

자동채점 기반 영어 말하기 시험 현황 분석

이진화 · 최윤덕 · 성민창 · 김혜영*

Lee, Jin-Hwa, Choi, YunDeok, Sung, Min-Chang, & Kim, Heyoung. (2023). Analysis of English automated speaking scoring tests. *English Teaching*, 78(2), 223-244.

This study investigated the feasibility of adopting an automatic scoring system (ASS) in a domestic English-speaking education context. Scope, test items, assessment criteria, scoring methods, and reporting strategies of six overseas English-speaking tests utilizing ASSs were examined. Moreover, a comparative analysis was conducted to identify disparities between ASS-based and non-ASS-based speaking tests. Findings were: 1) some ASS-based tests utilized ASS technology throughout the assessment, while others adopted a hybrid scoring system involving human raters; 2) compared to non-ASS-based tests, ASS-based tests used more test items targeting low-level skills such as sound and forms but fewer test items targeting conversation and discourse level skills; 3) pronunciation, fluency, and vocabulary were widely employed as evaluation criteria with sparse use of organization, content, and task completion in most ASS-based tests; 4) differences were minimal in assessment criteria application and score calculation between ASS-based and non-ASS-based tests; and 5) some ASS-based tests provided criteria-specific results and feedback with total scores and proficiency levels.

Key words: English speaking tests, automated speaking scoring (ASS), test items, assessment criteria, scoring, test report/영어 말하기 평가, 자동말하기채점, 평가 문항, 평가 준거, 채점, 평가리포트

*First Author: Jin-Hwa Lee, Professor, Department of English Education, Chung-Ang University

Corresponding Author: YunDeok Choi, Assistant Professor, Department of English Education, Chungnam National University; 99 Daehak-ro, Yuseong-gu, Daejeon 34134, South Korea; Email: yundeokchoi@cnu.ac.kr

Coauthor: Min-Chang Sung, Assistant Professor, Department of English Education, Gyeongin National University of Education

Coauthor: Heyoung Kim, Professor, Department of English Education, Chung-Ang University

Received 31 March 2023; Reviewed 17 April 2023; Accepted 30 May 2023



© 2023 The Korea Association of Teachers of English (KATE)

This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0, which permits anyone to copy, redistribute, remix, transmit and adapt the work, provided the original work and source is appropriately cited.

1. 서론

국내에서 인공지능 기술을 영어교육분야에 적용하려는 시도가 최근 급증하였고, 특히, 그동안 여러가지 제약으로 인해 미흡했던 영어 말하기 교육에 이를 활용하는 방안에 대한 관심이 높아지고 있다. 이들 인공지능기반 영어 말하기 교육 연구는 챗봇 등을 활용한 말하기 연습(Chu & Min, 2019; Kim, 2016; Kim & Lee, 2020), 수행평가 활용 방안(Choi, 2020; Song & Shim, 2020), 발음의 정확도 분석(Choi, 2021) 등의 영역으로 확대되었으나, 여전히 인공지능기반 말하기 시험 및 채점 방안에 관한 연구는 찾아보기 어렵다. 새로운 기술을 통한 말하기 교육의 성공적인 정착에 있어서 최우선 과제는 바로 평가이다. 지금까지 공교육에서 말하기 평가는 시행하기 매우 어려운 영역으로 여겨져 왔는데, 주요 이유로 영어교사들은 채점 결과에 대한 신뢰도 확보, 채점 소요 시간과 업무부담, 평가 문항 및 채점 기준 개발의 어려움 등을 지적하고 있다(Lee, Yun, Kim, & Joo, 2015). 따라서 영어 말하기 평가의 채점 단계에서 인공지능 기술을 활용한다면 시행의 어려움을 줄일 수 있을 것이다.

해외에서는 자동말하기채점(Automated Speaking Scoring, ASS) 기술의 개발 연구가 최근 빈번히 소개되고 있으며(예: Chen et al., 2018; Khabaazbashi, Xu, & Galaczi, 2021; Wang et al., 2018; Xu, Jones, Laxton, & Galaczi, 2021), 기존 영어 말하기 평가에 일부 도입되거나 새로운 평가 프로그램을 개발하여 운영 중이다. 이러한 자동채점 기술을 국내 영어 공교육에 도입할 수 있다면, 채점의 객관적 근거로 신뢰도를 높이고 채점에 소요되는 시간과 업무량을 획기적으로 줄여 학교 현장의 영어 말하기 교육을 촉진할 수 있을 것으로 기대된다. 하지만 자동채점기술 연구는 초기 단계이며, 국내에서의 연구가 미비하여 앞으로 이를 도입하기까지 다각적인 노력이 필요하다. 그 첫 단계로 현재의 자동채점기술과 이를 적용한 최신 자동채점 시험에 대해 면밀히 검토해 볼 필요가 있다. 고비용의 장기적 연구개발로 탄생된 새로운 말하기 시험들에 대한 기술적, 언어평가적 측면의 분석은 국내 자동평가시스템 개발의 기초자료가 될 수 있을 것이다.

본 연구의 목적은 인공지능기술 발달과 더불어 부상한 자동채점기술의 국내 영어 교육 도입 가능성을 타진하기 위하여, 이 기술이 접목된 해외 영어 말하기 시험에 대한 기술적, 언어평가적 정보를 제공하는 것이다. 이를 위해 현재 자동채점시스템을 채택하고 있는 주요 해외 영어 말하기 시험을 대상으로, 자동채점이 어느 수준까지 이루어지고 있는지 살펴보았다. 또한 이들이 채택하고 있는 문항 유형, 평가준거, 채점 방식, 결과 제시 방식을 차례로 분석하고, 이러한 요소들이 기존의 수동채점 방식을 고수하고 있는 영어 말하기 시험과 어떤 차이가 있는지 비교하였다. 구체적으로 본 연구는 다음과 같은 질문을 제시한다.

- 1) 자동채점시스템 기반 영어 말하기 시험에서 자동채점 기술이 적용된 영역과 방식은 어떠한가?

- 2) 자동채점시스템 기반 영어 말하기 시험에서는 어떠한 문항 유형, 평가 준거, 채점 방식, 결과 제시 방식을 사용하고 있는가?
- 3) 이러한 요소들은 자동채점시스템을 사용하지 않는 영어 말하기 시험과 어떠한 차이를 보이는가?

이를 통하여 현재의 자동채점기술 수준에서 가능한 말하기 평가의 가능성과 한계를 확인하여, 앞으로 국내 공교육에서 채택 가능한 기술과 자동채점시스템의 개발 방향에 대해 실질적인 정보를 제공할 수 있을 것으로 기대한다.

2. 자동 말하기 채점(Automated Speaking Scoring) 기술

높은 수준의 영어 말하기 능력에 대한 사회적 요구는 자연스럽게 영어 말하기 평가에 대한 요구를 촉발시켰다. 하지만 그동안 영어 말하기 평가는 높은 비용과 시간 소요, 그리고 무엇보다도 채점의 타당도와 신뢰도 문제로 지속적인 어려움을 겪어왔다. 최근 자동 음성 처리 기술의 발달은 이러한 문제의 새로운 대안으로 떠올랐다. 자동 음성 처리 기술을 활용하면 그림 1과 같이, 수험자의 음성 발화를 인식하여 자동으로 텍스트로 전환할 수 있고, 여기에서 말하기 능력과 관련된 자질들을 추출하여 점수를 부여함으로써 자동채점이 가능해진다.

FIGURE 1

Architecture of a Speech Automarker (Xu et al., 2021, p. 412)

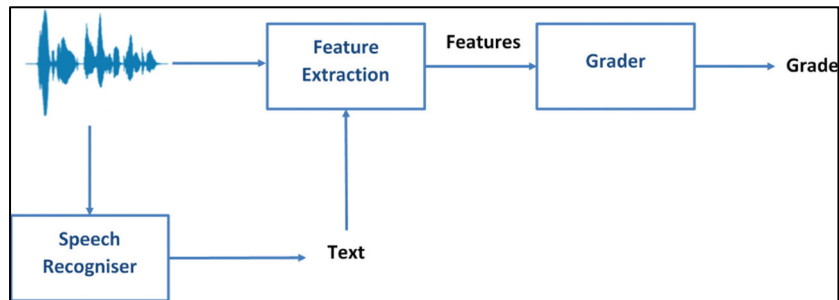


그림 1의 첫번째 단계인 자동음성인식(automated speech recognition, ASR) 단계에서는 음성 신호를 음소나 어휘 텍스트로 바꾸어주는 음향 모델과 말뭉치를 기반으로 어휘 배열의 가능성을 계산하는 언어 모델이 작동한다. 현재 자동채점시스템의 음성인식기의 텍스트 전사(transcription) 정확도는 Linguaskill General Speaking Test의 경우 대략 77%, ETS의 SpeechRaterSM의 경우 80% 정도로 보고되고 있다(Xu et al., 2021).

음성 자료가 텍스트로 전환이 되면, 다음 단계에서는 평가 대상이 되는 자질들을 추출하게 된다. 말하기 능력을 평가하는 데에 어떤 자질들을 추출할 것인지는 각 시험에서 설정한 평가 구인과 관련되기 때문에 시험마다 상이하다. 초창기 자동채점 시스템은 발음 평가에 국한되어 있었다. 주어진 스크립트를 수험자가 소리내어 읽으면 이를 디지털화 한 후 원어민 화자의 발음과 비교하는 방식이었다. 국내에서도 2003년 개발된 Dr. Speaking부터 2021년 초등용 ‘AI 팽톡’에 이르기까지 음성 인식으로 추출한 음성 파형을 원어민과의 발음과 비교하여 유사도를 측정하는 자동발음평가 프로그램이 다수 소개된 바 있다. 그러나 이러한 방식의 자동채점은 제시된 단어 나 문장에 대한 발음 정확도 평가에 불과하여 학습자의 실질적인 영어 말하기 능력을 측정한다고 보기 어렵다. 이에 2000년 초반부터 평가의 범위를 즉흥적인 발화(spontaneous speech)로 확대하고, 발음뿐만 아니라 유창성, 어휘, 문법, 내용, 담화 등의 다각적인 측면을 측정하려는 노력이 지속되어 왔다. 현재 ETS의 자동채점시스템인 SpeechRaterSM의 경우, ‘수행(발음과 유창성)’, ‘언어 사용(어휘와 문법)’, ‘주제 전개(내용, 담화 일관성)’에 대한 자동채점을 시행하고 있다. 그러나 ‘주제 전개’와 관련된 자질을 살펴보면 담화 연결어 유형 수, 대명사 수, 접속사 수를 세는 수준에 그치고 있고, 내용 적합도나 일관성에 대한 측정은 아직 제대로 이루어지지 못하고 있으며, 화용론적 능력, 대화 능력 등에 대한 평가 역시 미진한 실정이다(Wang & Evanini, 2019; Xu et al., 2021). 이러한 한계 때문에 일부 연구자들은 자동채점시스템이 아직은 저위험도 시험에 적합하다고 보고 고위험도 시험에서는 인간채점과 함께 사용할 것을 권장하고 있다(Wang & Evanini, 2019). 실제로 현재 시행되고 있는 자동채점 기반 영어 말하기 시험들 가운데 일부는 자동채점과 인간채점을 결합한 “하이브리드 채점 모형 (Hybrid marking model)”(Xu et al., 2020)을 사용한다.

일련의 평가 자질이 추출되면 다음 단계에서는 채점기(grader)가 채점 모형에 따라 수험자의 점수를 예측한다. 채점기 훈련은 통상적으로 사전에 인간 채점자의 점수가 부여된 대량의 학습자 발화 자료를 통해 이루어진다. 채점기의 신뢰도는 인간 채점자 점수와 얼마나 일치된 점수를 생성하는지를 통해 확인된다. 이러한 일련의 과정은 시험마다 대동소이하지만, 구체적으로 어떤 채점 모형과 알고리즘을 사용하는지는 시험마다 다르고, 채점 결과를 어떤 구인에 대하여 얼마나 상세하게 제공하는지 역시 상이하다.

요약하면 자동채점시스템 기반 영어 말하기 시험은 대체로 그림 1과 같은 구조를 가지고 있고, 최근의 자동음성인식 기술의 향상으로 상당히 높은 수준의 음성인식 정확도를 보이고 있다. 그러나 음성인식 이전 단계에서 어떤 문항을 사용하여 어떤 발화 유형을 수집할 것인지, 음성인식이 이루어진 후에는 수집된 발화에서 어떤 평가 자질을 추출하여 무슨 구인을 측정할 것인지, 추출된 평가 자질에 대해 어떤 채점 모형과 알고리즘을 사용하여 점수를 부여할 것인지, 채점 결과에는 어떤 정보를 얼마나 상세히 포함시킬 것인지 등의 구체적인 양상은 시험마다 상이하다. 따라서 이에 대한 세밀한 분석이 필요하다.

더불어 이러한 자동채점 기반 영어 말하기 시험의 제반 특성들이, 말하기 능력에 대한 이론을 토대로 오랜 기간에 걸쳐 평가 구인과 모형을 개발, 구축해 온 기존의 인간채점 말하기 시험과 비교했을 때 어떤 유사점과 차이점을 보이는지 살펴볼 필요가 있다. 이를 통해, 자동채점 말하기 시험들이 기존 시험에서 측정하고자 했던 다각적인 말하기 능력을 충분히 포괄적으로 측정하고 있는지, 또 자동채점기술의 활용을 통해 얻을 수 있는 장점과 한계는 무엇인지 등을 파악할 수 있을 것이다. 이에 본 연구에서는 현재 사용되고 있는 주요 자동채점 기반 영어 말하기 시험들을 면밀히 분석하고 그 결과를 기존의 수동채점 기반 영어 말하기 시험들과 비교하고자 한다.

3. 연구 방법

3.1. 분석 대상

본 연구에서는 표 1과 같이, 국내외에서 높은 인지도와 수험율을 보유하고 있고 상세 자료가 공개되어 있는 영어 말하기 시험, 총 12종을 분석 대상으로 삼았다. 이 가운데 자동채점기술이 접목된 시험은 6개로, TOEFL® Practice Online과 Duolingo English Test는 미국에서 개발하였고, EnglishScore Speaking Test, Linguaskill General Speaking Test, PTE Academic, Versant English Test는 영국에서 개발하였다. 이들 자동채점 기반 시험과의 비교를 위해 기존의 인간채점 기반 말하기 평가 6종을 선정하였다. 이중 국내에서 개발된 시험은 총 2종으로, 서울대학교 언어교육원에서 고등학생부터 성인을 대상으로 개발한 TEPS와 한국교육과정평가원에서 고등학생을 대상으로 개발했던 국가영어능력평가(NEAT) 2, 3급이다. 국가영어능력평가¹의 경우 현재는 운영되지 않지만, 우리나라 학교 영어교육을 기반으로 개발된 시험이라는 점에서 추후 학교 교육용 말하기 평가에 참조할 만하여 분석에 포함시켰다. 해외에서 개발된 수동채점 영어 말하기 평가로는 대만 교육부가 개발한 GEPT, 미국 ETS에서 개발한 TOEIC, TOEFL Junior, TOEFL iBT²가 분석에 포함되었다. OPIc과 IELTS 시험은 현재 널리 사용되고 있으나 문항 유형이 고정해져 있지 않고 수험자의 응답에 따라 질문이 바뀌는 인터뷰 방식을 취하고 있어 본 연구의 분석 대

¹ 국가영어능력평가지험은 듣기, 읽기, 말하기, 쓰기 각 영역을 평가하는 인터넷 기반시험(IBT)으로서, 성인 대상의 1급은 토익, 토플 등의 해외 영어시험 의존도를 낮추기 위하여, 고등학생 대상의 2, 3급은 수능영어시험을 대체함으로써 실용영어중심의 영어교육을 유도하기 위하여 개발되었다. 2007년 7월 기본계획이 발표된 이후 총 7회의 예비시행을 거쳐 2012년 2회 공식 시행되었다. 그러나 2014년 1월 교육부가 국가영어능력평가의 폐지를 발표하고 7월에 전산시스템을 모두 해체하면서 7년간 추진되어 온 정책이 종결되었다(Chae, 2015)

² 본 연구는 기존 수동 채점 방식을 토대로 분석하였으나, 최근 TOEFL iBT 말하기 영역의 채점 절차가 SpeechRaterSM와 인간 채점자를 결합한 하이브리드 방식으로 변경되었다(Educational Testing Service, 2023)

상에서 제외하였다.

TABLE 1
English Speaking Tests for Analysis

	Tests	Institutions	Country
ASS*	TOEFL® Practice Online	ETS	USA
	Duolingo English Test	Duolingo	USA
	EnglishScore Speaking Test	British Council	England
	Linguaskill General Speaking Test	Cambridge Assessment English	England
	PTE Academic	Pearson	England
	Versant English Test	Pearson	England
Non-ASS	TEPS	Seoul National University Language Institute	Korea
	NEAT (Levels 2 & 3)	Korea Institute of Curriculum and Evaluation (KICE)	Korea
	GEPT	Taiwan Language Training & Testing Center (LTTC)	Taiwan
	TOEIC	ETS	USA
	TOEFL Junior	ETS	USA
	TOEFL iBT	ETS	USA

Note. * Automatized Speaking Scoring

3.2. 분석 절차 및 방법

본 연구의 분석 절차는 다음과 같다. 먼저 인터넷 검색 및 참고문헌을 통해 최근 시행되고 있는 자동채점 영어 말하기 시험들을 파악하였다. 이들 시험의 존재를 파악한 후에는 각 평가 사이트 및 관련 사이트를 방문하여 해당 평가에 대한 자료를 확인하였고, 이 가운데 본 연구의 분석 대상인 문항 유형, 평가 준거, 채점 방식, 결과 제시 방식에 관한 자료를 제공하고 있는 6종의 시험을 최종 분석에 포함하였다. 이들 시험은 홈페이지를 통해 시험에 대한 소개, 샘플 문항, 대비 가이드라인, 시험 결과 리포트 샘플 등 세부 정보를 상세하게 제공하고 있었다. 이후 비교 대상으로, 동일한 방식으로 국내외에서 개발된 수동채점 영어 말하기 시험 6종을 선정하고 자료를 수집하였다.

수집한 시험에 대한 분석은 다음과 같이 이루어졌다. 첫째, 6종의 자동채점 영어 말하기 시험에서 자동채점이 이루어지고 있는 영역을 문항별, 평가구인별로 분석하였다. 둘째, 12종 시험 전체를 대상으로 문항 유형, 평가 준거, 채점 방식, 결과 제시 방식을 분석하였다. 각 측면의 분석은 해당 시험 주관 기관에서 제공하는 정보들을 면밀히 살펴보면서 유형화하고 세부 분류 기준을 세운 후 이를 토대로 코딩하는 방식으로 이루어졌다. 가령, 문항 유형 분석의 경우 각 시험의 샘플 문항들을 하나씩 살펴보면 10개의 문항 유형을 도출하였고, 이를 분석틀로 삼아 다시 각 시험의 문항

을 코딩하였다. 각 측면별 상세한 분석 방법은 결과 섹션에 제시되어 있다. 분석 결과를 토대로 먼저 자동채점 기반 시험의 특징을 살펴보고, 이를 수동채점 시험의 특성과 비교하여 유사점과 차이점을 파악하였다.

4. 결과

4.1. 시험별 자동채점기술 적용 방식

6개의 자동채점 시험에 적용된 자동채점기술은 채점 방식이나 적용된 범위가 각각 달랐고, 주관사의 말하기 평가에 대한 이해와 기술 수준에 따라 문항 유형과 평가 준거도 다르게 설계된 것으로 분석되었다. 표 2는, 문항과 평가 준거에 대한 자동채점기술 적용 여부에 따라 1) 전체 자동채점(짙은 음영), 2) 자동채점과 수동채점의 혼합(얇은 음영), 3) 수동채점(음영 없음)으로 구분하여 표시한 것이다.

첫번째, TOEFL® Practice Online은 미국 ETS사가 주관하는 영어표준시험인 iBT TOEFL의 모의시험용 서비스이다. 본 시험과는 달리 SpeechRaterSM라는 자체 ASS 기술로 자동채점을 하고 전문인력이 각 영역의 채점 결과를 통합하여 0~30 척도로 점수를 부여하고 있어 혼합형으로 분류할 수 있다. 표 2에서 보는 것처럼 iBT TOEFL과 동일한 두 개 유형의 과업을 제시하고 음성인식 기술을 활용하여 응답에 대한 발음 분석과 텍스트 전사를 하여 평가를 하는데, 이 때 본 시험의 평가 요소보다 간소화된 발음, 유창성, 문법, 어휘의 일부 주요 요소만을 다루고 본 시험의 평가 준거인 ‘주제 전개’는 평가하지 않는다고 밝히고 있다.

두번째 Duolingo English Test는 미국 Duolingo사가 개발한 평가로, 말하기 시험은 총 3개 영역으로 구성되어 있다. 첫번째 영역인 ‘Read aloud’는 음성인식 기술로 처리 후 전 과정을 자동채점하며 음성(소리, 억양, 강세 등)과 유창성(속도)을 평가하는 것으로 소개하고 있다. 두번째 영역인 ‘Extended Speaking’의 3개 과업은 작문 영역(writing test)의 결과물과 합쳐 동일한 채점 방식을 적용하는데, 음성인식으로 전사한 텍스트를 통계분석하여 문법의 정확성, 문장의 복잡성, 어휘의 다양성 등을 평가하는 것으로 설명한다. 또한 전체 응답의 길이와 속도를 조합한 유창성, 음성적인 특성(강세, 억양)이 포함된다. 과업의 적절성은 관련 소개가 없어 자동채점 방식을 알기 어려우며, 내용 평가는 이루어지지 않는 것으로 보인다.

세번째는 EnglishScore Speaking Test로 영국 British Council이 개발한 모바일 기반 말하기 시험으로 인공지능기술 채점과 동료 채점을 결합한 독특한 채점 방식으로, 혼합형으로 분류할 수 있다. 자동채점으로는 형태의 정확성을 보고, 인간채점으로는 효과적인 소통을 판단하는 것을 원칙으로, ‘Read aloud’ 및 ‘Listen and Repeat’ 과업은 자동채점, ‘Guided Task’는 인간채점으로 이루어진다. 동료평가 방식으로 6명의 응시자에게 타인의 답변에 대해 과업 적절성과 내용 부분을 평가하게 한 후, 그 결과를 합쳐 전문인력이 점검하는 것으로 파악된다.

TABLE 2
Scope of ASS Technology Use in English Speaking Tests

Tests	Test Items	Criteria
TOEFL® Practice Online	1. Independent speaking 2. Integrated speaking (campus-related, academic)	Some features (forms only) of pronunciation, fluency, vocabulary, and grammar
Duolingo English Test	1. Read aloud 2. Extended speaking (picture description, independent speaking with a written prompt, independent speaking with an aural prompt) 3. Speaking sample (interview)	Spoken feature (acoustic and fluency feature) Grammatical accuracy, grammatical complexity, lexical sophistication, lexical diversity, task relevance, length, fluency & acoustic features Ungraded
EnglishScore Speaking Test	1. Read aloud 2. Listen and repeat 3. Guided task (open-ended task with visual/textual prompts)	Pronunciation (sound and stress), fluency Pronunciation, fluency Communication skills
Linguaskill General Speaking Test	1. Interview (questions about oneself) 2. Reading aloud 3. Presentation (on a given topic) 4. Presentation with visual information 5. Communication activity (opinion-focused questions related to a scenario)	Pronunciation, fluency, language resource Individual sound, stress, rhythm, and intonation Pronunciation, fluency, language resource, discourse management, task appropriateness
PTE Academic	1. Personal introduction 2. Read aloud 3. Repeat sentence 4. Describe image 5. Re-tell lecture 6. Answer short question (in a word)	Ungraded Content (form (word) correctness), oral fluency, pronunciation Correct/incorrect
Versant English Test	1. Reading (read aloud) 2. Repeat (listen & repeat a sentence) 3. Short questions (listen a short question and answer with a single word or a short phrase) 4. Sentence builds (listen to three short phrases and rearrange them) 5. Story retelling (listen brief stories and retell them in their own words) 6. Open Question	Pronunciation and reading fluency Mastery of sentence structure, pronunciation and fluency Vocabulary Mastery of sentence structure, pronunciation and fluency Mastery of sentence structure, pronunciation, vocabulary and fluency Ungraded

Note. Shading indicates the extent to which automatic scoring techniques have been used. Dark shading means entirely automatic scoring; light shading means a combination of automatic and manual scoring; no shading means entirely manual scoring.

네번째 Linguaskill General Speaking Test³는 Cambridge Assessment English사가 개발한 말하기 시험으로 자동채점과 수동채점의 결합인 혼합형인데, 앞선 EnglishScore와는 또 다른 채점 방식을 채택하고 있다. 먼저 자동채점의 경우 CASE라는 자체 음성인식 및 텍스트 통계분석 알고리즘 기술을 보유하여 형태적인 분석인 발음, 유창성, 언어자료(어휘, 문법) 등을 분석하고, 다양한 응시자의 답변 샘플로 사전 기계학습을 시켜 답변 유사성을 비교하여 채점한다. 이때 인공지능 기계학습이 여러가지 이유로 신뢰도가 낮을 수 있어 전문채점자를 훈련시켜 투입하는데, 이들의 역할은 음질 문제나 예측하기 어려운 응시자 답변을 모니터하고, 이를 다시 시스템에 입력하여 기계학습을 시키는 일이다. 기존 타 ASS 방식과 유사하나 과업 적절성 부분을 기계학습을 통해 자동채점을 시도한다는 점에서 주목할 만하다.

다섯 번째는 PTE Academic⁴으로 영국 Pearson사가 개발한 아카데미 용도의 시험으로 말하기와 쓰기 평가를 동시에 시행한다. PTE Academics는 전 영역 자동채점을 시행하며, Ordinate Technology라는 자체 ASS 기술을 보유하는데, 이는 분절적인 텍스트 통계분석이 아닌, 수십만 개의 샘플을 전문 인간채점자가 영역별로 채점한 것을 기계학습 방식으로 훈련시킨 것이다. 따라서 평가 준거가 문항별로 구분되어 있지 않고 모두 동일하게 적용된다. 기술적으로 가장 첨단 채점시스템을 개발한 것으로 평가할 수 있지만, 다른 한편으로는 총 5개의 문항(미채점 개방형 질문 제외)이 주어진 지문을 읽거나 따라하기, 그림 보고 묘사하기, 질문에 대한 답변하기 등으로 답변이 크게 제한된 도메인을 벗어나지 않아 응시자의 반응을 예측하기 용이한 점이 기술 정확도를 높였을 것으로 보인다. 평가 준거 역시 매우 단순화 되었는데 ‘Content’에서는 예상 답변의 단어가 빠짐이 없는지 채점하고, ‘Oral fluency’는 대화 속도와 중지 여부를 채점하며, ‘Pronunciation’은 음성인식을 통한 소리 특성 추출로 정확성을 평가하는 방식이다.

마지막으로, Versant English Test⁵는 전 자동채점 방식으로 소개하고 있으나 타 평가와 달리 이에 대한 자세한 설명이 없다. 다만 PTE Academic를 개발한 Pearson이 개발하였고 동일한 기술을 사용한다고 설명하고 있다. 자세히 살펴보면 PTE Academic와의 문항 유형이 한 개를 제외하고는 유사하지만, 평가 준거는 Versant English Test의 경우 각 문항별로 다르다. 예를 들어 어휘는 ‘Short question’과 ‘Story retelling’에서만 평가한다.

³ Cambridge English Language Assessment. (2023). *Linguaskill*. Retrieved on April 21, 2023, from <https://www.cambridgeenglish.org/exams-and-tests/linguaskill/>

⁴ Pearson. (n.d.). *PTE: English tests for study, work, and visas*. Retrieved on April 21, 2023, from <https://www.pearsonpte.com/>

⁵ Pearson. (2023). *English speaking test: Versant English test*. Retrieved on April 21, 2023, from <https://www.pearson.com/english/versant/our-tests/english-speaking-test.html>

4.2. 평가문항 유형

자동채점 영어 말하기 시험들이 영어 말하기 능력을 평가하기 위해 어떤 문항들을 사용하고 있는지, 또 이러한 문항 유형이 기존의 수동채점 말하기 시험과 차이가 있는지 살펴보기 위하여 각 시험의 문항 유형을 분석하였다. 최종 점수에 포함되지 않는 문항들(예: Duolingo English Test의 ‘Speaking Sample(interview)’은 분석에서 제외하였다.

대부분의 시험은 문항을 유형화하여 명칭과 함께 제시하고 있었다. 하지만 그 방식이 상이하여, 동일 문항 유형임에도 다른 명칭을 사용하거나, 반대로 명칭은 유사하지만 실제로는 다른 문항 유형을 지칭하는 경우도 있었다. 이에 본 연구에서는 각 시험에서 제시하는 문항 유형의 명칭을 분석 기준으로 삼는 대신, 실제 예시 문항을 하나씩 확인하면서 유형을 재분류하였다. 그 결과 총 10개의 문항 유형이 추출되었다. 표 3은 분석 대상 말하기 시험에서 발견된 문항 유형 10가지를, 측정하고자 하는 말하기 능력의 수준에 따라 배열한 것이다. 오른쪽으로 갈수록 길고 복잡한 발화를 요구하는 문항 유형이다.

처음 세 문항 유형은 소리 인식, 소리와 문자의 연결, 문법 능력 등과 같은 기초 수준의 말하기 능력을 측정한다. 문장을 듣고 그대로 반복하는 ‘듣고 따라하기(Listen and Repeat)’, 글로 주어진 문장을 음성으로 발화하는 ‘소리내어 읽기(Read Aloud)’, 무작위로 주어진 구를 문장으로 재구성하여 말하는 ‘구 배열하기(Order Phrases)’유형이 여기에 속한다.

다음의 세 문항 유형은 질문을 듣고 답하는 대화 형식을 취하고 있다. 가장 낮은 수준에서는 단어나 구 수준의 정답이 있는 ‘질문 듣고 정답 말하기(Listen and answer to closed questions)’ 문항이 있었다. ‘열린 질문 듣고 답하기(Listen and answer to open questions)’는 일상 대화와 유사한 형태로, 자신이나 일상 생활에 관한 단일 혹은 일련의 질문에 답하는 문항이다. ‘주어진 정보 기반 답하기(Answer using information provided)’는 음성이나 텍스트로 주어진 질문에 대해 그림, 도표, 포스터 등의 정보를 활용하여 답하는 유형으로 발화의 내용을 일정 수준 통제할 수 있다. 이 문항 유형의 경우 시험에 따라 도출하는 발화 유형이 조금 다르다. 가령, TOEIC의 ‘Respond to questions using information provided’ 유형의 경우, 회사의 월례 직원 세미나 일정표를 보면서 질문에 답하게 된다. 질문이 음성으로 제공되고 일련의 질문이 이어진다는 점에서 업무 상황에서의 대화와 상당히 유사하다. 하지만 Linguaskill General Speaking Test의 경우에는 교통수단에 대한 정보표와 함께 친구에서 교통 수단과 그 이유를 추천하라는 과제가 텍스트로 제시되어 좀 더 독백형에 가깝다.

다음의 세 문항은 비교적 긴 독백 답화 수준의 발화를 유도하는 문항이다. ‘그림 묘사하기(Describe pictures/images)’는 화면에 주어진 사진, 도표 등을 묘사하는 문항이고, ‘그림 연결하여 이야기하기(Tell a story with pictures)’는 주어진 여러 장의 그림이나 사진을 연결하여 하나의 이야기로 만들어내는 서술(narration) 말하기 문항이다. ‘확장형 말하기(Extended talk)’는 주제에 대한 생각이나 의견을 독백형으로 말하는

TABLE 3
Test Item Types in Automated vs. Non-Automated Speaking Scoring Tests

Type	Tests	Sounds & Forms				Conversation			Discourse		Integration
		Listen & Repeat	Read Aloud	Order Phrases	Listen & Answer to Closed Questions	Answer to Open Questions	Listen & Answer Using Information Provided	Describe Pictures/Images	Tell a Story with Pictures	Extended Talk	
Non-ASS	TOEFL® Practice Online										V
	Duolingo English Test		V					V		V	
	EnglishScore Speaking	V	V			V					
	Linguaskill General		V			V				V	
	PIU Academic	V	V		V			V			V
Non-ASS	Versant English Test	V	V	V	V						V
	Frequency	3	5	1	2	1	2	2	0	3	3
	TLPS		V			V		V	V	V	
	NEAT Level 2					V		V	V	V	
	Level 3					V	V		V	V	
	GEPT Elementary	V	V			V					
	Intermediate		V			V		V			
	High-Intermediate					V		V		V	
	TOEIC		V			V	V	V		V	
	TOEFL Junior		V						V		V
Non-ASS	TOEFL iBT										V
	Frequency	1	4(5) ^{**}	0	0	4(7)	2	4(5)	3(4)	5(6)	2

Note. * Automated Speaking Scoring; ** number of tests (number of tests counting sub-level tests of NEAT and GEPT)

문항으로, 좋아하는 대상이나 장소 추천하기, 문제 상황에서 대처 방안 말하기, 특정 주제에 대한 견해 말하기, 주장에 대한 찬반 의견 말하기 등이 여기에 속한다.

마지막으로 ‘통합형 말하기(Integrated speaking)’는 듣거나 읽은 후 그 내용을 기반으로 말하는 문항이다. 가령, 강의나 이야기를 듣거나 읽고 그 내용을 요약하거나 다른 사람에게 전달하는 식이다.

이상의 10가지 문항 유형 가운데, 자동채점 말하기 시험은 ‘그림 연결하여 이야기하기’를 제외한 9가지 유형을 사용하고 있었다. 가장 빈번하게 사용된 유형은 ‘소리내어 읽기’로 TOEFL® Practice Online을 제외한 나머지 시험에서 모두 채택하고 있었다. 그 다음으로 ‘듣고 따라하기’, ‘확장형 말하기’, ‘통합형 말하기’가 각 3종의 시험에서 사용되었다. 반면, ‘구 배열하기’와 ‘열린 질문 답하기’는 오직 1종의 시험에서만 발견되었다.

이와 같은 자동채점 시험의 문항 유형 분포가 수동채점 시험과 차이를 보이는지 확인하기 위하여 두 그룹 간 문항 유형 빈도를 비교해 보았다. 우선, 기초 수준의 영어 말하기 능력을 측정하는 문항 유형은 3종 모두 자동채점 시험에서 훨씬 더 빈번하게 사용되었다. ‘듣고 따라하기’의 경우 수동채점 시험에서는 오직 1종(GEPT 초급수준)에서만 사용되었으나, 자동채점 시험에서는 3종에서 사용되었다. 자동채점, 수동채점 시험 모두에서 가장 빈도가 높은 ‘소리내어 읽기’의 경우에도 수동채점 시험에서는 4종, 자동채점 시험에서는 5종에서 발견되었다. ‘구 배열하기’는 오직 자동채점 시험 1종(Versant English Test)에서만 사용되었다.

대화 능력과 관련된 문항 가운데, ‘질문 듣고 정답 말하기’는 오직 자동채점 시험 2종(PTE Academic, Versant English Test)에서만 발견되었다. Versant English Test의 예를 들면, “Would you get water from a bottle or a newspaper?”라는 질문을 듣고 “a bottle”이라고 답하는 문항으로, 기본적으로는 어휘 능력을 측정하기 위한 문항이다. 이는 수동채점 시험에서는 발견되지 않는 문항 유형이었다. 이와 반대로, 일상생활 대화와 가장 유사한 형태인 ‘열린 질문 듣고 답하기’는 수동채점 시험에서는 4종에서 빈번하게 사용되었으나 자동채점 시험에서는 2종에 그쳐 차이를 보였다.

독백 담화 수준의 문항들 역시 수동채점 시험에 비해 자동채점 시험의 사용 빈도가 낮았다. 특히 ‘그림 연결하여 이야기하기’는 수동채점 시험에서는 3종에서 사용하고 있었으나 자동채점 시험에서는 전혀 사용하고 있지 않았다. ‘그림 묘사하기’, ‘확장형 말하기’의 경우에는 일부 자동채점 시험에서 채택하고 있었으나 수동채점 시험보다는 채택 빈도가 낮았다.

4.3. 평가 준거

12종의 영어 말하기 시험에서 사용하고 있는 평가 준거를 분석한 결과 시험마다 준거의 범위와 용어가 상이했다. 예를 들어, 발음 영역에서 분절 발음과 초분절 발음을 분리하여 각각의 준거로 제시한 경우도 있고, 둘을 통합하여 ‘발음’이라는 상위개념에 포함시킨 경우도 있었다. TOEFL iBT의 경우에는 ‘전달(delivery)’이라는 준거 아래에 발음 요소를 포함시켰다. 용어와 관련해서도 ‘유창성(fluency)’ 대신 ‘전달(delivery)’을, ‘어휘(vocabulary)’ 대신 ‘표현(expression)’을 사용한 말하기 시험도 있었다. 본 연구에서는 총 8개의 주요 평가 준거를 도출하고 이를 기준으로 영어 말하기 시험들을 재분석하였고, 이 과정에서 특정 시험에서 단일 준거로 제시된 평가 내용이 복수의 준거로 분석되기도 하였다.

표 4는 자동채점 시험과 수동채점 시험에서 8개의 준거가 사용된 분포를 정리한 것이다. 6종의 모든 자동채점 시험에서 사용된 준거는 ‘분절 발음’과 ‘유창성’이었고, ‘초분절 발음’과 ‘어휘’도 5종의 자동채점 시험에서 사용되었다. 이들 네 준거 중 ‘초분절 발음’을 제외한 다른 준거들은 수동채점 시험에서도 5종 이상에서 사용하고 있었다.

TABLE 4
Assessment Criteria in Automated vs. Non-Automated Speaking Scoring Tests

	Tests	Seg-Pro*	Sup-Pro**	Fluency	Grammar	Vocabulary	Organization	Content	Task Completion
ASS	TOEFL® Practice Online	V	V	V	V	V	V	V	
	Duolingo English Test	V	V	V	V	V			
	EnglishScore Speaking	V	V	V					(V)***
	Linguaskill General	V	V	V	V	V	V		(V)
	PTE Academic	V	V	V		V		V	
	Versant English Test	V		V	V	V			
	Frequency	6	5	6	4	5	2	2	(2)
Non-ASS	TEPS	V	V	V	V	V	V		
	NEAT	V		V	V	V	V		V
	GEPT	V	V	V	V	V		V	
	TOEIC	V	V	V	V	V	V		
	TOEFL Junior	V	V	V	V	V		V	
	TOEFL iBT			V	V	V	V	V	V
	Frequency	5	4	6	6	6	4	3	2

Notes. *Seg_Pro stands for segmental pronunciation; **Sup_Pro stands for suprasegmental pronunciation; ***(V) means that the assessment of the criterion is not automated.

반면, 자동채점 시험은 ‘구성’, ‘내용’, ‘과업 완성’에서는 자동채점을 실시하는 데에 제한을 보였다. 다만 일부 자동채점 시험에서는 ‘구성’과 ‘내용’을 평가하고 있었는데, 이러한 차이는 현재 자동채점 기술이 발전 중이며 평가 기관이 보유한 기술력

과 데이터에 따라 평가의 타당도와 신뢰도가 상이할 수 있음을 의미한다. 더불어, ‘과업 완성’ 준거는 2종의 자동채점 시험에서도 수동채점 방식을 사용하고 있었는데 이는 자동채점 시험에서 앞으로 ‘과업 완성’의 채점이 주요 해결 과제임을 보여 준다. 마지막으로, ‘문법’은 모든 수동채점 시험에서 준거로 채택되었으나 2종의 자동채점 시험에서는 준거로 채택되지 않았다. 이 역시 앞으로 채점 알고리즘과 기술의 발달이 필요한 지점으로 볼 수 있다.

영어 말하기 시험에서 사용되는 8개의 준거가 어떠한 개념으로 구인화되고 있는지 살펴보기 위하여, 준거의 구체적인 내용을 분석하였다. 예를 들어, TEPS는 유창성(fluency)을 “자모음, 강세, 억양 등 발음의 이해도(comprehensibility)는 어느 정도인가?”, “말의 흐름이 머뭇거림 없이 얼마나 자연스럽게 진행되는가?”라는 두 가지 측면에서 규정하고 있는데 이는 이해가능도(comprehensibility)와 흐름(flow)이라는 두 개의 개념이 구인 설정에 사용된 것으로 보인다. 준거 구인 분석에서 유의할 점은 준거 구인화에 사용된 다양한 표현들이 용어의 차이가 있을지라도 개념적으로는 유사한 경우가 많다는 것이다. 예를 들어, ‘구성’ 준거를 구인화할 때, 응집력(cohesion)이 주요 개념으로 사용되었는데, 이와 유사한 개념으로 connectives, relationships, cohesive devices 등의 용어들이 여러 시험에서 사용되고 있었다. 본 연구는 표현의 차이에 구애 받지 않고 개념적 일치도 및 유사성에 주목하여 준거의 개념을 분석하였다.

표 5는 각 준거의 관련 개념들을 말하기 평가에 사용된 빈도 순으로 제시한 것이다. 각 준거별로 0~3개의 핵심 개념이 자동채점 시험에서 공통적인 구인 요소로 사용되고 있었다. 그 외 개념들은 자동채점 시험의 목적과 과업 유형, 채점 방식 등에 따라 선택적으로 사용되었다.

준거별 주요 개념을 살펴보면 다음과 같다. 우선 ‘분절 발음’의 핵심 개념인 이해가능성(intelligibility)은 청자가 쉽고 명료하게 이해할 수 있는지에 초점을 두고 있다. 이해가능성이 요구하는 발음의 명료성은 자동으로 평가하기 용이한 내용이므로(예: Read Aloud 문항에서 제시된 문장과 수험자 발음을 전사한 문장을 비교) 모든 자동채점 시험이 준거 구인에 이해가능성의 개념을 활용하고 있었다. 비슷한 견지에서, 분절 발음 정확성(segmental accuracy)이 3종의 자동채점 시험에서 활용되고 있었다.

TABLE 5
Major Concepts of the Assessment Criteria
in Automated vs. Non-Automated Speaking Scoring Tests

Criterion	Concept	ASS	Non-ASS
Segmental Pronunciation	*Intelligibility (listener effort, clear)	6	4
	*Segmental accuracy (correct)	3	1
	Accentedness (language influence)	2	1
	Natural (continuous, assimilation and deletions)	1	1
	Lapse	0	1
Supra-segmental Pronunciation	*Stress (word, sentence)	5	2
	*Rhythm	4	0
	*Intonation (rising and falling)	4	4
	Shades of meaning (emphasis, pauses, differentiate)	1	1
Fluency	*Flow (hesitation, fluid, breakdown)	6	5
	Speed (speech rate)	2	4
	Repair	2	0
	Intelligibility (comprehensibility, listener effort, clear)	1	4
	Insignificant errors	1	3
	Sustainment (continuity)	1	1
	Chunking (grouping)	1	1
	Communication	0	2
Grammar	*Range (for coherent ideas, effective)	3	5
	Structural accuracy	2	3
	Morphological accuracy (Insignificant errors)	1	4
	Insignificant errors	1	2
Vocabulary	*Range/Variety	4	3
	Appropriateness (adequacy, content)	2	5
	Sophistication	2	0
	Insignificant errors	1	2
	Accuracy	0	4
	Effective	0	2
	Style (rhetoric)	0	1
Organization	Cohesion (connectives, relationships, progression)	2	4
	Argumentation (logic, rhetorical structure)	2	3
	Coherence	1	3
	Sustained (sufficient, details)	0	1
Content	Appropriate (relevant)	1	2
	Key information	1	1
	Idea connection	1	1
	Details	1	0
Task	Performance (fulfill, function)	0	2
Completion	Relevance	0	1

* Key concepts for the criteria of ASS, with the frequency of 3 and above.

두 번째 준거인 ‘초분절 발음’의 경우, 강세, 리듬, 억양이 자동채점에서 핵심 개념으로 사용되고 있었는데, 특히, 강세와 리듬에서 자동채점과 수동채점 시험 간 차이가 두드러졌다. 두 개념이 자동채점 시험에서는 각각 5종과 4종에 사용된 반면, 수동채점 시험에서는 2종과 0종에 사용되었다. 이는 강세 음절의 위치와 간격이 자동 측정과 분석에 용이하며, 강세박자 언어인 영어(강세의 일정한 간격으로 리듬을 형성)에서는 강세 정보가 리듬 분석 및 평가에 활용될 수 있기 때문으로 보인다.

세 번째 준거인 ‘유창성’의 경우, 흐름(flow), 즉 끊기지 않게 발화할 수 있는 능력이 모든 자동채점 시험에서 구인에 활용되고 있었는데, 이는 말소리의 연속과 단절이 자동 측정 및 평가에 용이하기 때문인 것으로 보인다. 반면 인간채점자의 판단을 요구하는 이해가능도(intelligibility), 소소한 오류(insignificant errors), 의사소통(communication) 등은 자동채점 시험에서는 활용도가 낮았다.

네 번째 준거인 ‘문법’은 자동채점 시험과 수동채점 시험 모두 범위(range), 즉 다양성에 더 큰 비중을 두는 경향을 보인다. 다양한 형태로 풍부하게 의미를 표현하는 것이 말하기 능력에서 중요한 개념으로 다루어 지는 셈이다. 한편, 형태론적 정확성은 4종의 수동채점 시험에서 활용된 것과 대조적으로 1종의 자동채점 시험에서만 사용되고 있었는데, 이는 현재 기술로서는 말하기 데이터에서 형태론적 오류를 자동으로 판별하는 것이 어렵기 때문인 것으로 보이며, 추후 채점 알고리즘과 기술의 발달로 개선될 수 있을 것이다.

다섯 번째로, ‘어휘’는 다양성(range)이 자동채점의 핵심개념으로 사용되었다. 반면, 수동채점 시험은 적절성과 정확성에 주목하는 경향을 보였다. 어휘 다양성은 빈도 분석과 참조 어휘 난이도별 목록과의 비교를 통해 비교적 수월하게 자동 분석될 수 있는 반면, 어휘 적절성 및 정확성은 의미와 관련된 개념으로서 현재의 기술로는 분석하기 쉽지 않은 것으로 보인다.

그 외의 준거인 ‘구성’, ‘내용’, ‘과업 완수’는 자동평가 시험에서 활용도가 낮았으나, 해당 구인과 관련하여 일부 자동 측정 및 평가가 용이한 개념(예: 연결사, 대명사)이 일부 자동채점 시험에서 활용되고 있었다.

4.4. 채점 방식

채점 방식에 대한 예비 분석 결과, 평가 준거 적용 방식과 점수 산출 방식에 있어서 몇 가지 유형이 존재함을 확인할 수 있었다. 우선, 평가 준거를 적용하는 데에 있어서, 모든 문항에 대해 동일한 준거를 적용하는 방식과 문항에 따라 상이한 준거를 적용하는 방식이 있었다. 점수와 관련해서는 평가 준거별로 점수를 산출하는 방식과 문항 유형별로 점수를 산출하는 방식이 있었다. 이 두 가지 측면에서 각 시험을 분석한 결과 표 6과 같았다.

TABLE 6
Applying Assessment Criteria and Calculating Scores

Score Calculation Unit	Evaluation Criteria	Assessment Criteria Application Across Item Types	
		Same	Different
		TOEFL® Practice Online, PTE Academic*, NEAT, TOEFL iBT*	<i>EnglishScore Speaking Test, Versant English Test, TEPS, GEPT Elementary</i> <i>Linguaskill General Speaking Test, Duolingo English Test, TOEIC, TOEFL Junior</i>

Note. Automated speaking scoring tests are denoted by italics; Tests marked with an asterisk (*), basically utilizing the same evaluation criteria, allow minor variations in the details based on the item types.

자동채점 시험 가운데 4종(67%)은 문항별로 상이한 평가 준거를 적용하고 있었다. 가령, Versant English Test의 ‘Part A. 읽기 문항’의 평가 준거는 ‘유창성’과 ‘발음’인 반면에 ‘Part C. Short answer question’는 ‘어휘’로 상이하였다. 반면, TOEFL® Practice Online과 PTE Academic은 모든 문항에 대해 동일한 준거를 적용하고 있었다. 수동채점 시험의 경우에도 문항별로 다른 평가 준거를 사용하는 방식이 4종(67%)으로, 동일한 준거를 적용하고 있는 시험보다 많았다. 따라서 자동채점 유무에 따라 평가 준거 적용 방식이 다르다고 보기는 어렵다.

다음으로 점수 부여 방식을 살펴보면, 자동채점 시험 4종(67%)은 문항별로 각각 점수를 부여한 후 합산하는 총점 산출 방식을 채택하고 있었다. 예를 들어, Linguaskill General Speaking Test는 다섯 개의 문항 점수에 동일한 비중을 부여하여 합산한 결과를 총점으로 사용하고 있다(Xu et al., 2021). 그러나, EnglishScore Speaking Test는 발음과 유창성이라는 평가 준거에 점수를 부여한 후 합산하는 방식을 채택하고 있었다. 수동채점 시험의 경우에도, 문항별 점수 합산 방식을 사용하는 시험이 4종(67%)으로, 준거별 점수 합산 방식을 채택한 시험보다 많았다.

EnglishScore Speaking Test를 제외한 나머지 자동채점 시험들은 모두 채점 척도(점수 범위) 정보를 제공하고 있었다. TOEFL® Practice Online, PTE Academic, Linguaskill General Speaking Test의 경우 모두 최저점은 0, 최고점은 4~6점이었다. 그러나 Duolingo English Test와 Versant English Test는 최소 점수가 공통적으로 10점이고 최고 점수는 각각 160점과 90점으로 다른 3종의 시험에 비해 높은 점수대를 사용하고 있었다.

점수 범위를 공개하고 있는 자동채점 시험들은 거의 대부분 문항에 상관없이 동일한 채점 척도를 사용하고 있는 것으로 드러났다. 예를 들어, TOEFL® Practice Online은 0~4점, Linguaskill General Speaking Test는 0~6점 척도를 모든 문항에 동일하게 적용하고 있었다. 다만 PTE Academic은 문항에 따라 0~3 또는 0~5점의 척도를 달리 사용하고 있었다. 수동채점 시험들 역시 TOEIC을 제외한 나머지 시험들이 동일한

척도를 활용하고 있었다.

4.5. 평가 결과 제공 방식

각 시험별 평가 결과 리포트 양식을 수집하여 분석한 결과 표7과 같았다.

TABLE 7
Score Reporting in Automated vs. Non-Automated Speaking Scoring Tests

	Tests	Numeric Scores	Levels	Criterion-Specific Results
ASS	TOEFL® Practice Online	V	V	
	Duolingo English Test	V		
	EnglishScore Speaking Test	V	V(CEFR)	V(Communication, Fluency, Pronunciation)
	Linguaskill General Speaking Test	V	V(CEFR)	
	PTE Academic	V		V(Grammar, Oral Fluency, Pronunciation, Vocabulary)
	Versant English Test	V	V(CEFR)	V(Sentence Mastery, Vocabulary, Fluency, Pronunciation)
Non-ASS	TEPS	V	V	
	NEAT(Levels 2 & 3)		V	
	GEPT(Elementary)	V	V(CEFR)	
	TOEIC	V	V	
	TOEFL Junior	V	V(CEFR)	
	TOEFL iBT	V	V	

자동채점 시험과 수동채점 시험 간 비교에서 가장 두드러진 차이는, 수동채점 시험의 경우 합산 점수와 통합적인 등급만을 제시하는데 그치는 반면, 자동채점 시험의 경우 절반에 해당하는 3개의 시험이 평가 준거별 세부점수를 제공하고 있다는 점이다. 예를 들어, Pearson사의 시험인 PTE Academic과 Versant English Test는 모두 동일한 4개 영역(문법, 어휘, 유창성, 발음)에 대한 세부 채점 결과를 그래프와 함께 자세한 설명을 곁들여 제시하고 있다. 이어서 각 영역별 점수가 무엇을 의미하는지에 대한 상세한 해설과 함께 향상 방안(“Tips to improve”)을 제시해 준다. EnglishScore Speaking Test 역시, 총점, CEFR 등급과 함께 3개 영역(커뮤니케이션, 유창성, 발음)에 대한 세부 점수를 공개하고, 상당히 구체적인 피드백을 제시한다.

한편 통상적으로 말하기 시험의 결과 리포트 방식은 결과를 점수로 제시하느냐 등급으로 제시하느냐 여부로 유형화 할 수 있는데 여기에서는 자동채점 시험과 수동채점 시험 간의 별다른 차이가 나타나지 않았다.

5. 논의 및 결론

본 연구는 국내 교육용 자동채점 영어 말하기 평가 가능성을 탐색하기 위하여, 현재 자동채점 기술을 사용하고 있는 해외 영어 말하기 시험 6종을 대상으로 자동채점 기술의 활용 정도, 문항 유형, 평가 준거, 채점 방식, 결과 제시 방식을 분석하였다. 또한 이러한 특성을 수동채점 시험과 비교하여 차이점을 분석하였다.

먼저 6개 자동채점 시험에서 자동채점기술의 활용 정도를 살펴본 결과, 예상보다 더 다양하고 많은 영역의 채점자동화가 이루어져 있었으며, 평가 기관마다 그 방식은 다양하였다. 공통적으로 적용된 기술은 음성 특성 분석 평가로 개별 발음과 강세, 억양, 휴지 등을 파악하고 그 외는 텍스트를 자동으로 전사한 후 통계분석을 통해 어휘, 문법, 유창성을 평가하는 방식이다. 그러나 Linguaskill General Speaking Test나 PTE Academic의 경우는 한 단계 더 나아가 방대한 실제 답안 채점 데이터로 기계화된 시스템에서 통계분석 없이 전문채점자의 결과와 거의 일치하는 수준의 전자동채점을 일부 (혹은 전체) 적용하고 있었다. 그러나 조사한 모든 자동채점 시험에서 내용 평가나 과업 적절성에 대한 평가는 여전히 제외하거나 회피 전략(동료평가, 개방형 문항 유형 제외 등)을 사용하고 있었다. 또한 훈련된 전문인력을 통한 혼합채점 시에도 내용 평가보다는 시스템 고도화나 등급 판정, 오류 제거 등의 기술적인 모니터링으로 운영되고 있어, 고비용에 즉각적인 결과 도출이 어려운 수동채점 방식을 채택하지 않는 방향으로 변화하고 있음을 알 수 있다.

다음으로 자동채점 시험에서 채택하고 있는 문항 유형을 살펴본 결과, 대체로 자동채점 시험은 수동채점 시험에 비해 기초 수준의 말하기 능력을 측정하는 문항은 더 많이 사용하는 반면, 대화, 담화 수준의 문항은 덜 사용하는 양상을 보이고 있었다. 이처럼 자동채점 시험에서 기초 수준의 평가 문항을 선호하는 이유는 자동채점 기술 수준과 관련된 것으로 보인다. 주어진 입력의 재생 및 재구성, 혹은 정답을 요구하는 문항들은 응답 내용이 고정되어 있어 형태 위주의 자동채점이 용이하다. 또한, ‘듣고 따라하기’나 ‘소리내어 읽기’ 문항의 경우에는 최근의 음성인식 기술을 활용하면 인간채점보다 더 정밀하고 객관적인 채점 결과를 제공할 수 있어 자동채점의 이점이 높다. 반면, 열린 대화나 독백 담화 수준의 문항은 내용면에서 가능한 응답의 범위가 넓고, ‘그림 묘사하기’나 ‘그림 연결하여 이야기하기’와 같이 시각적인 자료를 통해 내용을 어느 정도 통제하는 경우에도 언어적 측면에서 여전히 다양한 응답이 가능하기 때문에 이를 사전에 모두 예측하여 자동채점하기가 어렵다. 이러한 결과는 자동채점 시험의 편리성에도 불구하고 대화 능력과 담화 능력을 포함한 실제적인 말하기 능력을 측정하는 데에 있어서 여전히 보완이 필요함을 보여준다.

자동채점 시험의 평가 준거와 관련하여, 모든 자동채점 시험이 ‘유창성’과 ‘초분절 발음’을 평가 준거로 채택하고 있었고, ‘분절 발음’, ‘어휘’도 5종에서 사용하고 있었다. 반면, ‘구성’, ‘내용’, ‘과업 완수’ 준거의 경우 수동채점 시험보다 사용 빈도가 낮았는데 이들 평가 준거는 문장 단위의 형태적 수준을 넘어서 매크로 수준에서 주관적 판단을 요하는 만큼 자동채점에 제약이 있는 것으로 보인다. 평가 준거에 사용된

구체적인 구인들은 자동채점 시험과 수동채점 시험 간에 상당한 유사점이 있었다. 예를 들어, 분절 발음에서는 ‘이해가능성(intelligibility)’이, 유창성에서는 ‘흐름(flow)’이, 문법에서는 ‘다양성(range)’이 두 유형의 시험에서 공통적으로 준거 구인화에 활용되고 있었다. 자동채점 시험과 수동채점 시험이 평가를 위해 수집 및 측정하는 정보는 일부 다를 수 있으나, 영어 말하기 능력의 준거를 설정하고 그 준거를 정의하는 데에 있어서는 일반적으로 통용되는 기준을 공통적으로 따르고 있는 것으로 보인다. 다만, ‘초분절 발음’의 경우에는 자동채점 시험이 수동채점 시험보다 강세와 리듬을 사용하는 경향이 강한 반면, ‘어휘’에서의 적절성이나 정확성 등은 수동채점 시험에 비해 덜 사용하고 있었다. 따라서 자동채점 시험에서 평가 준거 및 구인 개념의 선택은 통상적인 말하기 능력 구인을 기반으로 하면서도 자동채점기술의 한계를 반영하여 선택, 운영되고 있는 것으로 보인다.

자동채점 시험의 채점 방식은 수동채점 시험과 큰 차이가 있다고 보기 어려웠다. 대부분의 자동채점 시험과 수동채점 시험이 문항별로 상이한 준거들을 적용하고 있었고, 문항별로 점수를 도출하여 최종 점수를 산출하고 있었다. 이는, 문항 유형에 따라 보다 용이하게 관찰, 측정할 수 있는 언어 특질이 달라지기 때문에 평가하고자 하는 구인이 무엇인지에 따라 문항을 결정해야 한다는 제 2언어 평가 전문가들의 주장을 반영한 것이라 볼 수 있다(Carr, 2011; Ockey & Li, 2015). 예를 들어, ‘그림 묘사하기’보다 ‘소리내어 읽기’가 수험자의 발음을 평가하기에 더 적절한 문항이라 할 수 있다. 최종 점수 합산의 단위로 준거별 점수가 아니라 문항별 점수를 더 빈번히 사용하는 경향은, 평가하고자 하는 구인이 수험자가 지닌 추상적인 말하기 능력(또는 세부 구인들)이라기보다는 다양한 의사소통 상황에서 자신의 영어 말하기 능력을 사용하여 주어진 과업을 얼마나 성공적으로 성취할 수 있는가에 더 초점을 두고 있음을 보여준다(Bachman, 2007).

마지막으로 자동채점 시험의 결과지에는 수동채점 시험에서처럼 총체적인 점수가 제시되고 추가적으로 CEFR과 같은 표준 등급이나 해당 구간의 말하기 능력 수준에 대한 설명을 제공하는 경우도 있었다. 한 가지 주목할 점은 일부 자동채점 시험의 경우 평가 준거별 상세 정보를 제공하고 있다는 것이다. 이는 자동채점을 통해 음성 데이터를 텍스트로 전환하여 언어분석값을 도출함으로써 세부 항목별 점수나 피드백 제공이 가능해졌기 때문이다. 이러한 자동채점 알고리즘의 도입은 말하기 평가의 객관성 및 신뢰성 확보는 물론 교육적 측면에서 추후 학습의 방향을 구체적으로 제시해줌으로써 학교 말하기 수업에서 유용하게 활용될 수 있을 것으로 기대된다.

본 연구는 우리나라 학교 영어 말하기 평가에 자동채점기술의 적용 가능성을 탐색하기 위하여 해외의 자동채점 기반 영어 말하기 시험을 상세히 분석하였다. 다만, 본 연구는 평가 기관이 자체적으로 제공하는 정보만을 분석하였다는 점에서 한계가 있다. 또한 자동채점기술이 하루가 다르게 향상되고 고도화되고 있는 만큼 본 연구에서 지적한 한계들은 내재적이고 고착화된 것이라기보다 기술의 진보와 함께 개선될 가능성이 높다. 이러한 한계에도 불구하고 본 연구는 자동채점 기반 영어 말하기

시험의 현주소를 확인함으로써 추후 자동채점 말하기 시험의 효율적인 활용 및 개발에 의미있는 참고자료가 될 것으로 기대한다.

Applicable levels: Primary, secondary, tertiary

REFERENCES

- Bachman, L. F. (2007). What is the construct? The dialectic of abilities and contexts in defining constructs in language assessment. In J. Fox, M. Wesche, D. Bayliss, L. Cheng, C. Turner, & C. Doe (Eds.), *Language testing reconsidered* (pp. 41-71). Ottawa, Canada: University of Ottawa Press.
- Carr, N. (2011). *Designing and analyzing language tests*. Oxford, England: Oxford University Press.
- Chae, J. K. (2015). A time lag approach to causes of the failure in the policy of National English Ability Test. *The Journal of Politics of Education*, 22(3), 1-31.
- Chen, L., Zechner, K., Yoon, S.-Y., Evanini, K., Wang, X., Loukina, A., ... Gyawali, B. (2018). *Automated scoring of nonnative speech using the SpeechRater v. 5.0 Engine*. (ETS Research Report No. RR-18-10). Princeton, NJ: Educational Testing Service. Retrieved on May 14, 2023, from <https://onlinelibrary.wiley.com/doi/full/10.1002/ets2.12198>
- Choi, W. (2020). Process-oriented speaking assessment of primary English using AI chatbots: Possibilities and limitations. *Primary English Education*, 26(1), 131-152.
- Choi, W. (2021). An application of an AI chatbot automatic pronunciation scoring system to elementary school students. *English Language Assessment*, 16(2), 167-185.
- Chu, S. Y., & Min, D. G. (2019). A study of using task-based artificial intelligence (AI) chatbot for further interaction in English and the analysis of students' production. *Primary English Education*, 25(2), 27-52.
- Educational Testing Service (ETS). (2023). *TOEFL iBT® Test*. Retrieved on April 21, 2023, from <https://www.ets.org/toefl/test-takers/ibt/about/content/speaking.html>
- Khabbazzashi, N., Xu, J., & Galaczi, E. D. (2021). Opening the black box: Exploring automated speaking evaluation. In B. Lantaigne, C. Coombe, & J. D. Brown (Eds.), *Challenges in language testing around the world* (pp. 333-344). Singapore: Springer.
- Kim, N. (2016). Effects of voice chat on EFL learners' speaking ability according to proficiency levels. *Multimedia-Assisted Language Learning*, 19(4), 63-88.

- Kim, S., & Lee, S. (2020). An analysis of a primary school English speaking lesson using an AI-powered conversational English learning application: A focus on student engagement as well as learners' and teachers' perceptions of the class. *Primary English Education*, 26(3), 177-202.
- Lee, M., Yun, J., Kim, S., & Joo, H. (2015). *School English speaking performance assessment support system utilizing ICT*. (ORM 2015-50-21). Seoul: Korean Institute for Curriculum and Evaluation.
- Ockey, G., & Li, Z. (2015). New and not so new methods for assessing oral communication. *Language Value*, 7(10), 1-12.
- 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.
- Wang, X., & Evanini, K. (2019). Features measuring content and discourse coherence. In K. Zechner & K. Evanini (Eds.), *Automated speaking assessment: Using language technologies to score spontaneous speech* (pp. 138-156). New York: Routledge.
- Wang, Y., Gales, M. J. F., Knill, K. M., Kyriakopoulos, K., Malinin, A., van Dalen, R. C., & Rashid, M. (2018). Towards automatic assessment of spontaneous spoken English. *Speech Communication*, 104, 47-56.
- Xu, J., Brenchley, M., Jones, E., Pinnington, A., Benjamin, T., Knill, K., ... & Geranpayeh, A. (2020). *Linguaskill building a validity argument for the speaking test*. 2020. Cambridge, UK: Cambridge Assessment English. Retrieved on May 14, 2023, from <https://www.cambridgeenglish.org/Images/589637-linguaskill-building-a-validity-argument-for-the-speaking-test.pdf>
- 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.