Korean Journal of English Language and Linguistics, **Vol 25, November 2025, pp. 1496-1517**
DOI: 10.15738/kjell.25..202511.1496

KOREAN JOURNAL OF ENGLISH LANGUAGE AND LINGUISTICS

ISSN: 1598-1398 / e-ISSN 2586-7474

http://journal.kasell.or.kr

자동채점 기반 영어 말하기 성적 보고서에 대한 학습자와 교사의 이해도 및 인식 탐색*

YunDeok Choi (Chungnam National University) · **Heyoung Kim** (Chung-Ang University) · **Jin-Hwa Lee** (Chung-Ang University) · **Min-Chang Sung** (Gyeongin National University of Education)

This is an open-access article distributed under the terms of the Creative Commons License, which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Received: August 19, 2025
Revised: September 11, 2025
Accepted: October 10, 2025

Choi, YunDeok (First author)
Assistant Professor, Department of English Education,
Chungnam National University
Email: yundeokchoi@cnu.ac.kr

Kim, Heyoung (Corresponding author)
Professor, Department of English Education, Chung-Ang University
Email: englishnet@cau.ac.kr

Lee, Jin-Hwa (Co-author)
Professor, Department of English Education, Chung-Ang University
Email: jinhlee@cau.ac.kr

Sung, Min-Chang (Co-author)
Associate Professor, Department of English Education,
Gyeongin National University of Education
Email: mcsung@ginue.ac.kr

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

ABSTRACT

Choi, YunDeok, Heyoung Kim, Jin-Hwa Lee and Min-Chang Sung. 2025. An exploratory study on the development of an online score report for automated English speaking assessment. *Korean Journal of English Language and Linguistics 25*, 1496-1517.

Score reporting plays an important role in supporting the validity of interpretations and uses of assessment results; however, it remains an underexplored topic particularly in domestic research contexts. This exploratory study examines how two user groups-students and teachers-understood and perceived a score report designed for an automated speaking assessment (ASA) system currently under development for Korean public education. A total of 224 Korean secondary school students completed an online survey measuring their understanding and perceptions. Five English teachers participated in online semi-structured interviews. Descriptive statistics, as well as group-independence chi-square and Mann-Whitney U tests, on the survey responses showed that the students overall held low levels of understanding but positive perceptions with no significant differences between school levels and negligible effect sizes. A thematic analysis of the interview transcripts revealed that teachers appreciated individualized feedback on the analytic aspects of speaking performance presented through visual graphs. However, they expressed concerns that the written prose feedback was too general and some terminology lacked specificity for students, providing recommendations for improvement. These findings have implications for the development of a user-friendly score report that enhances the validity of assessment interpretations and uses.

KEYWORDS

score reporting, understanding, perception, automated speaking assessment, secondary school

© 2025 KASELL All rights reserved

1. 서론

영어 말하기 능력은 구두 언어를 활용하여 타인과 효과적으로 의사소통하는 능력으로서(Fulcher 2003), 의사소통 중심 영어교육의 핵심 요소로 자리매김하고 있다. 이러한 말하기 능력을 효과적으로 향상하기 위해서는 학습자의 강점과 약점을 명확히 파악하고, 교수·학습 방향을 조정할 수 있는 체계적인 평가가 필수적이다. 이에 따라 최근에는 학습자의 실제 수행(actual performance)을 기반으로 한 직접 평가(direct tests) 방식이 다양한 교육적 맥락에서 널리 활용되고 있다(Ginther 2012).

국내 영어교육에서도 말하기 능력을 포함한 의사소통 능력의 함양이 교육과정의 핵심 목표로 제시되고 있으며, 이를 실현하는 방안으로 표현 영역에 대한 직접 평가의 활용이 강조되고 있다(교육부 2022). 이러한 방향은 학교 현장에서 수행평가의 형태로 구현되고 있으나, 실제로는 주로 초등학교와 중학교에서 교사 인터뷰, 역할극, 발표 등의 제한된 과업 중심으로 시행되고 있으며, 평가 결과 또한 점수 위주의 형태로 제시되는 경향이 강하다(김나예, 김정열 2018, 박선영, 민찬규 2018). 현재와 같은 방식으로 운영되는 수행평가는 본연의 취지를 살려 학습자의 말하기 능력을 세부적으로 진단하고 향후 학습 방향을 안내하는 데에 한계가 있다(김나예, 김정열 2018).

한편, 최근 영어교육 환경의 디지털 전환이 가속화되면서, AI 디지털 교과서와 AI 기반의 말하기 학습 프로그램이 학교 교육 현장에 점차 도입되고 있다(심창용 2024). 이에 따라 학습자에게 평가 결과를 단순히 점수로 전달하는 데 그치지 않고 다양한 정량적·정성적 정보가 포함된 성적 보고서(score report)를 제공하려는 시도가 활발해지고 있다(이진화 외 2023, Gu and Davis 2019, Gu et al. 2021). 이렇게 전달 정보의 범위와 깊이가 확대된 성적 보고서는 평가 결과를 학습자, 학부모, 교사 등 다양한 이해당사자가 정확히 해석하고 이를 기반으로 향후 학습 계획 및 교수 방향을 수립할 수 있도록 지원한다(Hambleton and Zenisky 2013). 이러한 점에서 성적 보고서는 단순한 평가의 결과물이 아니라, 시험 점수의 해석과 활용을 뒷받침하는 타당성의 핵심 구성 요소로 간주된다(AERA et al. 2014, O'Leary et al. 2017). 따라서, 성적 보고서는 사용자의 관점에서 정확한 정보를 쉽게 이해할 수 있도록 체계적으로 개발되어야 한다(AERA et al. 2014).

이러한 성적 보고서의 중요성에 대한 인식이 높아지면서 최근에는 해외를 중심으로 성적 보고서 개발에 관한 이론적 모형과 모범적 실천 방안이 제시되고 있다(예: Goodman and Hambleton 2004). 또한, 저부담 시험(Brown et al. 2019, Vezzu et al. 2012, Zapata−Rivera et al. 2018) 뿐만 아니라 고부담 시험(Gu and Davis 2019, Gu et al. 2021, Hsieh 2023) 맥락에서 실증 연구가 활발히 진행되고 있는데, 이들 연구의 공통된 발견 중 하나는 다양한 사용자들이 성적 보고서를 긍정적으로 인식하면서도, 그에 포함된 정보를 정확히 해석하고 이해하는 데 어려움을 겪는다는 점이다. 이러한 결과는 성적 보고서 개발 과정에서 사용자의 이해도와 인식을 파악하고 충분히 반영하는 체계적인 접근이 필요하다는 것을 시사한다.

그러나 기존 연구는 대부분 해외 사례를 중심으로 수행되었기 때문에, 국내 중등 영어교육 환경에 직접 적용하기에는 한계가 있다. 더욱이 국내에서는 중등학교 학습자를 대상으로 성적 보고서를 개발하고, 이에 대한 이해도와 인식을 실증적으로 조사한 연구가 매우 부족한 실정이다.

© 2025 KASELL All rights reserved

현재 교육과정에서 강조하고 있는 학습자 맞춤형 교육 및 평가(교육부 2022)를 실현하기 위해서는, 성적 보고서가 학습자에게 정보를 얼마나 효과적으로 전달하며, 그것이 교육적으로 어떤 의미를 지니는지를 확인하는 연구가 시급히 이루어져야 하는 상황이다.

이에 본 연구는 국내 영어 말하기 공교육의 질적, 양적 향상을 도모하기 위해서 현재 개발 중인 인공지능 기반 진단-학습-평가 통합 시스템에 포함된 학습자용 성적 보고서를 설계하고, 이에 대한 중등학교 학습자와 교사의 이해도 및 인식 수준을 실증적으로 탐색하고자 한다. 나아가, 사용자 친화적인 성적 보고서 설계 요소를 도출하여, 학습자의 이해와 인식을 제고하고 평가 결과의 해석과 활용의 타당성을 향상시키고자 한다. 이러한 과정을 통해 성적 보고서가 교수와 학습에 긍정적인 환류 효과(washback effects)를 제공할 수 있도록 하여, 효과적인 영어 말하기 교육의 실천에 이바지하고자 한다(Taylor 2005).

2. 이론적 배경 및 선행 연구

2.1 성적 보고서의 정의 및 유형

성적 보고서는 사용자에게 시험 결과에 대한 정보를 평가의 목적에 따라 수치화 된 점수, 서술형 텍스트, 도표 등의 다양한 형식으로 전달하는 도구로서, 시험 개발자와 점수 사용자 간의 핵심적인 의사소통 수단이다(Brown et al. 2019, Hambleton and Zenisky 2013, Zenisky and Hambleton 2016). 이러한 성적 보고서는 일반적으로는 크게 두 부분으로 구성된다. 첫 번째는 시험(평가) 명, 응시 일자, 성적표의 명칭과 목적 등 결과 해석에 필요한 맥락을 제공하는 기술적(descriptive) 영역이다. 두 번째는 점수와 수행 수준 등으로 구성된 평가 결과 영역으로 수험자의 수행을 총체적으로 요약하고, 각 하위 영역별로 세부적인 결과를 제공한다(Zenisky and Hambleton 2016).

또한 성적 보고서는 대상 사용자와 결과 전달 방식에 따라 다양한 유형으로 나뉜다. 먼저, 대상 사용자 기준에 따른 유형을 살펴보면 우선 학습자와 그 가족, 교사를 대상으로 개인의 평가 결과를 전달하는 개인용과 학교, 지역(district) 등의 단위로 학습자들의 평가 결과를 교사, 학교 관리자 또는 대중에게 공개하기 위한 집단용으로 구분된다. 개인용 성적 보고서에는 일반적으로 수치화 된 점수, 숙련도 수준(예: 상, 중, 하), 등수 등이 포함된다. 반면에 집단용 성적 보고서는 주로 특정 교육 프로그램 평가나 그 내부의 교수 방향 설정 등의 목적으로 사용된다(Zenisky and Hambleton 2016).

또한 전달 매체에 따라 국내 수능시험 성적표와 같은 종이 인쇄형(paper-based)과 온라인형(online-based)으로 나뉜다. 온라인형은 모든 사용자에게 동일한 정보를 제공하는 정적(static) 또는 온라인 웹 환경에서 필요에 따라 정보를 취사선택하고 학습 자료에 접근할 수 있도록 설계된 동적인(dynamic) 유형이 있는데, 후자의 대표적인 예로는 미국 칼리지보드(College Board)의 SAT 성적표 제공 시스템을 들 수 있다(Brown et al. 2019, Zenisky and Hambleton 2016). 이러한 동적 시스템에서는 사용자가 수험자의 학년을 포함한 다양한 기준을 선택하여 평가 결과를 열람할 수 있는 차별된 기능을 장착하고 있다.

© 2025 KASELL All rights reserved

특히 고부담(high-stakes) 시험의 경우, 점수의 해석이 사회적 결정에 중대한 영향을 미치기 때문에, 단순한 결과 전달을 넘어 점수 해석을 지원하는 다양한 해석 자료(interpretive materials)가 성적 보고서와 함께 제공되는 것이 일반적이다(Kim et al. 2020). 해석 자료는 주로 타당한 점수 해석과 점수 사용 방식 등에 대한 정보를 포함하여, 사용자가 시험 결과를 더 정확하게 이해하고 활용할 수 있도록 돕는다(Kim et al. 2020)

2.2 성적 보고서 개발의 이론적 기반

전통적인 심리측정 중심의 논의와 비교해서, 성적 보고서를 평가 타당도의 일환으로 간주하는 관점은 비교적 최근인 1990년대 후반부터 본격적으로 대두되었다. 이 시기부터, 이 분야의 연구자들은 정보 시각화(information display)에 관한 문헌(예: Wainer 1992)을 성적 보고서 설계에 접목하고, 효과적인 평가 결과 보고를 위한 모델 및 실천 방식을 다각도에서 제시하였다. 이 중 Hambleton과 Zenisky(2013), Zenisky와 Hambleton(2016)이 제안한 4단계 7절차의 성적 보고서 개발 모델은 설계부터 개발, 적용, 유지 및 보완에 이르기까지 일련의 과정을 포괄적으로 설명하고 있다. 4단계는 기초 수립(groundwork establishment), 보고서 개발(report development), 시범 사용(field test)과 평가와 유지(evaluation and maintenance)의 주요 핵심 단계를 의미하며, 이 전체 과정에 총 7개의 세부 절차가 포함되어 있다. 시험 목적 설정, 대상 사용자 정의, 요구분석, 관련 문헌 검토 등의 사전 작업에서 시작하여, 시안(prototype) 설계를 거쳐 실제 환경에서의 시험 적용을 시행한다. 그 결과로 사용자의 이해도와 반응에 대한 피드백을 수집하고, 이를 반영하여 최종본을 완성한다. 마지막으로 성적 보고서 출시 이후에도 지속적인 평가와 개선 작업이 이어진다. 이 모델은 절차 중심의 처방적(prescriptive) 성격을 지니고 있으면서도, 다양한 평가 상황에 유연하게 적용될 수 있도록 설계되었고, 검토에 활용할 수 있는 평가 기준을 함께 제시하고 있다.

한편, Hattie(2009)는 온라인 기반의 성적 보고서 설계를 위한 15가지 원칙을 제안하였다. 예를 들어, 설계 원칙 10에서는 온라인 성적 보고서의 스크롤 사용을 최소화하고, 화면을 단순하게 유지하며, 읽히는 것(read)보다 보이는 것(seen)을 극대화해야 한다고 강조하였다. 이 원칙은 가독성과 정보 해석력 향상에 중점을 두고 있으며, 이를 위해 충분한 여백 확보, 선명한 그래픽 사용(예: 줄 간격 최소 2pt), 높은 화면 해상도, 적절한 글꼴과 글자 크기 사용(예: 성인의 경우 Times New Roman 12pt), 줄당 적절한 글자 수 유지(예: 성인 기준 75-100자) 등의 구체적 방안을 제시하고 있다. 아울러 이러한 원칙들이 뉴질랜드 교육부 주도로 개발된 컴퓨터 기반 국가 수준 평가 시스템, 즉 교수·학습 지원형 평가 도구(Assessment Tools for Teaching and Learning, asTTle)의 개발 과정에서 어떻게 적용되었는지를 다양한 사례 연구를 통해 구체적으로 소개하였다.

성적표 개발과 활용에 대한 이러한 이론적 틀은 사용자 중심의 성적표를 체계적으로 개발하는 데 필수적인 방향성을 제시하며, 나아가 점수 해석과 활용의 타당성을 높이는데 기여할 수 있는 기반을 마련해준다. 본 연구에서는 이와 같은 이론적 근거를 바탕으로 성적 보고서를 설계하였으나, 설계 과정 그 자체보다는 결과물에 대한 학습자와 교사의 반응 분석과 개선 방안 모색에 초점을 두고 있다.

© 2025 KASELL All rights reserved

2.3 성적 보고서에 관한 실증 연구

실증 연구는 크게 기존 성적 보고서에 대한 현황 분석, 사용자 인식 탐색, 또는 새로운 개발 시도의 세 범주로 나누어 볼 수 있다. 먼저 현황 분석은 대규모의 표준화된 시험에서 제공하는 성적표의 특징을 분석하여, 효과적인 피드백 도구로서의 가능성과 한계점 및 개선 방향에 대해 논의한다(이진화 외 2023, Goodman and Hambleton 2004). 예를 들어, Goodman과 Hambleton(2004)은 미국과 캐나다의 K-12 학력 평가 시험 성적 보고서에서 점수 해석을 지원하는 안내 자료와 시각적 도표의 활용 등과 같은 긍정적인 특징들을 발견하였다. 그러나 전문 용어 사용으로 인하여 사용자 이해도가 저하되고 구체적인 학습 방향의 제시가 부족하다는 점을 한계로 지적하였다. 한편, 이진화 외(2023)는 채점 방식의 차이에 따른 성적표의 특징을 분석하였다. 자동채점 시스템을[1] 도입한 6종의 시험과 전통적인 인간 채점 방식을 사용하는 6종의 시험을 포함한 총 12개의 국내외 주요 시험을 비교 분석하여, 자동 채점 기반 시험의 경우 평가 구인별로 점수와 상세한 피드백을 제공한다는 특징을 발견하였다. 이는 자동 평가 시스템이 학습자의 말하기 능력을 더욱 정밀하게 평가하고 학습을 지원할 수 있음을 시사한다.

이에 반해 사용자 인식 연구는 다양한 사용자를 대상으로 기존에 사용되고 있는 성적 보고서의 장단점을 파악하고, 이를 개선하기 위한 목적으로 수행되어 왔다(Hsieh 2023, Kim et al. 2020, Sawaki and Koizumi 2017). 예컨대, Hsieh(2023)는 일본인 수험자를 대상으로 토익 듣기와 읽기의 성적 보고서에 대한 이해도, 사용 목적, 그리고 영어 능숙도와 성적표 이해도 간의 관련성을 조사하였다. 그 결과 평가나 언어학과 관련된 전문 용어 사용으로 인해 수험자들이 내용을 잘못 해석할 가능성이 있으며, 이는 부정적 인식으로 이어질 수 있음이 드러났다. 또한, 영어 능력이 높을수록 성적표 이해도도 높다는 점이 확인되었다.

최근에는 온라인 기반 평가 시스템의 성적 보고서 개발 과정에 관한 연구들이 보고되고 있다(Brown et al. 2019, Gu and Davis 2019, Gu et al. 2021, Vezzu et al. 2012, Zapara-Rivera et al. 2018). 이들 연구는 대체로 이론적 틀을 바탕으로 사용자 요구 분석, 시안 설계, 사용자의 평가와 피드백을 통한 수정이라는 체계적이고 순환적인 개발 절차를 따르고 있다. 예를 들어, Gu et al.(2021)는 TOEFL iBT의 연습용 시험인 TOEFL Practice Online(TPO)의 수험자와 영어 교사를 대상으로, 성적 보고서에 포함될 수 있는 세 가지 유형의 진단 피드백에 대한 사용자 인식을 조사하였다. 그 결과, 수험자들의 그래프나 용어의 이해를 돕기 위한 추가적인 설명이 필요하며, 정확한 해석을 위해 교사의 안내와 지원이 중요하다는 점이 밝혀졌다. 반면에, Vezzu et al.(2012)는 인지 기반 학습 평가 프로그램(Cognitively Based Assessment of, for, and as Learning [CBAL]™ initiative)의 일환으로 중학교 학습자들의 내용 이해와 몰입도를 향상하기 위해 탑재된 안내형 활동(guided-instructional activity)을 특징으로 하는 상호 작용형 학습자 성적 보고서의 개발 과정을 소개하였다.

이러한 해외 중심의 선행 연구들은 성적 보고서 개발에 대한 중요한 시사점을 제공하지만, 국내 교육 환경에 직접적으로 적용하기에는 한계가 있다. 따라서 국내 학교 교육 맥락에서 학습자의

[1] 최신 영어 말하기 자동 채점 기술과 타당성 관련 논의는 이진화 외(2023)와 Zechner와 Evanini(2019)의 편저서 Automated Speaking Assessment: Using Language Technologies to Score Spontaneous Speech 참조.

© 2025 KASELL All rights reserved

실제 이해도와 인식을 반영한 실증적 성적 보고서 개발 연구가 절실히 요구된다.

이에 본 연구는 우리나라 공교육 학습자를 대상으로 개발 중인 인공지능 기반 영어 말하기 진단-학습-평가 통합 시스템에 탑재된 말하기 레벨 테스트(Speaking Level Test)의 성적 보고서에 대한 두 사용자 집단(학습자와 교사)의 이해도와 인식을 탐색하고자 한다. 특히, 성인 학습자를 중심으로 한 기존 연구(Gu et al. 2021, Hsieh 2023)와 달리, 본 연구는 영어 학습 및 평가 경험, 인지적·정서적 발달 수준이 서로 다른 학령기 학습자를 대상으로 성적 보고서에 대한 이해와 인식을 분석하고자 한다. 또한 성적 보고서의 또 다른 주요 사용자인 교사들을 대상으로 성적 보고서의 활용 및 개선 방안에 대한 의견을 조사하였다. 이를 통해 중등학교 학습자의 특성과 요구에 맞는 맞춤형 성적표 설계의 근거를 마련하고, 평가 결과 제시 방식의 개선을 통하여 사용자의 이해도와 인식 제고에 기여하고자 한다. 이에 따라 본 연구는 다음과 같은 연구 질문을 설정하였다:

1) 말하기 레벨 테스트의 성적 보고서에 대한 국내 중등학교 학습자들의 이해도는 어떠한가?
2) 말하기 레벨 테스트의 성적 보고서에 대한 국내 중등학교 학습자들의 이해 용이성, 결과의 타당성, 학습적 유용성에 대한 인식은 어떠한가?
3) 말하기 레벨 테스트의 성적 보고서의 활용성과 개선 방안에 대한 교사의 인식은 어떠한가?

연구 질문 2)에 사용된 주요 개념은 다음과 같이 정의된다. '이해 용이성'은 성적 보고서에 제시된 내용이 학습자에게 얼마나 쉽게 해석될 수 있는지를 의미한다. '결과의 타당성'은 성적 보고서에 포함된 평가 결과가 자신의 실제 영어 말하기 능력을 얼마나 적절하게 반영하고 있는지를 나타낸다. 한편, '학습적 유용성'은 성적 보고서의 정보가 영어 말하기 학습에 있어 학습 방향 설정이나 자료 제시 등의 측면에서 학습자에게 얼마나 도움이 되는지를 가리킨다.

3. 연구 방법

본 연구는 혼합 연구 설계(mixed-methods design) 중 삽입형 설계(embedded design)에 주로 기반하고 있으며, 일부 측면에서는 설명적 순차 설계(explanatory sequential design)의 특성도 포함하고 있다(Creswell and Plano Clark 2011). 즉, 정량적 자료(학습자 설문 응답)의 분석을 통해 연구 질문 1과 2를 다루었고, 정성적 자료(교사 면담)를 활용해서는 연구 질문 3을 탐색하였다. 아울러, 교사 면담 자료는 연구 질문 1과 2의 설문 결과를 해석하는 데 보완적인 설명 자료로도 활용하였다.

3.1 연구 대상

인공지능 기반 영어 말하기 진단-학습-평가 통합 시스템의 시범 사용에 대한 연구팀의 SNS 공고를 보고, 다수의 교사가 학교장의 허가를 받아 자발적으로 참여를 신청하였다. 시범 참여 후 학교 일정을 고려하여 설문조사 실시가 가능했던 서울특별시에 있는 중학교 두 곳과 충청남도 소재 일반계 고등학교 한 곳이 본 연구의 대상 학교로 선정되었으며, 총 244명의 재학생이

© 2025 KASELL All rights reserved

학습자로 참여하였다. 설문에 참여한 학생들은 모두 말하기 레벨 테스트에 응시한 뒤, 그 결과로 제공된 성적 보고서(3.2.1 학습자 개인용 성적 보고서 참조)를 제공받은 후 이를 바탕으로 설문에 응답하였다. 학습자의 성별은 남학생 129명(52.9%)과 여학생 115명(47.1%)으로 유사한 분포를 보였다. 표 1은 연구에 참여한 학생들의 학교급과 학년 분포를 요약한 것이다.

표 1. 학교급과 학년별 학습자 분포

	1학년	2학년	3학년	합
중학교	72	69	51	192
고등학교	0	52	0	52

학습자와 더불어, 시범 사용에 자발적으로 참가한 교사들도 본 연구에 참여하였다. 표 2에 제시된 바와 같이 참여 교사는 총 다섯 명으로, 여성 네 명과 남성 한 명으로 구성되어 있었으며, 교직 경력은 평균 10.4년으로 나타났다. 이들이 근무하는 학교는 서울, 세종, 공주, 대구에 소재하고 있었으며, 담당 학년은 중학교 1학년부터 고등학교 2학년까지로, 고등학교 3학년을 제외한 전 학년에 걸쳐 고르게 분포되어 있었다. 중학교 교사들은 본 연구의 설문에 응답한 학생들의 수업을 실제로 담당하고 있었으며, 고등학교 교사 세 명 중 한 명도 설문에 참여한 학생들의 수업을 담당하고 있었다. 나머지 두 명은 말하기 레벨 테스트 프로그램의 체험 수업에만 참여하였다. 이들 교사 참여자는 영어 수업 시간에 ChatGPT, 팅커벨, Pang 등 다양한 디지털 기반 프로그램을 활용하고 있는 것으로 나타났다.

표 2. 참여 교사 배경 정보

교사	성별	교직 경력(년)	학교 소재지	지도 학년	학습자 설문 참여 여부
A	여	13	서울	중학교 1, 3	참여
B	여	7	세종	고등학교 2	참여
C	여	4	서울	중학교 1, 2	참여
D	남	14	공주	고등학교 1	미참여
E	여	14	대구	고등학교 1	미참여

3.2 연구 도구

3.2.1 학습자 개인용 성적 보고서

본 연구에서 실험한 성적 보고서는 연구진이 직접 개발하고 있는 인공지능 기반 영어 말하기 진단–평가–학습 시스템에 탑재된 말하기 레벨 테스트용으로 설계되었다. 이 시스템은 공교육 환경에서 학습자의 영어 말하기 능력을 보다 객관적이고 정밀하게 진단하고, 기초학력 보장을 위한 학습자 맞춤형 교육 지원 방안을 마련하는 것을 목적으로 개발되었다. 성적 보고서는 영어 말하기 자동 채점 기술이 적용된 정적 온라인 기반 개별 학습자용(static online individual score report)의 형태로 제작되었으며, 말하기 레벨 테스트를 완료할 때마다 각 학습자의 기기 화면에

© 2025 KASELL All rights reserved

자동으로 생성되며, 학습자의 계정에 누적되어 축적된 결과를 통해 변화를 확인 할 수 있다.

이 성적 보고서는 Hambleton과 Zenisky(2013)가 제안한 4단계 7절차의 사용자 중심 설계 원칙을 반영하여 설계되었다. 그림 1은 전체 개발 절차의 개요를 도식화한 것이며, 본 연구는 첫 세 단계에 중점을 두고 있다. 마지막 개발 단계는 시스템의 정식 출시 이후 수행될 예정이다.

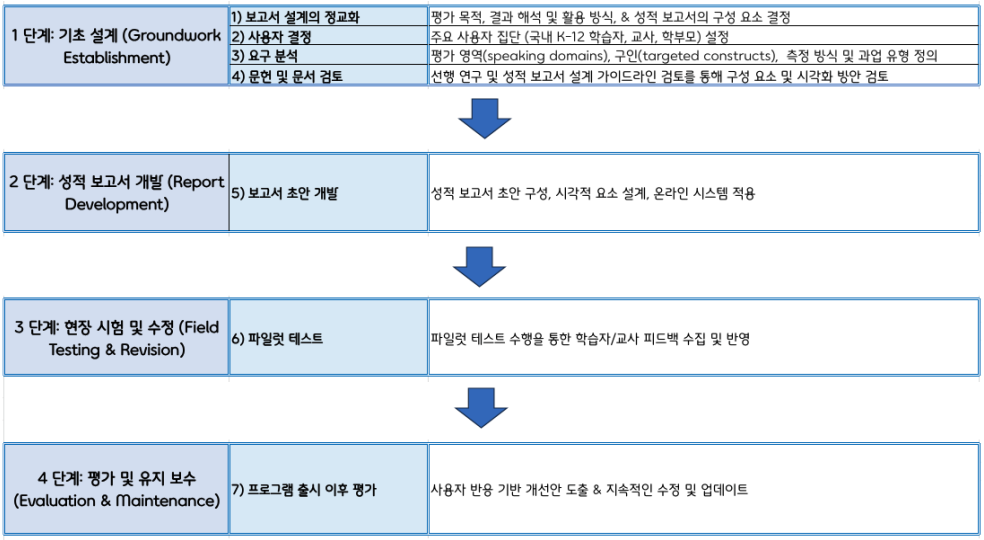

그림 1. 4단계 7절차 모형에 따른 성적 보고서 개발 과정 도식화

본 성적 보고서는 학생들의 영어 말하기 수행을 바탕으로 강점과 약점을 요약하고, 실천할 수 있는 학습 활동과 함께 개인화된 피드백을 제공하는 진단 평가용으로 제작되었다(Hattie and Timperley 2007). 이러한 목적을 효과적으로 달성하기 위해서 선행 연구에서 제시된 평가 결과 보고의 모범적 실천(best reporting practices) 방안을 참고하여, 사용자 친숙성(user-friendliness), 학습 방향 제시, 그리고 학습 동기 향상이라는 세 가지 구체적인 목표를 설정하였다(Brown et al. 2019, Goodman and Hambleton 2004, Hattie 2009, Hattie and Timperley 2007).

그림 2에서 볼 수 있듯 설계된 성적 보고서는 크게 네 부분으로 구성되었으며, 부분별 주요 내용은 다음과 같다. 첫째, 기술 영역은 평가 관련 정보(평가 명, 평가일, 소요 시간), 수험자 정보(이름, 학교명), 그리고 총점에 따라 수여된 트로피 개수를 포함한다(① 참조). 둘째, 종합 점수 및 피드백 부분은 응시한 시험 등급, 세 영역(구술 능력·대화 능력·발표 능력)의 점수를 합산한 총점(100점 만점 기준), 영역별 강·약점 요약 및 학습 방향을 제시하는 서술형 피드백을 포함한다(② 참조). 셋째, 세부 영역별 진단 결과 부분은 각 영역 내 하위 구인별 점수를 세 수준(Needs Work, Good, Excellent)으로 분류하고, 결과를 누운 막대그래프 형태로 시각화 하여 제시한다(③ 참조). 마지막으로, 추천 학습 및 연습 활동에서는 Needs Work로 평가된 구인에 대해, 학습 활동 페이지로 직접 연결되는 인터페이스를 제공함으로써 후속 학습을 유도하였다(④ 참조). 학습자가 성적 보고서의 내용을 제대로 이해하고 활용하기 위해서는 총점 뿐만 아니라 서술형 피드백, 세부 영역별 결과 그래프를 읽고 통합적으로 해석한 후 제시된 추천 및 연습 활동

© 2025 KASELL All rights reserved

이미지를 클릭하는 일련의 과정이 필요하다.

그림 2. 말하기 레벨 테스트의 성적 보고서 예시 이미지

3.2.2 온라인 설문지

본 연구의 목적은 말하기 레벨 테스트의 결과로 제공되는 성적 보고서에 대한 학습자의 이해와 인식을 탐색하는 것이며, 이를 위해 평가 직후 성적 보고서가 자동으로 제공되는 레벨 테스트를 시범 운영한 후 온라인 설문조사를 실시하였다. 온라인 설문지는 총 30문항으로 구성되었고, 이 중 다섯 문항은 학교 급, 학년, 성별 등 학습자의 배경 정보를 확인하기 위한 것이었고, 18개 문항은 말하기 레벨 테스트 사용에 대한 학습자들의 전반적인 경험과 인식을 측정하기 위해 설계되었다. 나머지 일곱 개 문항이 바로 성적 보고서에 대한 학습자의 이해도 및 인식을 측정하기 위해 고안되었으며, 본 논문에서는 이 일곱 개 문항에 대한 응답만을 분석하였다. 이 중 네 개의 사지선다 문항은 성적 보고서의 총점에 대한 해석(문항 27 '하은솔 학생의 *종합평가 점수 88점*은 어떤 말하기 능력들을 종합적으로 평가한 결과인가요?')과 구술(문항 28 '하은솔 학생의

© 2025 KASELL All rights reserved

*종합 평가 점수 및 피드백*을 읽고, 구술 능력을 향상시키기 위해 알맞은 방법을 고르세요'), 대화(문항 29 '하은솔 학생의 *종합 평가 점수 및 피드백*과 *세부 영역별 정밀 진단 결과*를 보고, 발표 능력에서 특히 더 노력해야 할 부분을 고르세요'), 발표(문항 30 '*세부 영역별 정밀 진단 결과*를 보고, 하은솔 학생의 대화 능력에 대해 알 수 있는 점은 무엇인가요?') 영역별 세부 피드백에 대한 이해를 측정하도록 고안되었다(자세한 사항은 부록 참조). 더불어, 5점 척도(1 = 매우 그렇다, 5 = 전혀 그렇지 않다) 기반의 3개 문항은 성적 보고서의 내용 해석에 대한 이해 용이성(문항 20 '시험 결과지의 내용을 이해하기 쉬웠다.'), 결과의 타당성(문항 21 '시험 결과는 나의 전반적인 말하기 능력을 잘 설명해 준다.'), 학습적 유용성(문항 22 '시험 결과지는 앞으로 말하기 능력을 향상하기 위해서 내가 무엇을 해야 하는지 잘 설명해 준다.')에 대한 학습자의 인식을 측정하도록 설계하였다. 설문 문항들은 본 연구의 목적에 따라 관련 선행 논문(Hambleton and Zenisky 2013, Hsieh 2023, Kim et al. 2020)을 고려하여 고안되었으며, 최종 설문지 문항은 구글 폼(Google Forms)의 공유 링크를 담당 영어 교사를 통해 학습자들에게 배포하였다.

3.2.3 교사 면담 질문

교사 면담은 반구조화(semi-structured)된 일대일 형식으로 진행되었고, 평균 면담 시간은 약 60분(32분~78분)이었다. 이러한 시간의 편차는 교사의 응답 방식과 후속 질문(follow-up questions)의 개수 등에 기인하였다.

면담은 말하기 레벨 테스트의 운영 방식, 장단점과 개선 방향 등을 파악하기 위한 목적으로 진행되었다. 하지만, 면담 문항 중 일부는 레벨 테스트의 결과로 제시되는 성적 보고서에 대한 학습자의 반응(이해도와 인식), 성적 보고서의 활용 가능성과 향후 개선 방안에 대한 교사의 견해를 심층적으로 탐색하기 위해 별도로 설계되었다. 본 논문에서는 전체 면담 자료 중에서 성적 보고서와 관련된 응답만에만 분석을 수행하였다.

3.3 연구 진행 절차

본 연구는 말하기 레벨 테스트 초기 모델을 실제 학교 현장에서 시범적으로 사용하고, 사용자의 경험에 대한 자료를 수집하는 순서로 진행되었다. 연구 착수 전 대학 생명윤리위원회(IRB)의 승인을 받아 모든 참여자가 동의서에 서명을 한 뒤 연구에 참여하도록 하였다. 시범 사용은 2025년 3월부터 6월까지, 서울특별시, 경기도, 충청남도, 경상북도에 소재한 초·중·고등학교 여덟 곳에서 실시되었다.

시범 사용 기간 중, 학생들은 영어 수업 시간에 교사의 지도하에 컴퓨터, 태블릿 PC 또는 휴대전화를 사용하여, 자신이 재학 중인 학년 군에 해당하는 레벨의 테스트에 적어도 한 차례 응시하였다. 시험 직후, 성적 보고서는 각자의 기기 화면을 통해 즉시 확인할 수 있었으며, 학습자 계정에 저장되어 이후에도 열람이 가능하게 하였다. 교사는 담당하는 학생들의 응시 여부와 성적 보고서를 확인할 수 있었다.

말하기 레벨 테스트에 응시한 후, 그 결과로 제공된 성적 보고서를 확인 후, 학습자들은 무기명 온라인 설문조사를, 영어 교사들은 반구조화된 개인 면담에 참여하였다. 모든 면담 내용은 사전

© 2025 KASELL All rights reserved

동의를 받고 녹음되었다. 녹음된 음성 파일은 이후 분석을 시행하기 위해서 한국어 전사에 최적화된 네이버 클로바노트(무료 버전)[2]를 활용해 자동 전사하였다. 전사본의 정확성은 주저자가 원본 녹음을 다시 청취하며 검토 및 수정하였다.

3.4 자료 분석 방법

학습자의 성적 보고서에 대한 이해도(연구 질문 1)는 전체 학습자의 전반적인 이해도를 파악한 후 학교급(중학교와 고등학교)에 따른 차이를 비교하는 방식으로, 크게 세 단계로 분석을 진행하였다. 먼저 이해도를 측정하는 각 객관식 문항에 대한 응답을 정답은 1점, 오답은 0점으로 처리한 뒤, 전체 학습자를 대상으로 문항별 정답률을 빈도와 백분율로 산출하였다. 다음으로, 학교급에 따른 문항별 정답률(빈도와 백분율)을 비교 분석하였다. 이후, 학교급별 정답률 차이가 통계적으로 유의미한지를 검증하기 위해 집단 독립성 카이제곱 검정 (group-independence chi-square test)을 실시하였다. 검정에 앞서 응답 데이터가 해당 통계의 기본 가정을 충족하였음을 확인하였다(Larson-Hall 2010).

성적 보고서의 이해 용이성, 결과의 타당성, 학습적 유용성에 대한 학습자 인식(연구 질문 2) 역시 전체 학습자의 인식 수준을 살펴본 후, 학교급 간 차이를 비교하는 방식으로 분석하였다. 우선 전체 학습자를 대상으로 인식의 세 가지 측면에 대한 응답의 평균과 표준 편차를 계산하였다. 그 다음 각 측면에 대해 학교급별로 평균과 표준 편차를 비교하여, 두 집단의 인식 경향을 비교하였다. 이어서, 학교급 간 인식 수준의 차이에 대한 통계적 유의미성 유무를 확인하였는데, 데이터가 정규성 가정을 충족하지 않아 비모수 두 집단 간 순위 비교 검정(Mann-Whitney U test)을 시행하였다(Larson-Hall 2010). 모든 통계 분석은 IBM SPSS version 29.0을 사용하여 수행하였다.

성적 보고서의 활용성과 개선 방안에 대한 교사의 인식(연구 질문 3)은 면담 전사 자료를 Braun과 Clarke(2012)의 주제 분석(thematic analysis) 절차에 따라 분석하였다. 주제 분석은 면담 자료를 반복적으로 읽으며 중요한 의미를 가지거나 반복되는 표현에 코드(code)를 부여하고, 유사하거나 관련성 있는 코드들을 묶어 주제(theme)를 도출하는 방식으로 수행하였다. 아울러, 전사 자료 중에서 학습자의 성적 보고서에 대한 이해와 인식에 관한 교사의 진술은 연구 문제 1과 2의 정량적 결과에 대한 심층적 해석을 보완하는 질적 자료로 활용하였다. 예를 들어, 학습자들이 이해 용이성을 긍정적으로 평가한 경우, 교사의 진술을 통해 이러한 인식에 영향을 미친 요인을 파악하였다.

[2] https://clovanote.naver.com

© 2025 KASELL All rights reserved

4. 연구 결과 및 논의

4.1 학습자의 성적 보고서에 대한 이해도

표 3은 성적 보고서 이해도를 묻는 문항에 관한 결과를 전체 학습자와 학교급(중학교, 고등학교)별로 제시한 것이다. 전체 학습자 기준으로 문항 27(총점)의 정답률이 가장 높았지만 64% 수준에 그쳤고, 문항 28–30(구술, 대화, 발표 영역에 대한 피드백)의 정답률은 50% 미만으로 낮게 나타났다. 이러한 경향은 학교급별 분석에서도 유사하게 나타났다. 한편, 학교급 간 이해도를 비교 분석한 결과, 문항 27(총점)과 문항 30(대화 영역 피드백)에서는 고등학생이 중학생보다 높은 정답률을 보였지만, 문항 28(구술 영역 피드백)과 29(발표 영역 피드백)에서는 중학생의 정답률이 높았다.

표 3. 학습자 이해도: 정답의 빈도수와 백분율

	N	문항 27 (총점)	문항 28 (구술 영역)	문항 29 (발표 영역)	문항 30 (대화 영역)
전체	244	156(64%)	118(48%)	104(43%)	119(49%)
중학교	192	120(63%)	94(49%)	83(43%)	93(48%)
고등학교	52	36(69%)	24(46%)	21(40%)	26(50%)

Note. 문항 번호 옆 괄호 안의 표현은 해당 문항과 관련된 성적 보고서 결과 또는 피드백 항목을 명시한 것임.

하지만, 표 4에 나타나 있듯이, 모든 문항에서 중학생과 고등학생 간 차이는 통계적으로 유의미하지 않았으며, 효과 크기 또한 모두 매우 작은 수준이었다. 이는 학습자의 총점 및 세부 영역별 평가 결과에 대한 이해도가 학교급에 따라 유의미하게 달라지지 않으며, 학교급은 이해도의 약 1~6%만을 설명하는 데 그친다는 점을 의미한다.

표 4. 학교급별 성적 보고서 이해에 관한 집단 독립성 카이제곱 검증 결과

	χ^2	*df*	*p*	φ
문항 27(총점)	.804	1	.370	.057
문항 28(구술 영역 피드백)	.242	1	.623	.032
문항 29(발표 영역 피드백)	.135	1	.713	.024
문항 30(대화 영역 피드백)	.040	1	.842	.013

이와 같이 성적 보고서에 대한 낮은 수준의 학습자 이해도는 기존 연구들에서도 지속적으로 지적된 바 있다(Hsieh 2023, Sawaki and Koizumi 2017). 주로 성인 학습자를 대상으로 한 이들 연구에 더하여, 본 연구는 중등학교 학습자에게도 유사한 문제점이 나타난다는 사실을 추가로 확인하였다.

성적 보고서에 대한 이해의 어려움이 구체적으로 어디에서 비롯되는지 알아보기 위해서 성적

© 2025 KASELL All rights reserved

보고서의 구성과 관련 문항을 분석한 결과, 학습자들이 다음의 세 가지 측면에서 어려움을 겪었을 가능성이 있는 것으로 나타났다. 첫째, 성적 보고서에 제시된 점수의 계층적 구조에 대한 배경지식의 부족한 경우, 점수 간의 관계를 올바르게 이해하지 못할 수 있다. 예를 들어, 문항 27('하은솔 학생의 *종합평가 점수 88점*은 어떤 말하기 능력들을 종합적으로 평가한 결과인가요?')은 총점의 구성 체계를 이해해야 정답을 선택할 수 있는 문항인데, 총점은 세 영역(구술, 대화, 발표) 점수의 합이며, 각 영역 점수는 다시 하위 평가 요소(예: '구술 능력'의 경우 '흐름'과 '유창성')들의 점수로 구성된다. 그러나 이러한 층위적 점수 구조에 대한 배경지식이 부족한 학습자의 경우, 명확한 설명이 제공되지 않는 한, 그 관계를 정확히 파악하기 어려웠을 수 있다. 예컨대, '정확성'을 '구술 능력'의 하위 요소로 이해하기보다는 총점에 독립적으로 반영되는 별개의 요소로 오인했을 수 있다는 의미이다.

둘째, 서술형 피드백의 진술을 자세히 읽고 분석하지 않는 경우, 의미를 정확하게 해석하지 못하고 모호한 용어에 기반해 직관적으로 판단할 위험이 있다. 예를 들어, 문항 28('하은솔 학생의 *종합 평가 점수 및 피드백*을 읽고, 구술 능력을 향상시키기 위해 알맞은 방법을 고르세요')은 '구술 능력'에 대한 피드백인 '속도를 조절하며 반복적으로, 큰소리로 읽기'라는 진술을 바탕으로 정답을 도출해야 하는 문항이다. 그러나 피드백이 줄글의 형태로 제시되어 있어 전체 내용을 주의하여 읽지 않고, '구술 능력'이라는 다소 모호한 용어에만 근거하여 의미를 직관적으로 판단했을 가능성이 있다.

셋째, 시각적 정보(그래프)와 서술적 피드백을 통합적으로 해석하지 못하고, 한 가지 유형의 정보에만 의존하는 경우에도 결과 해석에 어려움을 겪을 수 있다. 문항 29('하은솔 학생의 *종합 평가 점수 및 피드백*과 *세부 영역별 정밀 진단 결과*를 보고, 발표 능력에서 특히 더 노력해야 할 부분을 고르세요')는 '발표 능력'에 대한 서술형 피드백과 그래프 결과(Needs Work로 표시된 항목)를 연계하여 해석해야 정답을 유추할 수 있다. 하지만 두 정보를 통합적으로 활용하지 않고 시각적으로 접근이 용이한 그래프만 참고하고 서술형 피드백은 간과하는 경우 해석의 오류를 범할 수 있다. 이와 유사하게 문항 30('*세부 영역별 정밀 진단 결과*를 보고, 하은솔 학생의 대화 능력에 대해 알 수 있는 점은 무엇인가요?')의 경우 서술형 피드백과 그래프를 통합적으로 이해해야 할 뿐만 아니라, 그래프에 제시된 '상호작용'과 '과제 완성도'라는 용어를 정확히 이해해야 정답에 도달할 수 있다. 해당 용어가 서술형 피드백에 설명되어 있지만, 이 부분을 충분히 읽지 않은 학습자의 경우 그래프의 해석에 어려움을 겪었을 수 있다. 반대로, 용어에 대한 개념적 이해는 있지만, 시각 자료의 분석 역량이 부족하여 정확한 해석에 이르지 못했을 가능성도 존재한다.

종합적으로 보면 학습자들은 총점이나 그래프에 비해 서술형 피드백에는 접근을 회피하거나 해석에 어려움을 느끼는 경향이 있는 것으로 보이며, 이는 성적 보고서 내 다양한 정보 유형을 통합적으로 이해하는 데 저해 요인으로 작용했을 수 있다. 이러한 가능성은 아래에 제시한 교사 A의 면담 발췌 내용에서도 일부 확인된다.

> *…애들이 점수 딱 나오면은 애들은 일단 점수에 꽂혀 있잖아요. 그 세부적인 것들을 보는 것보다…보긴 보는데 이제 그게 무슨 의미인지를 잘 모르니까 그거를 조금 잘 설명해 주면 좋을 것 같고 그게 아무리 거기 리포트에 뭐라고 써져 있어도 일단 줄글로 있으면은 별로*

© 2025 KASELL All rights reserved

안 읽으니까…(교사 A)

요컨대, 본 성적 보고서에 대한 낮은 이해도는 대상 학습자의 특성과 요구를 충분히 반영하지 못했으며, 이론적 근거에 기반한 체계적인 설계가 반드시 사용자의 정확한 이해로 직결되지 않을 수 있음을 시사한다. 따라서 설계 결과물을 학습자가 실제로 해석하는 방식을 면밀히 검토하고, 그 과정에서 도출된 문제점을 개선하는 접근 방식을 통해 해석의 타당성을 반드시 확보할 필요가 있다.

4.2 이해 용이성, 결과의 타당성, 학습적 유용성에 대한 학습자 인식

표 5는 학습자들의 성적 보고서에 대한 인식 수준을 문항별로 분석한 결과이다. 전체 응답 결과를 살펴보면, 모든 문항에서 약 4점('그렇다'에 해당하는 수준)의 비교적 긍정적인 응답이 나타났다. 문항 간 점수 차이는 크지 않았으나, '이해 용이성'에 대한 평균 점수가 가장 높았으며, 그 다음으로 '결과의 타당성', '학습적 유용성' 순으로 나타났다. 이러한 경향은 학교급별 분석에서도 유사하게 확인되었다.

학교급 간 비교에서는 '이해 용이성' 항목에서 고등학교 학습자의 점수가 다소 높았고, '결과의 타당성' 항목에서는 중학교 학습자가 상대적으로 더 높은 점수를 부여하였으며, '학습적 유용성'에 대해서는 두 집단 간 점수 차이가 거의 없었다.

표 5. 학습자의 인식 분석 결과: 문항별 평균(표준 편차)

문항 내용	전체 학습자	중학교	고등학교
Q20. 이해 용이성	4.08(0.83)	4.05 (.85)	4.21 (.75)
Q21. 결과의 타당도	3.90(0.86)	3.92 (.84)	3.85 (.92)
Q22. 학습적 유용성	3.82(0.90)	3.82 (.89)	3.81 (.91)

Note. 전체 학습자 수는 244명이며, 이 중 중학생은 192명, 고등학생은 52명이었음.

표 6은 학교급에 따른 응답 점수 차이를 비모수적 방법으로 비교한 결과이다. 세 문항 모두에서 중학생과 고등학생 간 통계적으로 유의한 차이는 나타나지 않았으며, 효과 크기 또한 1~7%로 매우 낮은 수준이었다. 이러한 통계적 분석 결과는 본 연구에서 개발한 성적 보고서가 적어도 인식의 측면에서는 중학교와 고등학교 학습자의 특성과 요구를 비교적 일관되게 충족시켰음을 시사한다.

표 6. 학교급별 학습자 인식에 대한 비모수 두 집단 간 순위 비교 결과

	U	*p*	*r*
Q20. 이해 용이성	4,492.5	.238	−0.07561
Q21. 결과의 타당도	4,828	.699	−0.02478
Q22. 학습적 유용성	4,951.5	.924	−0.00608

© 2025 KASELL All rights reserved

각 인식 측면별로, 성적 보고서의 어떤 특성이 학습자들의 긍정적 수용에 기여했는지에 관해 자세히 살펴보도록 하겠다. 먼저, 이해 용이성과 관련된 학습자의 긍정적 인식은 성적 보고서를 설계할 당시, 선행 연구에서 제안한 점수 보고서에 대한 설계 지침(예: Zenisky and Hambleton 2013)을 반영하여, 평가 결과를 종합 점수, 영역별 점수, 시각적 그래프, 서술형 피드백 등 다양한 형태로 제공한 데 기인한 것으로 해석할 수 있다. 교사 면담에서도 성적 보고서의 구성 방식이 실제로 효과적이었다는 것을 확인할 수 있었는데, 다수의 교사가(교사 A, B, C) 학습자들이 점수와 그래프를 눈여겨본다는 사실을 언급하였다. 이러한 결과는 특히 간결한 수치 정보인 점수와 시각적 자료인 그래프가 성인뿐 아니라 중등학교 학습자에게도 직관적이고 접근성이 좋다는 선행 연구들(Gu et al. 2021; Hsieh 2023, Sawaki and Koizumi 2017)과도 일치한다.

그러나 긍정적 인식이 학습자의 실제 이해도와 일치하지 않았다는 점에 주목할 필요가 있다. 이는 학습자가 자신에게 직관적으로 해석할 수 있어 보이거나 선호하는 정보의 형태(예: 점수, 그래프)만 선택적으로 참고하여 평가 결과에 대한 주관적 해석을 도출하고, 이를 정확하다고 믿는 경향이 있는 것으로 해석된다. 예를 들어, 학습자들은 100점 만점 중 90점이면 '영어 말하기를 잘한다'라는 단순한 인상적 결론을 내리는 데에 그칠 뿐, 해당 점수의 구체적 의미를 충분히 파악하여 학습 계획을 세우고, 실천하는 데에는 한계가 있음을 시사한다. 따라서, 학습자가 단순히 인상적 해석을 넘어 결과의 구체적인 의미와 활용 방안을 이해할 수 있도록 추가적인 안내와 지원을 제공해야 하겠다.

평가의 타당성과 학습적 유용성에 대한 학습자의 인식 역시 전반적으로 긍정적인 것으로 나타났으며, 이는 본 성적 보고서의 주요 설계적 특성에서 기인한 것으로 해석된다. 먼저, 학습자의 전반적인 능력을 잘 설명해 준다는 타당성에 대한 긍정적 인식은 성적 보고서가 관련 문헌(예: Gu et al. 2021, Hsieh 2023, Sawaki and Koizumi 2017, Zenisky and Hambleton 2004)을 바탕으로 학습자가 세부 영역별 평가 결과 부분에서 막대그래프의 길이를 통해 자신의 상대적 강점과 약점을 직관적으로 한눈에 파악할 수 있도록 설계되었기 때문으로 판단된다. 또한, 학습적 유용성의 경우 'Needs work'로 평가된 구인에 대해서 링크를 통해 추천 학습 활동 화면으로 쉽게 이동할 수 있도록 성적 보고서가 설계되었는데, 이러한 후속 학습으로의 연계 기능이 학습자의 긍정적 인식에 영향을 미쳤을 가능성이 높아 보인다.

4.3 성적 보고서의 활용성과 개선 방향에 대한 교사의 인식

4.3.1 성적 보고서의 활용성

연구에 참여하여 성적 보고서를 검토한 교사들은 현재 채점 시스템이 충분히 엄격하지 않다는 점에 모두 공감하였다. 이는 학습 프로그램이 아직 개발 중으로, 시범 운영 단계에서 학생들의 긍정적 경험과 학습 동기 향상을 고려해 채점을 관대하게 설정했기 때문이다. 하지만, 향후 해당 문제가 개선된다면 본 성적 보고서가 학교 현장에서 다양한 교육적 목적으로 활용될 가능성에 대해서도 언급하였다.

무엇보다도 교사들은 모두 학생들의 영어 말하기 능력 향상을 위한 학습의 중요성을 충분히

© 2025 KASELL All rights reserved

인지하고 있었으나, 실제 수업 현장에서는 진도 부담과 다수의 학생을 동시에 지도해야 하는 현실적 제약으로 인해 개별적인 평가와 피드백 제공이 어렵다는 문제점을 토로하였다. 이런 측면에서 성적 보고서가 학습자 맞춤형 피드백을 제공한다는 점 자체에 큰 의의가 있으며, 평가 결과를 다양한 유형으로 풍부하게 제시한다는 점에 대해서도 전반적으로 만족감을 나타냈다(교사 A, B, C). 특히 중학교 교사 C는 학기 초에 지필평가로만 실시하던 진단 평가에 성적 보고서를 추가로 활용할 수 있을 뿐만 아니라, 수행평가나 수업 계획 수립, 생활기록부 교과 세부능력 및 특기사항(과세특) 작성에도 유용하게 활용될 수 있을 것으로 평가했다.

> *저희가 분반 작업할 때 같이 보지 않을까 생각은 들어요…저희가 단순히 그냥 지필평가로 봤던 그 진단 점수가 똑같이 높다면 그런 학생들을 조금 더 골고루 배치해서…아니면 그거를 보고 저희가 수행평가를 계획할 때 활용하기 좋을 것 같기는 해요. 이제 해마다 아이들의 실력이나 반응이 조금씩 다르다 보니까 뭔가 그런 진단 평가가 대체로 높게 나오는 학년이다 하면 수행평가를 할 때 저희가 예년보다 조금 더 욕심을 내서 이것도 추가해 볼까라고 한다던가 아니면 뭔가 조금 잘 못한다 하면 수업 시간에 이걸 좀 더 활용해서 활동할 시간을 좀 더 줘야겠구나라는 지표로 사용할 것 같아요.* (교사 C)

반면 고등학교 교사들(교사 B, D)의 경우, 대부분의 학교 내신 평가가 대학 입시와 직결되기 때문에, 본 성적 보고서에 담긴 정보를 공식적인 평가 도구로 사용하기에는 한계가 있다고 인식하였다. 대신, 학습자에게 학습 방향을 안내하거나 수행평가에서 보조적으로 활용할 수 있는 평가 및 학습 지원 도구로서의 가능성을 언급하였다. 예를 들어, 교사 B는 성적 보고서가 수행평가 상황에서 교사를 대신하여 학습자에게 어떤 부분을 연습해야 하는지를 제안해 줄 수 있는 피드백 제공 도구로서의 활용 가능성에 대해 긍정적으로 평가하였다.

4.3.2 개선 방향

면담 결과 교사들은 본 성적 보고서가 개별 피드백 제공 등 학습적으로 긍정적인 측면이 있지만, 실제 교육 맥락에서 사용되기 위해서는 중등학교 학습자들의 특성과 요구를 충분히 반영할 수 있도록 보완이 필요하다고 의견을 모았으며, 다양한 관점에서 구체적인 개선 방안을 제안하였다.

첫째, 교사 A, B, C는 공통으로 성적 보고서에 사용되는 용어가 학습자에게 친숙하지 않을 수 있음을 지적하며, 결과를 더 쉽게 이해할 수 있도록 성적 보고서에 사용되는 용어를 친숙한 표현으로 대체하거나 추가 설명을 함께 제공해야 한다는 점을 강조했다. 예를 들어, 교사 A는 다음과 같이 언급하였다:

> *… 유창성도 사실 유창하다는 게 뭐지 그러니까 같은 시간 내에 조금 더 빨리 빠르고 유창하게 말할 수 있다는 뜻이야라고 한다든지 조금 그 자료에 들어 있으면…저는 이제 일일이 애들한테 설명해 주면서 돌아다녔지만…예를 들어 어휘 다양성이 조금 낮게*

© 2025 KASELL All rights reserved

나왔다는 거는 반복되는 어휘를 썼다든지 뭐 그렇게 이제 얘기를 해줬거든요.(교사 A)

이와 관련해 일부 교사들은 학습자용 결과 해설 자료를 별도로 제작하여 참고할 수 있도록 하고, 교사가 구두 설명을 제공하거나 하이퍼링크 형태로 설명을 제공하는 방안을 제안하였다. 이러한 제안은 이전 연구들(Gu et al. 2021, Zenisky and Hambleton 2004)에서도 논의된 사항들이다. 또한, 선행 연구에서 밝혀진 바와 같이, 학습자가 성적 보고서 내용에 관한 질문에 응답하고 즉각적인 피드백을 받는 게임 기반 상호작용 활동을 포함하는 것 역시 이해를 높이는 데 효과적인 방안이 될 수 있을 것으로 보인다(Vessu et al. 2012).

둘째, 교사 B와 D는 채점 기준을 명확히 제시하고 이를 학습자의 실제 수행 데이터 분석 결과와 연계하여 더욱 구체적인 피드백을 제공해야 한다고 제안하였다. 성적 보고서 내에서 링크를 통해 관련 학습 활동으로 이동할 수 있는 기능이 제공되고 있기는 하나, 이에 더해 학습자의 관점에서 결과를 어떻게 해석하고, 어떤 부분을 어떻게 보완해야 하는지를 구체적으로 설명하는 내용이 추가될 필요가 있음을 시사한다. 이러한 관점은 교사 B의 면담 내용에서 확인할 수 있다.

> *대화 능력에서…상호작용 부분에서 negotiation이 이제 잘 통과해서 이렇게 15점인 건지 아니면은 실패를 거듭해서 상호작용을 많이 했기 때문에 15점인 건지가 드러나지 않는 점이 좀 아쉽다…*(교사 B)

셋째, 교사 A, B, C는 학습자들이 줄글 형태의 서술형 피드백보다는 시각적 그래프를 선호하는 경향이 있기 때문에 세부 영역별 서술형 피드백을 먼저 제시하기보다는 그래프 형태의 결과를 먼저 제공하고 이후 피드백은 줄글보다는 핵심 키워드를 중심으로 제시하여 학습자가 장단점을 한눈에 파악할 수 있도록 하는 방안을 추천하였다. 키워드 제시에서도 대조적인 색상을 활용해 시각적 구분을 강화하는 것이 효과적일 것이라는 추가 의견도 있었다.

이러한 교사들의 다양한 개선 의견은 향후 성적 보고서 개선 과정에 적극적으로 반영되어, 성적 보고서가 학습자의 특성과 요구를 충실히 반영하는 맞춤형 평가 도구로 기능할 수 있도록 해야겠다.

5. 결론 및 제언

본 연구는 국내 학교 영어교육의 제고를 목적으로 개발 중인 영어 말하기 자동 평가 시스템에 포함된 성적 보고서 개발 과정에 대해 사용자의 이해도와 인식에 중점을 두고 소개하였다. 주요 연구 결과는 다음과 같다. 첫째 성적 보고서에 대한 학습자들의 전반적인 이해도가 학교급과 관계없이 상당히 낮았다. 둘째, 성적 보고서의 세 가지 측면(이해 용이성, 결과의 타당성, 학습적 유용성)에 대한 학습자들의 인식은 대체로 긍정적이었다. 셋째, 교사들은 세부 영역별로 개별화된 피드백을 제공한다는 측면에서 긍정적으로 평가하였지만, 학습자들에게 친숙하지 않은 용어에 대한 설명의 부재, 학습자들이 선호하지 않는 서술형의 피드백 제시 방식 등을 주요한 문제점으로

© 2025 KASELL All rights reserved

지적하였다. 개선 방안으로는 학습자 친화적인 표현을 활용한 용어 해설의 제공, 그리고 결과를 그래프로 우선 제시하고 이후 피드백은 핵심 키워드를 중심으로 간결하게 제공하는 방식 등을 추천하였다.

주요 연구 결과 중 특히 주목할 점은 성적 보고서에 대한 실질적 이해도와 이해 용이성에 대한 인식 간의 괴리이다. 이는 평가 결과가 학습자에게 직관적으로 보이거나 선호하는 방식(예: 점수, 그래프)으로 제시되는 경우, '이해하기 쉬운 보고서'라는 긍정적인 인식을 유도할 수는 있지만, 실제 이해도와는 차이를 보일 수 있음을 시사한다. 이러한 괴리를 극복하기 위해서는 학습자들이 올바른 해석을 할 수 있도록 다양한 교육적 지원이 제공되어야 한다. 예를 들어, 평가 시행 이후 수업 시간에 성적 보고서에 대한 교사의 구두 해설과 함께 학습자용 설명 자료를 제공하거나, 학습자가 스스로 이해도를 점검하고 피드백을 받을 수 있는 퀴즈 기능을 성적 보고서 내에 탑재하여(Vessu et al. 2012), 이를 수업 내 학습 활동으로 활용할 수 있겠다. 나아가 성적 보고서는 학습자의 특성과 필요에 맞게 지속적으로 개선되어야 하며, 단순히 평가의 결과가 아니라 교실 내 평가-피드백-학습 순환의 일부로 자리 잡아야 한다. 이러한 방향은 2022 개정 영어과 교육과정에서 제시한 여러 평가 방향 중 하나인, 학습자가 평가를 학습 활동의 일부로 받아들이고, 이를 통해 자신의 학습 과정과 성과를 성찰할 수 있도록 평가를 설계해야 한다는 원칙과도 일치한다(교육부 2022).

이러한 평가 기반 활동이 학교 현장에서 활발히 운영되기 위해서는 교사 연수 개발 등의 제도적, 정책적 연계가 마련되어야 한다. 교사 연수 과정에서는 교사들이 성적 보고서의 내용을 학습자의 눈높이에 맞게 설명하고 지도할 수 있는 역량을 갖출 수 있도록 성적 보고서 해석 지도 전략, 성적 보고서 기반 수업 활동 연계 방안과 교실 내 적용 사례 소개 등이 포함될 수 있겠다.

본 연구를 통해 학습자 개인용 성적 보고서 설계와 관련해 도출할 수 있는 시사점은 다음과 같다. 성적 보고서는 평가 결과를 학습자가 목적에 맞게 해석하고 활용하는 데 핵심적인 역할을 하므로, 단순한 결과 전달 문서가 아니라 평가 도구의 필수적인 구성 요소로 개발되어야 한다. 또한, 성적 보고서는 이론적 근거뿐만 아니라 실제 학습자의 특성과 요구를 반영하여 체계적인 과정을 통해 맞춤형으로 개발되어야 한다. 예를 들어, 발표 능력의 하위 평가 항목인 '유창성'의 경우, 학습자가 개념을 보다 쉽게 이해할 수 있도록 성적 보고서 내에서 해당 용어를 클릭하면, 유창성이 '1분 동안 말한 단어의 수'를 의미한다는 간단한 설명이 제공되도록 보고서의 설계를 학습자 친화적으로 개선할 수 있다. 이와 더불어 평가 결과를 기준값과 함께 게이지 그래프 형태로 제시함으로써, 학습자가 자신의 상대적 성취 수준을 직관적으로 이해할 수 있도록 지원하며, '1분 동안 50단어 이상을 말할 수 있도록, 조금 더 빠르게 말하는 연습을 해 보세요' 와 같은 학습자가 쉽게 이해할 수 있는 구체적인 피드백 문구를 함께 제공하고, '1분' '50단어 이상', '빠르게 말하는 연습'과 같은 핵심 키워드에는 강조 색상을 적용할 수 있겠다. 이러한 개발 접근 방식을 통해 평가 타당성 즉 결과의 해석과 활용을 극대화할 수 있겠다.

본 연구는 개발 단계에 있는 인공지능 기반 영어 말하기 평가 시스템에 관한 탐색적 시도로서, 방법론적 측면에서 다음과 같은 한계가 있으며, 이는 향후 연구에서 보완되어야 할 과제이다. 우선 다소 후한 채점 기준 적용은 불가피한 선택이었지만, 이번 시범 운영을 통해 수집된 자료를 기반으로 채점 시스템을 고도화하고, 이를 통해 평가 결과의 정확성과 신뢰성을 확보해야 한다. 또한, 본 시스템이 K-12를 대상으로 개발되고 있으나, 연구의 참여자는 중등학교(특히 중학교)

© 2025 KASELL All rights reserved

학습자에 편중되어 있고, 초등학생 학습자나 교사는 포함되지 않아 이들에 관한 연구 결과의 대표성이 결여되어 있다. 따라서, 향후 연구에서는 고등학생들의 표집 수를 확대하고, 초등학생들과 초등 교사들을 연구에 포함해서 다양한 사용자 집단에 관한 결과를 제시할 필요가 있다. 아울러, 이해도를 객관적으로 측정하는 데 있어 정답-오답 기준의 이분법적 채점을 활용하였기 때문에, 학습자의 부분적 이해나 다양한 해석 과정에 대한 심층적인 탐구가 필요하다. 특히 본 연구에서 제시한 학습자의 낮은 이해도에 관한 결과 해석은 연구자들의 추론에 기반한 것으로, 향후 연구에서는 학습자 면담과 같은 질적 자료 수집을 통해 그 원인을 실증적으로 파악할 필요가 있다. 한편, 소수의 교사들을 대상으로 수행한 면담을 통해 성적 보고서의 개선 방안에 대한 구체적인 전문가 의견을 수집할 수 있었지만, 교사 집단에 관한 결과를 일반화하기 위해서는 보다 큰 표본을 대상으로 설문조사 등의 정량적 연구를 수행하는 것이 바람직할 것이다.

이러한 방법론적 한계에도 불구하고, 본 연구는 지금까지 국내 연구에서 상대적으로 등한시되어온 성적 보고서의 개발 과정을 체계적으로 탐구한 최초의 시도로서, 영어 말하기 평가 및 교육 연구의 새로운 방향을 모색했다는 점에서 학문적 의의를 찾을 수 있다.

참고 문헌

교육부(Ministry of Education). 2022. 『2022 개정 영어과 교육과정(교육부 고시 제2022-33호 [별책 14])』(*2022 Revised English Curriculum [Ministry of Education Notification No. 2022-33*, Appendix 14]). 교육부(Ministry of Education).

김나예・김정렬(Kim, N.-Y. and J.-R. Kim). 2018. 초등영어 수행평가에 대한 영어전담교사들의 인식조사(A Perception Study of Elementary School English Teachers on the English Performance Assessment). ≪교사교육연구≫(*Teacher Education Research*) 57(4), 529-538.

박선영・민찬규(Park, S.-Y. and C.-K. Min). 2019. 2015 개정 교육과정 기반 중학교 1학년 영어 말하기 수행평가 시행 방안에 대한 델파이 조사 연구(A Delphi study on the implementation of English-speaking performance assessment for middle school 1st graders based on the 2015 revised English curriculum). ≪영어교과교육≫(*Journal of the Korea English Education Society*) 18(1), 25-45.

심창용(Sim, C.-Y.). 2024. AI 디지털교과서 활용 영어교육의 기초(The basics of AIDT-powered English education). ≪*ESP Review*≫ 6(2), 137-154.

이진화・최윤덕・성민창・김혜영(Lee, J.-H., Y. Choi, M.-C. Seong and H. Kim). 2023. 자동채점 기반 영어 말하기 시험 현황 분석(An analysis of the current status of automated scoring-based English speaking tests). ≪영어교육≫(*English Teaching*) 78(2), 223-244.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. 2014. *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Braun, V. and V. Clark. 2012. Thematic analysis. In H. Cooper, P. M. Camic, D. L. Long, A. T.

© 2025 KASELL All rights reserved

Panter, D. Rindskopf, and K. J. Sher, eds., *APA handbook of research methods in psychology, Vol. 2: Research designs: Quantitative, qualitative, neuropsychological, and biological* (pp. 57−71). American Psychological Association.

Brown, G. T., T. M. O'Leary and J. A. Hattie. 2019. Effective Reporting for Formative Assessment: The asTTle case example. In D. Zapata−Rivera ed., *Score reporting research and applications* (pp. 107−125). Routledge.

Creswell, J. W., and V. L. Plano Clark, 2011. *Designing and conducting mixed methods research* (2nd ed.). Sage.

Fulcher, G. 2003. *Testing second language speaking.* Longman.

Goodman, D. P. and R. K. Hambleton. 2004. Student test score reports and interpretive guides: Review of current practices and suggestions for future research. *Applied Measurement in Education* 17(2), 145−220.

Gu, L. and L. Davis. 2019 Providing SpeechRater feature performance as feedback on spoken responses. In K. Zechner and K. Evanini, eds., *Automated speaking assessment* (pp. 159−175). Routledge.

Gu, L., L. Davis, J. Tao and K. Zechner. 2021. Using spoken language technology for generating feedback to prepare for the TOEFL iBT test: A user perception study. *Assessment in Education: Principles, Policy and Practice* 28(1), 58−76.

Hambleton, R. K. and A. L. Zenisky. 2013. Reporting test scores in more meaningful ways: A research−based approach to score report design. In K. F. Geisinger ed., *APA handbook of testing and assessment in psychology (Vol. 3): Testing and assessment in school psychology and education* (pp. 479−494). American Psychological Association.

Hattie, J. 2009. Visibly learning from reports: The validity of score reports. *Online Educational Research Journal*, 1−15.

Hattie, J. and H. Timperley. 2007. The power of feedback. *Review of educational research* 77(1), 81−112.

Hsieh, C.−N. 2023. Evaluating the use and interpretation of the TOEIC listening and reading test score report: Perspectives of test takers in Japan (ETS Research Report Series No. RR−23−01). *Educational Testing Service.* Available online at https://doi.org/10.1002/ets2.12364

Kim, A. A., M, Chapman, A. Kondo and C. Wilmes. 2020. Examining the assessment literacy required for interpreting score reports: A focus on educators of K−12 English learners. *Language Testing* 37(1), 54−75.

Larson−Hall, J. 2010. *A Guide to Doing Statistics in Second Language Research Using SPSS.* Routledge.

O'Leary, T. M., J. A. Hattie and P. Griffin. 2017. Actual interpretations and use of scores as aspects of validity. *Educational Measurement: Issues and Practice* 36(2), 16−23.

Sawaki, Y., and R. Koizumi. 2017. Providing test performance feedback that bridges assessment and instruction: The case of two standardized English language tests in Japan. *Language*

© 2025 KASELL All rights reserved

Assessment Quarterly 14(3), 234−256.

Taylor, L. 2005. Washback and impact. *ELT Journal* 59(2), 154−155.

Vezzu, M., W. VanWinkle and D. Zapata−Rivera. 2012. Designing and evaluating an interactive score report for students (Research Memorandum No. ETS RM−12−01). *Educational Testing Service.* Available online at https://www.ets.org/Media/Research/pdf/RM−12−01.pdf

Wainer, H. 1992. Understanding graphs and tables. *Educational Researcher* 21, 14−22.

Zapata−Rivera, D., P. Kannan, C. Forsyth, S. Peters, A. D. Bryant, E. Guo and R. Long. 2018. Designing and evaluating reporting systems in the context of new assessments. In D. Schmorrow and C. Fidopiastis, eds., *Augmented cognition: Users and contexts. AC 2018. Lecture notes in computer science* (Vol. 10916, pp. 143−153). Springer.

Zenisky, A. L., and R. K. Hambleton. 2015. A model and good practices for score reporting. In S. Lane, M. R. Raymond, and T. M. Haladyna, eds., *Handbook of test development* (2nd ed., pp. 585−602). Routledge.

예시 언어(Examples in): Korean
적용 가능 언어(Applicable Languages): English
적용 가능 수준(Applicable Level): Secondary & Tertiary

© 2025 KASELL All rights reserved

부록

학습자의 이해도 관련 설문 문항

문항	선택지	측정 영역
27. 하은솔 학생의 *종합 평가 점수 88*점은 어떤 말하기 능력들을 종합적으로 평가한 결과인가요?	1) 구술 능력, 정확성, 발표 능력 2) 대화 능력, 과제 완성도, 흐름 **3) 구술 능력, 대화 능력, 발표 능력*** 4) 발표 능력, 유창성, 상호작용	종합 평가 점수 및 서술형 피드백 이해
28. 하은솔 학생의 *종합 평가 점수 및 피드백*을 읽고, 구술 능력을 향상시키기 위해 알맞은 방법을 고르세요.	1) 다른 사람과 천천히 말하기 **2) 너무 빠르지 않게 큰소리로 끊어 읽기*** 3) 정확한 강세와 억양으로 읽기 4) 여러가지 단어를 사용해서 말하기	구술 능력 관련 서술형 피드백 이해
29. 하은솔 학생의 *종합 평가 점수 및 피드백*과 *세부 영역별 정밀 진단 결과*를 보고, 발표 능력에서 특히 더 노력해야 할 부분을 고르세요.	1) 문장을 정확하게 말하기 2) 적절한 단어와 표현 사용하기 **3) 말의 양과 속도 늘리기*** 4) 과제에 필요한 표현 사용하기	발표 능력 그래프 및 서술형 피드백 이해
30. *세부 영역별 정밀 진단 결과*를 보고, 하은솔 학생의 대화 능력에 대해 알 수 있는 점은 무엇인가요?	1) 대화를 잘 이어가지 못했고, 미션(과제)도 잘 해내지 못했다. 2) 대화를 잘 이어갔고, 미션(과제)도 잘 해냈다. **3) 대화를 잘 이어가지 못했지만, 미션(과제)은 잘 해냈다.*** 4) 대화를 잘 이어갔지만, 미션(과제) 잘 해내지 못했다.	대화 능력 그래프 및 서술형 피드백 이해

Note. 선택지의 * 표시는 해당 문항의 정답을 의미함.

© 2025 KASELL All rights reserved