교육과정평가연구
The Journal of Curriculum and Evaluation
2025, Vol. 28, No. 4, pp.153~187
DOI: https://doi.org/10.29221/jce.2025.28.4.153

논술형 평가에서 AI 협업 기반 채점자 협의와 전통적 채점자 협의의 양상 비교[1)]

김지수 (첨단고등학교 교사)*
최진영 (일산동고등학교 교사)**
김형성 (카이스트 부설 한국과학영재학교 전임교원)
송보라 (김해대청고등학교 교사)

요약

본 연구는 논술형 평가 상황에서 AI 협업 기반 채점자 협의와 전통적 채점자 협의를 비교·분석하여 채점 신뢰도 확보와 채점자 부담 완화 방안을 실증적으로 탐색하는 데 목적이 있다. 이를 위해 설명적 순차 설계에 따라 먼저 양적 연구를 실시하고, 그 결과를 토대로 질적 연구를 진행하였다. 양적 연구에서는 개별 채점 결과를 확인한 뒤, 협의 후 결과를 Rasch 분석하여 집단별 채점자의 엄격성과 일관성을 비교하고 협의 시간과 난도를 확인하였다. 질적 연구에서는 반구조화된 심층 인터뷰를 통해 평가 과정의 어려움, 협의 특징, AI 활용 인식을 탐색하였다. 연구 결과, 전통적 협의 집단은 채점 기준 해석 차이에 따른 불일치와 관계적 긴장을 어려움으로 인식하였으나, 타 채점자의 관점 공유와 채점 경향 성찰의 긍정적 효과도 확인되었다. 반면 AI 협의 집단은 AI 점수가 합의의 기준점으로 작용해 초기 기준 설정과 채점자 간 중재를 용이하게 했으나, 과도한 의존으로 협의가 점수 중심으로 축소될 위험성이 나타났다. 또한 AI 협의 집단은 전통적 협의보다 시간을 더 길게 소요하고 난도를 높게 인식하였는데, 이는 AI의 채점 근거와 피드백이 추가적 쟁점을 제공해 논의의 심층성을 높이는 동시에 채점자 부담을 가중시킨 결과였다. 종합하면, AI 협업 기반 채점자 협의는 전통적 협의를 대체하기보다 보완할 때 채점 신뢰도와 협의의 심층성 제고에 기여할 수 있음을 확인하였다. 이를 바탕으로 학교 현장에서 AI 협업 기반 채점자 협의 모델 도입 시 고려사항을 제언하였다.

주제어: AI 협업 기반 채점자 협의, AI 자동채점, 논술형 평가, 채점자 협의, 작문 평가

1) 본 논문은 제64회 한국작문학회 학술대회(2025.9.20.)에서 발표한 것을 수정·보완한 것임.

* 제1저자, soo979@naver.com
** 교신저자, fakemilk@naver.com

I. 서 론

AI 등의 디지털 기술의 발전에 따라 창의성, 문제해결력 등의 미래 역량을 함양할 수 있는 학생 평가의 방법으로 서·논술형 평가가 강조되고 있다. 특히 최근 교육부의 〈미래 사회를 대비하는 2028학년도 대학입시제도 개편 확정안(2023.12.27.)〉과 2022 개정 교육과정, 시·도 교육청 차원의 정책을 보면, 앞으로 서·논술형 평가는 정책적으로 더 확대될 예정이다.

그러나 학교 현장에서는 서·논술형 평가의 확대 필요성에 공감하지만, 채점 신뢰도 확보와 교사 부담 가중을 주요 문제로 지적한다(김경희 외, 2024; 김형성, 2025; 박혜영 외, 2019). 특히 논술형 평가는 다양한 응답을 통해 학생의 사고 과정을 심층적으로 평가할 수 있다는 장점에도 불구하고, 채점 기준 해석 차이와 채점자의 일관성 부족으로 신뢰도 확보가 어려워 교사의 부담이 크다는 어려움이 존재한다(박종임 외, 2024). 이에 따라 다수의 선행연구에서는 채점자 간의 기준 공유와 합의를 통한 채점자 협의의 중요성을 강조해 왔으나, 채점자 협의 과정 자체가 채점자에게 또 다른 부담이 되기도 한다.

이러한 배경에서 최근에는 채점의 신뢰도를 보완하면서 채점자의 부담을 완화할 수 있는 방안으로 AI 자동채점 기술이 주목받고 있다. 선행연구들은 AI가 일정 수준의 신뢰도와 효율성을 보이며 채점 시간 단축과 채점자 부담 완화에 효과가 있음을 보고하였고(신상윤·강신천, 2024; Jansen et al., 2024), 교육청 차원에서도 관련 시스템 개발이 추진되고 있다. 그러나 AI 자동채점은 오류 가능성과 고차원적 역량 평가의 한계가 지적되어, 논술형 평가를 단독으로 대체하기에는 부족하다(박소영 외, 2024; 함은혜 외, 2024). 이에 최진영 외(2025)는 논술형 평가 상황에서 채점 신뢰도를 보완하면서 채점자의 부담을 줄일 수 있는 방안으로 AI 자동채점 결과와 채점자 협의를 결합한 ‘AI 자동채점 기반 평가자 협의 모델’을 개발하였으며, 이는 인간의 채점자 협의 상황에서 AI를 활용한 사례라 볼 수 있다.

본 연구는 탐색적 성격의 소규모 연구로 최진영 외(2025)에서 개발한 모델을 실제 논술형 평가 상황에 적용하여, 전통적 채점자 협의와 비교했을 때 ‘AI 자동채점 기반 평가자 협의 모델’이 가지는 효과와 이에 대한 채점자의 인식을 규명하는 것을 목적으로 한다. 나아가 AI가 인간 채점자의 협의 과정에 어떻게 개입하고 조정 역할을 수행하는지를 분석하여, 교사와 AI의 협업을 통한 서·논술형 평가 신뢰도 확보 방안을 제시하고, 향후 교육평가에서의 AI 활용 가능성과 교육적 시사점을 도출하고자 한다. 이를 위해 설명적 순차 혼합 방법 설계를 적용하였으며, 양적 단계에서는 두 협의 집단 간 신뢰도와 부담의 차이를 분석하고, 질적 단계에서는 채점자의 경험과 인식을 탐색하였다. 구체

적인 연구 질문은 다음과 같다.

1. AI 협업 기반 채점자 협의 집단과 전통적 채점자 협의 집단에서 채점 신뢰도 및 채점 부담의 차이는 어떠한가?

2. AI 협업 기반 채점자 협의 집단(이하, AI 협의 집단)과 전통적 채점자 협의 집단(이하, 전통적 협의 집단)에서 채점 및 협의 과정에 대한 채점자의 인식은 어떠한가?

Ⅱ. 이론적 배경

1. 논술형 평가와 채점자 협의

논술형 평가는 기존의 선다형 평가가 지닌 한계를 극복하려는 시도로 국내 교육과정에서는 2009년부터 제도화되었다. 이어 2015 개정 교육과정은 역량 중심 교육과 과정 중심 평가의 중요성을 강조하며 서·논술형 평가의 확대를 요구하였고, 2022 개정 교육과정에서는 평가학적 측면에서 서·논술형 평가를 정의하기에 이른다. 특히 논술형 평가는 “주어진 문제에 대해 주장과 근거를 논리적으로 조직하여 작성하는 평가 방법”으로, 학습자의 자료 분석력, 논증 능력, 표현력과 같은 고차원적 사고 역량을 측정, 진단, 환류하는 평가 도구로 정의된다(교육부, 2022; 박종임, 2024; 최숙기, 2023).

논술형 평가의 이상적인 목적을 실현하기 위해서는 평가의 타당도와 신뢰도 확보가 필수적이다. 2022 개정 교육과정에서 문항의 명료성, 세부 채점 기준의 제시를 강조한 것 또한 이 같은 논술형 평가의 특성에서 비롯된 것이다. 그러나 타당하고 신뢰할 수 있는 문항, 채점 기준만으로 모든 문제가 해결되지는 않는다. 이는 글쓰기 평가의 본질과도 관련된다. 선행연구들은 채점자가 글을 읽고 판단하는 과정에 개인의 지식, 태도, 성별, 연령, 가치관 등 다양한 요인이 영향을 미친다고 보고한다(Engelhard, 1994; Linacre, 1989).

특히 중요한 채점자의 특성은 엄격성(severity)과 일관성(consistency)으로, 이는 채점자 신뢰도와 직결된다. 동일한 글이라도 채점자에 따라 점수를 높게 또는 낮게 부여하는 경우가 있는데, 이는 채점자의 엄격성(severity) 차이에서 비롯된다(McNamara & Adams, 1991; McNamara, 1996). 엄격성 차이는 채점자가 지닌 고유한 평가 특성으로 인정되어 왔으나, 지나친 엄격성의 차이는 채점 신뢰도를 떨어뜨리는 요인이 되기도 한다(설현수, 2010). 더 중요한 것은 일관성의 문제이다. 특히 채점 분량의 증대와 채점 시간의 제약은 채점자의 일관성을 무너뜨리는 중요한 요인으로 꼽힌다(Lunz & Stahl,

1990). 이는 결과적으로 동일한 능력 수준을 지닌 피험자에게 일관된 점수를 부여하지 않아 채점 신뢰도를 저하하는 주요 요인이 된다(박영민, 박종임, 2011).

채점자 협의(rater discussion)는 이 같은 문제 상황을 해결해 채점자 신뢰도를 확보하는 방안으로 도입되었다. 일반적으로 채점자 협의는 채점자들이 서로의 채점 결과와 근거를 공유하며 기준의 의미를 재확인하고, 불일치를 조정하는 과정이다(서수현, 2012). 이 과정에서 채점 기준 적용의 모호성을 감소시키고, 점수 부여 사례를 공유하면서 채점자의 엄격성을 조정하여, 채점 결과의 비일관성을 해소하는 효과가 있는 것으로 알려져 있다.

그러나 채점자 협의가 항상 긍정적으로 작동하는 것은 아니다. 실제로 Johnson 외(2005)의 연구에서는 채점자 협의 과정에서 특정 채점자의 의견이 우세하게 반영되는 채점자 지배(rater dominance) 현상이 관찰된 바 있는데, 특히 채점자들 간의 점수 차이가 클 경우 자신의 점수에 대한 더 강한 의견을 내세우는 경향이 있고, 덜 전문적인 채점자의 의견이 지배적일 경우 채점자 협의 과정에 부정적인 영향을 미칠 수 있음이 보고되었다. 최근 들어 채점자 협의의 대화 양상을 분석하여 효과적인 채점자 협의를 위한 전제 조건을 밝힌 연구가 있어 주목된다(강민석, 2025). 엄격성 측면에서 '평가 기준 일치 확인'과 '상대 근거 수용' 발화가 잦을수록 엄격성 격차가 줄어들었고, 반대로 '거부' 발화가 많을수록 엄격성 격차가 확대되었다. 일관성 측면에서 채점 기준을 명확하게 인지하려고 하거나, 타인의 채점 전략을 수용한 경우 일관성이 개선된 반면, 협의 자체에 소극적이거나 지나치게 자기 점검에 몰두한 경우에는 일관성이 악화되는 상황이 발생하였다. 이를 통해 채점자 협의에서 적극적인 협의 태도, 채점 근거에 대한 충분한 상호교류, 비교, 반박, 수용을 통한 설득적 합의, 반성적 자기 성찰 등이 중요함을 이해할 수 있다.

2. 인간-AI 협업과 자동채점 기반 채점자 협의

진일보한 AI 기술은 인간에 준하는 독립적 주체의 수준에 이르러 국내·외 교육 현장에 다각적 변화를 이끌고 있다(권경문, 최숙기, 2025; 최예영, 이남호, 2025; 함은혜 외, 2024; Fui-Hoon Nah et al., 2023; Stanford HAI, 2025). 그중 본 연구의 논점은 AI가 인간 친화적이면서도 고도화된 기능을 제공하리라는 예측에 있다(박소영 외, 2024; 최진영 외, 2025; 함은혜 외, 2024; Li et al., 2025). 그렇다면, '인간-AI 협업'의 개념은 무엇인가? 선행연구는 "AI가 교사와 공동 책임을 갖는 상호작용적 관계를 맺음으로써 수업 설계 및 평가 과정에 참여하여 교사의 의사결정을 보완하고 업무를 경감시키며 문항 생성 및 채점을 지원하여 학생들에게 개별화된 피드백을 신속하게 제공하는 총체

적 과정(최진영 외, 2025:1300)"으로 정의한다. 즉, 인간과 AI가 상호 보완적으로 평가의 질을 높여가는 과정이다(Tate et al., 2024; Wetzler et al., 2024).

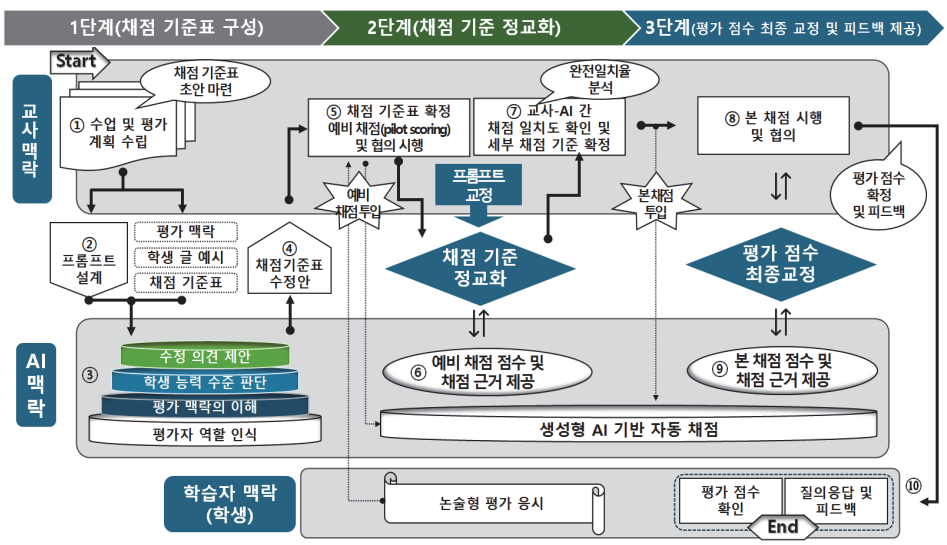
확실히 인간-AI 협업을 평가적으로 활용함은 교사의 평가 부담 완화에 긍정적인 측면이 있다. AI 자동채점은 채점 과정을 보조하여 채점 신뢰도를 확보하도록 돕기 때문이다(최진영, 2025; Attali & Burstein, 2005). 현재, AI 자동채점은 대규모 언어 모델(LLM)의 응용을 통해 다양한 채점 자질(scoring feature)이 학습되며 점차 안정된 기능을 기대할 수 있게 되었다(박종임 외, 2023; Kumar & Boulanger, 2021). 글의 형식뿐만 아니라 타당성, 통일성, 논리성 등 의미론적이고 맥락적인 차원까지 분석 가능하다는 의견들이 AI 자동채점의 효용 가치를 더욱 높인다(권경문, 최숙기, 2025; 최진영, 2025; Li et al., 2025; Wetzler et al., 2024). Tate 외(2024)는 실제 에세이 평가에서 GPT 기반의 생성형 AI(generative AI) 자동채점이 인간과 거의 차이가 없다는 결과를 보이기도 하였다.

다만, 인간과 AI의 협업을 긍정적으로 검토하면서도 한편으로는 AI 자동채점의 추가적인 실증 연구와 관찰이 필요하다. 신뢰성 여부, 인간 주도성 상실, 책임 윤리, 정보의 의존 등 선행연구들이 제시하는 우려를 외면할 수 없기 때문이다(Stanford HAI, 2025). 예컨대, Fui-Hoon Nah 외(2023)는 정보 편향, 과의존의 문제를 지적하며 적절함과 부적절함의 경계가 모호해짐을 지적하였다. 특히 AI 자동채점의 경우, 생성 결과의 진위를 알기 어렵고 평가 결과의 중간 과정이 불투명한 점도 한계로 남아 있다(Ercikan & McCaffrey, 2022). AI 자동채점의 긍정적 결론을 도출한 Tate 외(2024) 역시 완전 일치도(exact agreement)의 한계를 지적하며 AI의 관대한 채점이 문제가 될 수 있다고 보았다. 이렇듯, 상위 학교 진학이나 자격증 시험 등 엄밀함이 요구되는 고부담 평가에는 아직 AI 자동채점은 위험성이 남아 있다.

더하여, 채점자 협의 과정에서 AI 자동채점이 반드시 효율성과 편리성, 그리고 엄격성을 담보하는 것은 아니다. 최진영, 하민수(2023)는 교사들이 직접 AI 자동채점 결과를 활용하는 실험에서 오히려 채점 시간이 더 소요됨을 발견하였다. Ercikan & McCaffrey(2022)는 AI가 특정 어휘 및 문장 패턴의 통계적 상관성에 의존하므로 채점 기준 해석에 한계가 있다고 판단하였다. Wetzler 외(2024)도 ChatGPT의 자동채점 결과가 학생 성취도 수준에 따라 엄격성이 달라진다고 보았다. 즉, 단순 교사와 AI의 점수가 높은 상관관계를 보이는 것과 AI 자동채점의 신뢰성은 별개의 문제라는 것이다. 이에 인간과 AI의 협의는 중요하다.

인간과 AI의 협의는 채점 결과를 해석하고 피드백을 제공하는 측면에서 교사의 노력과 관점이 더 중요하다는 의견에 무게가 실린다. AI 자동채점의 효율과 편리가 주목받고 있지만, 실제 평가에서는 피드백까지 이어지는 평가 과정 전반의 질 개선에 의의를

두어야 할 것이다. 다시 말해, 평가와 피드백의 타당성, 신뢰성, 공정성 등을 높일 방법론으로 접근할 필요가 있는 것이다. 그 예시로 아래의 [그림 1]은 최진영 외(2025)의 연구 결과로, 일련의 'AI 자동채점 기반 평가자 협의 모델(이하, AI 협의 모델)'에 해당한다. 해당 모델은 '한 학기 동안 수행평가나 지필평가를 통해 논술형 평가를 시행하는 상황을 가정(최진영 외, 2025:1303)'한 것으로서 AI는 교사의 평가 계획, 채점 기준 선정, 채점 시행 과정에 대한 후행적 조언자의 역할을 갖는다.



[그림 1] AI 자동채점 기반 평가자 협의 모델(최진영 외, 2025:1304)

[그림 1]에 따르면 1단계는 AI 프롬프트 설계(prompt engineering)가 주를 이루고 2단계는 학습된 생성형 AI와 채점 기준을 조정하는 절차이다. 3단계는 본 채점에 해당하는 것으로서 교사는 논술형 평가로부터 수집된 학생 글을 모두 생성형 AI에 투입하고 함께 채점을 시행한다. 이후, 교사는 피드백과 채점 근거를 마련하며 AI 자동채점 결과를 참고할 수 있다. 즉, AI 협의 모델은 계획 구성, 예비 채점, 협의 결과 조정을 모두 포함한다. 교사는 각 단계에서 AI의 제안을 재검토하여 평가 과정의 질을 높이고, 역으로 AI를 반복 훈련시켜 교사와의 채점 일치도를 높일 수 있다. 해당 예시로 최진영 외(2025)의 프롬프트 설계를 [부록]으로 별도 제시하였다.

본 연구는 위의 3단계 절차를 중심으로 진행된다. 모델은 '교사가 다수인 경우, 상호 간의 채점 점수가 불일치할 때, 생성형 AI가 또 하나의 신뢰할 만한 의견으로 제시되어 평가자 간 점수 합의에 도움을 줄 수 있다(최진영 외, 2025:1308).'라는 전제를 갖는다. 이는 동 교과 논술형 평가에서 학급마다 다른 교사가 배치될 때, 학교 평가 규정에 따라 완전한 점수 일치가 필요한 경우(경기도교육청, 2025; 인천광역시교육청; 2025)를 말한

다. 교사는 여기서 AI 점수를 채점 결과의 또 다른 제안으로 수용하여 최종 협의에 AI를 참여시킬 수 있다. 연구에서는 AI 자동채점이 교사의 채점자 협의에 어떤 기능을 보이는지, 그리고 인간인 교사가 지닌 고유의 교육적 지위는 무엇인지 탐색한다. AI 협의 모델의 실증은 현재 변화하는 AI 기반 교육평가 패러다임에서 교사 가치의 중요성을 탐지하는 과정이라고 할 수 있다.

Ⅲ. 연구 방법

1. 연구 절차와 대상

가. 연구 절차

연구 절차는 연구 준비, 양적 연구(QUAN), 질적 연구(qual), 통합 및 해석의 총 4개 단계, 7개의 절차로 진행되었다. 다음 〈표 1〉의 설계는 설명적 순차 혼합 방법(QUAN → qual)을 활용한 것으로, 먼저 양적 자료를 수집·분석한 뒤 그 결과만으로 파악되지 않는 공백을 질적 자료를 통해 보완하는 연구 방법론이다. 이를 통해 초기의 양적 결과를 보다 심층적으로 해석할 수 있으며, 연구 결과의 확장과 깊이 있는 통찰을 가능하게 한다(Creswell & Creswell, 2018).

참고로, 채점자 협의는 예비 채점 과정에서 협의를 통하여 채점자 간 채점 기준의 공통된 이해도를 높이고 상호 영점 조정을 마친 후에 독립 채점으로 나아간다. 대신, 본 연구는 학교 현장 규정에 따라 지도 교사가 다수일 때 요구되는 공동 출제 및 공동 채점 원칙의 상황을 고려하였다. 이에 채점자 간 기준 조정은 완료되었다고 보며, 최종 점수 선정에 한정하여 '③ 본 채점 및 협의'의 결과를 분석하였다.

<표 1> 연구 단계 및 절차

단계	절차	세부 내용
연구 준비	① 모의 평가 및 채점자 구성	• 채점자 협의 실험을 위한 채점 기준표, 프롬프트 모델 구성 • 논술형 평가 경력을 보유한 현직 교사 4인 선정
▼		
양적 연구 (QUAN)	② 예비 채점 및 분석	• 참여자 채점 경향 파악 및 채점 엄격성, 일관성 등 측정 • 채점 결과에 따른 **'전통적 협의 집단'**과 **'AI 협의 집단'** 구성
	③ 본 채점 및 협의	• 연구 참여자의 독립적인 채점 진행 • 인간 채점자 중심의 전통적 협의와 인간 채점자가 AI를 참고할 수 있는 AI 협의를 병행적으로 진행

단계	절차	세부 내용
	④ 양적 자료 분석	• 양 집단의 엄격성 및 일관성 측정 • 양 집단의 개별 채점 및 협의 결과 비교 • 양 집단 협의 결과 및 부담 지표(채점 시간, 체감 난도) 분석
	▼	
질적 연구 (qual)	⑤ 심층 인터뷰 진행	• 양적 자료 분석 결과를 바탕으로 한 인터뷰 질문 설계 • 채점 결과, 협의 경험, 채점자 인식, AI 인식 등 주제 선정
	⑥ 질적 자료 분석	• 인터뷰 내용 해석을 통한 양 집단의 평가 인식 차이 분석 - 채점의 어려움, 협의 과정의 특징, AI 중재 효과 등
	▼	
통합 및 해석	⑦ 결과 통합 해석	• 양적 결과와 질적 결과의 통합 • 채점 결과, 협의 과정, 심층 인터뷰 내용 등의 종합적인 해석

나. 연구 대상

연구 대상인 채점자는 총 4명으로, 8년 이상의 교육 경력을 통해 논술형 평가 경험을 풍부하게 축적한 교사들로 구성하였다. 이에 실제 학교 현장에서 다양한 교사들과 채점자 협의를 거쳐 본 국어 교사를 연구 참여자로 선정함으로써 자료 해석의 타당성을 높이고자 하였다. 각 채점자의 구체적 교육 경력은 〈표 2〉와 같다. 이들의 채점 과정은 'AI 협업 기반 평가자 협의 모델'의 검증과 그 결과에 따른 개선 방향 탐색을 목적으로 두 개 집단으로 나누어 진행되었고, 그 결과를 비교하였다.

채점자들은 5개 문항에 대한 예비 채점을 통해 '전통적 협의 집단'과 'AI 협의 집단'의 구성으로 나뉜다. 집단 간 상호 협의가 활발하도록 예비 채점에서 친밀도와 채점 성향을 고려하였다. 상호 채점자 간의 사전 친밀도를 배제하고, 점수 배점의 성향, 다국면 Rasch 분석의 infit 값을 추가 확인하였다. 이에 채점 성향이 상반된 채점자가 짝을 이루도록 하였다.

<표 2> 연구 참여자 정보

채점자	전통적 협의		AI 협의	
	R1	R2	R3	R4
교육 경력(년)	11	8	16	14

2. 자료 수집

연구에 필요한 자료 수집은 앞의 Ⅲ-1절에서 제시한 〈표 1〉의 '③ 본 채점 및 협의'와 '⑤ 심층 인터뷰 진행'에 집중된다. 먼저, '③ 본 채점 및 협의'의 진행 절차와 자료 수집 과정에는 최진영 외(2025)의 'AI 자동채점 기반 평가자 협의 모델'의 실험적 검증

이 고려되었다. 따라서 필연적으로 선행연구의 채점 기준표와 AI 프롬프트 설계를 적용할 필요가 있었다. 대신, 선행연구가 'AI 자동채점 기반 평가자 협의 모델'의 전체 구성을 제시하고 타당성을 확보하는 과정이었다면, 이번에는 실제 전통적 협의와 AI 협의 양 집단의 평가 결과 비교를 통해 실증적 효과성을 탐색하는 연구라는 점에서 차별점이 있다.

가. 채점 기준표와 모의 답안

<표 3> 모의 평가 과제의 채점 기준표

성취기준	[12화작02-06] 청자의 특성에 맞게 내용을 구성하여 발표한다. [12화작03-02] 작문 맥락을 고려하여 자기를 소개하는 글을 쓴다.			
평가 주제	자기 진로를 소개하는 발표 준비를 위한 글쓰기			
평가 요소	채점 기준표			
	4점(우수)	3점(적정)	2점(부분부족)	1점(미흡)
내용	자신의 진로 목표와 관련된 핵심·보조 자료를 적절히 선별하여 내용을 정확·구체적으로 서술하고, 명시적인 주제를 끝까지 일관되게 유지하여 글의 통일성이 뛰어나다.	진로 목표에 맞춘 핵심·보조 자료를 선별해 대체로 정확하게 설명하였으나 일부 내용이 추상적이며, 주제는 명시되었지만 부분적으로 관련 없는 내용이 포함되어 통일성이 약간 저하된다.	진로 관련 자료를 일부만 선택해 설명했으며 부정확하거나 추상적인 부분이 있고, 주제가 늦게 제시되거나 암시적으로 드러나며 관련 없는 내용이 많아 통일성이 부족하다.	핵심·보조 자료를 거의 제시하지 못하거나 무관한 자료를 사용했고 주제가 불분명하거나 여러 번 바뀌어 글의 일관성과 결론 정리가 이루어지지 않는다.
조직	도입-전개-정리의 3단 구성이 명확하며 내용 전개 방식을 세 번 이상 다양하게 활용하여 논리를 효과적으로 이끈다.	3단 구성이 유지되지만 일부 단계가 빈약하고, 내용 전개 방식을 두 번 활용해 이해에 큰 무리는 없지만 완성도가 다소 떨어진다.	한두 단계를 생략하거나 나열·병렬식으로 전개되어 구조가 불완전하고, 내용 전개 방식을 한 번만 사용해 흐름이 매끄럽지 않다.	구조가 식별되지 않거나 전개가 거의 없으며 내용 전개 방식을 활용하지 못해 논리 전개가 이루어지지 않는다.
표현 및 매체	맞춤법·띄어쓰기 오류 없이 정확하며 질문·영상·사진 등 표현 전략을 세 번 이상 효과적으로 활용해 전달력을 높였다.	언어 오류가 세 건 이내로 제한(단순 오류, 경미한 오류 빈번, 중대한 오류 1번)되고 두 번의 표현 전략 사용으로 전달이 대체로 원활하다.	언어·맞춤법 오류가 4~10건 존재(중대한 오류가 빈번)하고 표현 전략을 한 번만 사용해 전달 효과가 제한적이다.	언어 오류가 10건을 넘어 이해가 어려우며 표현 전략을 전혀 활용하지 못했다.

〈표 3〉은 모의 평가 상황의 구성을 위해 꼭 필요한 것으로서, 연구에 참여하는 채점자들이 실제 학생 글을 채점할 때 사용할 채점 기준표이다. 해당 채점 기준표는 2025년 4월, 경기도 소재 일반 인문계 고등학교에서 3학년 1학기 [화법과 작문] 교과에서 이루

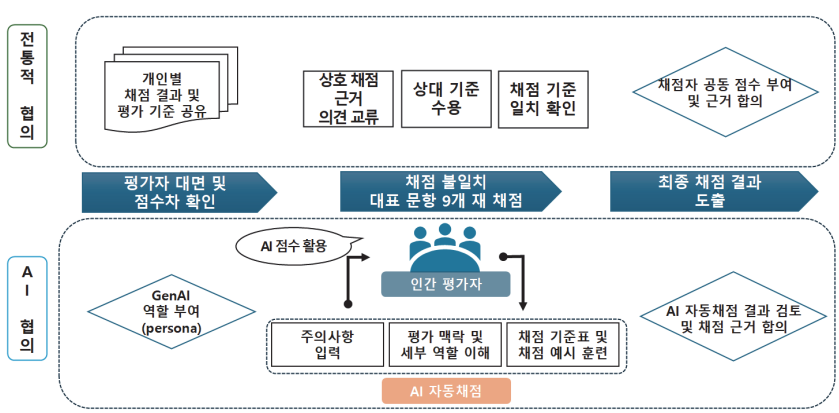
어진 수업 및 평가 일부를 가져와 실험적으로 재구성한 것이다[2]. 예비 채점과 본 채점에서는 실제 평가에 제출된 학생 글을 무작위로 20편을 추출하여 모의 답안을 구성하였다.

나. 채점 및 협의 결과 수집

채점자의 인지적 부담과 관련된 선행 연구(윤금준, 2021; 정분의, 2019)를 참고하여, '채점 시간 및 난도', '협의 시간 및 난도'를 측정하여 양적 분석 자료로 활용하였다. 각 자 학생 글의 채점 난도와 협의 난도를 6점 Likert 척도로 표시하였으며, 이 값들의 평균을 분석에 활용하였다.

우선 각 채점자 R1~R4는 네이버 폼을 활용하여 독립적으로 채점을 진행하였다. 이때 개별 학생 글에 대한 채점 부담 지표는 채점자의 난도 평가와 소요 시간을 기준으로 삼았다. 채점자는 학생 글에 점수를 부여하고, 개별 글마다 채점 난도를 평가하였다. 한편 연구자는 채점자의 인지적 부담을 정량적으로도 확인하기 위해 채점 화면을 녹화하여 채점 소요 시간을 산출하였다.

다음으로, 본 채점을 통한 협의 결과를 수집하는 과정은 [그림 2]로 요약되며 전통적 협의 집단과 AI 협의 집단은 각자 채점 결과를 바탕으로 채점자 협의를 진행하였다. 활발한 협의를 유도하기 위하여 연구자는 본 채점 결과로부터 평가 요소(내용, 조직, 표현·매체) 중 2개 이상의 요소에서 점수 차이가 나타나거나, 혹은 채점자 4명 중 2명 이상이 채점 난도를 4점(약간 어렵다) 이상으로 보고한 글을 선정하기로 하였다. 이에 총 9편의 학생 글이 선정되었다.



[그림 2] 전통적 협의 집단과 AI 협의 집단의 협의 자료 수집 절차

2) 본 연구는 최진영 외(2025)의 'AI 자동채점 기반 평가자 협의 모델'을 실험적으로 구현하는 과정이다. 최진영 외(2025)는 채점 기준표의 구성 과정은 제시하였지만, 전체 채점 기준의 완성은 설명하지 않았다. 본 연구는 해당 채점 기준표가 완성된 전체 구성을 제시함으로써 집단별 비교 실험과 모델의 개선 방향 탐색을 주요 목적으로 하였다.

[그림 2]의 양 집단별 협의 결과를 수집하기 위해 각각 합의된 채점 결과를 네이버 폼을 통해 기록하였다. 협의 과정은 온라인 화상회의 플랫폼인 Zoom을 통해 비대면으로 진행되었으며, 연구자는 전 과정을 녹화하여 협의 소요 시간을 측정하였다. 협의 종료 후 채점자들의 협의 과정에서 체감한 난도 역시 네이버 폼을 통해 수집하였다.

협의의 절차는 먼저 전통적 협의의 경우, 두 채점자가 개인별 채점 결과와 평가 기준을 상호 공유하는 것을 시작으로, 상호 채점의 근거와 의견을 교류하는 시간을 갖는다. 이후 상호 간의 기준을 수용할지를 대화로써 협의하고 채점 기준이 일치하는지 확인한다. 이들은 채점자 공동의 점수를 부여함으로써 최종 채점 결과를 도출하는데 이른다.

이어 AI 협의 집단의 경우, ChatGPT[3]와 같은 생성형 AI 도구에 [부록]에 제시된 프롬프트를 활용하여 공동 채점자 역할을 부여하는 작업을 먼저 거친다. 예를 들어, '주의사항', '평가 맥락 및 세부 역할의 이해', '채점 기준표와 채점 예시 훈련'으로 이어지는 채점 기준 정교화(최진영 외, 2025:1116)를 명령어로 입력한다. 이에 따라 AI 자동채점 결과가 도출되면 AI 협의 집단의 두 채점자는 채점 시 AI 자동채점 결과를 검토하여 상호 간의 채점 근거 합의를 시도하고 최종 채점 결과를 도출한다. 단, 이때 채점자들의 판단과 합의에 따라 AI 채점 결과를 활용하지 않을 수 있다.

다. 인터뷰 자료 수집

<표 4> '전통적 협의 집단'과 'AI 협의 집단' 질문 구성

항목		전통적 협의 집단	AI 협의 집단
전체 및 개별		1. 전체 평가 과정에서 가장 어려웠던 점은 무엇입니까?	
		2. 개별 채점 중 9, 10, 17번 학생 글의 채점 난도가 높다고 판단한 이유는 무엇입니까?	
채점자 협의	공통 질문	3. 채점자 협의가 채점자의 엄격성·일관성에 어떤 영향을 주었습니까?	
		4. 채점자 협의를 통해 채점 난도가 높은 답안의 채점이 수월해졌다고 보십니까?	
	집단별 질문	5. 협의 과정에서 주로 활용한 전략은 무엇입니까?	5. AI 채점 결과를 처음 받아본 느낌은 어떠했습니까? 협의 과정에서 어떤 역할을 했다고 생각합니까?
		6. 채점자 협의 과정에서 상대방의 어떤 특성이 채점 신뢰도에 영향을 미쳤습니까?	6. AI 채점 결과 중 '점수'와 '근거' 중 어느 부분을 더 중시·신뢰했습니까? 이유는 무엇입니까?
적용 방안		7. AI 자동채점 기술이 학교의 논술형 평가 상황에서 채점자 간 협의 강화를 위해 도입될 경우 고려할 사항이나 필요한 개선점은 무엇입니까?	

질적 자료의 수집은 개별 통화 및 Zoom을 활용하여 심층 인터뷰로 진행되었다. 〈표 4〉의 사전 질문지를 준비하되, 참여자 응답에 따라 탐색적 추가 질문을 유연하게 제시하였다.

3) 본 연구에서는 2025년 8월 새로이 배포된 GPT-5 모델이 적용되었다.

큰 범주의 질문은 양적 연구 결과를 토대로 채점 과정 전반에 대한 채점자의 실제 경험과 인식을 탐색하는 데 목적을 두었다. 주요 질문 범주는 채점 과정에서의 채점 난도, 협의 과정에서 경험한 인지적 부담, 전통적 협의와 AI 협의에 대한 상대적 인식 등을 포함하였다. 인터뷰 자료는 채점 및 협의 과정에서 도출된 정량적 결과를 심층적으로 해석하는 근거로 활용하였다.

이상에 따른 세부 질문 체계는 크게 두 가지 범주로 나뉜다. 우선 양 집단에 적용되는 질문은 채점 과정의 어려움과 특정 학생 글(9, 10, 17번)의 채점 난도가 높은 이유, 채점자 협의가 신뢰도에 미친 영향, 채점자 협의의 효과, 학교에서 AI 기술이 채점자 협의에 도입될 경우 고려해야 할 사항 등이었다. 이어 집단별 분리 질문의 경우, 전통적 협의 집단은 협의 과정에서의 전략 사용 유형과 채점 신뢰도에 영향을 미치는 채점자의 특성에 대해 질문하였다. AI 협의를 진행한 집단은 AI 채점 결과가 채점자 협의에서 수행한 역할, AI 채점 결과에 대한 채점자의 인식과 신뢰 수준, 특히 점수와 근거 중 어느 요소를 더 중시했는지를 질문하였다.

3. 분석 방법

양적 분석에서는 완전 일치도(exact agreement rate), 다국면 Rasch 모형을 활용하여 채점자의 엄격성(severity)과 일관성(consistency) 변화를 통계적으로 검증하였으며, 질적 분석에서는 양적 분석 결과를 토대로 개발한 질문을 중심으로 심층 인터뷰를 실시하여 채점자 협의 경험과 인식을 심층적으로 탐색하였다.

양적 분석에서의 완전 일치도는 동일한 답안에 대해 채점자들이 동일 점수를 부여한 비율로 채점 결과의 일치 정도, 즉 채점 신뢰도를 직관적으로 보여주는 지표로 Excel을 사용해 산출하였다. Rasch 분석은 문항반응이론(item response theory)에 기초하여 피험자, 과제, 채점자 등 평가에 영향을 주는 다양한 국면(facets)을 동시에 고려함으로써 요인별 효과와 채점 적합도를 추정할 수 있다(Linacre, 1989; Bond & Fox, 2015). 특히 본 연구에서 Facet 4.3.1 프로그램을 사용해 채점자의 엄격성을 logit으로, 채점자의 일관성을 내적합도(infit) 지표로 산출하였다.

우선 채점자의 logit이 높을수록 채점 경향이 엄격한 것으로 해석된다. Infit의 해석 기준은 연구자들에 따라 다소 차이가 있는데, 본 연구에서는 McNamara(1996), Lynch & McNamara(1998)의 기준에 따라 0.75~1.3을 적합한 범위로 제시하였다. Infit 값이 0.75 이하일 경우는 지나치게 비슷한 점수만을 부여하여 변별력이 부족한 과적합(overfit) 채점자로, 1.3 이상일 경우는 예측할 수 없는 불규칙한 점수를 부여하는 부적합(misfit) 채점자로 분류하였다(Wright & Linacre, 1994).

자기 보고식 설문 결과 및 영상 자료의 경우 채점자의 인지적 부담을 조사하는 데 활용하였다. 그리고 심층 인터뷰 자료는 질문 범주에 따라 연역적 방법을 활용한 분석 틀을 이용하여 초기 코드[4]를 설정하였다. 이후 개방 코딩을 통해 귀납적 하위 코드를 추가·정련하는 방식으로 진행하였다. 먼저 양 집단 담화의 내적 분석을 통해 각 채점자마다의 채점 맥락·해석·협의 전략을 서사적으로 재구성하고, 이어 지속적 비교법으로 두 집단 간 사례 간 비교를 수행하여 공통점과 차이점을 도출했다. 코딩은 3인 연구자가 독립적으로 자료를 분석한 뒤, 상호 검토를 통해 코드 체계를 반복 점검·수정하며 합의하였다. 연구자 간 일치 정도를 검토하여 신뢰성을 확보하였으며, 불일치 항목은 재논의를 통해 조정하였다.

IV. 연구 결과

1. 개별 채점 결과 분석

Rasch 분석을 통해 엄격성(logit)과 일관성(Infit)을 검증한 결과, 〈표 5〉에서 채점자 R2가 가장 엄격한 평가 경향(0.32)을, R3가 가장 관대한 평가 경향(−1.77)을 보이는 것으로 나타났다. 분리도 지수(separation)는 3.61로 나타나 채점자의 엄격성(logit) 차이가 통계적으로 구분이 가능한 것으로 나타났으며, 이에 따른 계층(strata) 값은 5.14로 산출되었다. 이는 네 명의 채점자가 다섯 수준 이상의 독립적인 집단으로 구분될 수 있을 정도로 채점 성향의 차이가 뚜렷함을 의미한다. 일관성(Infit) 측면에서 채점자 R2는 부적합 채점자(1.32)로, 채점자 R4는 과적합 채점자(0.57)로 분류되었다.

<표 5> 개별 채점의 Rasch 분석 결과: 엄격성(logit), 일관성(Infit) 및 완전 일치도 결과(n=240[5])

채점자(Rater)	logit (S.E.)	Infit	완전 일치도(exact agreement rate, %)			
			R1	R2	R3	R4
R1	-0.4(0.23)	1.16	-	55.0%	51.67%	61.67%
R2	0.32(0.23)	1.32	55.0%	-	45.0%	40.0%
R3	-1.77(0.24)	0.85	51.67%	45.0%	-	55.0%
R4	-0.45(0.23)	0.57	61.67%	40.0%	55.0%	-
Separation (Sample) = 3.61, Strata (Sample) = 5.14, Reliability (Sample) = .93						

4) 평가 과정 중 어려움의 이유, 채점 고난도 답안의 특징, 채점 신뢰도의 엄격성 및 일관성, 채점자 부담 측면에서 협의 시간과 협의 난도의 특징

5) 피험자 20명이 채점자 4명에게 3개의 평가 요소로 평가받았으므로, 20×4×3=240개의 관찰값이 생성됨.

본 연구에서 산출된 채점자 간 완전 일치도(exact agreement rate)는 전반적으로 40%에서 61.67% 사이에 분포하였다. 구체적으로, R1과 R4 간의 일치도가 61.67%로 가장 높게 나타났다. R1과 R2(55.0%), R3과 R4(55.0%) 역시 절반 수준의 일치율에 머물러 중간 정도의 합치도를 보였다. 반면, R2과 R3(45.0%), R2와 R4(40.0%)의 경우는 가장 낮은 일치율을 보여 두 채점자 간 채점 기준 적용 차이가 두드러지게 드러났다. 이 같은 엄격성 차이와 비일관된 채점 경향은 평가의 공정성을 위협할 수 있는 잠재적 요인을 내포한다. 동일한 피험자의 수행이라 하더라도 채점자에 따라 상이한 점수를 받을 가능성이 존재한다. 따라서 본 연구에 참여하는 참여자들이 채점자 협의의 과정을 거쳐 채점자 간 불일치를 해소할 필요성이 제기된 것이라 할 수 있다.

다음으로 개별 학생의 글을 채점하는 데 소요된 시간과 채점이 끝난 후 개별 글에 대한 채점 난도를 보고한 결과를 바탕으로 채점자의 인지적 부담 정도를 분석하였다. 결과는 다음 장의 〈표 6〉과 같다. 〈표 6〉에서 채점 난도와 채점 시간 간의 관계를 살펴본 결과, 전반적으로 채점 난도가 높게 평가된 글일수록 채점 시간이 길게 소요되는 경향을 보였다. 예를 들어, E9(평균 채점 난도 3.5, 평균 채점 시간 123.25초)와 E10(평균 채점 난도 4.0, 평균 채점 시간 112초), E17(평균 채점 난도 3.5, 평균 채점 시간 132.25초) 등은 상대적으로 높은 채점 난도를 보이며 채점 시간도 길게 나타났다. 반대로 E8(평균 채점 난도 1.5, 평균 채점 시간 69.5초)과 E18(평균 채점 난도 1.75, 평균 채점 시간 86.25초)은 채점 난도가 낮고 채점 시간 또한 짧게 산출되었다. 이는 채점 난도가 높은 글일수록 채점자가 더 많은 인지적 자원을 투입하여 판단을 내렸음을 뜻한다. 일반적으로 채점 난도가 채점자의 인지적 부담과 연결되어 채점 소요 시간에 영향을 미치는 것으로 추론할 수 있다. 다만 이러한 경향성은 통계적 검증을 거치지 않은 단순한 기술적 관찰에 기반한 것이므로 해석에 신중을 기할 필요가 있다.

<표 6> 채점 소요 시간 및 채점 난도

학생 글 (글자수)	채점 소요 시간(초)					채점 난도(6점 척도)				
	R1	R2	R3	R4	평균	R1	R2	R3	R4	평균
E1(1521)	117	196	140	126	144.75	2	2	2	3	2.25
E2(1173)	109	100	285	104	149.5	3	2	2	5	3
E3(1284)	94	138	119	108	114.75	2	3	4	4	3.25
E4(1986)	93	120	149	114	119	1	2	3	2	2
E5(538)	93	109	42	67	77.75	2	5	2	4	3.25
E6(1307)	95	74	66	85	80	1	2	2	4	2.25
E7(1768)	102	287	51	118	139.5	3	6	2	2	3.25
E8(1516)	55	119	42	62	69.5	1	2	2	1	1.5

학생 글 (글자수)	채점 소요 시간(초)					채점 난도(6점 척도)				
	R1	R2	R3	R4	평균	R1	R2	R3	R4	평균
E9(1627)	100	149	124	120	123.25	2	4	4	4	3.5
E10(1235)	108	123	154	63	112	4	4	4	4	4
E11(950)	48	90	86	73	74.25	2	2	3	4	2.75
E12(1644)	112	142	104	158	129	3	4	2	4	3.25
E13(1070)	77	80	92	112	90.25	3	2	3	5	3.25
E14(735)	66	83	132	52	83.25	2	4	3	2	2.75
E15(1425)	56	151	101	123	107.75	1	3	2	2	2
E16(639)	35	102	71	84	73	1	2	4	4	2.75
E17(1484)	167	134	112	116	132.25	4	3	3	4	3.5
E18(1719)	76	119	87	63	86.25	1	3	2	1	1.75
E19(714)	66	30	98	46	60	2	1	3	4	2.5
E20(754)	57	83	97	37	68.5	1	2	3	5	2.75
평균	86.30	121.45	107.60	91.55	101.73	2.05	2.90	2.75	3.40	2.78

2. 채점자 협의 결과 분석

본 절에서는 전통적 협의 집단과 AI 협의 집단 간의 협의 결과를 비교·분석하였다. 이를 위해 집단별 엄격성과 일관성, 협의 시간과 협의 난도, AI 활용 여부에 따른 변화 등을 살펴봄으로써 두 협의 방식의 차이를 구체적으로 확인하고자 하였다.

가. 채점 신뢰도 결과 분석

각 집단에서 나타난 채점자 협의의 효과를 살펴보기 위해 협의 결과로 도출된 합의 점수를 가상의 채점자 'R1+R2', 'R3+R4'로 상정하여 분석하였다.

<표 7> 집단별 Rasch 분석 결과

전통적 협의 집단			AI 활용 협의 집단		
채점자	엄격성(logit)	일관성(Infit)	채점자	엄격성(logit)	일관성(Infit)
R1	-0.81	1.13	R3	-2.76	1.48
R2	0.62	1	R4	0.17	0.86
R1+R2	0.72	0.81	R3+R4	-0.30	0.64

그 결과 전통적 협의 집단은 엄격성 차원에서 두 채점자의 엄격성이 평균화되지 않고 더 엄격한 쪽으로 기울었다. R1+R2 점수를 포함하여 분석한 결과 채점자 R1은 관대한

경향(－0.81), 채점자 R2는 엄격한 경향(0.62)을 보였으나, R1+R2는 0.72로 나타나 오히려 가장 엄격한 것으로 나타났다. 이는 전통적인 채점자 협의가 채점자 간 엄격성의 균형을 맞추기보다는, 상대적으로 엄격한 채점자의 점수 부여 경향으로 수렴되었음을 의미한다. 따라서 전통적 협의 집단의 협의가 채점자 간 엄격성 편차를 줄이는 조정 장치로 작동했다기보다, 평가 기준을 보다 엄격하게 적용하도록 하는 역할을 수행했다고 볼 수 있다. 이어 일관성을 분석한 결과 R1, R2 및 R1+R2 점수 모두 적합한 범위 내에 위치하였다.

다음으로 AI 협의 집단은 엄격성 차원에서 두 채점자의 엄격성이 평균(0)에 수렴하는 경향성을 보였다. R3+R4 점수를 포함하여 분석한 결과 R3는 매우 관대한 성향(－2.76), R4는 평균보다는 다소 엄격한 성향(0.17)을 보였고, R3+R4는 －0.30으로 나타났다. 이는 AI 기반 채점자 협의가 관대한 평가 경향을 지닌 채점자의 엄격성을 조정하여 평균적인 수준으로 수렴시키는 효과를 발휘했음을 의미한다. 그러나 전통적 협의 집단과 마찬가지로 비교적 엄격한 채점자의 점수 부여 경향으로 귀결되었다고 할 수 있을 것이다.

일관성 차원에서 R3, R4 채점 경향과 비교할 때, R3+R4의 일관성 지수가 0.64로 가장 낮게 나타나 과적합(overfit) 경향을 보였다. 이는 AI 협업 기반 협의를 통해 점수를 확정하는 과정에서 채점 변별력이 일부 축소되는 한계가 있었음을 의미한다. 이는 아래 제시될 〈표 8〉을 통해 나타나는 채점 척도의 제한적 사용으로 발생한 현상인 것으로 분석된다.

전통적 협의와 AI 활용 협의에서 부여한 개별 채점 점수 및 최종 협의 점수는 〈표 8〉과 같다. 두 집단은 동일한 세 편의 글(E9, E10, E17) 중에서 두 편의 글(E9, E10)에 대해 최종 합의 점수를 다르게 도출하였으며, 전통적 협의 집단의 점수 부여 경향이 보다 엄격했음을 알 수 있다.

<표 8> 집단별 개별 채점 및 협의 결과

전통적 협의 집단				AI 협의 집단				
학생 글 (글자수)	개별 채점 결과		협의 결과	학생 글 (글자수)	개별 채점 결과		AI 채점 결과	협의 결과
	R1	R2			R3	R4		
E3(1284)	2-3-3(8)	2-2-1(5)	2-2-1(5)	E1(1521)	4-4-4(12)	3-3-3(9)	**3-3-3(9)**	**3-3-3(9)**
E6(1307)	3-4-4(11)	3-3-3(9)	3-3-3(9)	E2(1173)	4-3-3(10)	2-3-2(7)	**3-3-3(9)**	**3-3-3(9)**
E7(1768)	2-4-4(10)	4-2-3(9)	2-3-3(8)	E9(1627)	4-4-3(11)	3-3-2(8)	**3-3-2(8)**	**3-3-2(8)**
E8(1516)	4-4-3(11)	3-3-2(8)	4-3-2(9)	E10(1235)	2-3-3(8)	2-2-3(7)	미활용	2-3-2(7)
E9(1627)	3-3-2(8)	2-2-1(5)	3-2-1(6)	E13(1070)	3-4-4(11)	2-3-3(8)	**3-3-3(9)**	**3-3-3(9)**
E10(1235)	2-2-2(6)	1-3-2(6)	1-2-2(5)	E14(735)	1-3-3(7)	2-2-2(6)	미활용	1-2-2(5)
E11(950)	4-3-2(9)	2-2-1(5)	3-2-1(6)	E15(1425)	4-4-3(11)	3-3-4(10)	미활용	3-4-3(10)
E17(1484)	2-3-2(7)	3-3-4(10)	2-3-3(8)	E16(639)	2-1-3(6)	2-1-2(5)	미활용	2-1-2(5)
E19(714)	2-3-1(6)	1-2-1(4)	1-2-1(4)	E17(1484)	3-3-3(9)	2-3-2(7)	미활용	2-3-3(8)

* '내용-조직-표현 및 매체' 순으로 점수 배열(괄호 안은 총점)

AI 협의 집단은 채점자 간 점수 차가 3점으로 가장 컸던 4편의 글(E1, E2, E9, E13)에서만 AI 채점 결과를 활용하였다. 특히 주목할 만한 점은 이 4편의 최종 합의 점수가 모두 AI 채점 결과와 정확히 일치했다는 사실이다. 이러한 결과는 채점자들이 협의 과정에서 AI 채점 결과를 단순히 참고하는 수준을 넘어 사실상 핵심 기준점으로 삼았음을 보여준다. 다시 말해, AI 채점 결과가 채점자 간 이견을 조율하는 과정에서 채점자 지배(rater dominance)와 같은 영향을 미쳤음을 추론할 수 있다.

반면, 전통적 협의 집단은 총 5편의 글(E3, E8, E9, E11, E17)에서 채점자 간 점수 차가 3점 이상 발생했음에도 불구하고, 채점자 간 상호작용을 통해 합의된 점수를 도출하였다. 채점자들은 '내용·조직·표현 및 매체'의 각 평가 요소에서 드러난 차이에 대해 서로의 채점 근거를 공유하며 성찰하였고, 모호한 지점에 대해서는 의견 조정을 거쳐 합의에 도달하였다. 예를 들어, E8의 경우 '내용'은 R1의 개별 채점 결과와, '조직'과 '표현 및 매체'는 R2의 개별 채점 결과와 일치하도록 점수를 조정하여 최종 합의 결과를 도출하였다. 이러한 협의는 결과적으로 점수 부여 경향을 더 엄격한 방향으로 수렴시키는 효과를 보였음을 알 수 있다.

나. 채점자 협의 부담 결과 분석

다음 〈표 9〉는 협의에 소요된 시간과 협의 난도, 개별 채점에서의 채점자 간 점수 차를 정리한 결과이다. 먼저 협의 시간과 협의 난도의 정량적 차이를 비교해 보면, 전통적 협의 집단보다 AI 협의 집단이 전반적으로 더 긴 시간을 투입하고, 더 높은 난도를 부여하는 경향을 보였다. 전통적 협의 집단의 총 소요 시간은 2,812초, 1편당 평균 협의 시간은 312.4초, 평균 난도는 1.6으로 나타났다. 반면 AI 협의 집단의 총 소요 시간은 4,260초, 1편당 평균 협의 시간은 473.3초, 평균 난도는 3.1로, 전통적 협의 집단보다 모든 부분에서 더 높은 수치를 기록하였다.

<표 9> 집단별 협의 결과

전통적 협의 집단						AI 협의 집단					
학생 글 (글자수)	협의 시간(초)	협의 난도			채점자 간 점수 차	학생 글 (글자수)	협의 시간(초)	협의 난도			채점자 간 점수 차
		R1	R2	평균				R3	R4	평균	
E3(1284)	479	2	3	2.5	3	E1(1521)	860	4	4	4	3
E6(1307)	475	2	2	2	2	E2(1173)	443	4	4	4	3
E7(1768)	516	2	2	2	1	E9(1627)	735	4	2	3	3
E8(1516)	639	1	1	1	3	E10(1235)	184	3	2	2.5	1
E9(1627)	170	1	3	2	3	E13(1070)	795	4	4	4	3

전통적 협의 집단						AI 협의 집단					
학생 글 (글자수)	협의 시간(초)	협의 난도			채점자 간 점수 차	학생 글 (글자수)	협의 시간(초)	협의 난도			채점자 간 점수 차
		R1	R2	평균				R3	R4	평균	
E10(1235)	174	1	1	1	0	E14(735)	300	2	2	2	1
E11(950)	119	1	2	1.5	4	E15(1425)	670	3	3	3	1
E17(1484)	170	1	1	1	3	E16(639)	118	2	3	2.5	1
E19(714)	73	1	1	1	2	E17(1484)	152	2	3	2.5	2
총 시간	2,812					총 시간	4,260				
평균	312.4	1.3	1.8	1.6	2.3	**평균**	473.3	3.1	3	3.1	2

* 소수점 둘째 자리에서 반올림

특히 두 집단 모두가 협의한 3편(E9, E10, E17)을 비교하면 이러한 차이가 더욱 뚜렷하게 드러난다. 전통적 협의 집단은 이 3편의 협의 시간이 평균 171.3초였던 반면, AI 협의 집단은 평균 357초로 약 두 배 더 길게 나타났다. 이는 동일한 글을 대상으로 하더라도 AI를 활용한 협의가 전통적 협의보다 상대적으로 더 많은 시간과 인지적 부담을 요구할 수도 있음을 보여준다. 이러한 결과는 교사가 AI 자동채점 결과를 활용하는 실험에서 오히려 채점 시간이 더 길어졌음을 보고한 최진영·하민수(2023)의 연구와도 맥을 같이한다.

다음으로 협의 시간과 협의 난도의 관계를 살펴보면, 개별 채점 단계와 달리 두 요소 간에 명확한 비례 관계가 나타나지 않았다. 예컨대 전통적 협의 집단에서 E8은 협의 시간이 639초로 가장 길었음에도 불구하고 협의 난도는 1로 낮게 평가되었다. 또한 AI 협의 집단에서 E2는 협의 시간이 평균(473.3초)보다 짧은 443초였으나 협의 난도는 4로 높게 평가되었다. 이는 협의 시간이 곧 협의 난도의 직접적인 지표가 아님을 보여주며, 채점자 간의 논거 제시, AI 채점 결과 활용 여부 등 다양한 상호작용적 요인이 협의 난도에 영향을 미쳤음을 보여준다.

또한 전반적으로 채점자 간 점수 차가 큰 글일수록 협의 난도도 높게 평가되는 경향이 나타났다. 예를 들어, 전통적 협의 집단의 E3은 채점자 간 점수 차가 3점으로 비교적 큰 편이었고, 협의 난도 또한 평균 2.5로 높게 나타났다. 또한, AI 협의 집단에서도 채점자 간 점수 차가 3점으로 가장 컸던 E1, E2, E13은 협의 난도 역시 가장 높은 4로 평가되었다. 즉, 채점자 간 점수 차가 크다는 것은 곧 채점자 간 이견의 폭이 크다는 것을 의미하며, 이러한 불일치가 협의 과정의 난도를 높이는 요인으로 작용한 것이다.

<표 10> 집단별 협의 결과의 평균 수치

전통적 협의 집단의 평균 수치				AI 협의 집단의 평균 수치			
학생 글	협의 시간(초)	협의 난도	채점자 간 점수 차	학생 글	협의 시간(초)	협의 난도	채점자 간 점수 차
E3	479	2.5	3	E1, E2, E13	699.3	4	3
E6, E7, E9	987	2	3	E9, E15	702.5	3	2
E11	119	1.5	4	E10, E16, E17	152.3	2.5	1.3
E8, E10, E17, E19	264	1	2	E14	300	2	1

* 소수점 둘째 자리에서 반올림

〈표 10〉의 평균 수치를 기준으로 살펴보면, 채점자 간 점수 차가 협의 난도 결정에 중요한 영향을 미친다는 사실이 보다 명확하게 드러난다. 특히 AI 협의 집단에서 이러한 경향이 두드러졌는데, 채점자 간 점수 차가 커질수록 협의 난도 역시 높아지는 양상이 뚜렷하게 나타났다. 실제로 점수 차가 1→1.3→2→3으로 커짐에 따라 협의 난도도 2→2.5→3→4로 높아지는 흐름을 보였다. 반면 전통적 협의 집단은 난도 1.5 수준에서 예외가 존재했지만, 나머지 사례들을 살펴보면 채점자 간 점수 차가 협의 난도에 일정 부분 영향을 미쳤음을 알 수 있다. 이러한 결과는 채점자 간 채점 기준 해석의 차이와 점수 불일치에 따라 채점자 간 협의 난도가 높아진다고 한 선행연구(박지수, 2021; Schauber et al., 2024)의 논의와 부합한다.

아울러 〈표 10〉의 평균 수치는 협의 시간과 협의 난도 간에 비례 관계가 성립하지 않는다는 사실 역시 분명하게 보여준다. 전통적 협의 집단의 경우 난도가 평균 2인 글(E6, E7, E9)의 평균 협의 시간은 987초로, 난도가 2.5인 E3의 479초보다 오히려 더 길게 나타났다. AI 협의 집단에서도 난도 평균이 2.5인 글(E10, E16, E17)의 협의 시간이 152.3초로 가장 짧아, 난도가 반드시 협의 시간의 증가로 이어지지 않음을 알 수 있다.

다만 두 집단 모두 협의가 뒤로 진행될수록 협의 시간이 점차 단축되는 경향을 보였으며, 이는 채점자들이 협의 절차와 채점 기준에 점차 익숙해지면서 합의의 효율성이 개선되어 갔음을 보여준다. 이러한 양상은 선행연구에서도 보고된 바 있다. Leckie & Baird(2011)는 대규모 쓰기 평가에서 채점자들이 채점 초기에는 기준 해석과 적용에 더 많은 시간을 소요하지만, 시간이 지남에 따라 점차 채점 기준이 내면화되면서 효율성이 향상된다고 밝혔다. 특히 전통적 협의 집단에서는 협의가 진행될수록 소요 시간이 점차 단축되는 경향이 더 두드러졌다. 협의 초기 단계에서 채점 기준의 의미를 명확히 하고 의견을 조율하는 과정에 상당한 시간이 소요되었으나, 일정한 합의 기준이 형성된 이후에는 이를 바탕으로 빠르게 합의가 이루어졌기 때문으로 보인다.

반면 AI 협의 집단은 AI 활용 여부도 협의 시간과 협의 난도에 유의미하게 작용하였

다(〈표 11〉). 앞서 확인했듯 채점자 간 점수 차가 3점으로 가장 큰 글에서만 AI 채점 결과가 활용되었고(E1, E2, E9, E13), 해당 글들의 평균 협의 시간은 709초, 평균 협의 난도는 3.8이었다.

<표 11> AI 협의 집단의 AI 활용 여부에 따른 차이

AI 활용						AI 미활용					
학생 글 (글자수)	협의 시간(초)	협의 난도			채점자 간 점수 차	학생 글 (글자수)	협의 시간(초)	협의 난도			채점자 간 점수 차
		R3	R4	평균				R3	R4	평균	
E1(1521)	860	4	4	4	3	E10(1235)	184	3	2	2.5	1
E2(1173)	443	4	4	4	3	E14(735)	300	2	2	2	1
E9(1627)	735	4	2	3	3	E15(1425)	670	3	3	3	1
E13(1070)	795	4	4	4	3	E16(639)	118	2	3	2.5	1
						E17(1484)	152	2	3	2.5	2
평균	709	4	3.5	3.8	3	평균	284.8	2.4	2.6	2.5	1.2

* 소수점 둘째 자리에서 반올림

이러한 결과는 AI가 협의에 추가적인 근거와 쟁점을 제공하여 협의의 심층성을 높이는 동시에, 협의 시간과 난도를 높이는 요인으로 작용하였음을 보여준다. 즉, 채점자 간 이견이 큰 글일수록 AI 채점 결과가 협의의 핵심 기준으로 활용되면서 협의가 심층화되고, 그만큼 시간과 난도도 상승하는 경향이 나타난 것이다. 반대로 AI 채점 결과를 활용하지 않은 5편의 글은 평균 점수 차가 1.2점으로 상대적으로 작았으며, 평균 협의 시간은 284.8초, 평균 협의 난도는 2.5로 나타났다. 이는 AI가 제공하는 추가적인 근거와 쟁점이 협의 과정에서 배제되면서 협의가 상대적으로 단순화되었기 때문으로 보인다. 종합하면, AI 협의 집단은 협의의 진행 단계, 채점자 간 점수 차, 그리고 AI 활용 여부 등 다양한 요인이 복합적으로 작용하며 협의 시간과 난도에 영향을 미친 것으로 보인다.

3. 전통적 협의 집단과 AI 협의 집단의 채점자 인식 비교

가. 평가 전체 과정과 채점 고난도 답안에서의 어려움: 채점 기준 해석의 차이

논술형 평가는 학생의 지식·개념·사고 과정을 종합적으로 평가할 수 있다는 장점이 있지만, 채점 기준의 해석과 적용 과정에서 채점자 간 차이가 발생할 수 있다는 구조적 한계를 지닌다. 이러한 한계는 특히 2인 이상의 다중 채점이 요구되는 상황에서 더욱 두드러진다. 여기에서는 심층 인터뷰를 통해 양적 분석에서 확인된 엄격성(logit) 차이와 낮은 일치율이 어떠한 과정적 어려움에서 비롯되는지를 확인하였다.

심층 인터뷰에서 연구 참여자들은 논술형 평가의 전체 과정에서 채점 기준 해석의 차이가 가장 큰 어려움이었다고 진술하였다. 이는 논술형 평가가 2인 이상의 채점을 전제로 하고 복합적인 역량을 측정한다는 점에서 불가피하게 나타나는 문제이다.

R1: 채점 기준을 해석하는 게 제일 힘들었어요. 저는 R2 선생님보다 전반적으로 내용을 조금 더 박하게 점수를 주는 편이었고, 반대로 조직이랑 표현 및 매체 부분은 후하게 점수를 줬어요. R2 선생님은 조직, 표현 및 매체를 저보다 더 깐깐하게 살피셨죠.

이 진술은 동일한 채점 기준표를 공유하더라도 채점자마다 적용의 기준점이 다르게 설정되어 불일치가 발생함을 보여준다. 즉, 한 채점자는 내용 요소를 엄격히 보았지만 다른 채점자는 조직과 표현·매체 요소를 더 엄격히 평가하는 방식으로 접근하면서, 해석의 차이가 곧 점수 편차로 이어진 것이다.

R3: 제가 직접 과제를 구상하고 제시한 교사가 아니기 때문에 쓰기 맥락에 대한 이해가 충분히 되지 않았던 것 같아요. 그래서 의도적으로 상향 평가한 것도 사실이고…

여기서 드러나듯, 과제 설계에 직접 참여하지 않은 상태에서 채점에 임할 경우 과제의 의도와 학생 글쓰기 맥락을 충분히 파악하기 어려워 채점 기준 적용이 흔들릴 수 있다. 이러한 해석 차이는 채점자 간 합의 과정을 지연시키는 동시에, 채점자 개인에게 인지적·정서적 부담을 가중시키는 원인으로 작용하고 있었다.

채점자들이 평가 전체 과정에서 어려움으로 지적한 지점들은 채점 난도가 높은 학생 글과도 연결되었다. 두 집단 모두 채점 기준을 명료하게 적용하기 어려운 글에서 채점 난도가 높았다고 진술하였으며, 특히 이러한 글일수록 인지적 부담이 크고 채점 시간이 오래 걸린다고 보고하였다.

R1: 10번 글은 문체가 이상하거든요. 구어체도 문어체도 아니면서 문장의 호응도 자연스럽지도 않아요. 글의 내용을 이해하기가 쉽지 않았는데, 그렇다고 내용이 아예 없지도 않잖아요? 그래서 어디서든 점수를 줄 만한 부분이 없는지 꼼꼼히 찾아야 했어요.

R4: 애초에 글이 너무 짧아서 틀릴 게 없는 경우에는 오류가 하나도 없는데 그럼 표현은 만점인가? 반대로 천 자 쓴 학생은 오류가 몇 개 있어요. 이런 경우 어떻게 판단해야 할지가 어려웠습니다.

이 두 진술은 문체나 호응처럼 해석이 애매한 요소뿐 아니라, 답안의 길이와 같은 형식적 특성이 채점 난도를 높이는 요인임을 보여준다. 즉, 답안이 모호하거나 분량에 따라 오류 발생 가능성이 달라질 경우 채점 기준의 명확한 적용이 어려워지고, 채점자는 판단 근거를 찾기 위해 더 많은 인지적 노력을 기울여야 하며, 이는 곧 채점 부담으로 이어진다.

이상의 내용을 종합하면, 두 집단 모두 채점 기준 해석의 차이를 가장 큰 어려움으로 인식했다. 채점 고난도 답안은 기준 적용이 애매한 글로, 채점자마다 엄격성과 일관성 차이를 드러냈다. 따라서 채점자 협의는 채점 단계의 절차로 한정하지 않고, 과제 설계 단계부터 채점 기준을 공유·합의하는 구조로 운영할 필요가 있다.

나. 채점자 협의 집단별 채점 경향의 비교: AI 중재 효과와 채점자 지배 현상

채점자 협의 과정에서 AI의 활용 여부의 차이는 채점자 협의 과정의 양상을 변화시켰다. 채점자 협의 결과에 따른 채점 신뢰도와 관련한 엄격성, 일관성, 그리고 채점자 부담과 관련한 협의 난도와 협의 시간에 대한 채점자의 인식을 확인하였다.

우선 채점 신뢰도 가운데서도 엄격성 조정 측면에서 두 협의 집단은 뚜렷한 차이를 보였다. 이에 대한 채점자들의 구체적인 인식을 살펴보면, 전통적 협의 집단의 채점자들은 협의 과정이 자기 성찰과 조율의 계기가 되었고, 결과적으로 채점 신뢰도를 높이는 데 기여했다고 진술하였다. 채점자들은 협의를 통해 개별 채점에서 나타난 불일치나 경향성을 점검하고, 지나치게 엄격하거나 관대한 채점을 조정할 수 있었다.

R1: 협의를 하다 보니 제가 '표현 및 매체' 부분에서 초반 글과 뒤쪽 글에 서로 다른 기준을 적용했다는 걸 알게 되었어요. 결국 협의 과정에서 일관성이 부족했던 부분을 발견하고 수정할 수 있었고, 또 너무 엄격했던 부분도 조율할 수 있었어요.

R2: "이런 경우에는 이렇게 통일합시다."라고 내부적으로 합의했어요. 그 기준을 세우니까 고민이 줄고 채점자 협의가 도움이 됐죠.

R1의 진술은 협의 과정이 채점자의 무의식적 편차를 점검하는 기회로 작용했음을 보여준다. 즉, 개별 채점에서는 인식하지 못했던 불일치가 협의를 통해 드러나고, 이를 수정함으로써 채점의 일관성과 신뢰성을 확보할 수 있었던 것이다. 또한 R2는 협의가 단순히 의견을 교환하는 수준을 넘어, 명시적인 합의 기준을 마련함으로써 채점 부담을 줄이고 결과의 일관성을 높이는 장치로 작용하였다고 진술하였다. 이는 채점자 협의가 단순한 소통 과정을 넘어 채점의 질을 개선하는 핵심 기제로 기능하고 있음을 보여준다.

반면, AI 협의 집단에서는 AI 채점 결과가 협의 과정에 직접적인 영향을 미치는 양상이 나타났다. 채점자마다 AI 활용에 대한 인식 차이는 있었지만, 대체로 AI 결과를 객관적 기준으로 받아들이며 자신의 채점 경향을 수정하거나 조정하는 데 활용하였다.

R3: AI와 동료가 같은 점수를 주면 저는 그걸 '3자 협의의 2:1'로 받아들였어요. 그래서 제 판단이 후하다는 걸 깨닫고 조정했죠.

R3: AI는 프롬프트에 제시된 기준만 확인해 정량적이고 표준적으로 판단한다는 느낌이었고, 그런

근거를 저희도 학습해서 뒤쪽 협의는 AI 없이도 진행할 수 있었어요.

이 진술은 AI 결과가 제3의 채점자로 기능하면서, 채점자가 스스로의 엄격성이나 관대함을 성찰하고 조정하는 계기가 되었음을 보여준다. 실제로 엄격성이 −2.76 로짓으로 가장 관대한 채점자였던 R3는 AI 결과를 '객관적 근거'로 삼아 협의 과정에서 자신의 채점 기준을 수정하였다. 또한 AI가 제시한 근거를 학습함으로써 채점자들이 후반부 협의를 AI 없이도 수행할 수 있었던 점에서, AI는 단순한 합의 도구를 넘어 학습을 통한 자기 성찰과 함께 채점 방식 개선의 도구로 확장될 수 있음을 보여준다.

R4: 처음 결과를 받았을 때 제 점수와 같아서 기분이 좋았습니다. 그런데 저도 모르게 그 점수를 '정답'처럼 인식하게 되더군요.

이처럼 AI 집단의 채점자는 AI 결과와 자신의 판단이 일치할 때 안도감을 느끼거나, 무의식적으로 AI 점수를 '정답'으로 받아들이는 경향을 보였다. 따라서 AI 협의 집단에서는 표면적으로 합의가 쉽게 이루어지는 효과가 있었지만, 동시에 협의가 점수 일치 여부에 과도하게 의존하는 경향도 드러났다.

다음으로 채점 신뢰도의 일관성 측면에서 두 집단의 협의는 서로 다른 양상을 보였다. 전통적 협의 집단은 상호작용을 통해 채점 기준의 의미를 조율하며 일관성을 높여 갔던 반면, AI 협의 집단에서는 AI 채점 결과가 강력한 기준점으로 작용하면서 채점자 간 점수 차이의 조정이 AI 채점 결과로 수렴하는 결과로 나타났다. 이와 함께 채점자들이 자신의 해석을 고수하기 어려워지고, 협의가 심층적 논의보다는 표면적 합의로 흐르는 경향이 나타났다.

R4: 결국 AI 점수가 협의의 중심이 되다 보니, 내가 처음 생각한 기준을 고수하기 어렵더라고요.

R4: AI 점수가 정답처럼 인식되니까 한쪽이 맞고 다른 한쪽은 틀린 것처럼 돼요. 협의가 아니라 합의, 즉 타협으로 변질되는 느낌이었어요.

이 진술은 AI 결과가 정답처럼 받아들여지는 순간, 협의가 본래의 목적을 상실하고 단순한 점수 일치의 장으로 축소될 위험이 있음을 보여준다. 즉, AI 활용이 협의의 효율성을 높일 수 있다는 장점과 달리, 채점자 간 상호 설득과 조율 기능을 약화시키는 구조적 한계도 드러난 것이다. 특히 AI 협의 집단에서는 AI 점수가 일정한 구간에 집중되면서 채점 변별력이 축소되고, 협의 과정에서 AI 결과가 사실상 기준점이 되어 특정 채점자의 의견이 우세하게 반영되는 채점자 지배(Johnson et al., 2005)가 AI에 의해 나타날 수 있음을 보여준다.

R3: 사실 계속해서 AI 채점 결과를 확인하는데 약간의 점수 부여가 좀 획일화된 '3-3-3'으로만 나오더라고요. 그래서 인공지능 채점이 너무 중앙에 몰려 있는 게 아닌가 하는 생각이 들었어요.
R4: 저희가 넣었던 글들이 '올 3점'으로 나와서 중심 경향성이 있나, 제대로 채점하는 게 맞나 하는 생각이 들었어요.

이 진술들은 AI 채점 결과가 중앙값에 과도하게 수렴해 점수 분포의 다양성을 약화시키고, 그 결과 학생들의 수행 수준 차이가 충분히 드러나지 못했음을 보여준다. 동시에 AI 점수가 협의의 기준점으로 작용하면서, 협의는 상호 논의의 장이라기보다 AI가 주도하는 표면적 합의 과정으로 변질될 위험을 내포한다. 이는 AI 협의가 효율성을 제공하는 동시에, 채점 변별력 축소와 채점자 지배 현상이라는 구조적 한계를 동반함을 보여준다.

이처럼 R3와 R4는 채점자 협의 과정에서 AI에 대한 인식에 차이가 존재하였다. R3가 비교적 채점자 협의 과정에서 AI 협업을 긍정적으로 인식했으며, 이에 대한 수용적인 자세가 드러났다. 반면, R4는 AI 협업이 오히려, 채점자 간의 협의를 위축시킨다고 보아 AI 협업을 부정적으로 인식하였다. 이는 채점자 협의 과정에서 채점자가 AI 활용을 어떻게 인식하느냐에 따라 채점자 협의에 영향을 줄 수 있음을 보여준다. 따라서 사전에 채점자 협의 과정에서 AI의 역할 내지 활용 방안을 명확히 확립하고 이를 토대로 채점자 협의 과정에서 AI와 협업할 필요가 있다.

다음으로 채점자 부담 측면에서, 전통적 협의 집단은 관계적 부담은 있었으나 비교적 짧은 시간 안에 합의에 도달할 수 있었다. 반면 AI 협의 집단은 AI 점수를 읽고 해석하며 검토하는 절차가 추가되면서 협의 시간이 길어지고 협의 난도도 높아졌다. 우선 이에 대한 전통적 협의 집단의 채점자 인식을 살펴보면, 1:1 협의에서 관계적 부담이 존재했지만, 상호작용과 대화를 통해 초기 해석 차이를 조율하며 점차 합의의 틀을 형성했다.

R1: 서로의 채점 기준을 이야기하면서 '이 정도는 이렇게 하자'로 맞춰 갔고, 뒤로 갈수록 합의가 훨씬 쉬웠습니다.
R2: 점수 차이가 조금 나면 왜 이런 식으로 매겼는지를 얘기하는 과정에서 직접적으로 말하지는 않지만, 그런 것을 표현하기 위한 완곡어법을 신경 쓰면서 말을 했어요.

두 발화는 채점자 협의가 단순한 채점 기준 공유를 넘어, 채점자 간의 관계를 고려하면서 채점 일관성을 높여 가는 과정임을 보여준다. 즉, 초반에는 채점 기준 해석의 차이와 관계적인 심리적 부담이 크지만, 상호 설명, 조정, 합의의 순환을 거치며 후반부로 갈수록 협의 부담이 줄어드는 점을 확인할 수 있다. 이는 채점자 협의가 채점자 간의 심리적 관계와 의사소통 방식에 따라 질적으로 달라질 수 있음을 뜻한다(강민석, 2025).

반면 AI 협의 집단은 전통적 집단에 비해 협의 시간이 길고 협의 난도가 높은 것을 당연한 결과라고 반응하였다. 즉, 논술형 평가의 경우, 채점 해석의 차이로 채점자 협의

가 이루어지는데 AI 채점 결과가 있으니, 이에 대한 읽기, 분석, 조율 등의 과정을 거치면서 협의 시간과 협의 난도는 AI를 활용하지 않은 전통적 협의보다 올라갈 수밖에 없다고 판단하였다. 다만 전통적 협의와 비교했을 때, AI 채점 결과까지 더해지면서 2:1의 상황이 되다 보니이에 따른 심리적 부담은 낮았으나, 평가 논의가 심층적으로 이루어지는 과정에서는 관계적 측면 등이 작용하여 채점자의 성향에 따라 앞서 언급한 점수 합의 중심의 협의로 마무리되는 경향이 나타났다.

R3: AI를 통해 채점에 대한 근거를 확인하면 합의가 잘 되기도 하는데, 그 과정이 더 힘들었어요. AI 점수를 설득력 있게 받아들이지 못하면 이야기가 막히는 거예요.

R4: 저희도 알다시피 현장에서는 편함을 추구하는 분들이 많기 때문에, 결국엔 결과가 나온 순간 '내가 틀렸구나' 하고 2:1 상황처럼 받아들이게 된다는 게 문제죠.

즉, AI 협의 집단은 협의 과정이 AI의 개입으로 채점자의 인지적 부담이 증가했으나, 동시에 AI가 협의의 기준점을 제공함으로써 불필요한 갈등을 예방하는 효과도 있었다. 그러나 AI의 활용이 표면상 일부 채점자의 부담을 주는 측면이 있으나, 그 이면에는 오히려 AI를 활용함으로 학생 글에 대한 분석과 채점자 간의 조율 과정이 전통적 협의에 비해 좀 더 심층적으로 이루어질 수 있음을 보여준다. 그러나 AI의 역할에 따라 채점자 협의가 점수 합의 방식으로 흐를 우려도 있음을 확인할 수 있었다. 따라서 AI 활용의 효과를 단순히 부담의 증감으로 제한해서 볼 것이 아니라, 협의 과정에서 AI가 어떻게 활용되어 학생 글을 얼마나 타당하게 평가하는지에 초점을 두고 확인할 필요가 있다.

이상의 내용을 종합하면, 전통적 협의 집단은 상호작용과 조율을 통해 엄격성이 강화되면서 일관성을 확보하며 짧은 시간 안에 합의에 도달했다. 반면 AI 협의 집단은 AI 점수를 기준으로 차이를 조정하는 중재 효과가 있었고, 이를 통해 채점 근거를 공유하며 심층적 논의가 이루어질 가능성도 확인되었다. 그러나 동시에 표면적 합의, 협의 시간 증가, 난도 상승 등의 한계가 나타났다. AI 점수가 정답처럼 인식될 경우 특정 해석이 지배할 위험이 있으므로, AI 결과는 참고 자료로 제한하고 협의는 근거 중심 구조와 명확한 역할 분담 아래 운영할 필요가 있다.

V. 결론 및 제언

본 연구는 논술형 평가 상황에서 AI 협업 기반 채점자 협의와 전통적 채점자 협의를 실시한 뒤, 개별 채점 신뢰도 및 인지적 부담, 채점자 협의 결과 및 인지적 부담, 채점자

협의에 대한 인식의 차이를 밝히는 데 목적이 있다. 이를 통해 학교 현장에 AI 협업 기반 채점자 협의 모델을 도입하는 과정에서 고려해야 할 교육적 시사점을 제언하고자 하였다.

우선, 연구의 주요 결과를 정리하면 다음과 같다. 첫째, 개별 채점 단계에서 채점 신뢰도 차이가 뚜렷하게 드러났으며, 이는 Rasch 분석을 통해 구체적으로 확인되었다. 채점자 간 신뢰도인 엄격성은 뚜렷한 차이를 보였으며, 채점자 내 신뢰도인 일관성 측면에서 부적합, 과적합 채점자가 나타났다. 원점수 기반의 완전 일치도 또한 40~60% 수준에 머물렀다. 이는 논술형 평가를 비롯한 글쓰기 평가 상황에서 채점자의 다양한 요인이 채점 신뢰도에 영향을 미치며(박영민, 박종임, 2011; 설현수, 2010; Engelhard, 1994; Linacre, 1989; Lunz & Stahl, 1990), 이 같은 차이를 해결하기 위해 채점자 협의가 도입되었다는 기존의 연구 결과(서수현, 2012; Johnson et al., 2005)를 뒷받침한다. 따라서 본 연구 또한 논술형 평가의 신뢰도를 확보하기 위해 채점자 간 협의 과정이 필수적임을 증명했다고 할 수 있다.

둘째, 전통적 협의 집단과 AI 협의 집단에서 도출한 점수의 엄격성과 일관성을 분석한 결과 뚜렷한 차이가 나타났다. Rasch 분석 결과, 전통적 협의 집단은 최종 점수가 다소 엄격한 방향으로 수렴했으며, 일관성은 적합한 범위 내에 위치하였다. 그러나 AI 협의 집단의 최종 점수는 Rasch 모델에서 상정하는 평균적인 수준(0)으로 수렴하였으며, 일관성은 사용하는 점수의 변별력이 약화되어 과적합으로 분석되었다. 특히 AI의 채점 결과 및 피드백을 참고하여 부여한 4편의 최종 점수는 모두 AI의 채점 결과와 일치했는데, 이는 AI가 협의 과정에서 일종의 '정답'처럼 작용했음을 보여준다. 이 같은 결과는 AI가 채점자 간 불일치를 조정하는 중재자로 기능할 수 있으나, 그 결과가 채점자 협의를 표면적인 점수 합의로 축소시키는 채점자 지배(rater dominance) 현상의 연장선에 있음을 의미한다.

셋째, 전통적 협의 집단과 AI 협의 집단이 인식한 채점자 협의에 대한 부담을 협의 시간, 협의 난도 측면에서 비교한 결과, AI 협의 집단의 부담이 더 높은 것으로 나타났다. 전통적 협의 집단보다 AI 협의 집단이 전반적으로 더 긴 시간 협의하고, 협의 난도를 높게 인식하였다. 이는 AI 채점 결과와 피드백이 협의 과정에서 추가적인 근거와 쟁점을 제공하여 논의의 심층성을 높이는 동시에, 협의 부담을 가중시키는 요인으로 작용했음을 보여준다.

넷째, 채점자 협의에 참여한 교사들의 심층 인터뷰를 분석한 결과 개별 채점 및 채점자 협의 과정에서 겪은 어려움의 원인과 양상이 도출되었으며, 두 집단 간 뚜렷한 차이가 나타났다. 특히 전통적 협의에 참여한 교사들은 협의를 통해 자신들의 채점 경향을 반성적으로 성찰하고 동료의 시각을 공유할 수 있었던 점을 긍정적으로 평가하였다. 그

러나 동시에 협의 과정에서 발생하는 긴장을 해소하기 위해 완곡어법을 사용하는 등 의사소통 방식을 달리하여 채점자 협의로 발생하는 관계적 부담을 피하려는 양상을 확인할 수 있었다. 반면 AI 협의 집단의 교사들은 AI 점수가 협의의 기준점으로 작용하여 다수의 판단을 존중하는 차원에서 협의가 이루어질 수 있었으나, AI 결과에 대한 과도한 의존, 관계적 측면을 고려한 점수 중심의 합의 등으로 채점자 협의가 형식적 절차로 축소될 수 있다는 우려를 표하기도 했다. 이러한 결과는 AI 기반 협의의 가능성과 위험성을 동시에 보여주는 것으로, AI 협의의 실제 적용 과정에서 중요한 시사점을 제공한다.

이상의 논의를 토대로, 본 연구는 향후 학교 현장에서 AI 협업 기반 채점자 협의 모델(이하, AI 협의)을 도입할 때 고려해야 할 몇 가지 교육적 시사점을 제언하고자 한다. 첫째, AI 협의는 협의 시간과 난도 측면에서 오히려 채점자의 부담을 다소 증가시킬 수 있으나, 동시에 AI가 제공하는 근거 중심 피드백이 채점자의 자기 판단을 점검하게 하는 긍정적 효과도 확인되었다. 따라서 여러 교사가 함께 참여하는 채점자 협의보다는 교사 단독으로 특정 과목을 맡아 수업과 평가를 병행해야 하는 상황에서 더욱 효과적으로 활용될 수 있다. 동료 교사와의 협의가 불가능한 상황에서 AI는 채점자의 '가상 협의 파트너'로 기능하며, 교사에게 대안적 관점을 제공함으로써 채점 신뢰도를 보완할 수 있을 것이다. 이를 위해서는 AI가 협의 과정의 어느 단계에 개입할 것인지 명확히 설계하고, 협의 초반에는 AI가 제시하는 예시 답안, 채점 기준 해석, 불일치 사항 등을 참고 자료로 활용하도록 하는 방안이 구체화될 필요가 있다. 또한 교사가 단독으로 평가를 수행하는 경우를 대비해, AI 협의 기능이 탑재된 온라인 평가 보조 도구를 개발하여 보급한다면 현장 활용도가 높아질 것이다.

둘째, AI 협의는 단순한 점수 합의에 머물 위험이 있으나, 동시에 심층적 논의를 촉발하는 출발점으로 기능할 수 있다. 따라서 다수의 채점자가 참여하는 대규모 협의 상황에서는 AI 채점 결과와 피드백을 협의 초반의 기초 자료로 활용하는 것이 효과적일 수 있다. 예를 들어, 협의 초반에는 AI의 결과를 참조하되, 중반부 이후에는 채점자 간 근거 중심 대화, 사례 공유, 기준 수정 논의를 중심으로 협의를 진행하도록 협의 단계를 구성할 수 있다. 이때 AI는 채점자 간 논의를 활성화하고 기준 해석의 다양성을 드러내는 도구로 의미를 가질 수 있을 것이다.

셋째, AI 협의의 교육적 가치는 결국 인간 채점자의 태도에 의해 좌우될 것으로 판단된다. 생성형 AI의 결과를 무비판적으로 수용할 경우 채점자 협의는 형식적 절차로 축소될 수 있으나, AI가 제공하는 결과를 비판적이고 성찰적으로 수용한다면 협의의 심층성과 타당성을 동시에 높일 수 있다. 따라서 교사들이 AI를 활용하되, 평가의 주체적 권한과 책임을 유지할 수 있도록 비판적 수용 역량을 함양하는 것이 무엇보다 중요하다. 이를 위해 교사 연수와 전문성 개발 과정에서 'AI 활용 채점자 협의 실습형 워크숍'을

토대로 AI 활용에 대한 교육적 성찰을 강화하는 노력이 병행되어야 한다.

본 연구는 몇 가지 한계를 지닌다. 본 연구는 탐색적 성격의 소규모 연구로서 Rasch 분석 결과의 통계적 해석을 일반화하기 어렵고, 연구 참여자가 편의 표집으로 선정되었다는 점에서도 결과 해석에 한계가 있다. 또한 고등학교 국어 교과의 논술형 평가 과제에 국한되었기 때문에, 학교급이나 다양한 교과, 과제 유형에서 동일한 결과가 나타날지는 추가 검증이 필요하다. 마지막으로 다른 생성형 AI 도구들과의 비교 검증이 이루어지지 못하였다. 후속 연구에서는 학교급별로 더 많은 교사와 과제를 대상으로 한 대규모 연구, 다양한 종류의 생성형 AI 비교 분석, 그리고 AI 협의의 장기적 효과를 검증하는 연구가 수행되어야 할 것이다.

그럼에도 불구하고 본 연구는 AI 협의와 전통적 협의의 양상을 비교, 분석하여 AI 협의가 지니는 가능성과 한계를 도출했다는 점에서 의의가 있다. 특히 AI 협의가 전통적 협의를 대체하기보다는 보완하는 방향으로 설계되어야 한다는 점을 분명히 함으로써, 향후 국어 교육뿐만 아니라 다양한 교과의 평가 장면에서 AI를 활용한 채점자 협의의 실행 가능성을 탐색할 수 있는 중요한 기초 자료를 제공했다는 점에서 의의가 있다.

참고문헌

강민석(2025). 한국어 쓰기 평가에서 채점자 간 협의의 양상 연구–채점 특성 변화와의 비교를 중심으로–. **Journal of Korean Culture, 68**, 93–136.

경기도교육청(2025). **고등학교 학업성적관리 시행지침**. 경기도교육청.

교육부(2022). **국어과 교육과정**(제2022–33호 [별책 5]). 교육부.

권경문, 최숙기(2025). AI 기반 한국어 글쓰기 자동 피드백의 질 평가 연구. **청람어문교육, 104**, 205–244.

김경희, 조지민, 김영은, 김혜숙, 나우열, 박지선, 송민호, 이동욱, 이민형, 임은영, 김선희, 박찬호(2024). **학생 역량 및 성장 중심의 학생평가 체제 개선 방안**(RRE 2024–8). 한국교육과정평가원.

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. **교과교육학연구, 27**(1), 31–50.

박소영, 이병윤, 홍유정(2024). ChatGPT를 활용한 AI 글쓰기 의사소통 역량 평가도구 개발 과정에 대한 연구: 기술 전문가와의 상호소통을 중심으로. **실천공학교육논문지, 16**(1), 21–31.

박영민, 박종임(2011). Rasch 모형을 활용한 국어교사의 채점 일관성 변화 양상 및 원인 분석. 우리어문연구, 39, 1–25.

박종미(2025). 문법 학습 비계 설정자로서 ChatGPT 활용을 위한 과제와 방향. **국어교육학연구, 60**(1), 83–109.

박종임(2024). 국어과 서·논술형 평가의 도입 현황 및 실행 상의 쟁점 탐색 연구. **청람어문교육, 101**, 273–307.

박종임, 최숙기, 박강윤, 김길재(2023). 채점자질을 활용한 인공지능 기반 글쓰기 자동채점 방안 탐색. **청람어문교육, 96**, 135–172.

박종임, 김유향, 박지선, 백승주, 정수진, 김진아, 권경문, 김형성, 사민선, 최숙기(2024). **서·논술형 평가 도구 개발의 방법과 사례 국어**(연구자료 ORM 2024–150–2). 한국교육과정평가원.

박지수(2021). 한국어 쓰기 평가의 채점자 요인 연구. 박사학위 논문, 서울대학교

박혜영, 김성숙, 김경희, 이명진, 김광규, 김지영(2019). **수업-평가 연계 강화를 통한 서·논술형 평가 내실화 방안**(RRE 2019-6). 한국교육과정평가원.

서수현(2012). 쓰기 평가 협의 과정에 나타난 쓰기 평가자의 인식 연구. **국어교육학연구, 44**, 335-367.

설현수(2010). 평정자간의 엄격성 차이정도가 피험자 총점산출 방법에 미치는 영향: 원점수, 표준점수, Facet점수 비교. **교육평가연구, 23**(1), 125-147.

송민호, 김동영, 김진구, 박상욱, 박종임, 정수진(2024). **교과 서·논술형 평가 자동채점을 위한 인공지능 모델 적용 방안 연구(Ⅰ)**(RRE2024-9). 한국교육과정평가원.

신상윤, 강신천(2024). 중등학교에서 생성형 AI를 학생 평가도구로 활용하기 위한 기초 연구: 중등교사의 생성형 AI에 대한 인식과 경험 분석에 기초하여. **컴퓨터교육학회 논문지, 27**(9), 1-14.

오세영, 박준용, 배재성(2013). 국어교사의 총체적·분석적 쓰기 채점과 평가자 협의 과정 연구. **국어교육, 141**, 183-230.

인천광역시교육청(2025). **고등학교 학업성적관리 시행지침**. 인천광역시교육청.

최숙기(2023). 국어과 서·논술형 수능 평가 문항 개발 방안 연구. **청람어문교육, 91**, 135-178.

최예영, 이남호(2025). 생성형 인공지능 활용 국어교육 연구의 동향 분석 - 체계적 문헌 분석과 LDA 토픽 모델링을 바탕으로. **청람어문교육, 104**, 143-174.

최진영(2025). 생성형 AI와 교사 협업 기반 글쓰기 피드백 방안 연구. 박사학위논문. 한국교원대학교.

최진영, 김지수, 김형성(2025). 논술형 평가 전문성 강화를 위한 AI 자동채점 기반 평가자 협의 모델 개발: 교사-AI 협업을 중심으로. **교육정보미디어연구, 31**(4), 1297-1329.

최진영, 하민수(2023). 국어과 읽기 영역 서술형 평가를 위한 비지도 기반 인공지능 채점 보조 프로그램(SAAI)의 성능과 활용도 탐색. **청람어문교육, 92**, 7-48.

함은혜, 박소영, 이병윤, 이성혜, 이유경, 홍유정(2024). GPT-4를 활용한 과학탐구역량 자동채점의 특성 분석. **교육정보미디어연구, 30**(3), 713-742.

Attali, Y., & Burstein, J.(2005). *Automated essay scoring with e-rater® v.2.0*(Research Report No. RR-04-45). Educational Testing Service.

Bond, T. G., & Fox, C. M. (2015). *Applying the rasch model: Fundamental measurement in the human sciences(3rd ed.)*. Routledge.

Broad, B.(2003). *What we really value: Beyond rubrics in teaching and assessing writing*. Utah State University Press.

Colombini, C. B., & McBride, M.(2012). "Storming and norming": Exploring the value of group development models in addressing conflict in communal writing assessment. *Assessing Writing, 17*(4), 191-207.

Creswell J. W., & Creswell, J. D.(2018). *Research design: Qualitative, quantitative, and mixed methods approaches(5th ed.).* **정종진 외 역(2022). 연구방법: 질적·양적 및 혼합적 연구의 설계**. 시그마프레스.

Engelhard, G., Jr.(1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement, 31*(2), 93-112.

Ercikan, K., & McCaffrey, D. F. (2022). Optimizing implementation of artificial-intelligence-based automated scoring: An evidence centered design approach for designing assessments for AI-based scoring. *Journal of Educational Measurement, 59*(3), 272－287.

Evmenova, A. S., Regan, K., Mergen, R., & Hrisseh, R.(2024). Improving writing feedback for struggling writers: Generative AI to the rescue?. *TechTrends, 68*(6), 790－802.

Jansen, T., Höft, L., Bahr, L., Fleckenstein, J., Möller, J., Köller, O., & Meyer, J.(2024). Comparing generative AI and expert feedback to students' writing: Insights from student teachers. *Psychologie in Erziehung und Unterricht, 71*(2), 80－92.

Johnson, R. L., Penny, J., Gordon, B., Shumate, S. R., & Fisher, S. P.(2005). Resolving score differences in the rating of writing samples: Does discussion improve the accuracy of scores?. *Language Assessment Quarterly, 2*(2), 117-146.

Kim, J., Yu, S., Detrick, R., & Li, N.(2025). Exploring students' perspectives on generative AI-assisted academic writing. *Education and Information Technologies, 30*, 1265-1300.

Kumar, V. S., & Boulanger, D.(2021). Automated essay scoring and the deep learning black box: How are rubric scores determined?. *International Journal of Artificial Intelligence in Education, 31*(3), 538－584.

Leckie, G., & Baird, J.-A.(2011). Rater effects on essay scoring: A multilevel analysis of severity drift, central tendency, and rater experience. *Journal of Educational Measurement, 48*(4), 399-418.

Li, Y., Raković, M., Srivastava, N., Li, X., Guan, Q., Gašević, D., & Chen, G.(2025). Can AI support human grading? Examining machine attention and confidence in short answer scoring. *Computers & Education, 228*, 105244.

Linacre, J. M.(1989, March). Objectivity for judge-mediated certification examination. *Paper presented at the Annual Meeting of the American Educational Research Association*, San Francisco, CA, United States.

Lumley, T.(2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters?. *Language Testing, 19*(3), 246-276.

Lunz, M. E., & Stahl, J.(1990). Judge consistency and severity across grading periods. *Evaluation and the Health Professions, 13*(4), 425-444.

Lynch, B. K., & McNamara, T. F.(1998). Using G-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing, 15*(2), 158-180.

McNamara, T. F.(1996). *Measuring second language performance.* Longman.

McNamara, T. F., & Adams, R. J.(1991, March). Exploring rater behaviour with Rasch techniques[Paper presentation]. *Annual Language Testing Research Colloquium, Princeton*, NJ, United States.

Schauber, S. K., Schloegl, M., Cichon, I., Tiffe, T., Resch, F., & Rotthoff, T.(2024). Inconsistencies in rater-based assessments mainly affect borderline performances. *BMC Medical Education, 24*(1), 382. https://doi.org/10.1186/s12909-024-04952-3

Stanford HAI(2025), *The Artificial Intelligence Index Report 2025.* CA: Institute for Human-Centered AI, Stanford University.

Stewart, J., Anthony, L., Batty, A. O., Nakamura, K., Nicklin, C., McLean, S., & Tomaru, K.(2025). Can we reliably score meaning recall vocabulary tests using AI? A comparison of human vs. AI scoring. *Computer Assisted Language Learning*, 1-23.

Wetzler, E. L., Cassidy, K. S., Jones, M. J., Frazier, C. R., Korbut, N. A., Sims, C. M.,

Bowen, S. S., & Wood, M.(2024). Grading the Graders: Comparing Generative AI and Human Assessment in Essay Evaluation. *Teaching of Psychology, 52*(3), 298-304.

Wright, B. D., & Linacre, J. M.(1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*(3), 370-371.

논문접수 : 2025.10.1. / 수정본 접수 : 2025. 10. 29. / 게재승인 : 2025. 11. 6.

ABSTRACT

Comparing AI-Collaboration-Based Rater Discussion and Traditional Rater Discussion in Essay Assessment

Jisoo Kim
Teacher, Cheomdan High School
Jinyoung Choi
Teacher, Ilsandong High School
Hyungsung Kim
Faculty, Korea Science Academy of KAIST
Bora Song
Teacher, Gimhae Daecheong High School

This study empirically explores ways to enhance scoring reliability and reduce rater burden by comparing AI-collaboration-based and traditional rater discussions in essay assessment. Using an explanatory sequential design, quantitative and qualitative methods were employed. In the quantitative phase, individual scores were analyzed using Rasch modeling to compare raters' severity, consistency, consultation time, and perceived difficulty. The qualitative phase involved semi-structured interviews on assessment challenges, group-specific consultation characteristics, and raters' experiences with AI use. Findings revealed that the traditional group identified discrepancies in rubric interpretation and relational tension but also valued sharing perspectives and reflecting on scoring. In contrast, AI-collaboration provided benchmarks for consensus and mediated differences but risked overreliance and score-centered negotiation. It also entailed longer discussion time and higher perceived difficulty, as AI feedback introduced additional evidence and points of contention, increasing cognitive load while deepening deliberation. The study concludes that AI-collaboration discussion should complement rather than replace traditional discussion, contributing to scoring reliability and reflective deliberation in school assessment. Based on these findings, considerations are proposed for implementing AI-collaboration-based rater discussion models in school assessment practices.

Key Words: AI-Collaboration-Based Rater Discussion, AI Automated Scoring, Essay Assessment, Rater Discussion, Writing Assessment

부록

<부록 – 표> 채점 기준 정교화를 위한 예비 채점 AI 프롬프트 설계(최진영 외, 2025:1312)

<table>
<tr><th>프롬프트 내용</th><th>설계 의도</th></tr>
<tr><td># GPT의 역할
{1. 현재 상황: GPT는 한국에서 고등학교 3학년 학생을 대상으로 국어를 가르치는 교사이고, 현재 〈화법과 작문〉이라는 과목을 가르치고 있다.
2. 학생 수행 과제: 과제는 학생들이 교과서에서 자신의 진로와 관련된 지문을 선정해 진로를 소개하는 발표 상황을 가정한 발표를 위한 글쓰기이다.
3. 학생 수행 과정 : 학생들은 지문과 진로의 관계를 먼저 밝힌 뒤, 주요 내용을 구성한다.
4. 평가 유형과 특징 : 이 과정에서 3단 구성, 효과적인 전개 방식, 적절한 표현 전략을 활용하고, 맞춤법을 준수해 내용을 전달하는 능력을 향상시키기 위해 논술형 평가를 시행하고 있다.
5. 요구사항 : 아래 〈주의 사항〉, 〈해야 할 일〉을 이해하고 〈채점 기준표〉와 〈채점 예시〉를 학습해서 결과를 도출하라.}</td><td>→ # : 구분자 사용
→ { } : GPT가 인식해야 할 명령
→ 1~n. : 역할 인식을 순차적으로 이해하도록 구성

→ < >: 강조</td></tr>
<tr><td># 주의 사항
{1. 채점 기준표에서 사용하는 용어는 다음과 같이 정의함.
– 도입 : 단순 도입 인사뿐만 아니라, 발표의 계기, 지문 선택 이유 등이 하나라도 포함이 되어야 함.
– 정리 : 단순 마무리 인사만 들어가는 것이 아니라, 요약, 의의, 제언, 포부 중 하나라도 포함이 되어야 함.
2. 구체성을 판단할 때는 작성 분량 또한 고려해서 판단할 것.
3. 문단을 구분할 때는 줄을 바꿔야 함. 너무 잦은 줄 바꿈은 문단이 제대로 이루어지지 않은 것으로 판단함.}</td><td>→ # : 2번째 인식
→ { } : GPT가 인식해야 할 명령
→ 1~n. : 채점의 정확도를 위하여 교사의 시각을 공유하는 채점 상의 유의점</td></tr>
<tr><td># 해야 할 일
{단계별로 생각해 보자. (let's think step by step)}
[1단계 : 우선 〈채점 기준표〉를 학습하라.]
[2단계 : 〈채점 예시〉에는 전체 학생의 글에서 '에크모'와 관련된 9편의 학생 글과 이 글에 부여한 점수, 점수 부여 근거가 제시되어 있다. 이 내용을 〈채점 기준표〉와 연계해 논리적으로 학습하라.]
[3단계 : 학생의 글 25편을 GPT가 학습한 〈채점 기준표〉, 〈채점 예시〉에 따라 분석하여 채점 근거를 명확하게 밝혀야 한다. 단, 이때 학생의 수준이 각기 다르다고 전제하고, 채점 기준표를 엄격하게 적용해서 학생 글에서 채점 근거를 찾아야 한다. 그리고 그 근거에 기반해서 평가 요소별 수준을 기술하라.]
[4단계 : 1~3단계의 분석 내용을 바탕으로 채점 기준별 점수를 산출하라.]
[5단계 : 채점 기준별 점수와 함께 총점을 계산해서 표로 제시하라.]</td><td>→ # : 3번째 인식
→ step by step: LLM 모델의 사고 과정을 강화하는 명령어 추가
→ [ ] : step by step에 따라 GPT가 인식해야 하는 단계별 수행 명령
→ < >: 강조</td></tr>
<tr><td># 채점 기준표
{ 1단계에서 완성된 채점 기준표 제시 }</td><td>→ # : 4번째 인식
→ { }: 채점기준표 삽입</td></tr>
<tr><td># 채점 예시
<table><tr><th>학생 글</th><th>내용</th><th>조직</th><th>표현</th></tr><tr><td>1</td><td>2점
진로와 선정 지문의 연관성이 부족</td><td>2점
결론 : 요약, 의의, 제언, 포부가 들어가야 마무리로 인정함.</td><td>2점</td></tr><tr><td>2</td><td>1점
자료를 활용하지 않고 자기 경험만 진술</td><td>1점
전체가 도입으로 작성되어 있음. 구조, 내용 전개 방식이 없음.</td><td>2점
매체 활용
+질문 제시</td></tr></table></td><td>→ # : 5번째 인식

→ 교사의 채점 점수 협의 결과와 그에 따른 근거 자료를 그대로 입력함.</td></tr>
<tr><td># 평가해야 할 학생 글
{채점 예시와 다른 25편의 학생 글 제시}</td><td>→ # : 구분자 사용
→ { }: 학생 글 삽입</td></tr>
</table>