Üniversitesi-Cerrahpaşa Rektörlüğü, 34320 Avcılar-İstanbul
Telefon:0212 404 03 00 Faks:0212 404 07 01
Web:https://www.iuc.edu.tr
Kep Adresi:iuc@hs01.kep.tr
Bilgi için: Lale KANER
Dahili: 53063

Bu belge,güvenli elektronik imza ile imzalanmıştır.

APPENDIX-18: INFORMED CONSENT FORM FOR EXPERIMENT GROUP (ENGLISH)

Informed Consent Form

Title of Research

The Effect of Virtual Reality on the Students' Speaking Performance

Details of the Research:
Dear Bilgi English Language Program Student,
You are cordially asked to take part in a research study that is part of a Master's program that is connected to Istanbul University-Cerrahpasa, which is carried out by Sümeyra M. Aydın. This study aims to see the impact of virtual reality on the students' speaking performance, so whether it increases, decreases, or has no proper impact. Also, it aims to understand the reasons for these results.

Participation Details:
As part of the experimental group, you will be taken to Santral Campus with the shuttle to experience learning and practicing English in a virtual reality environment that requires headsets for 6 weeks; once a week for 40 minutes for each student. Therefore, students will experience learning and practicing English in a virtual environment for 40 minutes. You will be taught how to use the virtual reality headsets if you don't know how to use them before you start learning and practicing English with them. The data will be collected through pre-test at the start of the study and post-test at the end of the study. Both of these will require audio recording. Also, you will be asked to answer questions of a perception scale and a semi-structured questionnaire after the post-test about your experience learning and practicing English in the virtual reality environment. Also, as part of the experimental group, it is crucial that you do not share that you are participating in this study with other students because it can affect the outcome of the research.

Voluntary Participation:
Participation is voluntary, so you decide if you want to participate or not. If you agree to participate at first, you can change your mind later on at any time during the research. There are no negative consequences of this.

Confidentiality:
Your identity will remain anonymous by using a pseudonym or a number. This along with the other data collected for this study will only be accessed by the researcher on a password-protected laptop. They will not be shared with any third parties. The collected data will be kept private, which will only be used for academic purposes such as publications or presentations at academic conferences or seminars.

Potential Risks and Discomforts in Research Participation:
There are no potential risks/discomforts of using low immersion virtual reality.

If you have any questions, you can ask the researcher or her supervisor any time you want.
Thank you for your participation.
Best regards,
Sümeyra M. Aydın
Contact information about the researcher:
Teaching Assistant Sümeyra M. AYDIN

Contact information about the supervisor:
Assoc. Prof. Tuncer CAN

Statement of Consent
I have been asked to participate in the research study titled "The Effect of Virtual Reality on the Students' Speaking Performance". I have read the information about the research above and I have had the chance to ask questions about it, which have been answered satisfactorily. I understand that I can't share that I'm participating in this study with other students because it can affect the results of the research. I consent voluntarily to being a participant in this study.

Name of the participant :_____
Email :_____
Phone :_____
Signature :_____
Date :_____
(Day / Month / Year)

One copy of the informed consent form will be kept by the participant and the second copy will be kept by the researcher. Your contact details will be kept private but separate from the data collected during the research.

APPENDIX-19: INFORMED CONSENT FORM FOR EXPERIMENT GROUP (TURKISH)

Bilgilendirilmiş Onam Formu

Araştırma Başlığı

Sanal Gerçekliğin Öğrencilerin Konuşma Performansı Üzerindeki Etkisi

Araştırma Detayları:
Sevgili Bilgi İngilizce Dil Programı Öğrencisi,
Sümeyra M. Aydın tarafından yürütülen İstanbul Üniversitesi-Cerrahpaşa'ya bağlı bir yüksek lisans programının parçası olan bir araştırma çalışmasına katılmanız rica olunur. Bu çalışma, sanal gerçekliğin öğrencilerin konuşma performansı üzerindeki etkisini, yani artırıp artırmadığını, azaltıp azaltmadığını veya uygun bir etkisi olup olmadığını görmeyi amaçlamaktadır. Ayrıca, bu sonuçların nedenlerini anlamayı amaçlamaktadır.

Katılım Detayları:
Deney grubunun bir parçası olarak, her öğrenci için haftada bir kez 30-40 dakika olmak üzere 6 hafta boyunca başlık gerektiren bir sanal gerçeklik ortamında İngilizce öğrenme ve pratik yapma deneyimi yaşamak için servisle Santral Kampüse götürüleceksiniz. Böylece her hafta 40 dakika olmak üzere sanal ortamda İngilizce öğrenme ve pratik yapma deneyimi yaşayacaksınız. Veriler, çalışmanın başlangıcında ön test ve çalışmanın sonunda son test yoluyla toplanacaktır. Bunların her ikisi için de ses kaydı gerekecektir. Ayrıca, son testten sonra bir algı ölçeğinde ve yarı-yapılandırılmış bir ankette sanal gerçeklik ortamında İngilizce öğrenme ve pratik yapma deneyiminizle ilgili soruları yanıtlamanız istenecektir. Ayrıca, deney grubunun bir parçası olarak, bu çalışmaya katıldığınızı diğer öğrencilerle paylaşmamanız çok önemlidir çünkü bu araştırmanın sonucunu etkileyebilir.

Gönüllü Katılım:
Katılım gönüllülük esasına dayalıdır, yani katılıp katılmayacağınıza siz karar verirsiniz. İlk başta katılmayı kabul ederseniz, daha sonra araştırma sırasında istediğiniz zaman fikrinizi değiştirebilirsiniz. Bunun hiçbir olumsuz sonucu yoktur.

Gizlilik:
Kimliğiniz bir takma ad veya numara kullanılarak anonim kalacaktır. Bu çalışma için toplanan diğer verilerle birlikte bu verilere yalnızca araştırmacı tarafından şifre korumalı bir dizüstü bilgisayar üzerinden erişilecektir. Herhangi bir üçüncü tarafla paylaşılmayacaktır. Toplanan veriler gizli tutulacak ve yalnızca yayınlar ya da akademik konferans veya seminerlerdeki sunumlar gibi akademik amaçlar için kullanılacaktır.

Araştırmaya Katılımda Olası Riskler ve Rahatsızlıklar:
Düşük düzeyli sürüklüyeciliğe sahip sanal gerçeklik kullanımının herhangi bir potansiyel risk veya rahatsızlık durumu bulunmamaktadır.

Herhangi bir sorunuz olursa, istediğiniz zaman araştırmacıya veya danışmanına sorabilirsiniz.

Katılımınız için teşekkür ederim.
Saygılarımla,
Sümeyra M. Aydın

Araştırmacı hakkında iletişim bilgileri:
Öğretim Görevlisi Sümeyra M. AYDIN

Danışman hakkında iletişim bilgileri:
Doç Dr. Tuncer CAN

Rıza Beyanı
"Sanal Gerçekliğin Öğrencilerin Konuşma Performansı Üzerindeki Etkisi" başlıklı araştırmaya katılmam istendi. Yukarıda araştırma ile ilgili bilgileri okudum ve araştırma ile ilgili sorular sorma şansım oldu ve bu sorular tatmin edici bir şekilde yanıtlandı. Araştırmanın sonuçlarını etkileyebileceği için bu çalışmaya katıldığımı diğer öğrencilerle paylaşamayacağımı biliyorum. Bu çalışmaya gönüllü olarak katılmayı kabul ediyorum.

Katılımcının adı :_____
E-posta :_____
Telefon :_____
İmza :_____
Tarih :_____
(Gün / Ay / Yıl)

APPENDIX 20: INFORMED CONSENT FORM FOR CONTROL GROUP (ENGLISH)

Informed Consent Form

Details of the Research:

Dear Bilgi English Language Program Student,
You are cordially asked to take part in a research study that is part of a Master's program connected to Istanbul University-Cerrahpasa, conducted by Sümeyra M. Aydın. The purpose of this study is to assess the speaking skills of B1+ students and explore factors influencing their speaking performance.

Participation Details:

As part of this study, you will participate in a speaking pre-test at the beginning of the study and a post-test at the end. Both of these assessments will require audio recording. As part of this study, you will also be participating in additional speaking activities for 40 minutes each week. This research will be 6 weeks long. The speaking activities may require pair work or group work. These activities will include tasks such as roleplay and if necessary other interactive speaking exercises designed to help practice and enhance your English speaking skills. These sessions will be designed to allow you to engage in realistic communication scenarios. Additionally, these sessions will be audio-recorded to provide as evidence that this research has been conducted for the defense of my thesis.

Voluntary Participation:

Participation in this study is voluntary. You are free to decide whether or not to take part. If you agree to participate but change your mind later, you may withdraw at any time without any negative consequences.

Confidentiality:

Your identity will remain anonymous, and a pseudonym or number will be used to protect your privacy. All collected data will only be accessed by the researcher, stored securely, and not

shared with any third parties. All data collected will be used for academic purposes, including publications or presentations at conferences or seminars. Your identity and all information provided will be kept confidential.

If you have any questions, feel free to contact the researcher or her supervisor.

Thank you for your participation.
Best regards,
Sümeyra M. Aydın

Contact information for the researcher:
Teaching Assistant Sümeyra M. AYDIN

Contact information for the supervisor:
Assoc. Prof. Tuncer CAN

Statement of Consent

I have been asked to participate in this research for an MA thesis. I have read the information provided about the study, and any questions I had have been answered to my satisfaction. I consent voluntarily to participate in this study.

Name of the participant :_____
Email :_____
Phone :_____
Signature :_____
Date :_____

(Day / Month / Year)

One copy of the informed consent form will be kept by the participant, and the second copy will be kept by the researcher. Your contact details will be kept private but separate from the data collected during the research.

APPENDIX-21: INFORMED CONSENT FORM FOR CONTROL GROUP (TURKISH)

Bilgilendirilmiş Onam Formu

Araştırma Detayları:

Sevgili Bilgi İngilizce Dil Programı Öğrencisi,
Sümeyra M. Aydın tarafından yürütülen ve İstanbul Üniversitesi-Cerrahpaşa'ya bağlı bir Yüksek Lisans programının parçası olan bu araştırmaya katılmanız rica edilmektedir. Bu çalışmanın amacı, B1+ seviyesindeki öğrencilerin konuşma becerilerini değerlendirmek ve konuşma performanslarını etkileyen faktörleri araştırmaktır.

Katılım Detayları:

Bu araştırmanın bir parçası olarak, çalışmanın başında bir konuşma ön testine ve sonunda bir konuşma son testine katılacaksınız. Her iki değerlendirme de ses kaydı gerektirecektir. Ayrıca, bu araştırma kapsamında haftada 40 dakika ek konuşma aktivitelerine de katılacaksınız. Bu araştırma 6 hafta sürecektir. Konuşma aktiviteleri eşli çalışma veya grup çalışması gerektirebilir. Bu aktiviteler, rol yapma ve gerektiğinde diğer etkileşimli konuşma egzersizleri gibi görevleri içerecek ve İngilizce konuşma becerilerinizi pratiğe dökmenizi ve geliştirmenizi amaçlayacaktır. Bu oturumlar, gerçekçi iletişim senaryolarında yer almanızı sağlayacak şekilde tasarlanacaktır. Ayrıca, bu oturumlar, bu araştırmanın gerçekleştirildiğine dair tez savunmamda kanıt olarak kullanılmak üzere ses kaydı yapılacaktır.

Gönüllü Katılım:

Bu çalışmaya katılım tamamen gönüllülük esasına dayanmaktadır. Katılıp katılmamaya karar vermekte özgürsünüz. İlk başta katılmayı kabul ederseniz ancak daha sonra fikrinizi değiştirirseniz, herhangi bir olumsuz sonuç olmaksızın istediğiniz zaman çalışmadan çekilebilirsiniz.

Gizlilik:

Kimliğiniz anonim tutulacak ve gizliliğinizi korumak için takma ad veya bir numara kullanılacaktır. Toplanan tüm veriler yalnızca araştırmacı tarafından erişilebilecek, güvenli bir şekilde saklanacak ve üçüncü kişilerle paylaşılmayacaktır. Toplanan tüm veriler akademik amaçlar doğrultusunda kullanılacak olup, yayınlar veya konferans ve seminerlerdeki sunumlar için kullanılabilir. Kimliğiniz ve sağladığınız tüm bilgiler gizli tutulacaktır.

Herhangi bir sorunuz varsa, araştırmacı veya danışmanıyla iletişime geçmekten çekinmeyin.

Katılımınız için teşekkür ederiz.

Saygılarımla,

Sümeyra M. Aydın

Araştırmacı İletişim Bilgileri:

Araştırma Görevlisi Sümeyra M. AYDIN

E-posta: sumeyra.aydin@bilgi.edu.tr

Danışman İletişim Bilgileri:

Doç. Dr. Tuncer CAN

E-posta: tcan@iuc.edu.tr

Onam Beyanı

Bu araştırmanın bir Yüksek Lisans tezi için yapıldığını kabul ettim. Araştırmayla ilgili sağlanan bilgileri okudum ve sormak istediğim sorular tatmin edici şekilde yanıtlandı. Bu çalışmaya gönüllü olarak katılmayı kabul ediyorum.

Katılımcının adı :_____

E-posta :_____

Telefon :_____

İmza :_____

Tarih :_____

(Gün / Ay / Yıl)

Bilgilendirilmiş onam formunun bir kopyası katılımcı tarafından saklanacak, ikinci kopya ise araştırmacı tarafından saklanacaktır. İletişim bilgileriniz gizli tutulacak ve araştırma sırasında toplanan verilerden ayrı olarak saklanacaktır.

APPENDIX-22: INFORMED CONSENT FORM FOR THE VOICE RECORDINGS OF THE STUDENTS FOR THE INTERRATER RELIABILITY PROCESS (ENGLISH)

Consent Form for Recording: Ensuring Assessment Reliability and Validity

Researcher: Sümeyra M. AYDIN
Institution: Istanbul Bilgi University – School of Foreign Languages

Purpose of the Recording:
You are invited to participate in a process to help ensure the accuracy and consistency of assessments in a research study. The purpose of this recording is solely to establish interrater reliability and its validity. Your recorded speaking performance will allow us to compare and standardize the ratings of different evaluators. The recording of your speaking performance will be reviewed by the researcher, her partner and other teachers at Bilgi English Preparatory Program who volunteer to assist in determining the validity of the interrater reliability process for this study. Please note that this recording is not part of the actual study and you will not be a participant in the research itself.

What You Will Be Asked to Do:

- Participate in a short speaking activity, which will be recorded.
- Your voice recording will be used to evaluate the reliability and the validity of the scores given by the raters for assessing your speaking skills.

Confidentiality:
Your identity will remain anonymous, and the recordings will only be used for the purpose of establishing interrater reliability and its validity. The data will be stored securely, and only the researcher and raters involved in this evaluation process (the reliability and validity of interrater assessments) will have access to it. All data collected will be used for academic purposes. No identifying information will be linked to your recording.

Voluntary Participation:
Participation is entirely voluntary. You may choose not to participate, or to withdraw at any time without any negative consequences.

Contact Information:
If you have any questions or concerns regarding this recording process, you may contact the researcher through email or visit her office to ask any questions.

Consent:
By filling in and signing the sections below, you confirm that you understand the purpose of this recording and agree to participate.

APPENDIX-23: INFORMED CONSENT FORM FOR THE VOICE RECORDINGS OF THE STUDENTS FOR THE INTERRATER RELIABILITY PROCESS (TURKISH)

Kayıt İçin Onay Formu: Değerlendirme Güvenilirliğini ve Geçerliliğini Sağlama

Araştırmacı: Sümeyra M. AYDIN
Kurum: İstanbul Bilgi Üniversitesi – Yabancı Diller Yüksekokulu

Kaydın Amacı:
Bir araştırma çalışmasında değerlendirmelerin doğruluğunu ve tutarlılığını sağlamak amacıyla bir sürece katılmaya davetlisiniz. Bu kaydın amacı yalnızca değerlendiriciler arası güvenilirliği ve geçerliliğini sağlamaktır. Kaydedilen konuşma performansınız, farklı değerlendiricilerin notlarını karşılaştırmamıza ve standartlaştırmamıza olanak tanıyacaktır. Konuşma performansınızın kaydı, araştırmacı, partneri ve bu çalışmada değerlendiriciler arası güvenilirlik sürecinin geçerliliğini belirlemeye gönüllü olan Bilgi İngilizce Hazırlık Programı'ndaki diğer öğretmenler tarafından incelenecektir. Bu kaydın asıl araştırmanın bir parçası olmadığını ve araştırmanın katılımcısı sayılmayacağınızı lütfen unutmayın.

Sizden İstenecekler:

- Kısa bir konuşma etkinliğine katılacak ve bu konuşma kaydedilecektir.
- Ses kaydınız, konuşma becerilerini değerlendiren değerlendiriciler arasındaki tutarlılığı incelemek için kullanılacaktır.

Gizlilik:
Kimliğiniz gizli kalacak ve kayıtlar yalnızca değerlendiriciler arası güvenilirliği ve geçerliliğini sağlamak amacıyla kullanılacaktır. Veriler güvenli bir şekilde saklanacak ve bu değerlendirme sürecinde (değerlendiriciler arası değerlendirmelerin güvenilirliği ve geçerliliği konusunda) yer alan araştırmacı ve değerlendiriciler dışında kimse erişemeyecektir. Toplanan tüm veriler akademik amaçlar için kullanılacaktır. Kaydınıza hiçbir kişisel bilgi eklenmeyecektir.

Gönüllü Katılım:
Katılım tamamen gönüllülük esasına dayanmaktadır. Katılmamayı tercih edebilir ya da herhangi bir olumsuz sonuç yaşamadan istediğiniz zaman katılımınızı geri çekebilirsiniz.

İletişim Bilgileri:
Bu kayıt süreci ile ilgili herhangi bir sorunuz ya da endişeniz varsa, e-posta adresinden iletişime geçebilirsiniz veya sorularınızı sormak için ofisini ziyaret edebilirsiniz.

Onay:
Aşağıdaki bölümleri doldurarak ve imzalayarak bu kaydın amacını anladığınızı ve katılmayı kabul ettiğinizi onaylamış olursunuz.

APPENDIX-24: RATER 2'S INFORMED CONSENT FORM

Informed Consent Form for Participation as a Second Rater in Speaking Performance Assessment

Title of Study: The Effect of Virtual Reality on Students' Speaking Performance in EFL Classes

Researcher: Sümeyra M. AYDIN
Institution: Istanbul Bilgi University School of Foreign Languages
Contact Information:

Purpose of the Study:
The aim of this study is to assess the speaking skills of intermediate (B1+) level students in the English Preparatory Program at Istanbul Bilgi University. The study will compare the speaking performance of two groups through pre-test and post-test after a 6-week virtual reality intervention to see its impact.

Role of the Second Rater:
As the second rater in this study, you will be responsible for evaluating the speaking performances of the participating students alongside the primary rater for the pre-test and the post-test stages. To ensure the reliability of the results, interrater reliability will be calculated. This will be done by listening to the audio-recordings of other L4 students who have given consent for this purpose and who won't be the actual participants in this study. The interrater process will be both audio and video recorded to provide as evidence during thesis defense. Additionally, the pre-test and post-test stages will require audio-recordings of the intermediate level students' speaking performances. For both the interrater reliability process and the pre-test and post-test stages, blind scoring will be utilized. The scoring process will be the same for the interrater reliability process and in the actual study (pre-test and post-test stage). Your assessment will contribute to the reliability of the results of the study.

Consent:
By signing this form, you agree to the following:

1. **Use of Personal Information:**
   - You consent to the use of your full name, professional title, and relevant background information in any reports, publications, or presentations resulting from this research.
2. **Participation in the Interrater Reliability Process:**
   - You agree to participate in the interrater reliability process, which will involve evaluating intermediate level students' speaking performances that are audio recordings. These students will not be the participants of the actual study. This process will ensure consistency in scoring and provide reliability to the study.
3. **Participation as the Second Rater in the Actual Research**

- You agree to participate as the second rater in the actual study, which will involve evaluating students' pre-test and post-test speaking performances. This process will ensure consistency in scoring and provide reliability to the study.

4. **Use of Audio and Video Recordings:**
   - You consent to the recording and the use of audio and/or video recordings of the interrater reliability process. These recordings may be used for research purposes, including in any reports, publications, or presentations.
   - You also consent to the recording and use of audio during the pre-tests and post-tests. This is necessary for reassessment of scores if needed and to provide as evidence.
5. **Confidentiality and Data Protection:**
   - All recordings and any personal information shared during this study will be kept confidential and stored securely. Your identity will only be revealed in publications or presentations with your prior consent.
6. **Voluntary Participation:**
   - Your participation in this study is voluntary. You may withdraw your consent at any time, and you won't face any consequences upon withdrawal.

If you have any questions, you can ask the researcher or her supervisor any time you want.

Thank you for your participation.

Best regards,

Sümeyra M. Aydın

Contact information about the researcher:

Teaching Assistant Sümeyra M. AYDIN

Contact information about the supervisor:

Assoc. Prof. Tuncer CAN

Declaration:

I have read and understood the information provided in this consent form. I agree to participate in the research study titled "The Effect of Virtual Reality on the Students' Speaking Performance" as a second rater, and I consent to the use of my personal and professional information and the audio and/or video recordings as outlined above.

Name (Second Rater): _____
Signature: _____
Date: _____

APPENDIX-25: 5 RATERS' INFORMED CONSENT FORM

Consent Form for Participation in Research Study

Title of Study: The Effect of Virtual Reality on Students' Speaking Performance in EFL Classes

Researcher: Sümeyra M. AYDIN
Institution: Istanbul Bilgi University School of Foreign Languages
Contact Information:

Purpose of the Study:
The purpose of this study is to assess the speaking skills of B1+ level students in the English Preparatory Program at Istanbul Bilgi University. The study will compare the speaking performance of two groups through pre-test and post-test after a 6-week virtual reality intervention.

Role of Teachers:
As a participating teacher, you are being asked to review and assign speaking scores to the 30 recorded speaking performances of students. Your scores, along with those of other experienced teachers, will contribute to setting a benchmark and be used to compare the scores provided by the researcher and their partner for grading as part of the interrater reliability process.

Additionally, basic information regarding the qualifications, experience, and knowledge of the teachers participating in this process will be included in the thesis to demonstrate their expertise. This information will help support the validity of the study during the thesis defense.

Voluntary Participation:
Your participation in this study is entirely voluntary. You may choose not to participate or withdraw from the study at any time without any negative consequences.

Confidentiality:
All responses and scores you provide will remain confidential. While your basic professional background (e.g., years of teaching experience, level expertise) may be shared to demonstrate your knowledge and experience, your identity will not be disclosed in any part of the research. All data will be used solely for research purposes and will be reported in aggregate form without individual identifiers.

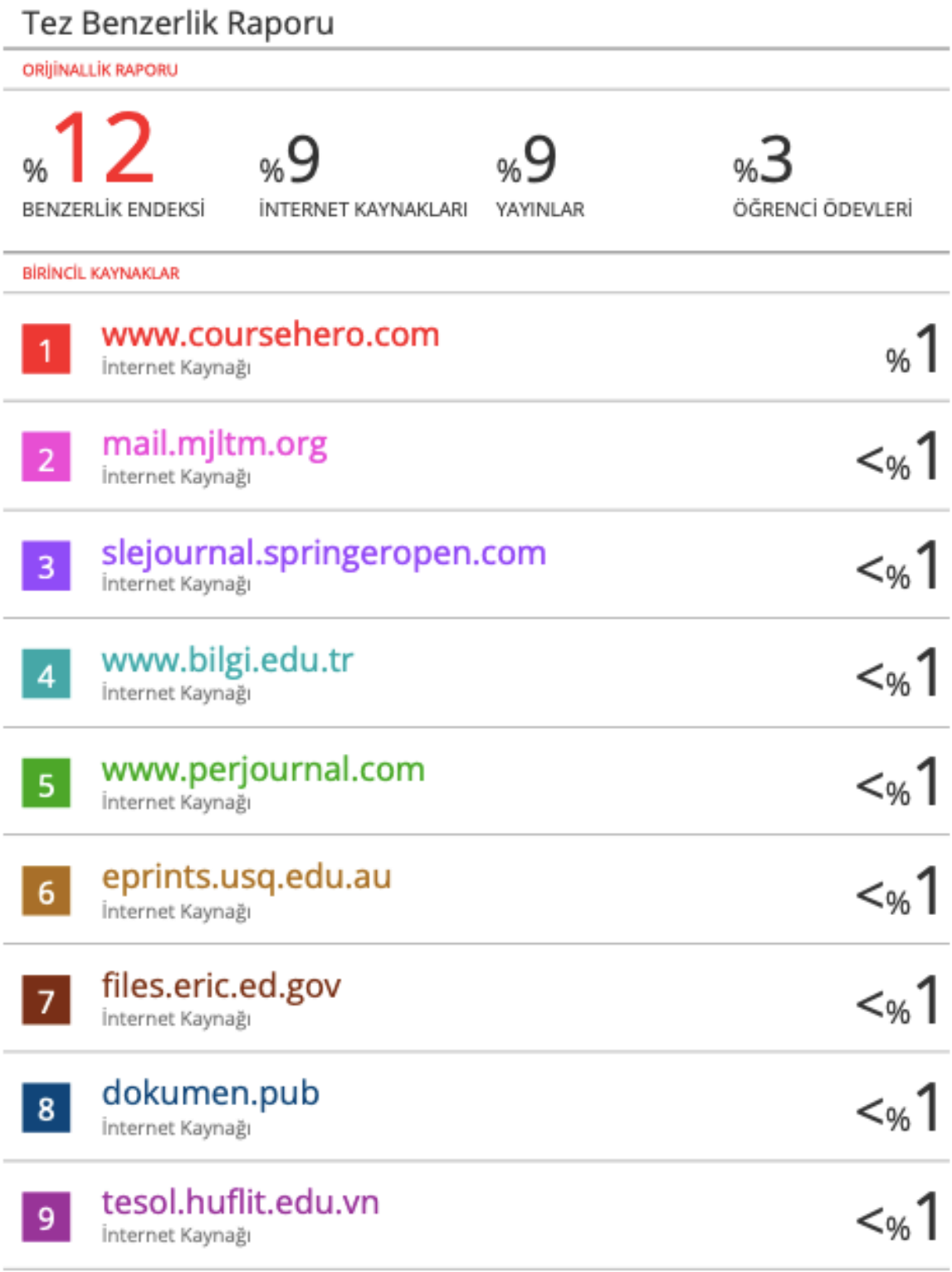
Consent Statement:
By filling in and signing the sections below, you acknowledge that you have read and understood the information provided above, and you agree to participate in this study as a contributing teacher.

Participant's Name: _____

Participant's Signature: _____

Date: _____

FIRST PAGE OF THE PLAGIARISM REPORT



Tez Benzerlik Raporu

ORİJİNALLİK RAPORU

%12	%9	%9	%3
BENZERLİK ENDEKSİ	İNTERNET KAYNAKLARI	YAYINLAR	ÖĞRENCİ ÖDEVLERİ

BİRİNCİL KAYNAKLAR

1	www.coursehero.com İnternet Kaynağı	%1
2	mail.mjltm.org İnternet Kaynağı	<%1
3	slejournal.springeropen.com İnternet Kaynağı	<%1
4	www.bilgi.edu.tr İnternet Kaynağı	<%1
5	www.perjournal.com İnternet Kaynağı	<%1
6	eprints.usq.edu.au İnternet Kaynağı	<%1
7	files.eric.ed.gov İnternet Kaynağı	<%1
8	dokumen.pub İnternet Kaynağı	<%1
9	tesol.huflit.edu.vn İnternet Kaynağı	<%1