

논술형 평가에서 자동 채점 및 피드백 시스템(AWE)은 인간 교사의 쓰기 피드백 유형을 변화시키는가

김형성 한국과학영재학교 인문예술학부 전임교원

- * 이 논문은 제66회 한국작문학회 학술발표대회(2026.3.28.)에서 발표한 것을 수정·보완한 것이다. 귀중한 토론문으로 더 나은 연구가 될 수 있도록 도와주신 청주대학교 최건아 교수님께 감사의 말씀을 드린다.

- I. 연구의 목적과 필요성
- II. 이론적 배경
- III. 연구 방법
- IV. 연구 결과
- V. 결론 및 논의

I. 연구의 목적과 필요성

고차원적 사고력과 문제 해결 능력을 배양하려는 목적으로 논술형 평가의 비중이 점차 확대되고 있다. 그러나 많은 연구를 통해 지적된 것처럼 논술형 평가는 필연적으로 교사에게 막대한 채점 및 피드백 부담을 유발한다(김경희·이명진, 2021; 김형성, 2025ㄴ; 남가영·김호정, 2023; 박혜영·김성숙·김경희 외, 2019). 이 같은 문제를 해결하기 위해 일부 시도교육청을 중심으로 인공지능(AI)을 활용한 자동 채점 및 피드백 시스템(Automated Writing Evaluation, 이하 ‘AWE’)을 도입하려는 움직임이 확산되고 있다.¹⁾ 이는 AWE가 논술형 평가 과정에서 발생하는 교사의 부담을 획기적으로 경감하고, 나아가 교실 평가의 패러다임을 변화시킬 것이라는 기대와 믿음에서 비롯된다.

1) 교육을비추다(2025. 12. 19.), 시범운영 시작, AI가 채점하는 시대...서울, 경기교육청, 선제적 도입 나서, 『교육을 비추다』, 검색일자 2026.1.3., 사이트 주소 <https://www.kyobit.com/news/articleView.html?idxno=3021>.

이 같은 기대와 믿음은 AWE와 관련된 다양한 연구로 확산되는 추세다. 대다수 연구는 AWE의 교육적 효과성을 검증하려는 목적으로 수행되고 있다. 가령 교사와 AWE 채점 결과의 높은 유사도 및 채점 성능 등을 바탕으로 AWE의 채점 신뢰도를 강조하거나(신윤범·김하정·원효현, 2025; 안해연, 2025; 오유경·김경선, 2025; 차준우·최서희·이기성, 2024), AWE가 제공하는 즉각적이고 빠른 피드백이 학습자의 전반적인 글쓰기 능력을 향상시킨다거나(권경문·최숙기, 2025; 권태현, 2024; 김소정, 2025; 김지연, 2024), AWE를 활용한 학습자의 고쳐쓰기 과정을 분석해 AWE 피드백의 교육적 활용 방안을 모색하는 연구(김선우·이경화, 2025; 백다빈·이경화, 2025; 최진영·김형성·송보라 외, 2023) 등이 이 같은 연구 경향에 해당한다.

이 같은 긍정적인 전망과는 달리 AWE가 인간 교사의 고유한 쓰기 평가 전문성과 교육적 역할을 단순히 대체할 수 없으리라는 우려 또한 제기된다(Jukiewicz & Wyrwa, 2026; Mariano, Taboréte, Gampal et al., 2025). AWE가 제공하는 채점 결과 및 피드백의 불완전성, 변동성 및 부정확성과 더불어 텍스트를 매개로 교사와 학생 간에 형성되는 정서적이고 인간적인 상호작용을 기술이 대체하는 것은 불가능하다는 이유에서다. 아무리 고도화된 기술 일지라도 교육 현장의 실제적인 교수·학습 맥락 및 교육적 의도와 유기적으로 통합되지 않는다면, 진정한 의미의 교육적 효과성을 창출하기는 어려울 수밖에 없다.

이 같은 맥락에서 교육부는 기억, 이해 수준의 저차원적 사고와 관련된 교육은 기술이 대신하되, 인간 교사는 적용, 분석, 평가, 창조 수준에 해당하는 고차원적 학습 경험을 유도하는 하이터치 하이테크(High-touch High-tech)를 이상적인 교수학습 모델로 강조한 바 있다(교육부, 2023; 대통령 직속 4차산업혁명 위원회, 2022). 하이터치 하이테크(High-touch High-tech)는 미래학자인 존 나이스비트가 언급한 개념으로 새로운 기술이 도입될 때마다 인간이 이에 대한 반작용으로 교감과 유대감을 더욱 갈망하게 된다는 것이다. 이는 기술이 고도화될수록 인간적 접촉이 더욱 중요하다는 점을 강

조하는 것이기도 하다(Naisbitt, 1982).

그러나 이러한 정책적 지향과 이론적 기대에 대한 우려도 존재한다. 기술의 물리적 도입이 인간 교사의 역할을 근본적으로 변화시킬 것이라는 전망이 충분히 검증되지 않은 낙관론에 불과하다는 것이다. 이 같은 우려는 기술이 저차원적 수준의 학습을 분담할 경우 교사가 고차원적인 학습 경험을 유도하는 방향으로 교육적 개입을 전환하는지를 실증적으로 분석한 연구가 현저히 부족하다는 한계에서 비롯된다.

본 연구는 이 같은 측면에서 AWE가 과연 인간 교사의 쓰기 피드백 유형을 의미 있게 변화시키는지 실증적으로 분석하고자 한다. 구체적으로 AWE의 활용이 교사의 피드백 부담을 줄여 글의 형식적, 표층적 오류를 지적하는 저차원적 쓰기 피드백이 아니라, 논술형 평가 본래의 취지에 걸맞은 고차원적 쓰기 피드백을 더 많이 제공하도록 유도할 가능성이 있는지를 탐색하고자 한다. 따라서 본 연구의 구체적인 연구 질문은 다음과 같다.

첫째, 초고와 AWE를 적용한 글(이하 'AWE 수정고')에 대한 교사의 채점 결과에 유의미한 차이가 존재하는가?

둘째, 초고와 AWE 수정고를 피드백할 때, 교사의 인지 부담에는 유의미한 차이가 존재하는가?

셋째, 초고와 AWE 수정고에 대한 교사의 글쓰기 피드백 유형에는 유의미한 차이가 존재하는가?

II. 이론적 배경

1. 인간 교사의 쓰기 피드백 특성

글쓰기 피드백의 교육적 효과성은 오랜 기간 다양한 연구를 통해 조명

되어 왔다(Matsumura, Patthey-Chavez, Valdes et al., 2002a; Olson & Rafeled, 1987; Patchan, Schunn & Correnti, 2016). 텍스트의 의미와 구조를 다루는 내용적 피드백이 학습자의 전반적인 글쓰기 능력을 향상시키는 데 기여하며, 적절한 수준의 쓰기 피드백이 긍정적인 글쓰기 동기를 형성하는 데 도움이 된다는 논의가 이를 뒷받침한다.

그러나 교사들이 제공하는 쓰기 피드백 양상을 분석한 연구들은 다소 다른 결과를 보고한다. 대다수 교사가 의미, 구조와 같은 고차원적 영역보다는 글의 형식적, 표층적인 부분에 초점을 맞추어 저차원적인 쓰기 피드백을 제공하는 경향이 있는 것으로 나타났기 때문이다(김경화, 2013; 김남북·권태현, 2022; 김정녀·유혜원, 2021; 박준범, 2014; Matsumura, Patthey-Chavez, Valdes et al., 2002b; Stern & Solomon, 2006). 이는 이상적인 쓰기 피드백의 방향과 실제 교실 환경에서 제공되는 쓰기 피드백 사이에 일정한 간극이 존재함을 의미한다.

일반계 고등학교 3학년 학생 30명의 글에 대한 중등 국어 교사 5인의 쓰기 피드백 2,862건을 분석한 결과, 어휘 수정 등의 ‘어휘’ 영역이나 띄어 쓰기, 문장 부호 등의 ‘어법’ 영역과 같은 글의 형식적 측면을 지적하는 피드백이 대다수를 차지하였다(김남북·권태현, 2022). 국외에서도 유사한 경향이 확인되는데, 41명의 교사가 학생 393명이 작성한 서사 텍스트(narrative text)에 제공한 쓰기 피드백을 분석한 결과, 대다수가 맞춤법 및 문법 오류를 교정하는 수준에 그치는 것으로 나타났다. 해당 연구는 교사의 쓰기 피드백이 과도하게 지엽적이고 표면적인 측면에 머무르고 있다는 점을 지적한다(Lucero, Fernández & Montanero, 2018). 교사들이 이처럼 저차원적 영역에 주목하는 주요 원인 중 하나는 다수의 학생 글을 읽고 평가하는 과정에서 수반되는 인지적 부담의 과도한 증가 때문일 것이다.

특히 국내에서 강조되는 논술형 평가에서의 쓰기 피드백은 학습자의 역량을 타당하게 진단하고 앞으로의 성장 방향을 안내하는 가장 핵심적인 요소로 강조되고 있다(김경희·이명진, 2021; 박혜영 외, 2019). 하지만 한 명의

교사가 다량의 답안을 읽고 개별적인 쓰기 피드백을 제공해야 하는 현실적인 제약을 고려할 때, 모든 학생에게 유의미한 수준의 쓰기 피드백을 제시하는 것은 물리적으로 쉽지 않은 과제임이 분명하다. 이러한 상황에서 교사의 쓰기 피드백이 형식적, 표층적인 오류를 교정하는 수준에 머무르게 된다면, 논술형 평가를 통해 달성되어야 할 고차원적 사고력을 제대로 측정, 진단, 지도하지 못하는 결과로 이어질 가능성이 있다.

2. 자동 채점 프로그램의 유형과 효과

자동 채점 프로그램은 오랜 시간 동안 기술의 발전을 거듭하며 다양한 용어로 일컬어져 왔다. 따라서 기술의 발전 양상에 따라 달라지는 용어의 차이를 이해하는 과정은 자동 채점 프로그램의 기능과 역할을 파악하는 데 필수적이라 할 수 있다.

자동 채점 프로그램의 시작이라 할 수 있는 자동 에세이 채점(Automated Essay Scoring, 이하 ‘AES’)은 1966년 고등학교 영어 교사였던 엘리스 배튼 페이지(Ellis Batten Page)가 교사의 채점 시간을 단축하기 위해 제안한 ‘PEG(Project Essay Grade)’ 시스템에서 출발하였다(Page, 1966). 초기 시스템은 교사의 채점 결과를 꽤 정확하게 예측했음에도 불구하고, 가정용 컴퓨터나 인터넷과 같은 기술적 인프라의 부족과 인간 채점자를 기계가 대체할 수 있다는 사회적 비판에 직면하여 교육계에서 제한적으로만 수용되었다. 또한 AES는 텍스트의 형식적, 표층적 특징 등을 중심으로 글을 채점하는 수준에 머물렀기에, 글의 의미와 구조를 온전히 평가하고 피드백하는 데에는 구조적인 한계를 지니고 있다는 지적을 받아왔다(Ramesh & Sanampudi, 2022; Weigle, 2013).

1990년대 이후 자연어 처리(NLP) 기술의 비약적인 발전과 인터넷 및 워드 프로세서의 대중화는 자동 채점 프로그램의 진화를 견인하였다. 특히 2010년대 이후 기계 학습(Machine learning), 심층 학습(Deep learning),

인공 신경망 알고리즘 등의 발전으로 자동 채점 프로그램은 단순히 점수를 부여하는 수준을 넘어 타당한 쓰기 피드백을 제공하는 수준으로 발전하게 된다. 따라서 이 같은 자동 채점 프로그램은 AES나 AEG(Automated Essay Grading)라는 용어에서 나아가, ‘평가와 진단’의 의미를 강조한 AWE(Automated Writing Evaluation) 또는 AEE(Automated Essay Evaluation)라는 용어로 대체되기 시작한다(Zupanc & Bosnić, 2017).

특히 최근에는 형성평가(formative assessment)의 맥락에서 AWE와 AWF(Automated Written Feedback)라는 개념이 널리 사용되고 있다. 두 용어는 학술적으로 혼용되는 경향이 있으나, AWE는 점수를 산출하고 피드백을 생성하는 인터페이스를 모두 포함하는 ‘인공지능 기반의 총체적 평가 시스템’을 지칭하는 반면, AWF는 AWE에 의해 학습자에게 최종적으로 제공되는 ‘피드백 텍스트 결과물’ 그 자체에 방점을 두어 설명될 때 주로 사용된다(Shi & Aryadoust, 2023).

AWE의 교육적 효과성은 다각적인 측면에서 연구되어 왔다. 그중 하나는 AWE가 제공하는 피드백이 전반적인 쓰기 결과물의 질을 개선하는 데 긍정적인 영향을 미친다는 것이다(Link, Mehrzad & Rahimi, 2022; Palermo & Thomson, 2018; Stevenson & Phakiti, 2014; Wade-Stein & Kintsch, 2004; Zhai & Ma, 2023). 그러나 AWE가 제공하는 쓰기 피드백이 항상 유의미한 글쓰기 능력 향상을 담보하는 것은 아니라는 한계 또한 보고되었다(Ware, 2014). AWE 피드백에 오류가 포함되어 있거나(Bai & Hu, 2017), 학습자가 AWE 피드백을 무시하는 등 수용을 거부하고, 수동적으로 대응할 경우 그 효과가 크게 반감될 수 있다는 것이다(Zhang & Hyland, 2018). 이는 AWE의 기술적 성능만큼이나 학습자 참여(Engagement), 학습자 주도성(Student Agency)과 글쓰기 피드백 리터러시(Writing feedback literacy)가 AWE의 효과성을 결정짓는 주요 변인임을 짐작하게 한다.

AWE의 도입이 인간 교사의 쓰기 피드백 양상에 어떠한 영향을 미치는지에 대한 탐구 또한 중요한 연구 주제로 다루어진 바 있다. Wilson &

Czik(2016)은 중학생들의 모어(L1) 쓰기 환경에서 ‘교사 단독 피드백 집단’과 ‘교사 및 AWE 결합 집단’을 비교 분석하였다. 분석 결과 두 집단 간 교사가 제공한 쓰기 피드백의 총량에는 차이가 없었다. 그러나 ‘교사 및 AWE 결합 집단’의 교사들이 ‘교사 단독 피드백 집단’에 비해 고차원적(higher-level) 쓰기 피드백을 더 많이 제공한 것으로 나타났다. 이는 AWE가 문법 오류나 철자 교정과 같은 저차원적(lower-level) 쓰기 피드백을 제공함으로써, 교사가 의미 구조나 논리적 조직과 같은 고차원적(higher-level) 쓰기 피드백에 인지적 자원을 할당하게 된다는 주장을 지지하는 결과로 해석된다.

그러나 AWE의 도입이 모든 교육 환경에서 교사의 고차원적 피드백을 증가시키는 결과로 직결되지 않는다는 연구 결과도 존재한다. L2 학습자를 대상으로 한 연구이기는 하나, AWE를 활용한 교사가 그렇지 않은 교사에 비해 저차원적 쓰기 피드백을 덜 제공하는 경향은 뚜렷하게 나타났으나, 고차원적 쓰기 피드백의 절대적인 양 자체가 유의미하게 증가하지는 않은 것으로 확인되었다(Link et al., 2022). 이러한 상반된 결과들은 교실 현장에 단순히 기술을 도입하는 것만으로는 교사의 교육적 초점과 쓰기 피드백 제시 전략이 자동적으로 고도화되지 않을 것임을 의미한다.

Transformer 알고리즘 기반의 BERT는 자연어 처리(NLP) 분야의 혁신을 이루었고, 이 과정에서 등장한 대규모 언어 모델(LLM) 기반의 생성형 인공지능(Generative AI)은 자동 채점 및 피드백 기술의 새로운 국면을 이끌고 있다(Shi & Aryadoust, 2023). 규칙 기반 체계에 머물렀던 과거의 AWE와 달리 생성형 인공지능은 문맥을 파악하여 논증의 타당성, 논리적 조직과 같은 고차원적 영역의 채점 및 피드백을 매우 타당하게 수행한다는 연구 결과가 국내외를 가리지 않고 보고되고 있다. 그러나 생성형 인공지능의 채점과 피드백이 지닌 편향성, 변동성, 부정확성, 불완전성, 비일관성 등은 여전히 문제로 지적되고 있다는 점에 주목할 필요가 있다.

중요한 것은 국외의 경우 기술의 발전에 따라 다양한 유형의 AWE가 교

육 현장에 점진적으로 적용되면서 이와 관련된 다양한 실험과 연구가 이루어졌다는 사실이다. 현재 국내 교육 현장은 AI 코스로 개발되어 간단한 조작만으로 사용이 가능한 AWE 프로그램과 더불어 생성형 AI 기반의 AWE가 동시다발적으로 사용되는 상황이다. 본 연구는 우선 기술의 발전 단계를 고려해 전자 측면, 즉 AI 코스로 개발된 AWE에 초점을 맞추어 진행되었다고 할 수 있다.

III. 연구 방법

1. 연구 대상

본 연구에 참여한 채점자 23명은 편의 표집(convenience sampling)을 통해 선정하였다. 채점자의 특성은 <표 1>과 같다.

<표 1> 채점자 정보

항목	구분	응답 수(명)	비율(%)
성별	남	7	30.4%
	여	16	69.6%
교직 경력	1-5년	1	4.4%
	6-10년	9	39.1%
	11-20년	13	56.5%
학교급	중학교	3	13%
	고등학교	20	87%

성별 기준으로 여성은 16명(69.6%), 남성은 7명(30.4%)이었으며, 교직 경력의 경우 11-20년이 13명(56.5%)으로 가장 많았으며, 6-10년 9명(39.1%), 1-5년 1명(4.4%) 순으로 나타났다. 소속 학교급은 고등학교 교사가 20명(87.0%), 중학교 교사는 3명(13.0%)이었다.

2. 연구 절차

1) 1단계 : 논술형 평가 문항 설계 및 답안 수집

학습자의 논술형 평가 답안을 수집하기 위해 2024년 12월 17일~23일에 A 지역에 소재한 고등학교 2학년 남학생 70명을 대상으로 논술형 평가를 실시하였다. 논술형 평가 문항은 다문서 기반의 논증적 글쓰기 형태로 설계되었으며, 주제는 ‘촉법소년의 연령 기준 하향’으로 설정하였다.

논술형 평가는 다양한 유형의 제시문과 <조건>을 바탕으로 자신의 견해를 합리적으로 논증하는 역량을 진단, 측정하는 평가 방법이다(교육부, 2022). 따라서 촉법소년의 법적 정의, 촉법소년의 범죄 추이 그래프, 소년 범죄 예방을 위한 인프라 조성 및 교화의 중요성, 청소년의 성숙도 변화, 국외 사례 및 UN 권고안 등 총 6가지의 자료를 제시문으로 구성하였다.

이어 <조건>을 통해 학습자의 답안 작성 방향을 상세하게 안내하였다. 구체적인 조건은 서론-본론-결론의 삼단 구성을 갖추고, 자신의 주장에 대한 타당한 이유와 근거를 자료에서 인용하여 제시하고, 자신의 견해에 대한 반론 및 재반박을 포함하여 1,000자 내외의 글을 작성하는 것이다. 논술형 평가 문항의 제시문과 <조건>은 <그림1>과 같다.

1. 최근 법무부가 촉법소년의 연령 기준을 낮추는 방안을 추진하면서 찬성과 반대 의견이 팽팽하게 맞서고 있습니다. (가)~(바)를 활용하여 '촉법소년의 연령 기준을 만 14세 미만보다 낮추어야 한다.'라는 논제에 대해 <조견>을 고려하여 논설문을 작성하세요.

촉법소년

14세 미만

형사처벌 불가

보호처분 가능

범법소년

14세 이상~14세 미만

형사처벌 가능

보호처분 가능

소년법은 19세 미만인 청소년에게 적용된다. 청소년 범죄자 중에서 만 10세 이상 ~ 14세 미만을 '촉법소년'이라고 한다. 이들은 범죄를 저질러도 형사처벌을 받지 않고 1호에서 10호에 이르는 보호처분을 받는다. 보호처분은 범죄 이력으로 기록되지 않는다. 가령 보호처분 3호는 사회봉사명령, 4, 5호는 보호관찰, 6, 7호는 소년보호시설 위탁, 8~10호는 소년원 송치에 해당한다.

(나) 지난해 보호처분을 받은 촉법소년 수는 7175명으로 1년 새 36.8% 늘었다. 전체 보호처분 수에서 차지하는 비율도 23.7%로 2년 연속 20%대를 넘겼다. 보호처분을 받은 촉법소년은 2014년부터 2020년까지는 3000명대 수준을 이어가며 증가·감소를 반복했다. 그러나 2021년 4142명으로 4000명을 넘어선 뒤 2022년 5245명으로 급증했다. 이들은 주로 우발적 범행을 저지르는 것으로 나타났다. 통계에 따르면 소년보호사건의 행위 유형별에서 우발적 범행이 46.1%로 가장 높았다. 이어 △호기심 34.9% △생활비 마련 8.3% △유혹 2.9% 등의 순이었다.

—촉법소년 1년새 37% 급증, 서울경제(2024.9.27.)

(다) 2017년부터 2023년 7월까지 촉법소년이 저지른 5대 강력범죄(살인, 강도, 강간·강제추행, 절도, 폭력)의 비율은 78.3%(5만 6766명)로 나타났다. 이 중 절도가 50.2%(3만 6378명)였고, 폭력 24.0%(1만 7403명), 강간·추행 4.0%(2904명), 방화 0.5%(293명) 순이었다. 김대권 한국형사법무정책연구원 법무정책연구실장은 "폭력·성폭력 범죄 등은 실제로 범죄 자체가 증가한 것일 수도 있지만, 사회 전반의 학교폭력에 대한 감수성·성인인 감수성 등의 확대에 따라 적극적인 고소·고발이 이뤄진 영향일 수 있다."라고 덧붙였다.

—'어차피 감방 안 가'...나이를 무기화하는 촉법소년, 노컷뉴스(2024.1.22.)

(라) "소년범죄가 1년에 6만건 정도 발생하는데 소년교도소와 소년원에 격리해 교화할 수 있는 수는 5000여명에 불과합니다. 일본이 소년교도소 7곳, 소년원 52곳을 운영하는 것과 대비되죠. 소년법은 심리 치유와 정신과 치료도 중요합니다. 하지만 국내에 이런 시설을 갖춘 곳은 대전소년원이 유일합니다. 인프라도 없는 상황에서 촉법소년 연령을 하향하면 범죄자만 양산합니다. 무엇을 잘못했는지 깨닫고 진심으로 반성하게 하는 게 가장 우선순위가 돼야 합니다."

—천종호 부장감사 인터뷰, 서울신문(2024.7.1.)

(마) "형사처벌을 받지 않는 촉법소년들이 법을 악용해 수사기관의 정당한 공권력 행사를 방해하고 치안유지 기능을 조롱하면서 이로 인한 국민의 공분이 매우 큰 상황입니다. 지금과 같이 14세를 기준으로 일괄적인 보호처분을 하는 것은 상황을 개선하지 못할 것입니다. 소년들이 과거보다 더 빨리 성숙해지고 세상을 받아들이는 속도도 빠르기 때문에 이에 따라 결국 촉법소년의 연령도 하향해야 합니다."

—국회의원 김병기 인터뷰, 세계일보(2023.12.09.)

(바) 주요 선진국들의 형사처벌을 받지 않는 연령 기준은 13세 미만인 경우가 많았다. 세부적으로 프랑스 13세 미만, 캐나다 12세 미만, 영국 10세 미만, 호주 10세 미만이다. 미국은 만 7세 등 주마다 다른 기준을 갖고 있지만 뉴욕주 등 대다수 주에서 13세 미만 기준을 채택하고 있다. 하지만 모든 선진국이 다 그런 건 아니다. 유엔(UN) 조사 결과에 따르면, 형사책임 연령을 14세로 규정한 나라는 33개국에 달한다. 덴마크·핀란드·스웨덴·이탈리아는 촉법소년 기준 연령이 15세로 우리보다 높다. 호주는 거꾸로 기존 10세 기준을 14세로 상향 조정하는 방안을 논의 중이다.

— [팩트체크] 촉법소년 만 14세에서 13세?, 파이낸셜뉴스(2023.1.25.)

<조견>

- 서론, 본론, 결론의 구조로 작성할 것
- 자신의 주장(claim)을 명시적으로 제시하고, 이를 뒷받침하는 타당한 이유(reason)와 근거(evidence)를 작성할 것. 단, 자료를 활용할 때 반드시 출처를 밝힐 것. (예 : (나)에 따르면~)
- 자신의 주장과 구분되는 반론을 제시하고, 이에 대한 반박을 논리적으로 작성할 것.
- 1000자 내외(+/-200자, 공백 포함)으로 작성할 것.

<그림 1> 논술형 평가 문항의 제시문과 조건

2) 2단계 : AWE를 활용한 고쳐쓰기 및 연구용 답안 선정

본 연구에서 학습자가 논술형 평가 답안을 고쳐쓰기 위해 활용한 AWE

는 투블럭 에이아이(TWOBLOCK Ai)가 개발한 키위(KEEWi, Korean Essay Evaluation with AI)다. 해당 프로그램은 한국어 전용 자동 채점 및 피드백 프로그램으로, 심층 학습 기반의 BERT(Bidirectional Encoder Representations from Transformers) 모델을 통해 글의 특징을 구조적으로 분석한다(차준우 외, 2024). 특히 이 프로그램은 미국의 ‘6 Traits of Writing’ 모델을 바탕으로 ‘문법의 정확성, 어휘의 풍부성, 의도의 표현, 문장 구사력, 조직과 전개, 주제의 명확성’이라는 6가지 지표를 통해 13가지 종류의 글쓰기 결과물의 품질을 평가한다는 특징이 있다. 해당 프로그램에서 제공하는 평가 지표별 세부 피드백 기준은 <그림 2>와 같다.

문법의 정확성  정확한 어문 규범 준수 · 모든 오류에 대한 수정 방안과 설명 제시 · 맞춤법, 띄어쓰기, 조사/어미/접사 오류 · 문체, 문장부호, 발음/임의 유사어, 존칭 오류 · 용언의 활용, 보조용언, 표현어, 입력 오류 등	어휘의 풍부성  다양한 단어의 정확한 의미 전달 · 단어 수준 오류에 대한 수정 방안과 설명 제시 · 약어 사용, 순화용어, 겸말, 사투리, 외래어 표기법 오류 · 문단내 단어 중복이 심할 경우 추천 단어 제시 · 글 전체에서 적절한 (25%) 단어 중복 권유	의도의 표현  독자가 쉽게 이해하도록 배려 · “같은 경우에” 등 독자를 어지럽히는 표현 지양 · 너무 어려운 단어는 쉬운 표현으로 변경 요청 · 한가지 표현을 여러 번 사용 지양 · 장문별 적절한 접속 표시 (그런데 등) 활용 권장
문장 구사력  원급을 조절하는 글 쓰기 · 중복된 문장 검토, 문장 부호, 문장 사이의 줄 나눔 · 단문과 복문의 균형, 지나치게 긴 문장 사용 지양 · 문장의 이해가 쉬운 정도(구밀 구조)를 평가에 반영	조직과 전개  목적에 맞는 글의 형태와 구조 갖추기 · 문단의 일부를 중복하여 글의 글자 수 늘리는 행위 금지 · 너무 긴 문단의 경우, 문단을 나누어야 할 위치 제시 · 문단 내 문장들이 쉽게 읽히는 정도를 평가에 반영	주제의 명확성  글을 하나의 주제에 집중하기 · 주제어 (3개) 선택이 쉬운 글을 권장 · 본문의 요약과 구조 분석이 쉬운 글을 권장 · 글의 내용이 주제어와 어울리는지 여부

<그림 2> 키위(KEEWi)의 6가지 평가 지표 및 세부 피드백 기준

이때 <그림 2>에 제시된 평가 지표 중 ‘의도의 표현’, ‘조직과 전개’, ‘주제의 명확성’은 언뜻 글의 내용이나 논리 구조에 대한 고차원적 피드백을 제공하는 것처럼 보일 수 있다. 그러나 세부 피드백 기준을 살펴보면, ‘의도의 표현’은 독자가 쉽게 이해하도록 특정 표현을 준수했는지를 점검하고, ‘조직과 전개’는 너무 짧거나 긴 문단의 존재 여부, 문단 간 연결어 사용 등 형식적 요소를 중심으로 판단하며, ‘주제의 명확성’ 역시 주제와 관련 없는 문장의 포함 여부를 탐지하는 방식에 가깝다. 따라서 이들 지표에 따른 피드백이 문장 단위의 표층적 수준에서 제시되고 있다는 점에 주목할 필요가 있다.

70명의 학생은 1차시(50분) 동안 논술형 평가 답안(이하 ‘초고’)을 작성한 뒤, 이후 1차시(50분) 동안 키위(KEEwi)를 활용하여 고쳐쓰기를 수행하였다. 학습자들은 키위(KEEwi)가 제시한 6가지 평가 지표별 점수와 구체적인 피드백을 실시간으로 확인하며 자신의 글을 수정한 뒤 최종적인 글(이하 ‘AWE 수정고’)을 완성하여 제출하였다.

수집된 130편(초고 70편, AWE 수정고 60편) 중 실험에 투입할 최종 답안 12편을 선정하기 위해 의도적 표집(Purposive Sampling)을 실시하였다.²⁾ 이때 AWE가 평가한 점수를 기준으로 초고(70편)와 AWE 수정고(60편)의 Z점수를 산출하여 다양한 수준의 글을 표집하였다. 이때 표집한 12편의 글은 모두 서로 다른 학생이 작성한 글이다.

이처럼 동일 학생의 초고와 AWE 수정고를 선정하지 않은 이유는 채점자가 특정 답안을 ‘동일 학생이 수정한 글’이라고 인지할 경우 발생할 수 있는 채점 편향을 차단하기 위함이다. 즉, 초고와 AWE 수정고의 점수를 Z점수로 표준화하여, 동일한 수준 내 서로 다른 학생의 답안을 교차로 선정함으로써 채점자가 모든 답안을 독립적인 결과물로 인식하게 하였다. 이 같은 표집 과정을 통해 실험 연구의 내적 타당성 요인 중 ‘피험자 선택(subject selection), 회귀(regression)’³⁾ 등의 문제를 최소화하고자 노력하였다(윤준채, 2023). 실험에 투입된 12편의 분량, 원점수, Z점수 등 구체적인 특성은 <표 2>와 같다.

2) 초고를 작성한 학습자 중 일부가 고쳐쓰기 과정을 수행하지 않아, 수집된 AWE 수정고는 총 60편이다.

3) 실험 연구에서 내적 타당도를 위협하는 요소는 다양하다. 피험자 선택은 실험집단 및 통제집단 구성원의 특성이 동질하지 않아 발생할 수 있으며, 회귀는 수행 능력이 매우 높거나 낮은 피험자들만이 실험에 참여하면서 발생한다.

〈표 2〉 글 특성

초고(M = 69.45, SD = 12.69)			AWE 수정고(M = 88, SD = 9.04)		
글	분량	원점수 (Z점수)	글	분량	원점수 (Z점수)
1번	676자	52.2 (-1.36)	3번	972자	83.2 (-0.56)
2번	581자	54.6 (-1.17)	4번	744자	84 (-0.47)
6번	1196자	61.1 (-0.65)	8번	945자	91.7 (0.38)
5번	638자	71.5 (0.16)	10번	920자	94.6 (0.70)
12번	1098자	74.9 (0.43)	11번	1147자	95 (0.74)
9번	1088자	79.2 (0.77)	7번	1446자	96.4 (0.90)

다만 답안 선정 과정에서 Z점수와 분량 변인을 완전하게 통제하지는 못하였다. AWE 수정고에서 Z점수가 -1 이하인 글은 총 7편이었으나, 해당 글은 학습자가 고쳐쓰기를 거의 수행하지 않아 사실상 ‘초고’에 가까운 글로 분석되어 선정하지 않았기 때문이다. 또한 실험에 투입한 초고 6편의 총 분량은 5,277자, AWE 수정고 6편의 총 분량은 6,174자로 나타났다. 이는 AWE 피드백을 반영해 글을 고쳐 쓰는 과정에서 내용이 구체화되는 등 고쳐쓰기의 자연스러운 특성이 반영된 결과이기도 하다. 따라서 본 연구는 글 수준 및 분량의 차이가 교사의 채점 과정, 인지 부담 및 쓰기 피드백 제시 과정에 미치는 영향을 완전히 통제하지 못했다는 한계가 있으므로 해석에 유의할 필요가 있다.

3) 3단계 : 교사의 채점 및 피드백, 인지 부담 측정

교사의 채점 및 피드백 과정에서 실험의 객관성을 유지하고 피험자 태도(Subject Attitude)⁴⁾에 따른 편향을 차단하기 위해, 실험 종료 전까지 연구

- 4) 피험자 태도는 실험에 참여하는 피험자들이 실험 대상이 되고 있다는 것을 알아차릴 때 실험에 대한 태도가 실험 결과에 영향을 주기 때문에 발생한다. 본 연구에서는 각주 3과 4

의 구체적인 목적을 밝히지 않는 단일 맹검(Single-blind) 방식을 채택하였다. 채점자 23명에게 연구의 목적을 ‘교사의 쓰기 피드백 분석 연구’라고 밝힌 뒤, 12편의 학생 답안, 채점 기준 및 해설, 피드백 방식 등이 상세하게 작성된 평가 안내문을 제공하였다. 채점 및 피드백은 2026년 1월 5일~16일까지 진행되었다.

채점 기준은 국립국어원의 ‘2023년 국민의 글쓰기 능력 진단 체계 개발’에서 활용된 채점 기준표를 수정하여 제작하였다(민병곤·남가영·김선희 외, 2023). 해당 채점 기준표는 내용(5개), 조직(2개), 표현 영역(2개)으로 나뉘며, 5점 척도를 사용한다. 그러나 본 연구에서는 내용, 조직, 표현에서 각 2개씩의 채점 기준만을 선별하였고, 기존 채점 기준표에 제시된 수행 수준의 체계성과 엄밀성을 수정, 보완하여 4점 척도로 변환해 사용하였다. 구체적인 채점 기준은 <표 3>과 같다.

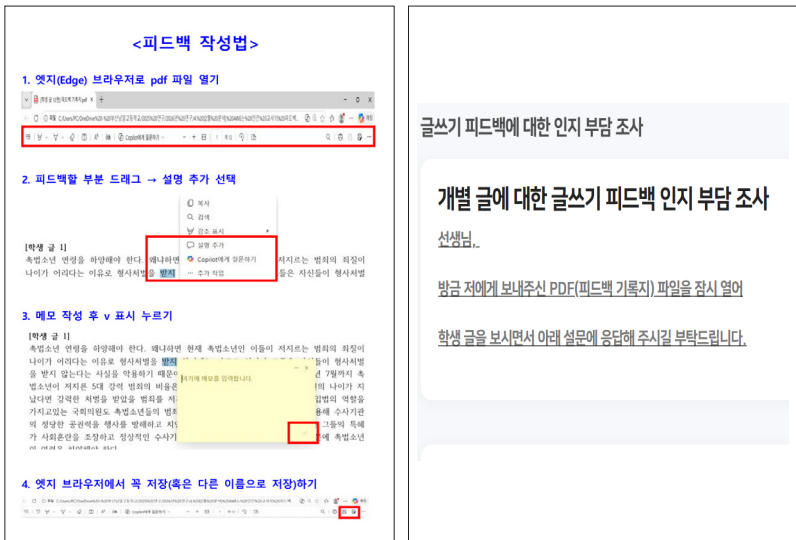
<표 3> 채점 기준표⁵⁾

영역	4점에 해당하는 수행 수준
내용의 타당성	주장과 이유, 이유와 근거의 타당성(관련성 및 충분성)이 높아, 주장을 효과적으로 뒷받침하였다.
예상 반론 및 반박의 타당성	예상 반론을 정확하게 언급하고, 예상 반론과 관련된 반박을 구체적으로 전개하였다.
형식 문단, 배열	문단 구분이 명확하여 서-본-결의 구조가 매우 적절하게 구분되고, 문단 배열이 매우 유기적이다.
내용 문단	모든 문단에서 완결성, 통일성, 일관성을 갖추었다.
문장, 어휘	문장과 어휘 사용이 자연스럽고 논증에 효과적이다.
어문 규범, 글쓰기 관습	어문 규범과 글쓰기 관습을 매우 정확히 준수하였다.

에서 설명한 내적 타당도를 위협하는 요소를 최소화하고자 노력하였다.

5) 분량 관계상 4점 척도에 해당하는 수행 수준만을 진술하였다.

채점자 23명의 피드백은 <그림 3>처럼 PDF 파일을 통해 수합하였다. 학습자의 글에서 피드백이 필요한 부분을 드래그한 뒤 ‘메모’ 기능으로 피드백을 작성하도록 안내하였다. 연구자는 참여자들에게 소인수 학급 학생들을 대상으로 1:1 개별 글쓰기 지도를 한다고 가정하고, 최대한 상세하게 쓰기 피드백을 작성해 줄 것을 요청하는 등 교사의 실제적인 반응을 최대한 유도하고자 노력하였다.



<그림 3> (왼쪽) 피드백 작성 안내문, (오른쪽) 인지 부담 설문지

다음으로 쓰기 평가자의 인지 부담을 측정하였다. 쓰기 평가의 인지 과정을 탐색해 온 선행 연구들은 사고 구술 프로토콜(Flower & Hayes, 1981), 시선 추적(이지원 · 박영민, 2015), 채점에 걸리는 시간(정분의, 2019) 등을 지표로 평가자의 인지 부담을 추론하였다. 그러나 사고 구술은 채점자가 자신의 인지 과정을 실시간으로 언어화하도록 하는 과정에서 본 연구의 분석 초점이 노출될 수 있고, 행동이나 시간 지표는 인지적 처리의 양적 변화를 포착하는 데에는 유용하나 채점자의 인지 부담을 질적으로 분석하려면 사고

구술을 병행해야만 한다.

본 연구의 관심은 채점자가 초고와 AWE 수정고를 평가할 때 발생하는 인지 부담의 차이를 분석하는 데 있다. 이를 위해 평가자가 채점과 피드백을 완료한 직후에 자기 보고식 설문에 응답할 것을 요청하였다. 즉, 어문 규범과 같은 표층적 오류 처리에 소요되는 인지 자원이 줄어들수록 인지 부담이 줄어든다고 인식하는지, 늘어난 인지 자원을 논리 구조 분석과 같은 고차원적 평가에 배분하여 사용하였는지를 살펴보고자 한 것이다.

이는 쓰기 평가자의 인지 자원이 고정된 것이 아니라 선택적으로 배분된다고 보고한 일련의 연구와도 관련된다. 전문 평가자는 글의 전반적 자질에 주목하며 하향식으로 채점하는 반면 초보 평가자는 표층적 오류 탐지에 주의를 집중한다(Barkaoui, 2010; Cumming, 1990; Wolfe, Kao & Ranney, 1998). 시선 추적 연구에서도 평가자의 주위가 채점 기준 영역별로 차등 배분되고 그 양상이 평가자 간 신뢰도와 연관된다는 결과가 나타난 바 있다(Winke & Lim, 2015). 이는 채점자가 제한된 인지 자원을 글의 표층적, 내용적, 구조적 측면에 따라 배분하여 할당함을 의미한다. 따라서 본 연구의 가설을 입증하기 위해 글의 표층적 오류가 수정된 AWE 수정고를 채점할 때 발생하는 인지 부담이, 초고를 평가할 때 발생하는 인지 부담보다 줄어들었는지를 채점 직후 확인하기 위해 <표 4>와 같은 설문 문항을 구성하였다.

채점자의 인지 부담은 <그림 3>처럼 네이버 폼(Naver Form)을 활용해 수합하였다. 설문조사 링크는 채점자들이 채점 결과(Excel)와 피드백 파일(PDF)을 연구자에게 전송하고 연구자가 그 결과를 확인한 다음 전송하였고, 가급적 짧은 시간 안에 응답할 것을 요구하였다. 이는 초고와 AWE 수정고를 피드백하는 과정에서 채점자들이 느낀 인지 부담을 정확하게 파악하기 위한 것이다. 설문을 완료한 참여자들에게만 최종적으로 연구의 실제 목적을 정확하게 설명하였다.

〈표 4〉 채점자의 인지 부담 측정 설문

	인지 부담 측정 문항 (5점 척도)
1번	기본적인 어문 규범 오류에 신경이 많이 쓰였다. ⁶⁾
2번	문장 가독성이 좋아 학생의 의도를 파악하는 것이 수월했다.
3번	문단 구분이 되어 있어 글 흐름을 파악하는 것이 수월했다.
4번	글의 전체적인 논리 구조를 분석할 여유가 있었다.
5번	전개 방식에 대해 구체적인 조언을 할 가치가 있다고 생각했다.

3. 분석 방법

1) 논술형 답안 채점 결과 및 교사의 인지 부담 분석

채점 결과의 신뢰도를 산출하기 위해 고전 검사 이론에 근거해 총점 기반의 크론바흐 알파(Cronbach's alpha) 계수를 산출하였다(최숙기, 2012). 이어 초고와 AWE 수정고의 채점 결과에 유의미한 차이가 있는지, 초고와 AWE 수정고에 대한 쓰기 피드백을 제시하는 과정에서 인지 부담의 차이가 발생했는지를 통계적으로 분석하기 위해 대응표본 T 검정(Paired Samples T-test)을 실시하였다.

대응표본 T 검정을 실시하기 전 각 요인(채점 기준 및 인지 부담 항목)별로 Shapiro-Wilk 정규성 검정을 선행하였으며, 정규성 가정을 충족하지 못하는 일부 항목에 대해서는 비모수 검정인 윌콥슨 부호 순위 검정(Wilcoxon signed-rank test)을 실시하였다.

2) 쓰기 피드백 유형에 따른 코딩 및 표준화

초고와 AWE 수정고에 대한 교사의 쓰기 피드백 유형에 차이가 발생하였는지를 분석하기 위해 쓰기 피드백 데이터를 범주별로 코딩하였다. 이때 교

6) 1번 문항은 역코딩하여 척도의 방향성을 통일하였다.

사의 쓰기 피드백을 분석한 대표적인 선행 연구라 할 수 있는 Cho, Schunn & Charney(2006), Hayes & Berninger(2010), Link et al.(2022), Wilson & Czik(2016)의 쓰기 피드백 코딩 체계를 참고하였다. 구체적인 절차는 다음과 같다.

먼저 23명의 교사가 작성한 모든 쓰기 피드백을 분석의 최소 단위인 ‘아이디어 단위(Idea unit)’로 분절하여 추출하였다. 아이디어 단위란 텍스트 내의 단일한 특징, 요소를 대상으로 하는 독립적인 의미 단위를 의미한다. 따라서 하나의 문장에 여러 가지 의미 단위를 내포한 피드백이 발견될 경우에는 각각의 의미 단위를 개별적으로 분리, 코딩하여 분석의 정밀도를 높이고자 하였다.

쓰기 피드백 코딩의 범주는 저차원적 쓰기 피드백(Lower-level Feedback)과 고차원적 쓰기 피드백(Higher-level Feedback)을 분리하여 코딩하였다. 다만 저차원적 쓰기 피드백에서 국어 글쓰기와 큰 관련성이 없다고 판단한 ‘대소문자 사용(Capitalization)’, ‘문법(grammar)’은 삭제하였고, 철자 및 문장 부호(Spelling, Punctuation)는 통합하였다. 또한 고차원적 쓰기 피드백에서 ‘자아(self)’와 관련된 내용은 학습자의 글에서 거의 드러나지 않는다고 판단하여 삭제하였다. 구체적인 쓰기 피드백 코딩의 범주는 <표 5>와 같다.

쓰기 피드백 데이터 분석의 객관성과 신뢰도를 확보하기 위해 교육 경력 12년의 국어 교육 전공 박사 학위를 소지한 현직 국어 교사와 교차 코딩 및 신뢰도 검증 과정을 거쳤다. 먼저 전체 쓰기 피드백 데이터 중 10%를 무작위로 추출하여 두 분석자가 코딩 범주에 따라 독립적으로 1차 코딩을 실시하였다. 이때 편향(bias) 발생을 차단하기 위해 채점자 이름, 학생 글 번호를 생략하고 쓰기 피드백만을 확인하고 코딩을 실시하였다.

1차 코딩 결과 불일치하는 항목에 대해서는 협의를 통해 이견을 조정하여 100% 코딩을 일치시켰다. 이후 남은 90%의 쓰기 피드백 데이터를 두 분석자가 분담하여 코딩을 진행하였다. 각자의 코딩이 완료된 후에는 서로의

〈표 5〉 쓰기 피드백 코딩 범주

수준	하위 범주	예시
저차원적 피드백 (Lower-level Feedback)	철자 및 문장 부호 (Spelling, Punctuation)	맞춤법 오류, 잘못된 띄어쓰기, 문장 부호 오류 등과 같은 정확한 문법적 표현 측면
	문장 구조 (Sentence structure)	주어와 서술어 호응, 목적어 서술어 호응, 과도한 피동표현 등 문장의 가독성 측면
	서식 (Formatting)	들여쓰기, 단락 구분 등과 같은 시각적 표현 측면
	단어 선택 (Word choice)	추상적이거나 부정확한 어휘(구, 절 포함), 속어 사용 지양과 같은 어휘(구, 절 포함) 선택의 적절성, 명확성, 구체성 측면
고차원적 피드백 (Higher-level Feedback)	조직/구성 (Organization)	단락 순서 조정, 서론 및 결론 포함 여부 등 논리적 조직과 구성 측면
	아이디어 및 상술 (Ideas and elaboration)	이유와 근거의 제시 여부, 출처 표시, 동일한 내용의 반복 지양 등과 같은 내용의 타당성, 신뢰성, 구체성 측면
	문체 (Style)	구어체, 주관적 및 감정적 표현 사용 지양 등과 같은 수사적 맥락(독자, 장르 등)에 대한 고려 측면

결과물을 전수 교차 검토(cross-review)하였으며, 이 과정에서 발견된 이견, 차이 등은 다시 한번 논의를 거쳐 조정하였다. 최종적으로 전체 피드백 코딩 결과에 대하여 100% 합의를 이루었다.

코딩을 완료한 후 글의 분량에 따라 쓰기 피드백의 빈도가 달라질 수 있다는 점을 고려하여 쓰기 피드백을 표준화(standardization)하는 과정을 거쳤다. 각 글에서 교사가 제시한 ‘전체 피드백 수’ 대비 ‘유형별 쓰기 피드백 수’의 상대적 비율(Proportion)을 산출하였다. 이러한 표준화 과정을 통해 텍스트 길이에 따른 쓰기 피드백 빈도의 차이를 통제하여, 쓰기 피드백의 구조적 비중 변화를 보다 정확하고 객관적으로 비교하고자 하였다.

이어 초고와 AWE 수정고 간에 쓰기 피드백 양상의 유의미한 차이가 나타나는지를 통계적으로 검증하기 위해 대응표본 T 검정을 실시하였다. 우선 Shapiro-Wilk 정규성 검정을 선행하여 데이터의 분포 특성을 확인한 뒤, 정규성 가정을 충족하지 못하는 변수에 대해서는 비모수 검정인 윌콕슨 부호 순위 검정(Wilcoxon signed-rank test)을 적용하였다.

IV. 연구 결과

1. 논술형 답안에 대한 채점 결과 및 인지 부담 분석

1) 채점 결과에 대한 기술 통계 및 대응 표본 T 검정

논술형 평가 답안에 대한 총점 기반의 신뢰도를 분석한 결과, 크론바흐 알파(Cronbach's alpha) 계수가 0.974로 도출되어 채점자 신뢰도가 매우 높은 수준임을 확인하였다. 전체 12편의 글에 대한 23인의 채점 결과에 대한 기술 통계는 <표 6>과 같다.

<표 6> 채점 결과에 대한 기술 통계 결과

기준 학생	내용의 타당성 M(SD)	예상 반론, 반박 M(SD)	형식 문단, 배열 M(SD)	내용 문단 M(SD)	어휘, 문장 사용 M(SD)	어문 규범, 관습 M(SD)	합계 M(SD)
2	1.91 (0.58)	2.17 (0.76)	1.09 (0.28)	1.65 (0.91)	2.35 (0.63)	2.57 (0.71)	11.74 (2.15)
5	2.35 (0.76)	2.48 (0.71)	1.74 (0.44)	2.22 (0.72)	2.30 (0.62)	1.91 (0.58)	13 (2.50)
4	2.22 (0.83)	1.00 (0.00)	2.26 (0.79)	2.61 (0.77)	2.39 (0.57)	2.87 (0.74)	13.35 (2.50)

6	2.22 (0.98)	1.96 (1.08)	3.00 (0.93)	2.87 (0.95)	2.22 (0.51)	1.87 (0.85)	14.14 (2.77)
1	3.00 (0.78)	2.57 (0.71)	1.65 (0.70)	2.61 (1.05)	2.57 (0.65)	2.48 (0.65)	14.88 (2.68)
3	2.48 (0.58)	1.22 (0.66)	3.26 (0.67)	2.78 (0.72)	2.61 (0.49)	2.65 (0.56)	15 (2.25)
12	2.48 (0.71)	2.70 (0.69)	2.48 (0.83)	2.70 (0.55)	2.52 (0.71)	2.17 (0.87)	15.05 (2.49)
9	2.96 (0.62)	2.57 (0.65)	2.61 (0.64)	2.91 (0.58)	2.96 (0.55)	3.17 (0.48)	17.18 (2.08)
10	2.61 (0.71)	3.22 (0.66)	2.57 (0.77)	2.74 (0.85)	3.30 (0.62)	3.22 (0.51)	17.66 (2.44)
8	3.09 (0.72)	2.96 (0.75)	3.26 (0.67)	3.04 (0.55)	3.17 (0.56)	3.04 (0.46)	18.56 (2.26)
7	3.39 (0.77)	2.96 (0.91)	3.30 (0.62)	3.57 (0.58)	3.13 (0.68)	2.96 (0.55)	19.31 (2.94)
11	2.78 (0.83)	3.04 (0.81)	3.65 (0.56)	3.39 (0.77)	3.43 (0.58)	3.04 (0.46)	19.33 (2.65)

기술 통계 결과를 구체적으로 살펴보면, 최고점은 11번 학생의 글(19.33점)로 확인되었으며, 최하점은 2번 학생의 글(11.74점)로 나타났다. 최고점을 기록한 11번 학생의 글은 AWE를 활용하여 고쳐쓰기를 수행한 ‘수정고’인 반면, 최하점인 2번 학생의 글은 AWE의 도움을 받지 않은 ‘초고’이다. 상위 6편 중 4편은 모두 AWE 수정고(7번, 8번, 10번, 11번)에 해당하는 글이며, 나머지 2편(9번, 12번)의 경우에도 Z점수를 기준으로 초고 내에서 비교적 높은 평가를 받은 글이라는 점이 특징적이다.

초고와 AWE 수정고 간의 채점 기준별 평가 결과에 유의미한 차이가 발생하는지 검증하기 위해 대응표본 T 검정(Paired Samples T-test) 및 윌콕슨 부호 순위 검정(Wilcoxon signed-rank test)을 실시하였다. 결과는 <표 7>과 같다.

〈표 7〉 초고와 AWE 수정고의 점수에 대한 검정 결과

채점 영역(기준)	구분	M (SD)	t	W
1. 내용의 타당성	초고	2.49 (0.55)		0.00 ***
	AWE 수정고	2.76 (0.52)		
2. 예상 반론 및 반박의 타당성	초고	2.41 (0.56)	0.09	
	AWE 수정고	2.40 (0.45)		
3. 형식 문단의 구분과 배열	초고	2.09 (0.37)		0.00 ***
	AWE 수정고	3.05 (0.40)		
4. 내용 문단의 적절성	초고	2.49 (0.55)	-5.36 ***	
	AWE 수정고	3.02 (0.42)		
5. 문장과 어휘 사용의 정확성	초고	2.49 (0.38)	-7.57 ***	
	AWE 수정고	3.01 (0.34)		
6. 어문 규범과 관습의 정확성	초고	2.36 (0.34)	-6.92 ***	
	AWE 수정고	2.96 (0.37)		
전체 총점	초고	2.39 (0.28)	-14.43 ***	
	AWE 수정고	2.87 (0.29)		

***p<.001

분석 결과 ‘예상 반론 및 반박의 타당성’ 항목을 제외한 모든 채점 기준 및 전체 총점에서 AWE 수정고가 초고보다 통계적으로 유의미하게 높은 평가를 받은 것으로 나타났다($p<.001$). 이는 AWE를 통한 고쳐쓰기가 글의 표층적 측면뿐만 아니라 내용의 타당성, 논리적 구조 등을 개선하는 데 효과적인 도구로 작용했음을 증명한다.

다만 ‘예상 반론 및 반박의 타당성’ 기준에서는 초고($M=2.41$)와 AWE 수정고($M=2.40$) 사이에 유의미한 차이가 도출되지 않았다. 이러한 결과는 해당 영역에서 AWE의 효과가 없는 것으로 해석될 여지가 있으나, 실제 텍스트를 분석한 결과 AWE 수정고에 해당하는 3번(1.22점)과 4번(1.00점) 글의 경우 ‘예상 반론 및 반박’에 관한 서술 자체가 거의 포함되지 않아 채점자들

이 최저점에 가까운 점수를 부여한 것으로 확인되었다.

이는 논술형 평가 문항의 <조건>에 예상 반론 및 반박을 반드시 포함할 것을 안내했음에도 불구하고 해당 글을 작성한 학생들이 해당 평가 요소를 자신의 글에 충분히 반영하지 않았기 때문인 것으로 판단된다. 이후 학생들이 고쳐쓰기 단계에서도 해당 평가 요소를 충분히 반영하지 않았고, 이에 대한 AWE의 피드백 또한 부재했거나 부족했던 것으로 보인다. 결과적으로 이는 개별 텍스트의 특수성에 기인한 것으로 제한된 표본 수, 특정 사례의 편향으로 인한 결과로 이해할 수 있을 것이다.

2) 교사의 인지 부담에 대한 대응표본 T 검정

다음으로 채점자가 초고와 AWE 수정고에 대한 글쓰기 피드백을 제시하는 과정에서 인지적 부담의 차이가 발생하였는지를 분석하였다. 결과는 <표 8>과 같다.

<표 8> 교사의 인지 부담에 대한 대응표본 T 검정 결과

인지 부담 항목	구분	M (SD)	t
1. 어문 규범 오류 (역코딩)	초고	2.92 (0.72)	-7.60 ***
	AWE 수정고	3.75 (0.63)	
2. 문장 가독성 및 의도 파악	초고	2.91 (0.49)	-7.17 ***
	AWE 수정고	3.71 (0.53)	
3. 문단 구분 및 흐름 파악	초고	2.92 (0.53)	-16.32 ***
	AWE 수정고	4.11 (0.45)	
4. 논리 구조 분석 여유	초고	3.59 (0.58)	-5.90 ***
	AWE 수정고	4.11 (0.57)	
5. 전개 방식 조연 가치	초고	3.70 (0.59)	-2.39 *
	AWE 수정고	4.07 (0.60)	

*p<.05, ***p<.001

분석 결과 교사가 텍스트의 표면적 오류를 처리하는 데 투입하는 인지적 자원인 ‘어문 규범 오류’ 항목에서 통계적으로 유의미한 차이가 확인되었다($t=-7.60, p<.001$). 이는 AWE가 글에 드러난 형식적, 표층적 오류를 수정하도록 제안하고 학습자가 이를 반영, 수정함으로써, 맞춤법 및 띄어쓰기 등에 대한 교사의 쓰기 피드백 부담을 어느 정도 경감시켰음을 보여준다.

다음으로 ‘문장 가독성 및 의도 파악’($t=-7.17, p<.001$)과 ‘문단 구분 및 흐름’ ($t=-16.32, p<.001$) 항목에서도 유의미한 차이가 관찰되었다. 이는 AWE가 학생의 글에서 잘못된 형식의 문장을 바로잡고, 형식 문단을 구분하도록 제안하고, 학습자가 이를 수정, 반영함으로써 글의 의미 구조를 빠르게 파악해야 하는 교사의 인지 부담을 어느 정도 경감시켰음을 의미한다.

마지막으로 ‘논리 구조 분석 여유’ ($t=-5.90, p<.001$), ‘전개 방식 조언 가치’ ($t=-2.39, p<.05$) 항목에서도 유의미한 차이가 관찰되었다. 이는 AWE가 형식적, 표층적 측면에서 글의 질을 개선하여 교사가 텍스트의 내용적, 구조적 측면에 집중할 수 있는 환경적 토대를 조성할 수 있음을 의미한다.

2. 교사의 쓰기 피드백 유형 분석

1) 초고 및 AWE 수정고의 쓰기 피드백 유형 분석

채점자가 초고와 AWE 수정고에 제시한 글쓰기 피드백의 개수에 차이가 발생하였는지를 분석하였다. 결과는 <표 9>와 같다.

<표 9> 초고 쓰기 피드백의 개수(개) 및 비율(%)

글	맞춤법	문장 구조	서식	단어 선택	조직구성	아이디어 상술	문체	합계
1	95 (33.33)	27 (9.47)	19 (6.67)	42 (14.74)	27 (9.47)	68 (23.86)	7 (2.46)	285 (100)
2	41 (16.87)	28 (11.52)	20 (8.23)	32 (13.17)	29 (11.93)	79 (32.51)	14 (5.76)	243 (100)

5	128 (41.97)	16 (5.25)	22 (7.21)	24 (7.87)	27 (8.85)	81 (26.56)	7 (2.30)	305 (100)
6	136 (33.25)	32 (7.82)	13 (3.18)	69 (16.87)	17 (4.16)	106 (25.92)	36 (8.80)	409 (100)
9	41 (17.37)	32 (13.56)	13 (5.51)	24 (10.17)	22 (9.32)	98 (41.53)	6 (2.54)	236 (100)
12	108 (33.75)	40 (12.50)	0 (0.00)	28 (8.75)	25 (7.81)	93 (29.06)	26 (8.13)	320 (100)
합계	549 (30.53)	175 (9.73)	87 (4.84)	219 (12.18)	147 (8.18)	525 (29.20)	96 (5.34)	1798 (100)

〈표 10〉 AWE 수정고 쓰기 피드백의 개수(개) 및 비율(%)

글	맞춤법	문장 구조	서식	단어 선택	조직구성	아이디어 상술	문체	합계
3	56 (19.86)	17 (6.03)	9 (3.19)	57 (20.21)	19 (6.74)	110 (39.01)	14 (4.96)	282 (100)
4	5 (1.91)	31 (11.83)	15 (5.73)	43 (16.41)	25 (9.54)	125 (47.71)	18 (6.87)	262 (100)
7	68 (28.10)	26 (10.74)	11 (4.55)	38 (15.70)	17 (7.02)	77 (31.82)	5 (2.07)	242 (100)
8	38 (19.49)	13 (6.67)	4 (2.05)	37 (18.97)	19 (9.74)	71 (36.41)	13 (6.67)	195 (100)
10	9 (4.33)	7 (3.37)	15 (7.21)	34 (16.35)	27 (12.98)	105 (50.48)	11 (5.29)	208 (100)
11	35 (21.47)	5 (3.07)	9 (5.52)	30 (18.40)	7 (4.29)	65 (39.88)	12 (7.36)	163 (100)
합계	211 (15.61)	99 (7.32)	63 (4.66)	239 (17.68)	114 (8.43)	553 (40.90)	73 (5.40)	1352 (100)

초고와 AWE 수정고에 부여된 쓰기 피드백의 개수를 비교한 결과, 초고 집단에 제시된 피드백의 총합은 1,798개인 반면, AWE 수정고 집단에 제시된 피드백은 1,352개로 나타났다. 이러한 쓰기 피드백의 감소는 학습자가

AWE를 활용하여 고쳐쓰기를 수행하는 과정에서 맞춤법, 띄어쓰기, 단순 문장 오류 등 글의 형식적, 표층적 문제를 스스로 점검하고 1차적으로 수정했기 때문인 것으로 해석된다. 그 결과 채점자가 피드백할 수 있는 오류의 빈도가 상대적으로 줄어들어, 쓰기 피드백 총량의 변화로 이어진 것이다. 이는 AWE가 쓰기 피드백에 대한 교사의 물리적, 인지적 부담을 줄여주는 보조 도구로 활용될 수 있음을 보여준다.

개별 텍스트 단위로 쓰기 피드백 발생 빈도를 살펴보면, 쓰기 피드백이 가장 많이 제시된 글은 초고 집단의 6번 글(409개)이었으며, 가장 적게 제시된 글은 AWE 수정고 집단의 11번 글(163개)이었다. 여기서 주목할 만한 점은 11번 글이 전체 글 평가에서 최고점(19.33점)을 기록한 글이라는 사실이다. 이는 텍스트의 질적 수준과 교사가 부여하는 쓰기 피드백의 양 사이에 일정한 상관관계가 성립할 가능성을 보여준다. 그러나 본 연구는 채점자 개인의 쓰기 피드백 제시 성향 등을 충분히 통제하지 못한 상황에서 다소 적은 표본으로 진행된 연구이므로 이 같은 해석에는 다소 유의할 필요가 있다. 가령 가장 많은 쓰기 피드백이 제시된 6번 글의 경우 다른 글(2, 4, 5번)보다 채점 점수가 높았음에도 100건 정도 더 많은 쓰기 피드백이 부여되었기 때문이다. 이는 쓰기 피드백의 양이 점수뿐만 아니라 개별 텍스트가 지닌 특성(길이, 내용 등)과 관련이 있을 수도 있음을 의미한다.

2) 채점 기준별 쓰기 피드백 유형 분석

초고와 AWE 수정고에 제시된 글쓰기 피드백 유형별 비율에 차이가 존재하는지를 분석하기 위해 대응표본 T 검정(Paired Samples T-test) 및 윌콕슨 부호 순위 검정(Wilcoxon signed-rank test)을 실시하였다. 결과는 <표 11>과 같다.

〈표 11〉 채점 기준별 쓰기 피드백 유형 비율(%)에 대한 검정 결과

피드백 유형	구분	M (SD)	t	W
1. 맞춤법	초고	30.48(10.47)	6.94 ***	
	AWE 수정고	16.16(6.82)		
2. 문장 구조	초고	9.92(6.67)		179.0
	AWE 수정고	7.35(5.20)		
3. 서식	초고	4.46(3.65)	0.46	
	AWE 수정고	4.16(4.28)		
4. 단어 선택	초고	12.15(5.00)	-4.24 ***	
	AWE 수정고	17.10(7.83)		
5. 조직 구성	초고	8.79(7.25)	-0.15	
	AWE 수정고	8.96(9.04)		
6. 아이디어 상술	초고	29.03(10.28)	-5.85 ***	
	AWE 수정고	41.02(11.94)		
7. 문체	초고	5.17(4.20)	-0.11	
	AWE 수정고	5.26(5.25)		

***p < .001

저차원적(Lower-level) 영역의 대표적인 항목인 ‘맞춤법’의 쓰기 피드백 양상을 살펴보면, 초고에서는 쓰기 피드백의 평균 비율이 30.48%였으나 AWE 수정고에서는 16.16%로 나타나 유의미하게 감소한 것으로 나타났다($t=6.94$, $p<.001$). 이는 AWE가 띄어쓰기, 문장 부호 등 규칙 기반의 표면적 오류를 사전에 점검하고 상당 부분 수정하도록 유도했기 때문인 것으로 해석된다.

반면 저차원적 영역에 속하는 ‘단어 선택’ 영역에서는 ‘맞춤법’과 대조적인 양상이 확인되었다. 초고($M=12.15$)에 비해 AWE 수정고($M=17.10$)에서 교사의 쓰기 피드백이 유의미하게 증가한 것이다($t=-4.24$, $p<.001$). 이는 텍

스트의 형식적, 표층적 오류가 어느 정도 해소된 단계에서, 교사의 주의와 인지적 초점이 텍스트의 질적 수준을 높이는 ‘어휘의 정확성, 적절성, 다양성’ 측면으로 이동했기 때문으로 추측된다. 텍스트를 더욱 세밀하게 읽으며 더 정교한 어휘 사용을 제안하고 지도하는 방향으로 쓰기 피드백 제시 양상이 달라졌으리라 추론할 수 있을 것이다.

고차원적(Higher-level) 영역에서는 ‘아이디어 및 상술’의 쓰기 피드백이 유의미하게 증가했다는 사실에 주목할 필요가 있다. 초고(M=29.03) 대비 AWE 수정고(M=41.02)에서 쓰기 피드백 비율이 가장 큰 폭으로 상승하였는데($t=-5.85$, $p<.001$), 채점자가 기초적인 오류가 수정된 텍스트를 읽으며, 주장, 이유, 근거의 타당성, 신뢰성, 구체성 등에 주목하게 된 것으로 추측된다.

마지막으로 저차원적 기능인 ‘문장 구조’와 ‘서식’, 그리고 고차원적 기능인 ‘조직 구성’과 ‘문체’ 영역에서는 초고와 AWE 수정고 간에 통계적으로 유의미한 쓰기 피드백 비율의 차이가 발견되지 않았다. 저차원적 피드백에 속하는 ‘문장 구조’와 ‘서식’의 경우 초고보다 AWE 수정고에서 비율이 낮아지기는 했으나, 통계적으로 유의하지는 않았다. 고차원적 피드백에 속하는 ‘조직 구성’ 유형에서 유의미한 변화가 나타나지 않은 것은, 앞서 살펴본 바와 같이 키워(KEEwi)의 평가 지표 중 하나인 ‘조직과 전개’ 피드백이 문장 단위의 표층적 수준에 머무르고 있다는 기술적 특성과 무관하지 않다. 즉, 현재의 AWE가 제공하는 피드백만으로는 글의 논리적 구조에 대한 학습자의 고쳐쓰기를 유도하지 못했음을 의미하는 것일 수 있다. 이 같은 결과는 AWE의 활용이 쓰기 피드백 유형을 일률적으로 변화시키지 않음을 의미한다. 물론 이는 본 연구의 제한된 표본 수와 단일 과제라는 특수성에서 기인한 결과일 수 있으므로, 향후 후속 연구를 통해 보다 장기적이고 다각적인 검토가 필요할 것이다.

V. 결론 및 논의

본 연구의 목적은 자동 채점 및 피드백 시스템(AWE)의 교육적 실효성을 실증적으로 검증하는 것이다. 구체적으로 AWE가 논술형 평가의 확대에 가중되는 교사의 채점 및 피드백 부담을 완화할 수 있는 가능성이 있는지를 확인하고자 하였다. 이를 위해 AWE의 도입이 교사의 쓰기 피드백 유형을 어떻게 변화시키는지 분석하였다.

연구 결과 AWE를 활용해 고쳐쓰기를 수행한 수정고는 초고와 비교해 ‘예상 반론 및 반박’을 제외한 모든 채점 기준에서 통계적으로 유의미하게 높은 점수를 받았다. 또한 교사들은 AWE 수정고의 분량이 초고보다 많음에도 불구하고, 해당 유형의 글에 피드백할 때 어문 규범 오류를 처리하거나 문장의 가독성 및 문단 흐름을 파악하는 데 드는 인지적 부담이 크게 감소했다고 응답했다. 이는 AWE가 학생 글의 형식적, 표층적 오류를 일차적으로 점검하고 교정하도록 유도함으로써, 교사가 텍스트의 논리 구조를 분석하고 전개 방식에 대해 조언할 수 있는 인지적 여유를 확보하는 데 기여했을 가능성을 보여준다.

교사의 쓰기 피드백 유형을 분석한 결과, 초고 대비 AWE 수정고에서 맞춤법과 같은 저차원적 쓰기 피드백의 비율은 유의미하게 감소한 반면, 아이디어 및 상술과 같은 고차원적 영역의 쓰기 피드백 비율은 뚜렷하게 증가하였다. 이는 다수의 교사가 글의 형식적이고 표층적인 오류 교정에 집중하는 경향이 있다는 기존의 한계(김남복·권태현, 2022; Lucero et al., 2018)를 AWE가 보완할 가능성을 짐작하게 한다(Wilson & Czik, 2016). 그러나 이와 동시에 저차원적 쓰기 피드백에 속하는 ‘문장 구조’, ‘서식’에는 변화가 없었으며, ‘단어 선택’의 경우에는 오히려 쓰기 피드백 비율이 증가하였다는 점, 고차원적 쓰기 피드백에 속하는 ‘조직’, ‘문체’는 유의한 차이를 보이지 않았다는 점에 유의할 필요가 있다.

위와 같이 부분적으로 나타난 차이에도 불구하고, 논술형 평가 상황에서 AWE의 도입은 교사의 채점 및 피드백에 수반되는 과도한 인지적 부담을 어느 정도 경감하는 도구가 될 수 있을 것으로 판단된다. 기계적인 오류 교정과 형식적인 쓰기 피드백은 AWE라는 ‘하이테크(High-tech)’에 위임하고, 인간 교사는 학습자의 글을 매개로 더 깊이 있게 소통하며 논증의 타당성이나 논리적 조직 등을 짚어주는 ‘하이터치(High-touch)’에 온전히 집중할 수 있다는 것이다. 이는 AWE가 비판적 사고력, 문제 해결 능력 등 논술형 평가가 지향하는 고차원적 사고력을 향상하기 위한 효과적인 수단임을 의미하는 것일 수도 있다. 그러나 해당 연구 결과를 바탕으로 ‘논술형 평가에서 AWE를 도입해야 한다.’라는 단선적인 결론을 추론하기 위해서는 보다 깊은 고민과 논의의 과정이 필요할 것으로 판단된다. 이는 본 연구의 여러 가지 한계와도 직결된다.

첫째, 연구 방법론적 측면에서 여러 한계가 존재한다. 편의 표집, 제한된 연구 대상과 글쓰기 장르, 부족한 표본 수 등으로 인해 연구 결과를 일반화하는 데에는 명백한 제약이 따른다. 또한 쓰기 피드백 분석이 양적 차원에 국한되었으며, 채점자의 인지 부담 측정 또한 자기보고식 설문에 의존하였다. 자기보고는 실제 인지 처리 과정에서 발생한 부담을 직접적으로 포착하지 못하므로 주관적 인상 평가의 한계에서 자유롭지 못하다. 더불어 본 연구의 결과는 특정 시점의 특정 기술이라는 조건과 맥락에 종속된 결과이기도 하다. AWE의 피드백 알고리즘은 빠르게 고도화되고 있으므로, 본 연구에서 확인된 효과와 제한점은 당시 시스템의 조건적 결과로 해석되어야 한다.

둘째, 본 연구에서 확인된 교사 쓰기 피드백 유형의 비중 변화가 교사가 제시하는 쓰기 피드백의 질적 향상을 담보하는 것은 아니라는 점이다. 쓰기 피드백의 질은 타당성, 구체성, 그리고 제시 방식의 적절성 등을 포괄하는 개념이다. 유의미한 쓰기 피드백이 갖추어야 할 여러 교육적 조건이 충족되고, 이러한 쓰기 피드백이 학습자에게 실질적으로 수용되어 내면화될 때 비로소 쓰기 능력의 신장으로 이어질 수 있다. 따라서 AWE의 개입으로 고차원적 쓰

기 피드백의 비율이 증가했다는 결과만으로 그것이 질적으로 우수한 피드백이 되리라고 단정하는 것은 성급한 판단일 수 있다. 따라서 향후 연구에서는 AWE의 개입으로 변화된 인간 교사의 쓰기 피드백의 질적 수준을 심층적으로 분석하는 작업이 이루어져야 할 것으로 판단된다.

셋째, 본 연구에서 저차원적 쓰기 피드백의 비율 감소를 긍정적 변화로 해석하였으나, 이것이 맞춤법, 문장 구조, 서식, 어휘 측면에서 제공되는 쓰기 피드백의 가치가 낮다는 것을 의미하지는 않는다. 이것들 또한 글쓰기 관습 준수, 정확한 의사소통을 위해 반드시 습득되어야 할 역량에 해당한다. 문제는 이러한 역량의 습득 경로에 있다. AWE를 통해 기계적으로 오류를 수정하는 경험이 반복된다 하더라도, 그것이 학습자의 장기기억에 부호화되어 독립적인 글쓰기 역량으로 전이될 수 있는지는 별개의 문제일 것이다. 즉, 교사의 피드백 부담을 경감한다는 실용적 이점이 곧 학습자의 학습 효과를 보증하는 것은 아니다. 따라서 AWE를 활용한 저차원적 쓰기 피드백의 수용과 활용이 어떠한 교육적 조건 하에서 실질적인 역량 전이로 이어질 수 있는지에 대한 후속 연구가 반드시 필요하다.

넷째, AWE의 활용이 글쓰기 결과물의 질을 표면적으로 개선할 수는 있으나, 이것이 학습자의 ‘글쓰기 능력’ 자체를 변화시켰다고 보기는 어렵다. 본 연구의 결과는 AWE의 도움을 받아 다듬어진 최종 텍스트의 완성도가 높아졌음을 보여줄 뿐, AWE를 배제한 상태에서도 학습자의 독립적인 글쓰기 역량이 향상되었는지를 검증한 것은 아니다. 따라서 향후 연구에서는 장기 간에 걸쳐 AWE를 활용한 쓰기 수업을 진행한 후, AWE의 지원 없이도 학습자의 글쓰기 능력이 유의미하게 성장했음을 실험적으로 증명해 내는 과정이 필수적이다.

마지막으로, 쓰기 피드백의 궁극적인 효용성은 쓰기 피드백을 주체적으로 수용하고 활용하는 학습자의 ‘피드백 리터러시(feedback literacy)’ 차원에서 더욱 깊이 있게 연구되어야 한다(김형성, 2025 ㄱ). AWE나 교사가 텍스트의 오류를 찾아내 해결책을 제시하는 ‘진단과 처방’ 측면의 일방적이고 기

계적인 방식으로 텍스트를 매개로 이루어지는 쓰기 피드백의 역동적인 측면을 온전히 포착할 수 없다. 따라서 향후 연구에서는 학습자가 다양한 층위의 피드백을 어떻게 인지하고 해석하며, 자신의 글쓰기 과정에 주도적으로 통합해 나가는지를 분석할 필요가 있다. 이 같은 후속 연구가 뒷받침되어야만 AWE를 활용한 쓰기 피드백 연구의 결과물들이 보다 온전한 의미를 획득할 수 있을 것이다.

최근 들어 교육 현장에는 다양한 AI 기술들이 도입되고 있다. 교사의 노력과 부담을 경감하려는 목표 아래, 교수 학습 현장에 AI 기술을 통합하려는 시도 또한 지속되고 있다. 본 연구는 이 같은 흐름 속에서 대표적인 AI 기술인 AWE의 효과성을 실증적으로 탐색하고자 하였다. 그러나 전술한 것처럼 본 연구는 여러 한계를 내재하기에 연구 결과에서 드러난 효과성만을 근거로 기술의 적극적인 도입과 활용을 주장하기에는 어려움이 있다. 오히려 본 연구가 AWE를 교수·학습 과정에 통합하는 과정에서 반드시 고려해야 할 다양한 쟁점을 제기하였다는 점에서 의의가 있을지도 모른다. 이러한 논의와 숙고의 필요성을 포함한 본 연구가 향후 AWE 기반 쓰기 교육 연구의 방향을 설정하는 데 하나의 기초 자료로 활용되기를 기대해 본다.

* 본 논문은 2026.04.29. 투고되었으며, 2026.05.13. 심사가 시작되어 2026.06.09. 심사가 종료되었음.

참고문헌

- 교육부(2022), 『2022 개정 국어과 교육과정(교육부 고시 제2022-33호)』, 세종: 교육부.
- 교육부(2023), 『AI 디지털 교과서 추진 방안』, 세종: 교육부.
- 권경문·최숙기(2025), 「AI 기반 한국어 글쓰기 자동 피드백의 질 평가 연구」, 『청람어문교육』 104, 205-244.
- 권태현(2024), 「ChatGPT를 활용한 쓰기 채점 및 피드백 방안 - 프롬프트 전략을 중심으로 -」, 『새국어교육』 141, 7-42.
- 김경화(2013), 「학생 글에 대한 교사의 반응 분석 - “꿀맛닷컴 중학사이버논술”의 첨삭 자료를 중심으로 -」, 『작문연구』 17, 101-153.
- 김경희·이명진(2021), 「교수학습과 학생평가 개선을 위한 서·논술형 평가 지침 활용 및 피드백 효과 제고 방안」, 『교육과정평가연구』 24(3), 27-51.
- 김남복·권태현(2022), 「고등학생의 설득적 글쓰기에 대한 교사 피드백 양상 분석」, 『학습자중심교과교육연구』 22(20), 1049-1076.
- 김선우·이경화(2025), 「생성형 인공지능을 활용한 초등학생 고쳐쓰기 양상 연구」, 『한국초등교육』 36(4), 1-15.
- 김소정(2025), 「AI 자동 피드백 프로그램 키위(KEEwi)를 활용한 글쓰기 수업 사례와 교육 효과」, 『한국과 세계』 7(3), 739-763.
- 김정녀·유혜원(2021), 「글쓰기 첨삭문의 텍스트적 성격과 도식 구조 분석 - 교수자 첨삭문 사례 비교를 중심으로 -」, 『한민족어문학』 92, 71-104.
- 김지연(2024), 「글쓰기 수업에서 AI 기반 글쓰기 자동 피드백 프로그램 활용의 효과와 한계 - 독서와표현 교과에서 키위(KEEwi) 활용 사례를 중심으로 -」, 『영주어문』 58, 355-386.
- 김형성(2025), 「고등학생의 글쓰기 피드백 리터러시에 대한 잠재프로파일 분석」, 『국어교육학연구』 60(2), 71-106.
- 김형성(2025), 「논술형 평가에 대한 국어 교사의 어려움 분석 연구 - 평가 리터러시(Assessment Literacy)의 ‘사회문화적 및 학교 맥락’을 중심으로 -」, 『청람어문교육』 104, 111-141.
- 남가영·김호정(2023), 「서술형·논술형 평가 실행에 관한 국어 교사의 최적화 행동 분석」, 『교과교육연구』 27(1), 31-50.
- 대통령직속4차산업혁명위원회(2022), 『디지털 대전환 시대 교육 정책 방향 제안(제27차 4차산업혁명위원회 심의안건 제3호)』, 서울: 대통령직속4차산업혁명위원회.
- 민병곤·남가영·김선희·장성민·이성준·권은선(2023), 『2023년 국민의 글쓰기 능력 진단 체계 개발』, 서울: 국립국어원.
- 박준범(2014), 「글쓰기 학습자의 첨삭논평 수용 양상에 관한 고찰」, 『교양교육연구』 8(6), 271-305.
- 박혜영·김성숙·김경희·이명진·김광규·김지영(2019), 『수업-평가 연계 강화를 통한 서·논

- 술형 평가 내실화 방안(RRE 2019-6)』, 충북: 한국교육과정평가원.
- 백다빈·이경화(2025), 「고쳐쓰기 단계에서의 인공지능과 교사의 피드백 비교 연구」, 『초등교육 연구』 38(4), 53-77.
- 신윤범·김하정·원효현(2025), 「국내 서·논술형 평가 자동 채점의 효용에 관한 메타분석」, 『교육평가연구』 38(4), 999-1020.
- 안혜연(2025), 「서·논술형 평가에서 생성형 AI 활용 가능성 탐색: 리젠트 시험 데이터를 중심으로」, 『교육평가연구』 38(3), 823-846.
- 오유경·김경선(2025), 「다국면라쉬모형을 활용한 ChatGPT-4o의 논술형 평가 자동채점 정확도 및 채점자 효과 분석」, 『교육평가연구』 38(3), 769-793.
- 윤준채(2023), 「읽기 연구에 대한 연구 방법론적 검토 - 실험 연구의 내적 타당도를 중심으로 -」, 『한국초등국어교육』 76, 384-399.
- 이지원·박영민(2015), 「채점 수 누적에 따른 쓰기 평가 과정의 눈동자 움직임 연구」, 『국어교육학연구』 50(4), 395-424.
- 정분의(2019), 「학생 글 채점 누적에 따른 국어교사의 인지 부담 변화」, 『작문연구』 42, 30-57.
- 차준우·최서희·이기성(2024), 「인공지능 기반 쓰기 평가 도구와 글쓰기 평가자의 비교 연구 - 독후감 텍스트를 중심으로 -」, 『국제어문』 102, 681-705.
- 최숙기(2012), 「국어과 수행 평가의 평가자 신뢰도 보고 방안 탐색 -고등학생 요약문 평가 결과 중심-」, 『작문연구』 14, 395-424.
- 최진영·김형성·송보라·김지수(2023), 「생성형 인공지능 ChatGPT를 활용한 고등학생의 작문 과제 고쳐쓰기 양상」, 『리터러시 연구』 14(6), 79-119.
- Bai, L. & Hu, G. (2017), "In the face of fallible AWE feedback: How do students respond?", *Educational Psychology* 37(1), 67-81.
- Barkaoui, K. (2010), "Variability in ESL essay rating processes: The role of the rating scale and rater experience", *Language Assessment Quarterly* 7(1), 54-74.
- Cho, K., Schunn, C. D., & Charney, D. (2006), "Commenting on writing: Typology and perceived helpfulness of comments from novice peer reviewers and subject matter experts", *Written Communication* 23(3), 260-294.
- Cumming, A. (1990), "Expertise in evaluating second language compositions", *Language Testing* 7(1), 31-51.
- Flower, L. & Hayes, J. R. (1981), "The pregnant pause: An inquiry into the nature of planning", *Research in the Teaching of English* 15(3), 229-244.
- Hayes, J. R. & Berninger, V. W. (2010), Relationships between idea generation and transcription: How the act of writing shapes what children write, In C. Bazerman, R. Krut, K. Lunsford, S. McLeod, S. Null, P. Rogers & A. Stansell (Eds.), *Traditions of writing research*, New York: Routledge.
- Jukiewicz, M. & Wyrwa, M. (2026), "Can ChatGPT replace the teacher in assessment? A review of research on the use of large language models in grading and providing

- feedback", *Applied Sciences* 16(2), 680.
- Link, S., Mehrzad, M., & Rahimi, M. (2022), "Impact of automated writing evaluation on teacher feedback, student revision, and writing improvement", *Computer Assisted Language Learning* 35(4), 605-634.
- Lucero, M., Fernández, M. J., & Montanero, M. (2018), "Teachers' written feedback comments on narrative texts in elementary and secondary education", *Studies in Educational Evaluation* 59, 158-167.
- Mariano, M. J., Taborere, J. J., Gampal, A., Dawisan, A., Maraganas, M. G., & Garcia, R. (2025), "Analysis of AI and teacher feedback in grading students' activities", *International Journal on Studies in Education* 7(4), 1028-1047.
- Matsumura, L. C., Patthey-Chavez, G. G., Valdes, R., & Garnier, H. (2002a), "Measuring instructional quality in accountability systems: Classroom assignments and student achievement", *Educational Assessment* 8(3), 207-229.
- Matsumura, L. C., Patthey-Chavez, G. G., Valdes, R., & Garnier, H. (2002b), "Teacher feedback, writing assignment quality, and third-grade students' revision in low- and higher-achieving urban schools", *The Elementary School Journal* 103(1), 3-26.
- Naisbitt, J. (1982), *Megatrends: Ten new directions transforming our lives*, New York: Warner Books.
- Olson, M. W. & Raffeld, P. (1987), "The effects of written comments on the quality of student compositions and the learning of content", *Reading Psychology* 8(4), 273-293.
- Page, E. B. (1966), "The imminence of... grading essays by computer", *The Phi Delta Kappan* 47(5), 238-243.
- Palermo, C. & Thomson, M. M. (2018), "Teacher implementation of Self-Regulated Strategy Development with an automated writing evaluation system: Effects on the argumentative writing performance of middle school students", *Contemporary Educational Psychology* 54, 255-270.
- Patchan, M. M., Schunn, C. D., & Correnti, R. J. (2016), "The nature of feedback: How peer feedback features affect students' implementation rate and quality of revisions", *Journal of Educational Psychology* 108(8), 1098-1120.
- Ramesh, D. & Sanampudi, S. K. (2022), "An automated essay scoring systems: A systematic literature review", *Artificial Intelligence Review* 55(3), 2495-2527.
- Shi, H. & Aryadoust, V. (2023), "A systematic review of automated writing evaluation systems", *Education and Information Technologies* 28(1), 771-795.
- Stern, L. A. & Solomon, A. (2006), "Effective faculty feedback: The road less traveled", *Assessing Writing* 11(1), 22-41.

- Stevenson, M. & Phakiti, A. (2014), "The effects of computer-generated feedback on the quality of writing", *Assessing Writing* 19, 51-65.
- Wade-Stein, D. & Kintsch, E. (2004), "Summary Street: Interactive computer support for writing", *Cognition and Instruction* 22(3), 333-362.
- Ware, P. (2014), "Feedback for adolescent writers in the English classroom", *Writing & Pedagogy* 6(2), 223-249.
- Weigle, S. C. (2013), "English language learners and automated scoring of essays: Critical considerations", *Assessing Writing* 18(1), 85-99.
- Wilson, J. & Cziki, A. (2016), "Automated essay evaluation software in English Language Arts classrooms: Effects on teacher feedback, student motivation, and writing quality", *Computers & Education* 100, 94-109.
- Winke, P. & Lim, H. (2015), "ESL essay raters' cognitive processes in applying the Jacobs et al. rubric: An eye-movement study", *Assessing Writing* 25, 38-54.
- Wolfe, E. W., Kao, C. -W., & Ranney, M. (1998), "Cognitive differences in proficient and nonproficient essay scorers", *Written Communication* 15(4), 465-492.
- Zhai, N. & Ma, X. (2023), "The effectiveness of automated writing evaluation on writing quality: A meta-analysis", *Journal of Educational Computing Research* 61(4), 875-900.
- Zhang, Z. V. & Hyland, K. (2018), "Student engagement with teacher and automated feedback on L2 writing", *Assessing Writing* 36, 90-102.
- Zupanc, K. & Bosnić, Z. (2017), "Automated essay evaluation with semantic analysis", *Knowledge-Based Systems* 120, 118-132.

논술형 평가에서 자동 채점 및 피드백 시스템(AWE)은 인간 교사의 쓰기 피드백 유형을 변화시키는가

김형성

본 연구의 목적은 자동 채점 및 피드백 시스템(AWE, Automated Writing Evaluation)이 논술형 평가 상황에서 인간 교사의 쓰기 피드백 유형을 변화시키는지를 탐색하는 것이다. 이를 위해 고등학교 2학년 학습자를 대상으로 논술형 평가를 실시하고 키위(KEEwi)를 활용한 고쳐쓰기를 진행한 뒤, 초고 6편과 AWE 수정고 6편을 선정하였다. 이후 국어 교사 23명에게 채점·피드백을 의뢰하고, 통계적 분석을 실시하였다.

분석 결과, AWE 수정고는 대부분의 채점 기준에서 초고보다 유의미하게 높은 평가를 받았으며, 교사들은 AWE 수정고를 피드백할 때 ‘어문 규범 처리’와 ‘문장, 문단 파악에 따른 인지 부담’이 감소하는 것으로 나타났다. 쓰기 피드백의 유형을 비교한 결과 ‘맞춤법’과 같은 저차원적 피드백 비율은 감소한 반면, ‘아이디어 및 상술’과 같은 고차원적 피드백 비율은 증가하였다. 본 연구는 AWE 시스템을 교수학습 과정에 통합하는 과정에서 고려해야 할 다양한 시사점을 제언한다.

핵심어 논술형 평가, AWE, 자동 채점, 쓰기 피드백, 피드백

Does Automated Writing Evaluation(AWE) Change Human Teachers' Writing Feedback Patterns in Essay - Type Assessment?

Kim Hyungsung

This study empirically examines whether automated writing evaluation (AWE) systems meaningfully change human teachers' writing feedback patterns in essay-type assessment. Seventy high school students completed essays and revised them using KEEwi, a Korean-language AWE program, and six initial and six AWE-revised drafts were selected through purposive sampling and were scored and given feedback by 23 Korean language teachers.

The results showed that AWE-revised drafts received significantly higher scores than initial drafts, and teachers reported reduced cognitive load when providing feedback. The proportion of lower-level feedback, such as feedback on spelling, decreased significantly, whereas the proportion of higher-level feedback, such as feedback on ideas and elaboration, increased markedly. This study has implications for integrating AWE systems into teaching and learning processes.

KEYWORDS Essay-type assessment, AWE, automated scoring, writing feedback, feedback