English Teaching, Vol. 80, Special Issue 2025, pp. 67–93
DOI: https://doi.org/10.15858/engtea.80.5.202512.67
https://journal.kate.or.kr

학교 영어 말하기 교육을 위한 맞춤형 AI 말하기 진단-학습-평가 시스템 설계 연구*

김혜영 · 성민창 · 이진화 · 최윤덕**

Kim, Heyoung, Sung, Min-Chang, Lee, Jin-Hwa, & Choi, Yundeok. (2025). A study on the design of a customized AI-based speaking diagnosis, learning, and assessment system for public English education. *English Teaching, 80*(5), 67–93.

The purpose of this study is to develop and implement a customized AI-based speaking diagnosis, learning, and assessment system, *SpeakMaster*, in order to overcome the lack of systematic evaluation and practice opportunities in school English speaking class. This system integrates automated speaking scoring to provide students with feedback on their speaking abilities across pronunciation, conversation, and presentation. This study adopts a design-based research methodology, demonstrating the development and implementation process. 1,451 students and eight teachers in elementary, middle, and high schools participated in the experiment. Data were collected through learning logs, teacher journals, interviews, and post-surveys. The findings indicate that the system design is appropriate for English class, promoting students' flow in engaging speaking practice. Students showed motivation and satisfaction while teachers found the system valuable for monitoring student progress and facilitating speaking assessments. Despite the challenges of improving chatbot performance and enhancing scoring reliability, the results suggest that *SpeakMaster* shows potential to enhance English speaking education.

Key words: AI-based diagnosis system, automated speaking scoring (ASS), speaking assessment/AI 기반 학습 진단 시스템, 자동 말하기 채점, 말하기 평가

---

*This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2023S1A5A2A03086569).

**First Author: Heyoung Kim, Professor, Department of English Education, Chung-Ang University Corresponding Author: Min-Chang Sung, Associate Professor, Department of English Education, Gyeongin Nat'l University of Education; 155 Sammak-ro Anyang-si Gyeonggi-do 13910, Korea; Email: mcsung@ginue.ac.kr Third Author: Jin-Hwa Lee, Professor, Department of English Education, Chung-Ang University Fourth Author: YunDeok Choi, Assistant Professor, Department of English Education, Chungnam National University

Received 31 August 2025; Reviewed 29 September 2025; Accepted 17 December 2025

1. 서론

1984년 *English Teaching* 27호에 게재된 “영어 말하기 교육의 저해요인”이라는 제목의 논문(Bae, 1984)의 서두에는 다음과 같은 내용이 있다.

> 가장 습득하기 어려운 기능이 말하기라는 점에도 일반적으로 공통된 의견을 갖고 있다. 엄격히 말하자면, 중·고등, 대학과정을 고루 이수한 한국인들 중에서 문교부가 마련한 중학교 1학년 영어과 교과과정이 요구하는 교육목표의 말하기 실력을 갖춘 사람은 드물다고 봄이 옳을 것이다. (p. 15)

Bae(1984)는 학교 말하기 교육의 저해요인으로 과밀학급으로 인한 환경상의 어려움, 교사의 영어 구사력, 수업의 소통매개로서 우리말을 사용하는 점, 문형 및 읽기 위주의 교재 및 교수법 등을 꼽았다. 40년이 지난 지금에도 우리의 말하기 교육이 동일한 문제를 그대로 안고 있다는 사실은 60주년 기념호에 논문을 투고하는 영어교육자로서 무거운 자책감을 느끼지 않을 수 없다. 그 많은 연구와 교육적 노력에도 불구하고 왜 학교 말하기 교육은 이토록 제자리 걸음을 벗어나지 못하는 것인가? 과연 영어 말하기 교육은 이대로 관심 밖으로 사라지고 말 것인가?

실제로 말하기는 학교 영어수업에서 점점 더 자리를 잃어가고 있다. 고등학교 교사 463명을 조사한 Lee(2018)의 연구에 따르면, 43.8%가 영어 회화 과목을 편성하지 않으며, 말하기 활동을 수행 평가 대비용으로만 활용하는 비율도 21.4%에 달하였다. 반면 학생들의 말하기 교육에 대한 요구와 인식은 정반대이다. Jung과 Cha(2014)의 990명 중고생 조사 연구에서도 수업 시간에 주로 다루는 기술은 문법(58.4%)이며 말하기는 9.5% 정도라고 응답한 반면, 자신이 영어학습에서 가장 중요하게 생각하는 것은 말하기(44.1%)라고 하였다. 최근 한국인 1천명 조사 결과(Macromillembrain, 2022) 역시 89.2%가 ‘영어를 잘하고 싶다’고 답하여 강한 요구를 드러냈으나, 응답자 80%가 자신의 현재 영어 실력을 ‘왕초보’ 혹은 ‘초보’ 수준이라고 평가하였다.

말하기 교육의 활성화를 저해하는 근본 원인은 평가에 있다고 보는 견해가 많다. 학교 영어 말하기 평가를 보면 첫째, 사전 준비형 수행 평가로 일방향 암기 위주이며, 현실의 소통에 가까운 즉흥적인 실제 상호작용 능력을 평가하지 못한다. 둘째, 루브릭 기반의 현행 채점 방식은 채점 기준 개발의 어려움, 채점 소요시간과 업무 부담, 채점 결과에 대한 신뢰도 문제, 적절한 도구의 부재 등으로 교사들로 하여금 말하기 평가를 꺼리게 한다(Koh, 2014; Lee, Yoon, Kim, Choo, & Kwon, 2015; Song & Shim, 2020). 평가가 어려운 영역은 수업에서 적극적으로 다루기 힘들다는 것은 누구나 주지하는 바이다.

최근 이러한 말하기 평가의 어려움을 획기적으로 극복할 수 있는 AI 기술이 바로 자동 말하기 채점(Automated Speaking Scoring, ASS)이다. 자동 발화 인식(Automatic Speech Recognition, ASR) 기술은 이미 수년 전 초등영어학습 앱 AI 펭톡이나, 최근

개발된 AI 디지털 교과서까지 도입되어 학습자 영어 발화의 음소, 단어, 문장의 발음과 강세 억양 등의 정확도를 평가해왔다.

반면, 최근 부상한 ASS 기술은 음성 특징뿐만 아니라 내용 및 기타 언어적 특징을 포함하여 데이터를 다각적으로 측정 평가한다. 즉 발음에 더하여 유창성, 문법적 정확성, 어휘 수준, 내용의 적정성 등을 자연어처리기술(NLP)을 통해 평가하게 되므로 총체적인 말하기 능력 평가라고 할 수 있다(Xu, Jones, Laxton, & Galaczi, 2021; Zechner & Evanini, 2019). Lee, Choi, Sung과 Kim(2023)에 따르면 TOEFL Practice Online 등 대부분의 해외 저명한 표준 말하기시험에서 이러한 ASS 기술을 전체 혹은 부분적으로 도입한 상태이며, 최근에는 대규모 언어 모델(Large Language Model, LLM)을 활용한 end-to-end 방식과 트랜스포머 기반 모델 등의 통합 채점으로 빠르게 진화하는 추세이다(Irshad et al., 2024; Kang et al., 2024; Wang, Evanini, Qian, & Mulholland, 2021). 이러한 ASS 기술을 적용하여 공교육에 적합한 시스템이 개발된다면, 말하기 연습과 평가가 수월하게 이루어질 수 있어, 말하기 수업의 저해요인이 상당 부분 해소되고, 나아가 학습자 능력에 따른 맞춤형 교육이 실현될 수 있을 것이다.

따라서 본 연구의 목적은 영어 말하기 교육을 활성화하고, 교육과정의 목표에 부합하는 초,중,고 학습자의 영어 말하기 능력을 진단평가하기 위하여, ASS 기술을 결합한 학습 시스템을 개발하여 학교에 적용해 보는 데에 있다. 이를 구현하기 위하여 본 연구는 스픽마스터(*SpeakMaster*)[1] 프로젝트를 다년간 추진하였고, 다음과 같은 순서로 연구를 진행해 왔다. 첫째, 맞춤형 영어 말하기 진단 학습 시스템의 개념과 구조도를 설계하였다. 둘째, AI 기반 웹플랫폼을 개발하였다. 여기에서는 말하기 활동 및 평가 과업과 ASS 기술을 결합한 진단평가 리포트와 자동 피드백 등을 제공한다. 셋째, 1차 개발된 시스템을 초, 중, 고 실험 학교 8곳을 통해 수업에 적용하고, 이를 보완하고 개선하기 위하여 데이터를 분석하고 성찰하였다. 본 연구의 초점은 바로 프로젝트의 세 번째 단계인 '1차 시행'으로 시스템의 시범 적용에 해당하므로, 다음과 같은 연구문제를 제시한다: AI 기반 맞춤형 영어 말하기 진단-학습-평가 시스템인 스픽마스터의 설계와 내용·기능은 공교육 영어 말하기 수업에 적용하는데 적합한가?

이를 위하여 다음의 세 가지 세부 연구문제를 제시한다.

1) 시스템의 각 영역별 내용 및 기술적 설계는 어떠해야 하는가?
2) 시스템의 적용시 사용 적절성과 만족도는 어떠하였는가?
3) 참여교사들이 인식하는 시스템의 활용 가치와 개선사항은 무엇인가?

[1] https://speakmaster.co.kr

2. 연구 방법: 설계기반연구(Design-based Research, DBR)

2.1. 연구 절차

스픽마스터 프로젝트는 교육환경의 복잡한 문제들의 해결방안을 찾기 위한 연구 기법인 설계기반연구(Design-based Research, DBR)를 채택하였다. DBR은 2,000년대 초반에 소개되기 시작한 질적 연구 방법으로, 교육용 테크놀로지 혹은 프로그램을 개발하는 과정에서 현장 경험을 접목하여 설계의 보완점을 찾아가는 검증 과정이다(Reeves, 2006; Reimann, 2013). DBR은 여러 학자에 의해 발전되었지만(Bannan-Ritland, 2003; Wang & Hannafin, 2005), 공통적으로 1) 문제점 인식, 2) 이론에 근거한 설계 방안 제안, 3) 프로그램의 개발 및 시행, 4) 평가 및 성찰의 단계를 거치는 점에서 질적연구 중 실행연구(action research)와 유사하다. 또한 이 과정은 순환적으로 설계의 보완 및 재시행의 과정을 지속한다. DBR은 컴퓨터보조 언어 학습의 영역에서 널리 적용되는 연구 방법론이다(Rodríguez & Ballester, 2013). Hoadley(2005)는 특히 최근 일반화되고 있는 온라인 학습의 설계에서 DBR의 연구방법론은 사람과 시스템을 혁신하는 데 유용하게 적용될 수 있다고 하였다.

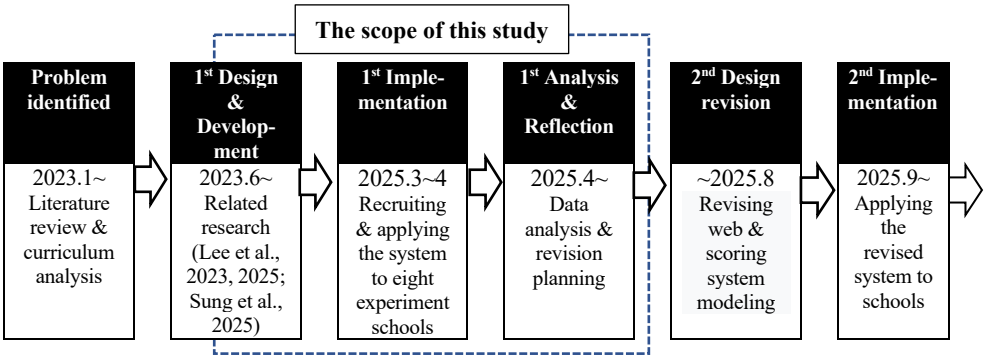
본 프로젝트는 그림 1과 같이 순환적인 DBR 연구과정의 틀을 따랐으며, 특히 본 논문은 빗금 친 영역에 해당한다. 즉 1차 설계 개발을 완료한 후, 첫 번째 실행과 결과 분석을 통한 성찰의 단계이다. 따라서, 본 연구의 절차 기술 방법은 다음과 같다. 서론에서 간략히 다룬 1단계 문제점 인식을 제외하고, 첫째, 2단계인 '1차 설계 개발'을 소개한다. AI 기반 맞춤형 영어 말하기 진단-학습-평가 시스템의 개념 정의와 스픽마스터의 구조도를 설명한다. 이를 위해 사전 연구로 ASS 기술 연구, 자동 채점을 토대로 한 국내외 말하기 테스트 현황 및 문항 분석 연구(Lee et al., 2023)와 해외 Common European Framework of Reference (CEFR) 지표를 근거로 한 우리나라 영어 말하기 성취기준 분석 및 이원분류(Lee, Choi, Sung, & Kim, 2025), 또한 ASS 채점 모델링을 위한 말하기 구인과 분석 알고리즘 등에 대한 연구(Sung, Lee, Kim, & Choi, 2025)가 진행된 바 있다. 둘째, 본 연구의 주요 초점인 3단계로 1차 실행이다. 8개의 실험학교를 운영하면서 1개월간 시스템을 수업에 적용하였고, 수업 관찰, 교사 인터뷰, 설문조사를 실시하였다. 본 연구의 범위는 1차 실행 결과로 수집된 언어 데이터, 레벨 테스트 결과, 수행 평가 결과, 참여자 피드백을 분석하여 시스템 설계의 적합성을 성찰하는 단계이다. 향후 본 연구를 토대로 시스템의 2차 개발을 시행하여 고도화를 추진하고, 각 학교에 보급하면서 순환적인 DBR 연구를 이어갈 예정이다.

본 연구의 첫번째 연구 문제는 이론적 배경과 사전 연구를 토대로 한 맞춤형 말하기 평가라는 시스템과 콘텐츠와 기술 설계의 논의이다. 연구결과의 1차 설계와 개발(그림 1의 두번째 단계)에서 주로 다루기로 한다. 두번째와 세번째 연구 문제는 적용에 대한 다각적인 결과 분석이다. 1차 시행 (그림 1의 세번째 단계)과 1차 분석 및 성찰(그림 1의 네번째 단계)에 해당하며 기능의 적절성, 참여자인 교사와 학생

인식, 만족도 등을 근거로 논의할 것이다.

FIGURE 1

The Research Procedure of the Entire Project and the Scope of the Current Study



2.2. 연구 대상

본 연구에는 전국 초, 중, 고 총 8개의 학교 1,451명의 학생과 8명의 교사가 참여하였다(표 1). 이들 교사는 연구자들의 개별 채널을 통한 공고를 보고 참여 의사를 밝혔다. 사전 워크숍을 하여 시스템을 소개하였고, 3월 둘째 주부터 순차적으로 교사의 일정에 따라 수업에서 시스템을 사용하였다. 이때 수업에서 레벨테스트와 수행 평가도구를 이용한 발표형 활동을 최소 각 1회씩 진행하고, 반별 수업일지, 사후 개별 인터뷰 등을 요청하였다. 실험 학교 학생들은 표 1과 같이 선택한 등급을 통해 레벨 테스트에 응시하였다. 등급은 국가교육과정에 맞춰 총 4개로 제시하고, Rookie(초 3–4), Semi-Pro(초 5–6), Pro(중), Master(고)로 명명하였다.

TABLE 1

Level Test Participation Status and Proficiency Levels of Student Participants in the Experimental Schools

Name	School	Location	*N*	Rookie	Semi-Pro	Pro	Master
A	Elementary	Gyeonggi	339	297	96	3	0
B	Elementary	Seoul	227	99	116	10	0
C	Middle	Seoul	130	3	2	116	0
D	Middle	Gyeonggi	152	1	0	116	1
E	High	Chungbuk	150	4	5	5	43
F	High	Gyeongbuk	199	4	0	0	130
G	High	Sejong	153	39	16	19	30
H	High	Seoul	101	0	0	30	46
	Total		1,451	447	235	299	250

2.3. 데이터 수집 및 분석 방법

본 연구는 3단계 DBR 과정에서 아래와 같은 네 가지 유형의 데이터를 수집하였으며 이를 양적, 질적으로 분석하였다. 또한 각각의 결과를 반복적으로 상호 교차 비교함으로써 분석의 다각화(triangulation)를 시도하여 보다 신뢰할 수 있는 결과를 도출하고자 하였다.

첫째, 학습운영시스템에 기록된 학생 학습 로그이다. 교사용 운영 시스템에는 반별 등록학생 명단과 학습 기록이 저장되어 있고, 각 교사가 생성한 과제와 이를 수행한 학생의 결과물을 확인할 수 있다. 교사 대시보드에 남아있는 기록과 학습 기록은 엑셀시트로 다운로드 받을 수 있으며, 이를 통해 과제 생성, 부여, 완료 횟수, 각 학생의 사용내역 및 결과 리포트, 레벨테스트 응시자 수, 응시 결과 등 다양한 사용 패턴을 분석하였다. 개인정보는 암호화하여 식별할 수 없도록 처리하였다.

둘째, 교사 일지와 SNS 소통기록이다. 교사 일지는 연구자가 제공한 양식에 따라 실험 학교 참여 교사가 진단 프로그램의 레벨 테스트를 실시할 때마다 1) 사용 내용(성공 건수), 2) 수업 상황 관찰, 3) 프로그램 의견을 간략히 작성하였다. 일지를 작성하여 제출한 교사는 총 7명(A–G교)이며, 이는 학교급별, 내용별, 개선사항 등의 세부 카테고리에 따라 분류하여 반복 검토되었고, 기술 수정요청사항 등은 목록화 되어 수시로 계속 반영되었다. 활용상에서의 학생들 반응, 교사의 생각 등은 주제별로 축코딩 분류하여 반복적인 패턴을 찾아 나갔다. 또한 실험 중에 실시간 SNS 오픈채팅방을 운영하며, 수업 도중 급한 도움 요청, 수업 후 문제점 지적, 실시간 반응 등을 수집하여, 기술 개선을 지원하였고, 모든 대화는 파일 저장하여 전 처리 후 키워드 빈도 분석하였다.

셋째, 전화 면접을 시행하였다. 실험 학교 운영기간이 종료된 후 5인의 교사(C, D, E, F, G교)와 1:1 전화 면접을 실시하였다. 전화 면접은 각각 30–70분 가량이 소요되었으며, 면접에서 다룬 내용은 1) 전반적인 소감, 2) 사용 일정과 시행 방식 및 수업 상황, 3) 레벨테스트에 대한 의견, 4) 파트별 과업 전반에 대한 상세 의견, 5) 등급/난이도/점수 타당도 및 학생들의 반응, 6) 성적 리포트에 대한 의견 등이었다. 모든 전화 면접은 참여 교사의 허가를 받고 녹음하여 자동 전사하였고, 반복적인 검토를 통해 축코딩 한 후 일지와 교차분석 하였다.

넷째, 사후 온라인 학생 설문조사를 시행하였다. 설문 문항의 이해 등에 있어 어려움이 예상되는 초등학교를 제외하고, 시스템 사용이 비교적 원활하게 이루어졌던 C, D, G 3개교를 선정하여 시행하였으며 총 244명이 응답하였다. 설문 문항은 1) 기초조사 5문항 2) 레벨테스트 이해도 7문항, 3) 인식 및 난이도 7문항 4) 레벨테스트 리포트 이해도 10문항 5) 개방형 문항 1개로 구성되어 있으며, 본 연구에서는 2), 3), 5)의 영역을 평균, 빈도 분포 등을 기술통계로 분석하여 전체 사용자의 인식 패턴을 검토하였다.

다양한 소스를 통해 수집된 데이터는 우선 다섯 개의 대주제인 ‘내용적절성’, ‘기술적절성’, ‘설계적절성’, ‘학습자인식’, ‘교사인식’으로 분류되었고, 각 주제별로

내용적절성은 ‘흥미도’와 ‘난이도’ 라는 하위의 생각단위(thinking unit)로 분류되었으며, 기술적절성과 웹설계적절성은 ‘편의성’, ‘시스템이해도’, ‘오류’, ‘불완전성’ 등의 하위 단위를 도출하였다. 또한 학습자인식은 ‘만족도’, ‘유용성’, ‘사용용의성’으로, 교사인식은 ‘활용가치’, ‘개선사항’의 하위 영역으로 나누어 패턴 분석되었다. 이는 3.3 성찰부분에서 상세히 다루기로 한다.

3. 연구 결과

3.1. 1 차 설계 개발

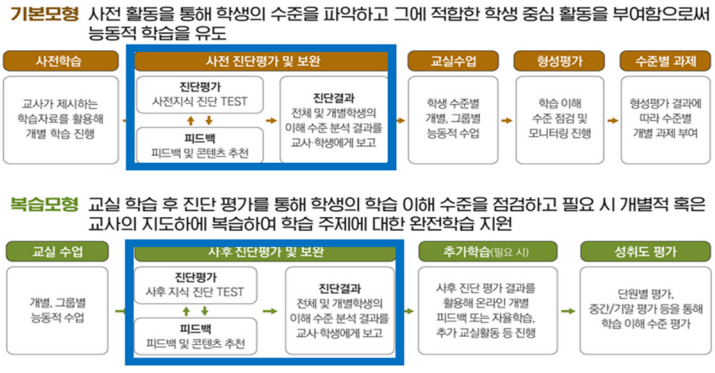
3.1.1. 맞춤형 영어 말하기 교육의 개념도 설정

맞춤형 교육의 개념은 상·중·하로 분류하여 수준별 학습을 제공하는 방식에서 탈피하여 학생 개개인의 차이를 존중하는 ‘개별화’ 학습으로 발전되었다(Jung, 2017; Jung & Hong, 2021). 맞춤형을 하는 방식은 학습자 차이를 고려해 학습 선택권을 부여하거나(Choi, Hong, & Park, 2019), 개인의 요구와 흥미, 능력, 학습 스타일에 맞추어 설계된 교육(Kim, Yoon, & Park, 2020)을 기본으로 한다. AI기술이 교육에 본격 소개된 이후 맞춤형은 주로 시스템, 데이터, 온라인, 인공지능, 효능감, 평가, 역량 등과 연결되어 왔다(Jung & Hong, 2021).

교육부는 2019년 이래 인공지능기반 교육을 확대하기 위한 다각적인 노력을 해왔으며, 2023년 2월 ‘모두를 위한 맞춤 교육’을 실현하기 위한 ‘디지털 기반 교육혁신 방안’이라는 기치 하에 그림 2와 같이 맞춤형 교수학습모델을 제시하였다. 맞춤형 교육에는 1) 사전 진단과 수업학습 이후의 사후 평가, 2) 개별 맞춤형 피드백 및 3) 수준별 학습 활동 추천이 핵심요소라고 할 수 있다.

FIGURE 2

Teaching-Learning Models Utilizing Digital Technology (Ministry of Education, 2023, p. 15)

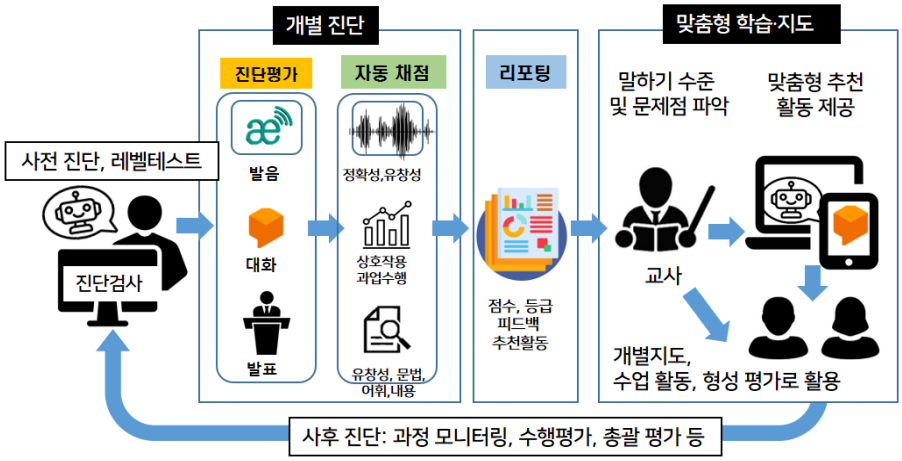


즉 효과적인 맞춤형 교육은 사전 진단을 통해 부족점을 개선할 수 있는 활동을 추천하고 연습하도록 하는 것이며, 이후 사후 진단을 통해 학습의 효과를 확인하는 것이다. 본 연구는 이를 토대로 영어 말하기 역량을 진단하기 위한 개념도를 그림 3과 같이 시각화 하였다.

첫째, 학습자는 수업 혹은 가정에서 자신의 총체적 말하기 능력을 진단한다. 진단평가는 발음 능력, 대화 능력, 발표 능력의 3부분으로 나누어 시행되며, 자동채점 기술을 통해 정확성, 유창성, 상호작용, 문법, 어휘, 내용 등 세부적인 부분을 측정한다. 둘째, 개별진단 직후 시스템에서 부여하는 총점과 영역별 등급 및 피드백을 확인한다. 성취기준 대비 자신의 실제 말하기 능숙도 수준을 파악하고, 부족한 점을 발전시킬 수 있는 활동을 추천받는다. 셋째, 교사는 학생들의 점수를 모니터링하며, 수업 활동을 선택한다. 교실에서 말하기를 지도하고 평가하는 데 시스템이 지원하는 다양한 등급별 말하기 과업을 수업 목표와 내용에 맞게 자율적으로 활용한다. 넷째, 사후평가로 말하기 능력의 향상도를 확인한다. 개별 진단은 수시로 가능하며, 교사의 수행 평가 혹은 총괄평가의 도구로 활용가능하다. 평가 리포트는 누적 기록되므로, 학습 발전 과정을 모니터링하는 데 효과적이다.

FIGURE 3

Cyclical Conceptual Model of a Personalized Diagnosis, Learning, Assessment System



3.1.2. 시스템의 구조 및 영역별 구성

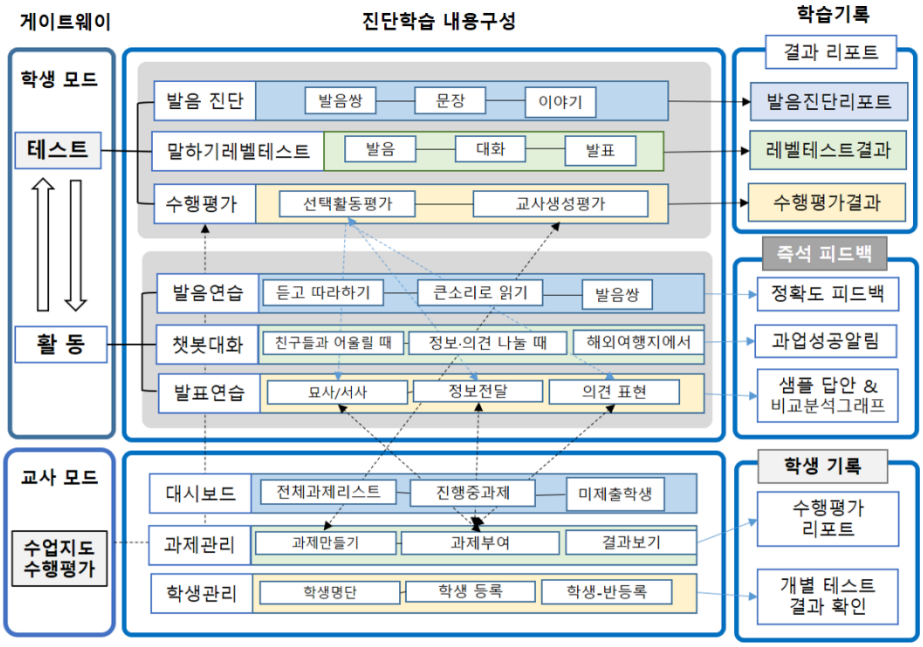
의사소통기능 표현 연습에 그쳤던 기존 교과서 활동을 총체적 말하기 능숙도 평가 및 활동으로 변화하는 것은 간단하지 않다. 본 연구에서는 Lee 등(2023)이 분석한 국제 말하기테스트의 기본 구인을 따르고, CEFR과 2022 개정 교육과정에서 제시하는 등급별 말하기 수준(Lee et al., 2025)을 고려하여, 영어 말하기 진단·학습·평가 시스템을 총 네 등급으로 설계하였다. 그림 4에서 보는 것처럼

학생모드에서는 사전 진단과 학습, 사후 평가라는 학습의 순환적인 활동을 위해서는 크게 '테스트'와 '활동'이라는 두가지 메뉴를 구성하였다.

먼저 '테스트'에는 세가지 하위 영역이 있다. 첫째, '발음 진단'에서는 의미를 좌우하는 음소 구별 여부, 발음쌍, 문장, 이야기 읽기 능력을 수준별로 제시하여 검사한다. 둘째, '말하기 레벨 테스트'는 말하기 총괄 능력 평가이다. 기존 말하기 표준평가 시험과 동일한 항목을 구성하여, 발음, 대화능력, 발표능력을 측정한다. 이를 위해 이원분류를 통한 등급별 말하기 과업수준을 개발하였고(Lee et al., 2025), 이를 토대로 평가 및 활동 과업을 개발하였다. 세 번째는 '수행 평가' 영역이다. 이미 만들어 놓은 평가 문항을 선택하여 수행 평가를 시행하거나, 교사가 직접 말하기 과제를 생성하도록 지원한다. 발음 진단, 말하기 레벨 테스트, 수행 평가는 응시가 종료되면 자동으로 결과 리포트가 생성되어 즉시 제공된다.

FIGURE 4

The Architecture of English Speaking Diagnosis, Learning, Assessment System, *SpeakMaster*



두 번째 '활동' 메뉴에는 '발음 연습', '챗봇 대화', '발표 연습'으로 나뉘어 말하기 영역을 고루 학습할 수 있도록 등급별로 다양한 AI 기반 과제를 제공한다. '발음 연습'에서는 단어, 문장, 이야기 수준에서 '듣고 따라하기'와 '소리내어 읽기' 활동, 그리고 음소구별 연습하는 '발음쌍' 활동을 제공한다. '챗봇 대화'에서는 챗봇과의 상호작용을 하며 주어진 상황에서 자신의 생각을 말하는 과업을 제공하였다. '발표 연습'에서는 짧은 자기소개부터 의견 주장에 이르기까지 최소 문장 이상을 발화하는

연습 과업을 제공한다. 세 가지 연습 영역 모두에서 학습자는 각 문항을 완성할 때마다 즉각적인 피드백을 받을 수 있다.

또한 교사 모드가 있는데 이는 학생들의 진단평가 결과를 모니터링하고, 말하기 과제를 부여하고 관리하는 영역이다. '대시보드'는 전체 과제관리와 학생들의 진도 등을 확인할 수 있으며, '과제관리'는 수행 평가 등 말하기 과업을 직접 생성하거나 제공되는 과업을 선택하여 학생들에게 부여하고, 보관할 수 있다. '학생관리'는 반 혹은 학생을 등록하고 명단을 관리한다.

3.1.3. 웹 화면 설계

웹화면은 모바일, 태블릿 PC, 노트북, 데스크탑 화면 모두에서 사용가능하도록, 최대한 단순하고 이해하기 쉽게 구성하였다(그림 5). 메뉴 등을 최소화 하고, 한눈에 파악할 수 있는 제목을 사용하였다. 또한 '챗봇 대화'와 '발표 연습' 화면 등의 상황에 대한 이해를 돕고, 생각의 단서가 될 수 있도록 상세한 삽화를 제공하였다.

FIGURE 5

The Screenshots of *SpeakMaster* Website

3.1.4. 자동 채점 결과 리포트 및 피드백

그림 6과 같이 학습 기록은 크게 결과 리포트와 즉석 피드백 방식으로 제공되며 총 6가지 형태가 제공된다. 먼저 결과 리포트는 테스트 메뉴의 모든 진단 검사의 결과지로 시험 응시 후 바로 확인할 수 있으며 학습자의 계정에 보관이 되어 자신의 향상도를 모니터링 하는 데 참고할 수 있다. 첫 번째, 발음 진단 평가 리포트로 그림 6에서 보는 것과 같이 전체 총점과 총평, 개별음(자음, 모음) 오류 분석, 초분절음 분석, 추천 활동과 회차별 종합진단 결과를 제시한다. 두 번째, 말하기 레벨테스트 결과 리포트는 등급내 총점과 총평, 말하기 개별 영역에 대한 3 등급평가, 추천 활동을 제시한다. 세 번째, 수행 평가 결과는 교사의 선택에 따라 레벨 테스트형과 발표 연습형을 선택할 수 있으며, 두가지 유형의 복합형으로 볼 수도 있다.

FIGURE 6

Examples of Result Report Formats: Level Test (upper left), Performance Assessment (lower left), and Pronunciation Diagnosis (right)

즉석 피드백 방식은 활동 메뉴에서 제시하는 모든 말하기 과업을 학습함과 동시에 제공된다(그림 7). 첫째로, 발음의 경우 정확도 피드백을 받게 된다. 각 음소별로 얼마나 정확한 발음을 했는지, 활동에 따라 속도, 초분절음 요소(강세, 억양 등)에 대한 피드백을 제시한다. 두 번째로, 챗봇 대화의 경우 각 과업을 성공적으로 수행하고 나면 과업 성공을 알리는 팝업창과 과업에 걸린 시간, 턴 수, 오류 개수 등을 보여준다. 셋째, 샘플 답안 비교분석 그래프로 원어민 모범 답안과 말하기 평가 영역별 샘플 답안과의 차이 등을 제공한다.

FIGURE 7

Examples of an Instant Feedback Method: Pronunciation (upper left), Chatbot Conversation (upper right), Presentation Practice (lower)

3.2. 시행 및 검토

참여교사에게 요청한 레벨테스트와 수행 평가 1회 시행결과를 중심으로 수업이 이루어졌으나, 교사에 따라서는 발음진단테스트와 기타 말하기 활동 영역도 수업에 활용하였다. 본 연구에서는 학습자 로그가 기록되는 레벨테스트와 수행 평가를 중심으로 결과를 설명하기로 한다.

3.2.1. 레벨테스트 시행 결과

실험 학교는 각 학교 일정에 맞게 순차적으로 수업에서 레벨테스트를 실시하였다. 총 시행 건수는 1,857회였는데, 이는 추가로 응시한 인원을 포함한 숫자이다. 교사 일지, SNS 로그, 교사 면접 등의 데이터를 분석 결과를 정리하면 표 2와 같다.

첫째, 초기 기술적인 문제가 해결된 후 시행에는 큰 어려움은 없었으며 다수의 학생들이 결과 리포트를 받는 데 성공하였다. 첫 주는 시스템의 적응기로 수많은 오류와 문의가 잇달았다. 한 주 먼저 시작하였던, A초, C중, E고, G고 참여 학생들의 경우 버퍼링, 음성인식 프로그램 기술 오류, 디바이스 출력 등의 문제로 진행상의 어려움을 겪었으나, 문제해결 기간을 거쳐 점차 안정화되었다.

가장 많은 학생들이 참여한 A초의 경우 첫 주 8개반 수업 성공률이 21.77%(248명중 54명 성공)에 불과하였으나, 그 다음주 8개반 수업에는 83%(231명중 193명 성공)의 학생이 성공적으로 레벨테스트를 마쳤다. E고 역시 동일한 양상을 보였는데, 첫

수업에서는 25명 중 0명, 그 다음 두개 반 71명 중 45명만이 성공을 하는 등 진행상에 어려움을 겪었지만, 한 주가 지난 후 총 7개반 171명 중 130명이 완료하여 76%가 진단 테스트를 성공적으로 시행하였다. 소요 시간도 성공한 경우 30–40분이었지만 이후 D중은 10-25분, G고는 7–13분만에 완료하였다고 보고하였다.

TABLE 2

The Result of *SpeakMaster* Level Test in Seven Schools

School	Attempted Level	Time Taken (min)	Attempts (*N*)		Completion (*N*)		Rate (%)	
			Total	2nd wk	Total	2nd wk	Total	2nd wk
A	R, S	N/A	653	231	355	193	54	83
B	R, S, P, M	?	365	365	252	252	69	69
C	S, P	35-45	128	0	43	N/A	33	N/A
D	S, P	10-25	152	152	105	105	69	69
E	M	30-40	292	171	177	130	60	76
F	M	30-40	150	0	37	N/A	25	N/A
G	R, S, P	7-13	117	117	84	84	71	71
Total			1,857	1,036	1,053	764	56	73

Note. R = Rookie; S = Semi-Pro; P = Pro; M = Master, representing grade levels by school stage

둘째, 기술 오류가 없어도 처음 시도하는 경우 10명중 평균 7명 정도의 학생만이 레벨테스트를 완료하는 것으로 드러났다. 기술 환경이 개선된 후 실시한 B초, D중, G고의 경우 응시 학생 모두 1회씩 시도하여 69%, 69%, 71%의 유사한 성공률을 보였다. A초가 기술 개선 후 83%가 성공하였으나, 거의 대부분의 학생들은 두 번 반복 시도한 결과인 것으로 분석되었다. 데이터 분석 결과, 약 30%의 학생이 완료하지 못한 이유로 가장 빈번하게 언급된 것은 학생의 영어능력 부족이었다. 선택한 테스트가 학생 수준에 비해 너무 어려워서 자유 대화와 즉흥 발표 영역에서 대답을 하지 못하는 경우가 많았다.

> 저희 학생들이 조금 레벨이 낮은 편이었어요. 학생들 수준으로는 좀 많이 어려워 보였거든요. (F 고 교사면접)

> 영어를 잘하지 못하는 친구들, 태블릿 오류가 있던 학생들 말고는 말하기 레벨 테스트, 수행 평가를 수월하게 진행함. (G 고 수업일지)

둘째, 피로감이나 좌절감으로 중간에 포기하는 경우가 있었다. 테스트의 길이가 길어서 피로감을 보이는 경우가 많았고, 자신이 한 말을 인식하지 못할 때(예를 들어, towel을 tower로 인식) 부정적 감정이 커짐이 목격되었다.

> minimal pair 발음 테스트가 끝난 뒤부터 생각보다 테스트가 길다고 느낀 학생들 위주로 중도포기자가 발생하기 시작. (G 고 수업일지)

> 반복적으로 tower 이라고 나와 넘어가게 되자 그 부분에서 흥미를 잃은 아이들이 많았습니다. 심지어 학급에서 영어를 제일 잘하는 편에 속하는 아이가 이 부분 때문에 레벨 테스트를 완료하지 못해 좌절하기도 하였습니다. (B 초 수업일지)

3.2.2. 수행 평가 시행 결과

F고를 제외한 모든 참여 교사가 수업시간에 수행 평가를 시행한 결과, 실험학교에서 교사가 부여한 과제 건수는 표 3에서 보는 것처럼 총 1,492건으로 집계되었다. 교사는 자율적으로 자신의 진도에 맞추어 말하기 수행 과제를 부여하여 시행하였으며, 이를 완료한 학생은 902명으로 60.4%였다.

수행 평가 활동에서 나타난 특성은 다음과 같다. 첫째, 교사는 스스로 말하기 과제를 생성하는 것을 선호했다. 스픽마스터에는 이미 80여개의 수행 평가용 과제가 제공되고 있었지만, 이를 선택하여 시행한 교사는 C중과 E고뿐이었으며, 나머지 학교는 모두 교사가 직접 만든 과제를 활용하였다. 둘째, 생성한 수행 평가 과제의 활용 방식이 매우 다양했다. A초의 경우는 3학년과 6학년 대상으로 수업내용을 기반으로 말하기 평가 과제를 만들었으며, 3학년의 경우 많은 학생이 100점을 획득하였다. B초 교사 역시 4가지 말하기 과제를 직접 만들어서 수업에 시행하였는데, 자기 소개하기와 같은 기존에 있는 말하기 과제와 매우 유사한 자기소개였지만, 그대로 사용하지 않고 레벨별로 2개씩 다시 제작하여 사용하였다. 한편, D중과 H고 교사가 생성한 말하기 과제는 스픽마스터에서 제공하는 과업과는 매우 다른 문법 표현과 정확성을 강조한 형태였다.

TABLE 3

The Result of *SpeakMaster* Performance Assessment in Seven Schools

School	*N* of Assign	*N* of Finish	*N* of Tasks	Level (task type)
A	329	225	2	R (created), S (selected)
B	227	117	4	R (created), S (created)
C	119	94	1	P (selected)
D	75	73	1	P (created)
E	396	303	2	M (created)
F	43	41	7	M (selected)
H	303	49	1	M (created)
Total	1,492	902		

3.3. 성찰

3.3.1. 내용 및 기술적 설계의 적절성

내용 적절성

SNS, 교사 일지, 교사 면접과 설문 등 다각적인 데이터 소스로 교차 분석한 결과, 흥미도와 난이도면에서 비교적 적절한 것으로 드러났다. 첫째, 참여 학생들은 말하기 프로그램에서 제공하는 말하기 평가 과업과 분석결과에 상당한 흥미를 보였다. 레벨테스트에서 가장 흥미를 끌었던 것은 파트2챗봇대화였고, 그 다음으로 발음 연습의 피드백과 리포트에도 상당한 관심을 보인 것으로 드러났다. 전 등급에 걸쳐 챗봇 대화에 관심이 많았지만, 초등, 중등이 보다 긍정적이고, 고등의 경우는 대화의 자율성이나 유연성에 대한 비판적 태도 또한 자주 목격 되었다.

> 챗봇 활동도 재미있어 하며 영어가 어려운 학생들은 파파고로 번역기를 사용하면서 활동함. (A 초 수업일지)

> 레벨 테스트할 때도 이제 애들이 챗봇이랑 대화하는 거를 되게 즐거워했었고 (C 중 교사면접)

> 애들이 레벨 테스트보다 챗봇 대화를 더 재밌어 했어요. 훨씬. (G 고 교사면접)

대화 영역을 특별히 좋아했던 이유를 면밀히 분석한 결과 학생들이 챗봇을 좋아하는 이유는 '상호작용'의 힘인 것으로 드러났다. 자신이 한말을 상대방이 알아듣고 반응을 보이며, 원하는 말을 주고받는 것에 대한 기쁨을 보였다.

> 모놀로그만 많이 익숙해져 있는데 사실 아이들의 요구는 진짜 뭔가 대화를 하고 싶어 하는 아이들이 되게 많거든요. (C 중 교사면접)

> 애들이 제일 좋아하는 것도 챗봇이고 제일 싫어하는 것도 챗봇이에요. 일단 내 말을 알아듣는 거에 너무 흥분해요. 애들이 "제가 이렇게 말했는데 알아들었어요." 하면서 "선생님 제가 이렇게 설명했더니 이렇게 막 애가 대답을 했어요." … 근데 동시에 "내가 이렇게 말했는데 왜 못 알아들어" 라고 분노를 해요. 너무 분노해요. (G 고 교사면접)

> 로봇과 티카타카가 되어 재미있었다. (학생 설문)

데이터 분석 결과, 학습자는 말하기 연습에 있어 상대 대화자와의 의미있는 내용의 교환에 대한 흥미와 몰입을 보인다는 점을 알 수 있었다. 한편, 발음 연습 활동의 경우 학습자들이 관심을 보인 이유는 영어 발음은 자신이 없고 개선하고 싶은 영역이며, 또한 제공된 활동이 어렵지 않다고 생각하기 때문인 것으로 보인다. H고의 경우, 자율적인 이용 후 남긴 사용자 후기에 장단점에 관한 개방형 문항에 답을 한 20명 중 11명이 발음에 대한 긍정 평가를 하였다.

> 저의 발음에 대한 프로그램의 진단이 생각보다 적절하였고, 이를 토대로 더 정확한 발음을 구사할 수 있게 되었습니다.

> 영어로 대화, 발표를 할 때 상대방이 내 발음을 안 좋게 들을까 걱정했었는데 이번에 활동을 하며 잘못된 발음을 수정할 수 있어서 좋았다.

> 발음이 일치하지 않는 단어를 다시 연습 시켜 주는 시스템도 아주 만족스러웠고, 정말 실력을 키울 수 있었습니다.

많은 학생들은 연휴기간에 단발성으로 진행한 말하기 활동만으로도 자신의 발음 실력이 향상되었다고 인식하였고, 이에 대한 만족감을 보였다.

> 발음이 안 좋은 애들이 되게 많아요 생각보다. 그래서 '듣고 따라하기' 그러니까 일단 되게 이제 inhibition 이 낮잖아요. 듣고 따라하는 거 너무 쉽잖아요. (C 중 교사 면접)

또한 교사들은 학생들이 흥미를 느낄 요소로서 '실제성(authenticity)'을 꼽는 경우가 많았다.

> 주제들이 되게 리얼 라이프와 맞닿아 있다라는 생각을 처음에 했고요… (중략) 진짜 실제적이어서 특히 마지막에 그 호텔 예약하는 것들은 동기부여가 될 수 있겠다. (D 중 교사면접)

> 그리고 교육과정에 기반해서 어쨌든 많은 이제 화행들이 들어 있으니까 수업시간에 한 거…그걸 이제 거기 프로그램 들어가서 또다시 진짜 리얼한 컨텍스트에서 다시 한 번 해보는 그런 거가 좋은 것 같아요. (C 중 교사면접)

둘째, 내용의 난이도 면에서는 학교급별, 학교별 편차가 큰 편이었으나, 대체로 난이도가 무난한 것으로 평가할 수 있다. 중, 고등학교 3개교 참여자를 대상으로 조사한 설문에서 '프로그램 내용은 쉬웠다' 혹은 '학교의 말하기보다 쉬웠다'는 평가를 하였다. 표 4에서 보는 것처럼 C중과 D중은 같은 중학교임에도 평균 3.14와 2.82로 레벨테스트 난이도를 느끼는 정도가 꽤 달랐다. G고의 경우는 평균 4.08로 중학교보다 더 높았다. 그러나 파트 1 발음 테스트의 경우 평균이 각각 4.04, 3.50, 4.04로 다른 영역보다 훨씬 높은 점수를 보이며, 쉽다는 사실에 대부분 동의하였다. 이는 앞서서 C교사가 발음 연습을 좋아하는 이유는 '시도하기가 쉬워서'라고 분석한 것과 맥을 같이한다. 챗봇 난이도에 있어서는 세 학교 모두 매우 비슷한 수준으로(M = 3.43, 3.29, 3.42), 난이도가 보통인 것으로 나타났다.

TABLE 4

Survey Results of Perceived Difficulty of *SpeakMaster*

N	Item	School (*N*)	Response Ratio (%)					*M*	*SD*
			1	2	3	4	5		
13	The level test was easier than the school speaking test.	C (51)	7.8	13.7	41.2	31.4	5.9	3.14	0.99
		D (141)	9.2	23.4	44.0	17.0	6.4	2.82	1.09
		G (52)	1.9	0.0	40.4	25.0	32.7	4.08	1.14
14	Part 1: Pronunciation was easy.	C (51)	2.0	2.0	15.7	51.0	29.4	4.04	0.83
		D (141)	2.1	10.6	40.4	27.0	19.9	3.50	1.00
		G (52)	0.0	1.9	28.8	32.7	36.5	4.04	0.96
15	Part 2: Conversation was easy.	C (51)	3.9	11.8	39.2	27.5	17.6	3.43	1.03
		D (141)	3.5	19.1	44.7	20.6	12.1	3.29	1.06
		G (52)	5.8	13.5	30.8	23.1	26.9	3.42	1.25
16	Part 3: Presentation was easy.	C (51)	11.8	21.6	29.4	25.5	11.8	3.04	1.19
		D (141)	10.6	22.0	41.8	18.4	7.1	2.99	1.09
		G (52)	13.5	21.2	42.3	11.5	11.5	3.08	1.18

가장 어렵게 느낀 것은 발표였는데, 즉흥적으로 시행하는 활동 경험이 거의 없어 더욱 어려움을 느끼는 것으로 파악되며, 단답형이나 짧은 문장이 아닌 비교적 긴 스피치를 해야한다는 점에서 이 결과는 당연하다고 볼 수 있다. 난이도를 느끼는 상황을 분석해본 결과, 몇 가지 반복적인 패턴이 있었다. 첫째, 참여 학생은 사전 표현 학습 없는 내용 중심 즉흥 말하기에 대해 익숙하지 않았다. 초등학생의 경우는 이미 축적된 학습데이터가 없어서 표현이 조금만 달라져도 대응할 수 없었다. A교사는 거의 매 실험 수업에서 교과서에 나온 표현으로 프로그램 수정을 요청했다. 예를 들어 how are you doing?은 how are you로, Can I take your order?는 May I help you?로 변경을 부탁했다. 또한 B교사는 How are you?는 교과서에서는 인사가 아닌 기분을 묻는 표현으로 소개되므로, 학생들이 I am happy라고 말하는 경향이 있으므로, 이러한 반응도 정답으로 인식해야 한다고 제안하였다. 고등학교에서는 학생 간 영어 능력 차가 심해서 챗봇의 성능을 비판하는 수준의 학생과 발음 연습 외에는 진행조차 못하는 학생으로 갈리기도 했다.

> 좀 극단적으로 반으로 갈려서 잘하는 친구들은 막 프리토킹 하면서 왜 내 프리토킹을 알아듣지 못하느냐 라고 화를 내는 상황이었고, 자기 공부하는 친구들은 약간 교과서스러운 대화하면서 잘 넘어갔고 그 아래에 있는 친구들은 …대화에서 많이 탈락들을 했어요. (G 고 교사면접)

> 사실 수업 고등학교 수업에서는 이제 말하기를 많이 다루지 않아서… 학교에서 배우는 것도 이제 다 수행 평가를 위한 준비 과정이라서 이런 걸 따로 배우지는 않아서 낯설 수는 있는데… 아이들이 많이 어려워했거든요. (F 고 교사면접)

둘째, 난이도는 과업의 모호성에서 비롯된다는 의견도 빈번히 반복되었다. 특히 챗봇 대화에는 구체적인 절차와 목표가 있었으나, 학생들은 이를 무시하거나, 놓치는 경향이 있었다. 이러한 이유로 많은 학생이 챗봇 미션에서 탈락을 하는 경우가 많았다. 일부 학생들은 아무 얘기나 가능한 자유 대화를 원했다.

> 또한, 아이들이 AI 와의 대화에서 기대하는 바는 '보기'에서 골라서 대화하는 것이 아닌 자유발화이다. '보기에 있는 내용을 반드시 사용하여'라는 말을 넣어야만 혼란이 없을 것 같다. (C 중 일지)

> 제 생각에는 여기서는 좀 명확하지 않다는 문제가 조금 있다고 느꼈어요. 그래서 과업 설정에서 좀 더 쉬운 걸로 설정해 두거나 힌트를 주면 어떨까 생각이 조금 들었고요. (G 고 교사면접)

기술의 적절성

약 한달 간의 기간동안 기술적인 오류와 개선사항은 끊임없이 쏟아졌다. 실험 학교 운영기간동안 오픈채팅방에 접수된 기술문제보고 혹은 개선 제안은 총 137건이었던 것으로 분석되었다. 사전 테스트 등을 수차례 거쳤지만, 첫 주에는 시스템 및 네트워크오류(버퍼링, 지연, 튕김, 33회), 음성인식 및 분석오류(오인식, 인식실패, 분석지연, 28회) 등의 문제가 발생하여 수업에서 활용하는 데 큰 차질을 빚기도 하였다. 그러나 이러한 부분은 두 번째 주부터는 대부분 확연히 줄어들며 안정화되었다. 그러나 그 외에도 매우 중요한 기술적인 개선 사항이 지적되었다.

우선 채점의 부정확성에 대한 보고가 많이 있었다. 실험 기간이라 다소 후한 점수를 받도록 임시적인 세팅을 하였는데, 거의 대부분의 참여자들은 자신의 점수가 높은 것에 대해 의구심을 표했다.

> 몇 개 활동을 스킵했는데도 97 점이라며 본인 말하기 점수가 너무 높게 나왔다고 했음. (E 고 수업일지)

> 4 학년은 100 점 많고 6 학년은 100 점 없다. (B 초 수업일지)

> 나는 못한 거 같았는데 점수를 후하게 준 거 같아 좋긴 했지만 좀 더 정밀하고 깐깐하게 테스트해줬으면 좋겠다는 생각이 들었다. (학생설문)

학생들은 점수에 매우 민감했으며, 기술 오류가 없는 상태에서 신뢰할 만한 점수를 제공하는 것이 매우 중요하다는 것을 알 수 있었다. 또한 복수의 교사들도 너무 후한 점수를 주는 것은 바람직하지 않다고 하였다.

둘째, 챗봇 대화의 불완전성이다. 챗봇 대화는 학생들의 예상치 못한 반응과, 미션의 절차를 숙지하지 못한 문제로 인해 대화의 단절이나 대화의 연결이 매끄럽지

않은 경우가 많아, 미션을 성공하지 못하는 사례가 많았다.

> 챗봇이 기대치가 이제는 GPT 에 맞춰져 있다 보니까 네 거기서 오늘 이제 조금 이거는 조금 아직 부족해요라고 느끼는 지점이 있는 것 같습니다. (G 고 면접)

> 말이 인식이 안 될 때가 있고 대화를 할 때 애가 원하는 대답이 아니면 이상한 답변 취급해서 대화가 잘 안된다. 애가 원하는 답변이 뭔지도 모르겠을 때가 많았다. (학생설문)

원활하고 자연스런 챗봇 대화를 위해서는 확보된 언어데이터 학습이 충분히 되어야 할 것이며, 대화의 단절이 일어날 때 이를 벗어날 수 있도록 해주는 챗봇 전략이 필요할 것이다.

셋째, 음성인식 정확도에 대해서는 대체로 만족하는 편이었다. 그러나 음성인식율이 지나치게 좋아서 발음을 틀리게 말하는 경우나, 심지어 한국말로 이야기할 때도 챗봇이 이해하는 경우가 있어서 이에 대한 개선이 필요했다. E교사는 음성인식 성능이 좋은 것이 교육적으로 좋은 것인지 모르겠다는 의견을 주기도 하였다. 또한 소음으로 인해 인식이 잘못되어 평가점수에 영향을 미칠 것에 대한 우려도 있었다.

> 근데 이상하게 녹음이 된 거예요. 그래서 봤더니 앞에 앉아 있는 애 소리가 중간중간 들어갔던 거예요. (C 중 교사면접)

> 아이들이 좀 다르게 발음한다든가 아예 발음을 안 한다거나 해도 발음을 했다고 넘어가는 것들이 아직은 있는 것 같아요. (D 중 교사면접)

소음의 문제는 수행 평가로 시스템을 사용하는데 장애물이 되었다. 다수의 교사가 소음 등의 이유로 시스템을 수행 평가로 사용하는 것에 대한 우려를 표하였다.

웹 설계의 적절성

스픽마스터 웹 화면의 설계에 대해서는 대체적으로 적절하다는 의견이 많았다. 학생들은 대체로 사용이 편리하고 직관적으로 이해하기 쉽다는 반응이었다. 표 5를 보면, 1번 문항에서 레벨테스트를 응시함에 있어서 기술적인 문제가 없었다는 학생의 평균이 각각 4.43, 3.52, 4.13으로 매우 높게 나왔다. D중은 가장 처음 레벨테스트를 진행하여 초기 안정화기간에 어려움을 겪었기 때문에 다소 낮게 평가한 것으로 추측되지만 대체로 어렵지 않게 완료한 것으로 보인다. 버튼이나 녹음기능들도 사용하기 용이했던 것으로 평가하였다(M = 4.38, 3.82, 4.04). 나머지 두 항목, 즉 챗봇 사용(M = 3.66) 이나 전체 프로그램 테스트 사용(M = 3.34)에 있어서는

이보다는 더 낮은 평가를 하였으나, 이 역시 비교적 긍정적인 평가로 볼 수 있다.

TABLE 5

Survey Results of Perceived Easy-to-use of *SpeakMaster*

No	Item	School (*N*)	Response Ratio (%)					*M*	*SD*
			1	2	3	4	5		
1	The speaking level test worked smoothly without any problems.	C (51)	2.0	0	11.8	25.5	60.8	4.43	0.84
		D (141)	6.4	14.9	25.5	26.2	27.0	3.52	1.25
		G (52)	0.0	5.8	15.4	38.5	40.4	4.13	0.81
2	The navigation and recording buttons were easy to use.	C (51)	2.0	2.0	9.8	29.4	56.9	4.38	0.88
		D (141)	2.8	7.8	19.9	40.4	29.1	3.82	0.99
		G (52)	0.0	1.9	28.8	32.7	36.5	4.04	0.96
3	Talking with the chatbot was simple.	C (51)	3.9	0.0	23.5	47.1	25.5	3.90	1.03
		D (141)	7.8	7.1	33.3	31.2	20.6	3.44	1.14
		G (52)	3.8	15.4	19.2	36.5	25.0	3.65	1.12
4	It was easy to understand what to do during the test.	C (51)	2.1	5.7	34.8	41.1	16.3	3.65	0.95
		D (141)	7.1	18.4	41.8	22.0	10.6	3.01	1.09
		G (52)	5.8	11.5	28.8	40.4	13.5	3.38	1.10

사용교사들 역시 프로그램의 사용의 용이성을 언급하였다. 기능의 위치나 구성에 대해 단순하고 어렵지 않다는 생각을 가지고 있었다.

> 심플하고 되게 직관적이어서 애들이 뭘 들어가서 그리고 메뉴가 그렇게 또 많지도 않잖아요. (C 중 교사면접)

3.3.2. 학생 만족도

사용만족도를 묻는 문항에서 '보통 이상'의 만족도를 보였다(표 6 참고). 실험 시기 등에 따른 편차가 약간 있었으나 대개는 긍정적인 느낌을 가진 것으로 드러났다. 레벨테스트를 보고 나니 영어공부를 열심히 하고 싶어졌다"(*M* = 3.45), "레벨테스트를 연습해서 다시 보고 싶다(*M* = 3.44)" 등에서 보통 이상의 호감을 보였고, 특히 "스픽마스터를 수업시간에 사용했으면 좋겠다"는 응답에는 문항 중 가장 높은 호감(*M* = 3.61)을 보였다. 다만 "스픽마스터로 학교 시험을 보면 좋겠다"는 것은 상대적으로 낮은 편이었는데(*M* = 3.42, 2.91, 3.33), 점수의 신뢰성의 문제와 더불어, 시험이라는 점에서 부정적인 인식을 한 것으로 추측해볼 수 있다.

또한 자율적인 사용방식으로 운영한 H고의 경우 전반적인 만족도가 평균 4.56으로 매우 높게 나왔으며, 사용 용이성(*M* = 4.32), 콘텐츠 적절성(*M* = 4.44) 등에서 모두 높은 만족도를 보였다. 3개 학교 설문 개방형 문항을 살펴보면, 작성자 127명 중 55명인 43.3%가 '재밌다' 혹은 '좋았다' 등의 느낌을 말하였다. 다음으로는 나를

진단할 수 있어 '유용하다'는 의견이 많았다(총 10명, 7%). 개방형문항에서도 "나의 말하기 레벨이 이 정도구나 알게 되었다." "부족한 점을 알게 되어서 좋았다." "이런 식으로 뭐가 문제인지 어떻게 고칠지 알려주면 영어실력이 늘 것 같다는 느낌을 받았다." 라고 답하였다.

TABLE 6

Survey Results of User Satisfaction with *SpeakMaster*

No	Item	School (*N*)	Response Ratio (%)					*M*	*SD*
			1	2	3	4	5		
1	After taking the test, I wanted to study English speaking more.	C (51)	3.9	9.8	37.3	25.5	23.5	3.55	3.45
		D (141)	5.7	10.6	50.4	21.3	12.1	3.33	
		G (52)	7.7	9.6	38.5	26.9	17.3	3.46	
2	I want to practice and take the speaking level test again.	C (51)	5.9	7.8	27.5	31.4	27.5	3.67	3.44
		D (141)	8.5	20.6	43.3	16.3	11.3	3.12	
		G (52)	7.7	11.5	32.7	21.2	26.9	3.52	
3	It would be good to use *SpeakMaster* in English class.	C (51)	3.9	3.9	21.6	33.3	37.3	3.96	3.61
		D (141)	5.0	12.8	51.8	17.7	12.8	3.23	
		G (52)	5.8	5.8	30.8	28.8	28.8	3.65	
4	It would be good to use *SpeakMaster* in school speaking test.	C (51)	9.8	5.9	39.2	21.6	23.5	3.42	3.22
		D (141)	10.6	17.0	43.3	19.9	9.2	2.91	
		G (52)	15.4	9.6	30.8	19.2	25.0	3.33	

3.3.3. 교사가 인식하는 활용가치 및 개선점

활용가치

첫째, 교사들은 프로그램이 레벨 진단 및 개별 추천에 도움이 되는 것으로 인식하였다. 학생들과 교사들 모두 말하기 레벨 진단에 관심을 보였고, 이를 통해 학생들에게 말하기 활동을 맞춤으로 선택할 수 있게 해주는 것에 활용가치를 느끼는 것으로 분석되었다.

> 이건 일단 개별 기기를 가지고 각자의 수준에 맞게 할 수 있으니까 개별화 시키는 건 되게 좋을 것 같아요. (F 고 교사면접)

> 일단 애들의 초창기에 애들의 실력을 파악하기에 너무 좋은 것 같고. (C 중 교사면접)

> 모든 과업에서 레벨별로 연습해 볼 수 있다는 게 가장 좋았던 것 같고요. 애들이 그냥 어려워서 못해라는 말을 할 수 없게끔 이거라도 해 봐라 할

> 수 있는 게 사실 교사 입장에서는 그게 정말 좋은 것 같아요. (D 중 교사 면접)

더 나아가 교사들은 실제 수업에서 학생들에게 맞춤 추천을 하거나, 학생 스스로 등급을 선택하면서 점점 자신감을 보이는 것을 확인하고 이러한 인식을 강화한다.

> 네. 처음에는 이제 저는 하나도 못해요. 루키 할래요 했더니…루키로 시작을 했다가 할 만하네. 다음 거 세미 프로 한번 해 볼래 나 프로도 갈래 나 마스터 100 점 한번 보여주겠어 하면서 점점 올라가더라고요. (G 고 교사면접)

> 학생들은 어려운 레벨 테스트를 진행하면서 좋은 자극을 받는 것이 느껴짐. (A 초 수업일지)

둘째, 프로그램의 활용이 말하기 수행 연습에 도움이 될 것이라고 인식하였다. 참여교사들은 스픽마스터의 활용가치 1순위로 말하기 수행 연습, 수행 평가전 연습용으로 활용하기를 희망하였다. 말하기 수행 평가는 한학기 1-2회의 민감한 시험으로 대부분의 교사들은 기존 방식을 고수하기를 희망하였고, 프로그램은 연습용으로 적합하다고 하였다.

> 수행 평가하기 전에 …그러니까 연습 단계에서 활용을 하면 선생님들이 개별적으로 애들이 다 연습할 시간 주고 피드백 주기 어려운데 거기 프로그램을 통해서 하면…분석이 되잖아요. 학생들이 자신의 테스트 수행 평가 준비하는 데 있어서는 도움이 될 것 같습니다. (F 고 교사면접)

> 실제 채점이라기보다는 애들이 학습을 하는 과정에서 연습용으로 써라 이렇게 될 것 같아요. (G 고 교사면접)

> 포트폴리오 평가로 … 사용할 수 있을 것 같습니다. (C 중 교사면접)

셋째, 실험 수업 이후 교사들은 학생들이 말하기의 목적은 의사소통에 있다는 사실을 이해하는데 도움이 되는 것으로 보았다. 사전 준비를 통한 형태 중심의 말하기 수행 평가에 익숙했던 학생들은 자기생각을 말하는 즉흥 발화에서 상당한 어려움을 느꼈다. 수업일지와 SNS를 분석해보면, 학년과 등급과 상관없이 파트 2 챗봇 대화에서 어떻게 시작할 지를 몰라서 입을 떼지 못하는 학생이 많았다. 이러한 학생들은 레벨테스트에서 시간이 지연이 되자 단계를 건너뛰고 싶어하거나 포기하는 경우가 자주 발생하였다. Rookie 등급 (초3-4)의 경우 모르는 단어나 표현이 하나가 나오면 해결하지 못하였다. A초 교사는 다음시간에 학생들에게 “한 단어라도

중요한 말을 해봐, 챗봇이 알아들을 거야”라고 안내하였더니 학생들이 부담을 내려놓고, 하고 싶은 말을 편안하게 하며 대화를 이어갈 수 있었다고 밝혔다. D중에서도 유사한 현상이 있었다.

> (챗봇 활동시) ..이제 한 두 단어를 말할 때 자기가 생각했을 때 중요한 단어를 말했는데 그게 이제 AI 가 잘 이해한다든가 그런 긍정적인 반응이 왔을 때 평소에는 그런 스피킹 활동을 잘 안 하는…학생들도 되게 호의적인 반응을 보이니까 이게 연습이 되고 이게 훈련이 되면 길게 말해야 하는 수행 평가에서도 아이들이 주눅 들지 않고. (D 중 교사면접)

마지막으로, 말하기 자동 평가 시스템은 교사 자신이 말하기 수업의 자신감을 얻게 하는데 도움이 되었다는 것이다. 분석결과 도출된 교사의 인식 변화는 미래 말하기 수업에 대한 자신감의 강화였다. 실험 수업 이후 지금까지는 기피하던 말하기 활동에 대한 동기부여가 되는 것으로 보인다.

> 고등학교는 확실하게 그런 걸(스피킹) 거의 안 하더라고요. 네 근데 그런 환경에 있을 때 아이들이 이미 개발된 것들을 가지고 어느 정도 연습을 하는 건 좋겠다. (G 고 교사면접)

> 학생들의 개인차도 있고 제가 많이 봐주고 싶은데 그게 어렵다고 생각해서 근 3 년 동안은 사실 제대로 된 스피킹 활동을 해볼 수가 없었는데 이렇게 뭔가 플랫폼이 체계적으로 있다 보니까 제가 좀 그런 부담을 덜면서 아이들이 조금 더 말을 할 수 있는 기회를 줄 수 있는 게 되게 좋았던 것 같아요. (D 중 교사면접)

개선 사항

첫째, 대부분의 교사들은 등급과 점수에 대해 조정이 필요함을 지적하였다. 특히, 후한 점수를 주는 것에 대한 우려를 나타냈다. 학년과 관계없이 많은 학생들은 후한 점수에 기뻐하기도 했지만, 의구심을 보였는데, 특히 고등학생들은 채점의 정확도 부족으로 간주하여 비판적 입장을 취하였다. 또한, 점수의 일관성에 대한 지적이 있었는데, 예를 들어 Rookie는 100점이 많은데, Semi-Pro는 100점이 안 나온 점, 전체 말하기 레벨평가는 점수가 후한데, 발음 진단 결과는 점수가 너무 낮다고 하였다. 점수에 대한 신뢰성은 수행 평가 채택여부와 관련이 있으므로 이에 대한 신중한 검토와 개선이 필요하다.

둘째, 힌트와 매뉴얼이 반드시 필요하다는 의견을 제시하였다. 프로그램을 전반적으로 사용하는데 있어서 사용의 용이성은 매우 높았지만, 레벨 테스트 특히 챗봇 대화에 있어서는 절차의 이해에 어려움이 종종 발생했다. 또한 교사들이 테스트 시행전과 중간에 알려주어야 할 사항에 대한 정보가 부족하다는 의견이 빈번하게

제기 되었다. 교사들은 '챗봇 대화에 힌트 기능을 도입해서 스피킹을 지도하기', '교사의 추가 설명 필요 지점에 대한 매뉴얼 추가' 등 다양한 방법으로 해결책을 제안하였다.

셋째, 학생들의 배려와 다양성 존중의 필요성을 제시하기도 하였다. 예를 들어 총점을 제시하는 방식에 대해 의문을 제기하는 교사도 있었다. 아라비아 숫자로 점수를 주는 것은 교육적인 의문점이 있다는 것이다.

> 종합 평가 점수가 일단 100 점일 때 완벽하다 라고 …해석을 할 수 없는 거고, 어떤 학생들은 저희 50 점 나왔는데 그 50 점이면 … 수우미양가로 보면 가인건데, 그렇게 판단할 수도 없는 거기 때문에 저는 이거를 점수로 하지 않고 … 이거 이제 육각형으로 많이 하거든요. (E 고 교사면접)

또한 다양한 형태의 가정이 있는 아이들에 대한 배려를 좀 더 해주기를 제안하였다. 말하기 대화에서 호텔 예약한 아버지의 성을 물어보는 것이었다.

> "아버지와 같이 살지 않는 경우, 혹은 아버지가 안 계시는 경우는 어떡하지?"라는 걱정이 들었음. 학생들이 자신의 이름으로 예약을 확인하도록 질문을 바꾸면 어떨까 라는 생각이 들었음. (D 중 수업일지)

4. 결론

DBR 연구 절차를 통해 본 연구가 개발한 영어 말하기 진단·학습·평가 시스템의 공교육 영어수업 적용가능성에 대해 탐색한 결과, 다음과 같은 결론을 얻을 수 있었다. 스픽마스터는 공교육 영어수업에 적용 가능하며, 영어 말하기 교육의 저해요인을 해결할 수 있는 다음과 같은 몇 가지 가능성을 보여주었다. 첫째, 프로그램은 내용적, 기술적으로 적절하였다. 참여자는 영어로 즉흥 대화를 나누고 자신의 말하기 결과의 피드백을 확인하는데 흥미를 보였으며, 사용편의성이나 난이도의 측면에서 보통 정도의 어려움을 느꼈다. 더욱이 학습자들은 자신의 말하기 수준이나 개선점을 알고 학습동기가 부여되거나 쉬운 등급을 도전한 후 스스로 더 높은 등급을 시도하는 등의 학습 주도성을 자주 보였다. 또한 짧은 사용 기간 동안 상당한 기술적인 오류를 겪었음에도 보통 이상의 만족감을 보여주었고, 기술 안정화 이후 시작한 학교에서는 매우 높은 호감을 보이기도 하였다. 무엇보다도 긍정적인 신호는 교사들의 인식 변화이다. 말하기는 단지 교과서 표현 연습과 암기 등의 평가로 그치던 것에서 발음, 대화, 발표 등 즉흥성과 내용 교환의 중요성을 새롭게 인지하기도 하였고, 프로그램을 통해 말하기 활동 지도에 긍정적인 자신감을 표하기도 하였다. 뿐만 아니라 수행 평가 혹은 평가 연습 도구로 적절하다는 인식을 모두 같이 하였다.

본 연구의 결과물이 공교육현장에서 말하기 문제를 해결하고 정확하고 신뢰할 수 있는 평가 도구로 자리매김하는 데까지는 아직 갈 길이 멀다. 시스템의 기술적인 오류를 최소화하고, 무엇보다도 챗봇 성능과 자동 채점의 신뢰도와 교육적인 적합도를 높이는 것이 가장 커다란 과제로 남아있다. 그러나, 본 연구를 통하여 확보된 데이터를 통한 고도화를 실현하고 사용자를 확대함과 동시에 DBR의 순환적 연구를 거듭함으로써 이러한 목표는 점차 성취될 것이라고 기대한다. 본 1차 실험에는 여러가지 제한점이 있다. 실험 중간에 프로그램 오류와 채점 방식을 조정하여, 문제 상황 전후의 결과 데이터를 분리하지 못하였고, 그 결과 학습자의 점수 데이터를 병렬 비교하기 어려워 프로그램 성능, 채점타당도에 대해 논의하지 못하였다. 설문 문항 이해에 어려움이 예상되었던 초등학교 학습자의 피드백 데이터가 부족했던 점도 앞으로의 연구과제로 남긴다. 마지막으로 본 연구에서 개발된 AI 영어 말하기 진단-학습-평가 시스템이 학생들의 영어 말하기 능력 향상에 미치는 장기적인 효과가 탐구될 필요가 있다.

Applicable levels: Elementary, secondary

REFERENCES

Bae, Y. J. (1984). Factors detrimental to learning speaking command of English. *English Teaching, 27*(1)*,* 15–31.

Bannan-Ritland, B. (2003). The role of design in research: The integrative learning design framework. *Educational Researcher, 32*(1), 21–24.

Choi, Y. I., Hong, S. J., & Park, J. H. (2019). A study on the development of a program for data-based differentiated instructional design for the Korean language trainee teachers. *Teacher Education Research, 58*(2), 221–236.

Hoadley, C. M. (2005). Design-based research methods and theory building: A case study of research with “SpeakEasy”. *Educational Technology: The Magazine for Managers of Change in Education, 45*(1), 42–47.

Irshad, U., Mahum, R., Ganiyu, I., Butt, F. S., Hidri, L., Ali, T. G., & El-Sherbeeny, A. M. (2024). UTran-DSR: A novel transformer-based model using feature enhancement for dysarthric speech recognition. *EURASIP Journal on Audio, Speech, and Music Processing, 2024*(54), 1–18.

Jung, H. Y., & Hong, H. J. (2021). An analysis on customized education research trends in the era of the 4th industrial revolution through text mining. *The Korean Journal of Educational Methodology Studies, 33*(3), 433–454.

Jung, J. Y. (2017). *Student-centred personalized learning as 'future education'. Happiness Education.* Retrieved on January 17, 2025, from https://happyedu.moe.go.kr/happy/bbs/selectHappyArticleImg.do?nttId=7073&bbsId=BBSMSTR_000000000191

Jung, Y., & Cha, K. (2014). An investigation of middle school and high school students' private English education in Korea. *The Research Institute of Korean Education, 32*(4), 141–160.

Kang, D., Li, X., Stoica, I., Guestrin, C., Zaharia, M., & Hashimoto, T. (2024, May). Exploiting programmatic behavior of LLMs: Dual-use through standard security attacks. In *Proceedings of the 2024 IEEE Security and Privacy Workshops (SPW)* (pp. 132–143). San Francisco, CA **:** IEEE.

Kim, M. H., Yoon, K. S., & Park, J. W. (2020). An exploratory research on learning competency based personalized learning in K university. *Journal of Practical Engineering Education, 12*(1), 49–60.

Koh, H. W. (2014). A study of perception on effectiveness and impracticability of English speaking performance assessment among secondary English teachers. *The New Korean Journal of English Language and Literature, 56*(3), 43–66.

Lee, J-H., Choi, Y., Sung, M., & Kim, H. (2023). Analysis of English automated speaking scoring tests. *English Teaching, 78*(2), 223–244.

Lee, J-H., Choi, Y., Sung, M., & Kim, H. (2025). Developing a curriculum-based speaking task taxonomy for school English instruction and assessment in Korea. *English Teaching, 80*(3), 23–45.

Lee, M. B. (2018). A study of high school English teachers' teaching English speaking and performance assessment. *Journal of the Korea English Education Society, 17*(1), 1–20.

Lee, M. B., Yoon, J. H., Kim, S. Y., Choo, H. W., & Kwon, S. K. (2015). *Development of an Internet-based high school English speaking performance assessment support system (Report No. RRI 2015–2).* Seoul: Korea Institute for Curriculum and Evaluation.

Macromillembrain. (2022). *English learning perception survey* (Research Rep. No. TK_202203_TR_1734). Seoul: Macromillembrain.

Ministry of Education. (2023). *Realizing customized education for all: Digital-based education innovation plan*. Sejong, Korea: Ministry of Education.

Reeves, T. (2006). Design research from a technology perspective. In J. van den Akker, K. Gravemeijer, S. McKenney & N. Nieveen (Eds.), *Educational design research* (pp. 64–78). New York: Routledge.

Reimann, P. (2013). Design-based research—Designing as research. In R. Luckin, S. Puntambekar, P. Goodyear, B. Grabowski, J. Underwood & N. Winters (Eds.), *Handbook of design in educational technology* (pp. 44–52). New York: Routledge.

Rodríguez, J. C., & Ballester, M. C. P. (Eds.). (2013). *Design-based research in CALL*. San Marcos, TX: CALICO.

Song, E. J., & Shim, K. (2020). A study on the development of AI-based assessment model for English speaking skills in primary school. *English Language Assessment, 15*(2), 53–76.

Sung, M., Lee, J-H., Kim, H., & Choi, Y. (2025). A study on grammatical diversity measures for automated English speaking assessment of Korean learners. *Korean Journal of English Language and Linguistics, 25*, 1048–1065.

Wang, F., & Hannafin, M. J. (2005). Design-based research and technology-enhanced learning environments. *Educational Technology Research and Development, 53*(4), 5–23.

Wang, X., Evanini, K., Qian, Y., & Mulholland, M. (2021, January). Automated scoring of spontaneous speech from young learners of English using transformers. Paper presented at the *2021 IEEE Spoken Language Technology Workshop (SLT)*. Shenzhen, China.

Xu, J., Jones, E., Laxton, V., & Galaczi, E. (2021). Assessing L2 English speaking using automated scoring technology: Examining automarker reliability. *Assessment in Education: Principles, Policy & Practice, 28*(4), 411–436.

Zechner, K., & Evanini, K. (Eds.). (2019). *Automated speaking assessment: Using language technologies to score spontaneous speech.* New York: Routledge.