교육과정평가연구
The Journal of Curriculum and Evaluation
2026, Vol. 29, No. 2, pp.141~164
DOI: https://doi.org/10.29221/jce.2026.29.2.141

# AI 기반 과정 중심 평가 시스템의 교육적 의미 탐색: 실행연구를 중심으로[1)]

정윤리*, 김유정, 권정인, 채상미, 신병철
(경기과학고등학교 교사)

**요약**

본 연구는 G영재학교에서 AI 기반 과정 중심 평가 시스템을 적용하는 과정에서 나타난 교육적 의미와 개선 과정을 탐색하였다. 교사 연구자 4인이 참여한 실행연구를 통해 자기 보고서, 반구조화 면담, 학생 및 교사 설문을 수집하고 분석하였다. 연구 결과는 다음과 같다. 첫째, AI 기반 평가 시스템은 맞춤형 피드백 제공과 반복적 채점 업무 경감을 통해 교사가 학생과의 개별적 상호작용에 집중할 수 있는 여건을 조성하는 교육적 가능성을 보여주었다. 둘째, AI의 채점 일관성 부족, 필기체 인식 오류, 고차원적·정의적 영역 평가의 한계가 확인되었다. 셋째, 연구 참여자들은 1, 2차 실행의 성찰을 통해 교과 특성에 따른 적용 범위의 선별적 설정, 정밀한 루브릭 설계, 교사 검증 체계 구축이라는 세 가지 현장 적용 조건을 도출하였다. 넷째, 연구 참여자들은 AI를 평가의 보조 도구로 위치시키고, 루브릭 설계자이자 비판적 검토자로서 교사 역할을 확장하며, 학생-교사 간 교육적 관계의 가치를 재확인하는 방식으로 전문성을 재구성하였다. 이러한 결과는 AI 기반 평가 시스템이 과정 중심 평가의 취지를 실현하는 데 기여할 수 있으나, 교사의 능동적 개입과 전문적 판단이 전제되어야 함을 시사한다.

주제어 : AI 기반 평가, AI 기반 과정 중심 평가, 과정 중심 평가, 학습을 위한 평가, 실행연구

1) 본 논문은 2025학년도 과학영재학교 경기과학고등학교 교원연구비지원에 의한 논문임.
* 제1저자 및 교신저자, jyr200104@gmail.com

# I. 서 론

우리나라의 학생 평가는 선발적 교육관에 기초하여 오랜 기간 학생을 선발하고 배치하는 기제로 활용되어 왔다(김성훈, 2008). 이러한 관행은 학생들로 하여금 교과 내용을 단순 암기하여 재생하도록 유도하고, 학습을 통해 무엇을 얼마나 배웠는지보다 성적을 얼마나 받았는지에 집중하게 함으로써, 학생의 성장과 발달이라는 교육의 본질을 점차 흐리게 하였다(김신영, 2015). 이에 교육부는 형성평가와 수행평가 개념을 통해 학생 평가의 질적 개선을 도모해 왔으며, 2015 개정 교육과정에서는 '과정 중심 평가'를 본격적으로 도입하여 평가 패러다임의 전환을 추구하였다(교육부, 2015; 박정, 2017).

과정 중심 평가는 학습 과정에서 학생에 관한 정보를 다각도로 수집하고 평가 결과를 환류함으로써 궁극적으로 학생의 성장과 발달을 지원하는 데 목적을 둔다(교육부, 2015). 이는 특정 평가 방법을 지칭하기보다, 평가의 결과를 점수화하는 데 머무르지 않고 학생이 자신의 학습 과정을 스스로 점검하며, 교사에게는 교수·학습의 질을 개선하는 자료로 활용하도록 하는 거시적 평가 정책이다(박정, 2019; 신혜진 외, 2017).

그러나 과정 중심 평가를 현장에서 실천하기에는 여러 어려움이 존재한다. 한 과목에서 수백 명의 학생 답안을 채점하는 데 많은 시간이 소요되고(Baker, 2021), 이로 인한 시간적, 물리적 부담은 교사들이 수업 준비와 개별 학생 지도에 투입할 자원을 감소시킨다(김유정 외, 2019; 반재천 외, 2018). 또한 주관적 채점에서 비롯되는 평가자 간 일관성 부족(Malouff & Thorsteinsson, 2016)과 평가 공정성에 대한 불신(Agarwal et al., 2023)도 현장 안착을 저해하는 요인으로 지적된다.

이러한 맥락에서 최근 교육 현장에서는 AI 기반 평가 시스템에 주목하고 있다. AI 기반 평가 시스템은 자동 채점과 데이터 분석을 통해 기존 평가의 구조적 문제를 개선할 수 있는 혁신적 대안으로 부상하고 있으며(Holmes et al., 2019), 특히 서술형 답안을 포함한 다양한 형식의 과제를 신속하게 처리하여 교사의 채점 부담을 실질적으로 경감하는 데 기여할 것으로 기대된다(Attali & Burstein, 2006; Baker, 2021). 또한 AI는 학생 개개인의 오류 패턴을 분석하여 즉각적이고 맞춤형인 피드백을 제공할 수 있어(Mehta & Sharma, 2023), 학습 격차 완화와 자기조절학습 역량 강화에도 기여할 수 있다(Luckin et al., 2017). 국내에서도 과정 중심 평가와의 연계 가능성을 탐색하는 연구가 이루어지고 있으며(이서교 외, 2024), AI 기반 평가가 교사의 업무 부담 경감과 학생 맞춤형 지도 측면에서 실질적인 도움이 될 것으로 전망되고 있다.

그러나 AI 기반 평가 시스템의 가능성에 대한 기대에 비해, 이를 실제 교육 현장에 구체적으로 설계하여 적용한 실증적 연구는 여전히 부족하다. 특히 교사가 의도한 평가 기

준을 AI가 충실히 반영하는지, 시스템의 기술적 한계는 무엇이며 이를 어떻게 개선할 수 있는지에 대한 체계적인 탐색이 요구된다.

이에 본 연구는 AI 기반 과정 중심 평가를 G영재학교의 교수·학습 및 평가 현장에 적용함으로써, AI 기반 과정 중심 평가의 교육적 의미를 탐색하고자 하였다. 특히 1차 실행과 2차 실행을 거치며 효과적인 현장 적용을 위한 조건은 무엇인지, AI가 과정 중심 평가를 어떻게 지원하는지, 그리고 이 과정에서 교사의 역할이 어떻게 재정립되는지를 분석함으로써, AI 기반 과정 중심 평가의 현장 안착을 위한 실천적 지식을 생성하고자 하였다. 구체적인 연구 문제는 다음과 같다.

(1) AI 기반 과정 중심 평가를 현장에 적용하기 위한 조건은 무엇인가?
(2) AI 기반 과정 중심 평가의 실행과 성찰의 과정에서 확인된 가능성과 한계는 무엇인가?
(3) AI 기반 과정 중심 평가 실행 과정에서 교사의 역할은 어떻게 재정립되는가?

## Ⅱ. 이론적 배경

### 1. AI 기반 평가 시스템의 기능과 한계

AI 기반 학습 평가 시스템은 기계학습, 자연어 처리(Natural Language Processing; NLP), 패턴 인식 등 다양한 AI 기술을 활용하여 학생의 학업 성취도를 평가하고, 이를 토대로 학습 피드백을 제공하는 체계를 의미한다(Luckin et al., 2017). 기존의 컴퓨터 기반 평가(Computer-Based Testing; CBT)가 미리 정해진 답안과 규칙에 따라 채점하는 방식이었다면, AI 기반 평가는 대량의 데이터를 학습한 알고리즘이 스스로 평가 기준을 적용하여 시험 답안 채점, 학습 분석, 성적 예측 등을 수행함으로써 보다 지능적인 평가를 가능하게 한다. 특히 자동화된 채점 기능을 통해 객관식이나 단답형뿐 아니라 서술형 답안까지 신속하게 처리할 수 있어, 평가의 효율성과 일관성을 높이고 교사의 업무 부담을 경감하는 데 기여할 것으로 기대된다(Attali & Burstein, 2006; Baker, 2021).

AI 기반 평가 시스템의 핵심 기능 중 하나는 개인 맞춤형 피드백 제공이다. AI는 학생 개개인의 답안을 분석하여 취약한 개념이나 오류 패턴을 식별하고, 즉각적인 피드백을 제공할 수 있다(Mehta & Sharma, 2023). 실시간 피드백을 통해 학생들은 기다림 없이 자신의 개선점을 파악하고 다음 학습 단계로 나아갈 수 있으며, 이는 학습 참여도

와 성취도 향상으로 이어진다. 또한 AI 기반 평가는 평가의 공정성 제고에도 기여할 것으로 기대된다. 알고리즘은 동일한 답안에 대해 언제나 동일한 채점 기준을 적용하므로 인간 평가자에게서 나타날 수 있는 피로도나 무의식적 편향 등의 요소를 구조적으로 배제한다(Malouff & Thorsteinsson, 2016). 국내에서도 한국교육과정평가원을 중심으로 서답형 문항 자동 채점 프로그램 개발과 적용 방안 연구가 이루어지고 있으며(송민호 외, 2024; 신병철 외 2024), AI 기반 평가가 교사의 업무 경감과 학생 맞춤형 지도 측면에서 도움이 될 것으로 전망되고 있다(이서교 외, 2024).

그러나 AI 기반 평가 시스템이 지닌 기술적·교육적 한계 역시 간과할 수 없다. 우선 기술적 측면에서, 현재의 AI 기반 평가 시스템은 명확한 구조와 어휘로 서술된 텍스트 기반 답안에서는 비교적 안정적인 성능을 보이지만, 수식·화학 구조식·도형 등 비정형적 표현 양식이나 손글씨 답안에 대해서는 인식 정확도가 불안정한 한계를 지닌다(Dikli, 2006). 또한 AI 알고리즘은 훈련 데이터에 내재된 편향 문제를 완전히 배제할 수 없어, 채점 결과의 일관성과 공정성에 대한 지속적인 점검이 요구된다(Agarwal et al., 2023). 교육적 측면에서는 더욱 본질적인 한계가 존재한다. AI는 여러 정보를 종합하여 의미를 도출해야 하는 복합적 추론 과제나 학습자의 독창적 접근의 가치를 판별하는 데 어려움을 보이며, 학습자의 태도·동기·정서적 반응 등 정의적 영역은 원천적으로 평가하기 어렵다(Luckin et al., 2017). 이는 AI가 교사의 평가 활동을 전면적으로 대체하기보다 특정 영역에서 교사를 보조하는 역할로 자리매김해야 함을 시사하며, 효과적인 AI 평가 시스템의 활용을 위해서는 무엇보다 교사의 전문적 판단과 세심한 루브릭 설계가 전제되어야 함을 보여준다(Xu et al., 2020).

## 2. 과정 중심 평가의 실천과 AI 기반 평가의 역할

2015 개정 교육과정을 기점으로 본격적으로 도입된 과정 중심 평가는 학생의 성장 과정에 주목하여 적절한 피드백을 제공하는 평가 방침이다. 이는 결과 중심적 관점을 지양하고 학생의 학습 과정을 면밀하게 관찰하고 평가하여 그 결과를 교수·학습 개선에 환류하는 것을 핵심으로 한다. 과정 중심 평가가 본래의 의도대로 실천된다면, 이는 Black & Wiliam(1998)이 제안한 학습을 위한 평가(Assessment for Learning; 이하 AfL)의 실현으로 이어질 수 있다. AfL은 평가를 학습의 종착점이 아니라 학습을 촉진하는 과정으로 바라보는 관점으로, 교사가 평가를 통해 수집한 학생 정보를 수업 개선과 학생 피드백에 적극 활용함으로써 궁극적으로 학생의 학습과 성장을 지원하는 데 목적을 둔다(Wiliam, 2011).

그러나 학교 현장에서는 과정 중심 평가 실천에 있어 여러 현실적인 어려움을 보고하

고 있다. 특히 중·고등학교에서는 교사의 평가가 대학 입시와 직결되는 구조적 특성으로 인해 객관성이 담보되는 선다형 평가 방식이 선호되는 경향이 있으며(박혜영 외, 2019; 조수영, 2017), 과정 중심 평가 실천에 수반되는 대량의 학생 산출물 분석, 개별 피드백 제공, 학교생활기록부 기술 등의 업무는 교사에게 상당한 시간적·물리적 부담을 지운다(이현주 외, 2020). 이러한 문제는 과정 중심 평가의 본질을 실현하기 위해 선행되어야 할 자원과 지원의 필요성을 잘 보여준다.

이러한 맥락에서 AI 기반 평가 시스템은 과정 중심 평가의 실천적 한계를 보완하는 유용한 도구로 기능할 수 있다. AI가 대량의 학생 산출물을 구조화하여 요약하고, 루브릭에 기반한 채점 결과와 피드백 초안을 제공함으로써 교사의 반복적·시간 소모적 업무를 경감하면, 교사는 확보된 시간을 학생과의 개별적 상호작용과 수업의 질적 개선에 투입할 수 있다. 이 과정에서 AI 기반 평가 시스템이 AfL의 핵심 요소인 즉각적 피드백과 학습 개선의 순환 구조를 실질적으로 지원할 수 있다면, 그동안 자원의 한계로 온전히 실현되지 못했던 과정 중심 평가의 이상에 더욱 근접할 수 있을 것이다(Hattie & Timperley, 2007).

다만 AI 기반 평가가 과정 중심 평가를 진정으로 지원하기 위해서는 교사의 역할 또한 재정립될 필요가 있다. 기존에 교사가 평가의 기술적 수행자로서 채점과 기록에 많은 시간을 할애했다면, AI와의 역할 분담이 이루어진 환경에서 교사는 평가 설계자이자 학습 촉진자로서의 역할에 보다 집중할 수 있다. 교사는 AI가 수행하기 어려운 복합적 사고 평가, 정의적 영역의 관찰, 학생과의 관계적 상호작용을 담당하고, AI 채점 결과를 비판적으로 검토하여 최종적인 교육적 판단을 내리는 전문가로 기능해야 한다(Luckin et al., 2017). 이처럼 AI 기반 평가 시스템의 도입은 단순한 기술적 혁신이 아니라, 교사의 전문성이 발휘되는 영역을 재구성하고 평가 문화 전반을 변화시키는 교육적 전환의 계기로 이해될 필요가 있다.

# Ⅲ. 연구 방법

## 1. 연구 설계

본 연구는 AI 기반 과정 중심 평가 시스템을 교육 현장에 적용하고, 그 과정에서 나타나는 교육적 의미를 탐색하기 위해 연구방법으로 실행연구(Action Research)를 선택하였다(Reason & Bradbury, 2001). 실행연구는 연구자인 교사가 자신의 교육 실천을 반

성적으로 검토하면서 계획-실행-성찰의 순환 과정을 통해 이론과 실천의 간극을 좁히고 실천적 지식을 생성하는 연구 방법으로(이형빈, 2023), 특정 맥락에서 AI 기반 평가 시스템의 가능성과 한계를 면밀히 탐색하는 데 적합한 접근이다. 특히 본 연구는 연구자 4인이 공동으로 참여하는 협력적 실행연구(Collaborative Action Research)의 성격을 지니며, 연구 설계의 전 과정에서 참여자들의 협의와 상호 피드백을 통해 연구의 신뢰성을 높이고자 하였다(Kemmis & McTaggart, 2000).

본 연구의 실행 절차는 이형빈(2023)이 제시한 실행연구의 5단계를 기반으로 구성하였다. 〈표 1〉과 같이 계획, 1차 실행, 실행 중 성찰, 2차 실행, 실행 후 성찰의 순으로 진행되며, 각 단계는 이전 단계의 성찰이 다음 단계의 실행으로 이어지는 순환적 구조를 갖는다.

**<표 1> 연구 절차**

| 단계 | 주요 내용 |
|---|---|
| 계획 | • 선행연구 조사 및 연구집단 형성<br>• 프로토타입 AI 기반 과정 중심 평가 시스템 학습 |
| 1차 실행 | • 2025학년도 1학기 교수·학습 실행<br>• 교과별 15주차 수업 진행, 과제물 평가 및 피드백 자료 수집 |
| 실행 중 성찰 | • 시스템 적용 과정에서의 기술적 문제점 및 교사·학생의 활용 양상 파악<br>• 시스템의 초기 효과와 한계를 분석하여 AI 기반 과정 중심 평가 시스템 개선 |
| 2차 실행 | • 2025학년도 2학기 교수·학습 실행<br>• 교과별 15주차 수업 진행, 과제물 평가 및 피드백 자료 추가 수집 |
| 실행 후 성찰 | • 1차 대비 2차 실행에서 교수·학습 및 평가 방식의 질적 변화 확인<br>• AI 평가 시스템의 교육적 함의 및 지속적인 발전 방향 모색 |

## 2. 연구 현장 및 연구 도구

본 연구는 연구자가 소속되어 있는 수도권 소재 G영재학교를 연구 현장으로 선정하였다. G영재학교는 학점제를 운영하며 모든 과목이 절대평가 방식을 채택하고 있어, AI 기반 과정 중심 평가 시스템을 통해 제공되는 평가 결과가 성적에 직접적으로 반영되지 않는다. 이러한 환경은 학생들이 점수 부담 없이 AI 피드백을 학습 도구로 활용할 수 있는 조건을 제공하며, 수식·화학 구조식이 포함된 과학 보고서, 프로그래밍 코드, 장기 학습 포트폴리오 등 다양한 산출물을 통해 AI 평가 시스템의 적용 가능성과 한계를 폭넓게 탐색하는 데 유리하다.

본 연구에서 활용한 AI 기반 과정 중심 평가 시스템은 대규모 언어 모델(Large Language Model; LLM)을 핵심 엔진으로 하는 서술형 자동채점 및 피드백 생성 플랫폼으로, G영재학교 자체 운영 체계를 통해 구현되었다. 시스템의 주요 작동 방식은 다

음과 같다. 교사는 교과별 과제와 함께 루브릭(채점 기준표)을 시스템에 입력하며, 학생이 제출한 서술형 답안, 보고서, 포트폴리오를 업로드하면 AI가 루브릭에 기반하여 채점 결과와 서술형 피드백을 자동 생성한다. 교사는 AI가 생성한 결과를 검토하고 수정한 후 학생에게 제공함으로써 최종 피드백의 교육적 타당성을 담보한다.

이 시스템은 나선형 프로세스 모델(Boehm, 1986)에 따라 점진적으로 개발하고 개선했다. 나선형 모델은 한 번에 완성된 시스템을 구현하는 것이 아니라 여러 번의 개발 주기를 반복하며 기능을 점진적으로 발전시키는 방식으로, 현장 교사들의 요구사항에 유연하게 대응할 수 있다는 장점이 있다. 연구자 A는 소프트웨어 개발자 1인과 협력하여 2024년 10월부터 2025년 2월까지 시스템의 목표와 제약조건을 설정하고 프로토타입을 제작하였으며, 1차 실행을 통해 파악된 기술적 한계를 바탕으로 산출물 인식 기능, 평가 루브릭 요소, 평가 결과 양식 등을 개선한 버전을 구축하여 2차 실행에 적용하였다. 시스템의 핵심 기능은 교사가 작성한 평가 루브릭을 토대로 AI가 학생 산출물을 채점하고 개별 피드백을 생성하고 교사가 이를 검토하여 제공하는 것이며, 이를 통해 과정 중심 평가의 실천을 지원하는 도구로 기능하도록 설계되었다.

## 3. 연구 참여자

본 연구의 주체인 연구자 집단 4인은 자신의 교육 실천을 탐색하고 개선해 나가는 연구자이자 실천가(researcher-practitioner)의 이중적 역할을 수행하였다. 〈표 2〉는 연구 참여자 정보를, 〈표 3〉은 AI 기반 과정 중심 평가 시스템 활용 현장을 정리한 것이다. 연구 참여자 4인은 연구 기간인 2025년 3월부터 12월까지 평균 월 1회 정기적인 협의를 통해 연구 설계의 전 과정, 시스템 적용 현황 공유, 자료 분석 및 해석, 다음 실행 주기 계획 수립에 이르기까지 모든 연구 활동을 협력적으로 수행하였다.

**<표 2> 연구 참여자 정보**

| 교사 | 성별 | 근무 경력(년) | 주요 역할 |
|---|---|---|---|
| A | 남 | 12 | • 본 연구의 총괄 책임자<br>• 연구의 전반적 설계 및 현장 적용 주도<br>• 2학년 교수·학습에 AI 평가 시스템 적용 |
| B | 여 | 20 | • 2, 3학년 교수·학습에 AI 평가 시스템 적용<br>• 사용자 경험 및 인터페이스 관련 피드백 제공<br>• 시스템 개선을 위한 심층적 피드백 제공 |
| C | 여 | 15 | • 3학년 교수·학습에 AI 평가 시스템 적용<br>• 사용자 경험 및 인터페이스 관련 피드백 제공<br>• 시스템 개선을 위한 심층적 피드백 제공 |

| 교사 | 성별 | 근무 경력(년) | 주요 역할 |
|---|---|---|---|
| D | 여 | 5 | • 1학년 교수·학습에 AI 평가 시스템 적용<br>• 사용자 경험 및 인터페이스 관련 피드백 제공<br>• 시스템 개선을 위한 심층적 피드백 제공 |

**<표 3> 연구 참여자의 AI 기반 과정 중심 평가 시스템 활용 현장**

| 구분 | 교사 | 과목 | 주요 활용 방식 |
|---|---|---|---|
| 1학기 | A | 기초화학및실험III | 자기노트, 포트폴리오 |
| | B | 객체지향프로그래밍 | 코드 구현, UML 기반 설계 |
| | C | 고급화학 I | 연습 문제 |
| | D | 기초화학및실험 I | 연습 문제 |
| 2학기 | A | 일반화학 I | 자기노트, 포트폴리오 |
| | B | 인공지능 | CNN, RNN/LSTM, Transformer |
| | C | 고급화학 II | 연습 문제 |
| | D | 기초화학및실험 II | 연습 문제 |

## 4. 자료 수집 및 분석

본 연구는 실행연구의 특성을 고려하여 질적 자료를 중심으로 다각도적인 자료를 수집하고 분석에 활용하였다. 2025년 3월부터 12월까지 수집된 자료의 종류와 수량은 〈표 4〉와 같다.

**<표 4> 자료 수집 개요**

| 자료 유형 | 수집 내용 | 수집 시기 | 수량 | 활용 목적 |
|---|---|---|---|---|
| 월 1회 협의록 | 시스템 적용 현황 공유, 개선 논의 | 2025.3~12 | 약 10건 | 실행 과정 추적 및 성찰 근거 |
| 자기보고서 | A~D 교사의 수업 적용 경험, 인식 변화, 성찰 | 실행 종료 후 | 4편 | AI 시스템 교육적 효과·개선점 분석 |
| 반구조화 면담 (연구 참여자) | B·C·D 교사 심층 면담<br>※ A교사는 총괄연구자 | 1차 실행 종료 후 | 3인, 각 약 60분 | 시스템 적용 경험, 한계, 개선 방향 |
| 학생 설문 (20문항) | AI 피드백 활용 역량·인식·정서 조절·지속 의향 (김욱태, 김영식(2024)) | 각 학기 말 (6월, 12월) | 1차 103명<br>2차 71명 | 삼각검증 보조자료 (학생 인식 확인) |
| 교사 설문 (44문항) | AI 기반 서술형 자동채점 수용 의도 및 인식 (신세인 외(2022)) | 각 실행 종료 후 (1차, 2차) | 4인 (A~D교사) | 삼각검증 보조자료 (교사 인식 확인) |
| 과제 및 산출물 | 학생 과제, AI 피드백 결과, 루브릭, 세부능력 및 특기사항 초안 | 실행 기간 중 (1·2차 전체) | 교과별 산출물 | AI 평가 적용 양상 및 학생 학습 변화 |

수집된 자료는 Braun & Clarke(2006)의 주제 분석(thematic analysis) 절차에 따라 전사 및 반복 정독, 주제 탐색, 하위범주 및 상위주제 도출, 주제 검토 및 정교화, 최종 주제 명명 및 보고의 순으로 분석하였다. 한편, 연구자 A는 수업 실천가이자 면담 진행 및 자료 분석을 겸하는 연구 총괄자로 이중적 위치를 지닌다. 이에 따른 편향 가능성을 최소화하기 위해, 면담 전에 연구 목적이 시스템 성능 평가가 아니라 실천 경험의 공유임을 참여자들에게 명확히 안내하였고, 분석 단계에서 연구자 A의 해석을 B, C, D가 검토하는 교차검토(cross-checking) 절차를 운영하였으며, 참여 교사들이 자기보고서를 통해 면담 맥락과 독립적으로 자신의 경험을 직접 서술하도록 하였다.

# Ⅳ. 연구 결과

본 연구의 결과는 세 가지 연구문제에 다음과 같이 응답한다. 1, 2절은 AI 기반 과정 중심 평가 시스템이 보여준 교육적 가능성과 기술적 한계를 분석함으로써 연구문제 (2)에 응답한다. 3절은 1·2차 실행의 성찰로부터 귀납적으로 도출된 현장 적용 조건을 제시함으로써 연구문제 (1)에 직접 응답하며, 4절은 이러한 경험을 바탕으로 교사의 역할이 어떻게 재정립되었는지를 기술함으로써 연구문제 (3)에 응답한다.

## 1. AI 기반 과정 중심 평가 시스템의 교육적 가능성

### 가. 학생을 위한 맞춤형 피드백 제공

AI 기반 과정 중심 평가 시스템은 학생의 학습 결과뿐만 아니라 문제 해결 과정 전반을 분석하여 피드백을 제공함으로써 과정 중심 평가를 실질적으로 지원하였다. 기존의 결과 중심 평가에서는 정답 여부가 핵심 기준이었으나, AI 평가 시스템을 활용한 수업에서는 수식 전개 방식, 변수 정의의 적절성, 개념 설명의 논리성 등이 세분화된 피드백 항목으로 제시되었다.

연구 참여자 B는 인공지능 수업에서 학생들이 자신의 풀이 과정과 설명 방식이 어떻게 평가되는지를 AI 피드백을 통해 구체적으로 인식하게 되었으며, 이후 과제에서 수식 전개와 개념 설명을 보다 명확히 서술하려는 학습 태도의 변화로 이어졌음을 보고하였다. 연구 참여자 D는 형성평가 후 즉각적 개별 피드백의 현실적 어려움을 다음과 같이 설명하였다.

> 형성 평가를 많이 진행하잖아요. 근데 그 형성 평가를 그 많은 애들한테 매번 즉각적으로 피드백 주기가 되게 어렵고 보통은 그래서 그냥 걷어서 제출을 했냐 안 했냐로 점수를 부여하는 경우가 대부분인데 AI를 활용하게 되면 어쨌든 개별적으로 피드백을 해 줄 수 있어요.

학생 설문 결과에서도 AI 피드백에 대한 긍정적 인식이 확인되었다([부록 1] 참조). AI 피드백 활용 역량 영역의 긍정 응답 비율은 1차(n=103) 평균 43.4%에서 2차(n=71) 50.7%로 상승하였으며, 향후 AI 피드백을 지속적으로 활용하고 싶다는 문항에서는 1차 M=3.33에서 2차 M=3.41로 소폭 증가하였다.

한편, AI 피드백의 교육적 효과는 피드백 자체의 질뿐만 아니라 학생의 피드백 수용성(feedback receptivity)에 따라서도 달라질 수 있다. Winstone et al.(2017)은 피드백의 효과가 실현되기 위해서는 학생이 피드백을 수동적으로 받는 데 그치지 않고 이를 능동적으로 해석하고 학습에 적용하는 '수용적 참여(proactive recipience)'의 태도가 필요함을 강조하였다. Carless & Boud(2018) 역시 학생이 피드백 정보를 처리하고 활용하는 능력인 피드백 리터러시(feedback literacy)를 갖추지 못할 경우, 아무리 세밀하고 즉각적인 피드백이라도 학습 개선으로 이어지기 어렵다고 지적하였다. 이러한 관점에서 학생 설문 결과를 살펴보면, 'AI 피드백 활용 역량' 영역의 긍정 응답 비율이 1차(43.4%)에서 2차(50.7%)로 상승하였으나 절반을 소폭 상회하는 수준에 머물렀고, '향후 AI 피드백을 지속적으로 활용하고 싶다'는 문항(M=3.41) 역시 중간값 이상이지만 높지 않은 수준을 보였다. 반면 '부정적 피드백을 수용하고 회피하지 않는다'는 정서 조절 관련 문항들은 두 차례 모두 80% 이상의 긍정 응답을 유지하였다. 이는 학생들이 피드백에 대한 정서적 개방성은 비교적 갖추고 있으나, AI 피드백을 실질적인 학습 전략으로 내면화하는 능동적 수용 역량은 아직 발전 중임을 시사한다. 따라서 AI 기반 피드백이 과정 중심 평가의 취지대로 학생 성장에 기여하기 위해서는, 피드백 제공 방식의 개선과 함께 학생의 피드백 리터러시를 함양하는 교육적 지원이 병행될 필요가 있다.

### 나. 교사의 평가 업무 부담 경감

AI 평가 시스템은 대량의 학생 산출물을 구조화하여 요약하고, 루브릭에 기반한 채점 결과와 피드백 초안을 제공함으로써 교사의 기록 업무를 효율화하는 데 기여하였다. 연구 참여자 B는 다음과 같이 표현하였다.

> 과정 중심 평가를 하게 되면 내야 되는 과제들이 엄청 양이 많아지는데, 그걸 요약하는 게 사실 AI가 제일 잘하는 일이에요.

특히 과목별 세부 능력 및 특기사항(세특)의 초안 마련 과정에서 AI의 효율성이 높게 평가되었다. 연구 참여자 C는 AI 피드백이 단순한 채점 부담 경감을 넘어, 학생이 틀린 부분을 정확하고 명확하게 제시하는 교정적 피드백의 역할에서 수용도가 더 높을 것으로 예상하였다. 교사 설문에서도 업무 부담 감소 문항은 1, 2차 모두 M=4.75(긍정 100.0%)로 높은 수준을 유지하였으며([부록 2] 참조), 교사 전문성 신장 관련 문항은 1차 M=3.33에서 2차 M=4.25로 상승하여 실제 적용 경험을 통해 AI 시스템이 전문성 신장에 도움이 된다는 인식이 강화된 것으로 나타났다.

## 2. AI 기반 과정 중심 평가 시스템의 한계

AI 기반 과정 중심 평가 시스템은 앞서 논의한 바와 같이 다양한 교육적 가능성을 보여주었으나, 실제 교수·학습 현장에 적용하는 과정에서 여러 기술적·개념적 한계가 드러났다. 특히 이 과정에서 한계의 상당 부분이 AI 시스템 자체의 결함이 아니라 교사의 루브릭 설계 방식과 긴밀하게 연결되어 있음이 확인되었다.

가장 빈번하게 지적된 문제는 채점의 일관성 부족이었다. 유사한 수준의 답안에 상이한 점수가 부여되거나, 동일한 답안을 재채점하였을 때 평가 결과가 달라지는 사례가 관찰되었다. 연구 참여자 D는 AI가 일관된 잣대로 평가하지 않는 경우가 많았으며 학생별 피드백 분량에도 편차가 있었음을 보고하면서, 이 문제의 근본 원인이 루브릭 작성의 모호성에 있었다고 성찰하였다.

> 유사한 답안임에도 불구하고 학생별로 점수 부여의 편차가 크게 발생하는 채점 일관성 문제도 두드러졌는데, 이 문제의 근본적 원인은 평가 기준인 루브릭 작성의 모호성에 있었습니다. 어떤 기준을 어떻게 설정해야 할지 명확하지 않아 루브릭을 매우 포괄적으로 설정하였고 이것이 채점 기준의 일관성을 저해하는 요인으로 작용하였다고 판단됩니다.

이에 연구 참여자 D는 개선책으로 모범답안을 업로드하는 기능 도입과 루브릭에 측정 가능하고 명확한 진술을 포함시킬 것을 제언하였다. 실제로 2차 실행에서 문제은행 방식으로 모범답안 기반 채점을 적용한 결과, 자기보고서에 따르면 채점 일관성이 이전보다 크게 향상되었음이 확인되었다. 이는 명확하고 구체적인 채점 기준이 AI 평가의 정확도를 결정하는 핵심 변수임을 보여주는 사례이다.

1차 실행 경험을 바탕으로 진행한 면담에서 연구 참여자 B 역시 루브릭 작성에 대한 인식 변화를 보고하였다.

> 처음에 쓰라고 했을 때는 루브릭 만들기도 생소하고 했는데 좀 이렇게 해보고 채점 받아보니까 요령이 생기더라고요. 루브릭을 좀 디테일하게 하고 해야 AI가 채점도 디테일하고 하겠구나가 이제 좀 생각이 들었고, 루브릭을 잘 해서 쿼리를 잘 던질 수 있게끔 하는 것까지도 우리의 전문성으로 같이 향상이 돼야 되지 않을까 생각을 했어요.

이와 같이 AI 평가 시스템의 성능은 루브릭의 구체성과 정밀성에 크게 의존한다. 교사가 평가 의도를 명확하게 진술하고, 측정 가능한 항목으로 루브릭을 구성할수록 AI는 교사의 의도에 부합하는 채점 결과를 산출할 수 있다. 이는 AI 평가 시스템 도입이 단순히 기술을 수용하는 것이 아니라, 교사의 평가 설계 역량 향상을 동시에 요구하는 실천임을 시사한다.

산출물 인식 오류 역시 중요한 기술적 한계로 확인되었다. 화학 교과에서는 수식, 화학 구조식, 오비탈 도형, 그래프 등 도식화된 표현이 핵심적인 역할을 하는데, AI는 이러한 양식을 정확하게 인식하는 데 어려움을 보였다. 연구 참여자 C는 손으로 작성한 수식이나 구조식, 궤도 모형에서 이미지 판독 오류가 발생하였으며, 확률밀도함수, 루이스 구조, 혼성화 판단, 전자 배치, 공명 구조 등 도식화가 핵심인 과제에서 기호, 화살표, 결합선의 인식 정확도가 불안정하여 잘못된 평가가 나타났음을 지적하였다. 연구 참여자 D 역시 필기체 인식 오류가 잘못된 채점 결과가 나타나는 가장 큰 이유였다고 보고하였다.

고차원적 사고 평가의 한계도 명확하게 드러났다. 연구 참여자 C는 기본적인 개념 확인 수준의 과제에서는 AI 채점이 효과적일 수 있으나, 여러 정보를 종합하여 추론하고 그 과정에서 의미를 도출해야 하는 과제에서는 교사가 직접 채점하는 것이 미묘한 차이를 읽어낼 수 있다고 언급하였다. 이는 AI가 명시적으로 구조화된 지식의 확인에는 강점을 보이지만, 학습자의 사고 과정에 내재된 논리적 비약이나 창의적 접근의 질을 판별하는 데에는 여전히 인간 교사의 전문적 판단이 필요함을 시사한다.

정의적 영역의 평가에서도 본질적인 한계가 확인되었다. 연구 참여자 D는 정의적 영역은 공교육에서 중요하게 평가하고 있는 부분이지만 그러한 영역까지 AI에게 평가권을 부여하는 것은 적절하지 않다는 입장을 밝혔으며, 연구 참여자 B는 학생이 과제를 수행하는 과정에서 보이는 비언어적 반응이나 태도, 정서적 변화는 AI가 판단하기 어려운 영역임을 지적하였다.

## 3. AI 기반 과정 중심 평가의 현장 적용 조건

앞서 확인된 가능성과 한계를 바탕으로, 연구 참여자들은 1, 2차 실행 과정의 성찰을

통해 AI 기반 과정 중심 평가를 현장에 효과적으로 적용하기 위한 실천적 조건을 구체화하였다. 이러한 조건들은 단일 실행의 산물이 아니라, 1차 실행의 시행착오가 2차 실행의 설계 변화로 이어지는 순환적 성찰 과정에서 귀납적으로 도출된 것이라는 점에서 실천적 지식으로서의 의미를 갖는다.

### 가. 교과 특성과 과제 유형에 따른 적용 범위의 선별적 설정

연구 참여자들의 실행 경험은 AI 기반 평가가 모든 과제 유형에 균등하게 적합하지 않으며, 적용 가능한 영역과 그렇지 않은 영역이 교과 특성과 과제의 성격에 따라 분명히 구분됨을 보여주었다. 정답이 명확하게 정해진 과제, 개념 이해 수준의 확인이 목적인 과제, 구조화된 서술형 답안 등에서는 AI 채점이 효과적으로 기능하였다. 반면, 복합적 추론을 요구하는 개방형 서술, 화학 구조식·수식·도형 등 비정형 표현이 핵심인 과제, 실험 수행 및 발표와 같이 과정 관찰이 필요한 활동, 그리고 정의적 영역의 평가는 여전히 교사의 직접적 판단이 요구되는 영역으로 확인되었다.

연구 참여자 A는 물질 명명처럼 정답이 명확한 화학 과제에서는 AI 채점이 매우 효과적이지만, 에세이나 개방형 서술 과제에서는 교사의 판단이 필수적임을 언급하였으며, 연구 참여자 D는 실험 수행·발표와 같은 활동은 AI가 대체하기 어려운 교사 고유의 평가 영역임을 강조하였다. 이러한 경험은 교사가 AI를 도입하기 전에 해당 과제가 AI 채점에 적합한 특성을 갖추고 있는지 사전에 검토하고, AI 평가와 교사 평가 간의 역할 분담을 명확히 설정하는 것이 현장 적용의 첫 번째 조건임을 시사한다.

### 나. AI 채점 정확도를 결정하는 정밀한 루브릭 설계

1, 2차 실행 전반에 걸쳐 반복적으로 확인된 것은 AI 기반 평가 시스템의 성능이 루브릭의 구체성과 정밀성에 결정적으로 의존한다는 사실이다. 연구 참여자 D는 1차 실행에서 루브릭을 포괄적으로 설정하였을 때 채점 일관성이 심각하게 저하되었음을 보고하였으며, 2차 실행에서 모범답안 기반 채점으로 전환한 이후 일관성이 크게 향상되었음을 확인하였다. 연구 참여자 B 역시 루브릭을 세밀하게 구성할수록 AI 채점 결과의 정밀도가 높아진다는 인식을 형성하였다.

### 다. 교사의 검증 과정을 전제한 운영 체계의 구축

현장 적용의 세 번째 조건은 AI 채점 결과를 최종 평가 결과가 아니라 교사의 검토를 위한 1차 분석 자료로 위치 설정하는 운영 체계의 구축이다. 연구 참여자들은 AI 피드

백을 학생에게 그대로 전달하기보다 교사의 확인 과정을 거쳐 제공하는 것이 바람직하다는 인식을 공유하였다. 연구 참여자 B는 AI 채점 결과를 학생에게 공개하는 순간 교사의 교육관이 투영되어야 함을 강조하였다.

이는 AI 평가 결과에 대한 교사의 최종 검토와 승인 과정이 운영 체계 내에 명시적으로 내재화되어야 함을 의미한다. 특히 AI 채점의 오류 가능성이 높은 비정형 표현 과제나 개방형 서술 과제에서는 모든 학생의 결과를 교사가 검토하는 단계를 필수적으로 설정해야 한다. 이러한 교사 검증 체계의 구축은 AI 기반 평가 시스템이 교육적 신뢰성을 확보하면서 과정 중심 평가의 취지를 실현하기 위한 핵심 운영 조건이다.

## 4. 교사의 역할 재정립과 전문성 재구성

앞서 2절과 3절에서 확인된 가능성, 한계, 현장 적용 조건에 대한 경험은 연구 참여자들로 하여금 AI 기반 과정 중심 평가 환경에서 교사가 어떠한 역할을 수행해야 하는지를 재정립하도록 이끌었다. 연구 참여자들은 1·2차 실행과 면담을 통해 교사의 역할 경계와 전문성의 방향에 대해 비교적 일관된 견해를 형성해 나갔다.

### 가. AI를 보조 도구로 위치 설정: 평가 주체로서의 교사

연구 참여자들에게 공통적으로 드러난 핵심 인식은 AI를 평가의 '주체'가 아닌 '보조 도구'로 자리매김해야 한다는 것이었다. 연구 참여자 D는 교사 자신이 평가의 주체가 되어 AI를 보조 도구로 활용하는 것이 바람직하다고 언급하였으며, 연구 참여자 A는 평가에 있어서 교사의 의도성과 목적성이 AI로 대체될 수 없음을 다음과 같이 분명히 하였다.

> AI가 주도권을 갖는다는 것은 어떤 일의 의도와 목적을 정의하는 것이라고 할 때, AI에게는 그러한 의미의 주도권이 없습니다. AI는 철저하게 도구로 규정해야 합니다.

연구 참여자 B는 AI를 보조 교사에 비유하면서, 채점 결과를 그대로 사용하지 않고 교사가 다시 한번 확인하는 피드백 과정이 전제된다면 AI 평가도 공정하다고 볼 수 있다고 언급하였다. 연구 참여자 B는 또한 교사가 AI의 채점 결과를 학생에게 공개하는 과정에서 자신의 교육관이 중심을 잡고 있어야 함을 강조하였다.

> AI가 하는 말이긴 하지만 내가 클릭을 했기 때문에 이거는 나의 모습도 같이 투영돼서 애들한테 나가고 있다라고 생각이 되더라고요. 교사가 정말 교육관을 가지고 내가 이 결과를 애들에게 오픈할

수 있을 만한지 아닌지를 중심을 잘 갖고 판단을 해야 한다고 그걸 클릭하는 순간 생각을 했어요.

AI 평가가 효과적으로 적용될 수 있는 영역과 그렇지 않은 영역에 대해서도 연구 참여자들은 비교적 명확한 구분을 하고 있었다. 연구 참여자 C는 화학에서 물질 명명과 같이 정답이 명확하게 정해져 있는 과제에서는 AI 채점이 매우 효과적이지만, 에세이나 개방형 서술 과제의 평가에서는 여전히 교사의 판단이 필요하다고 언급하였다. 연구 참여자 D 역시 실험 수행이나 보고서 작성, 발표와 같은 활동에 대한 평가는 AI가 수행하기 어려운 고유한 영역이라고 보았다.

'AI에 의존하는 교사, AI를 활용하는 교사, AI와 협력하는 교사' 중 바람직한 교사상을 묻는 질문에서 연구 참여자 대부분은 현재 시점에서는 'AI를 활용하는 교사'를 선택하였다. 연구 참여자 D는 사용하다 보면 활용 단계를 넘어 의존 단계로 이행될 위험이 있다며 적절한 거리두기의 필요성을 강조하였다. 다만 연구 참여자 B는 현재 시점의 '활용'에서 장기적으로는 'AI와 협력하는 교사'로 발전해 나갈 것이라는 전망을 제시하기도 하였으며, 이는 AI와 교사의 관계가 기술의 발전과 교사의 경험 축적에 따라 점진적으로 변화할 수 있음을 시사한다.

### 나. 교사의 역할 확장과 전문성 재구성

연구 참여자들은 AI 평가 시스템을 활용하는 과정에서 교사에게 요구되는 역할이 단순 채점자가 아니라 루브릭 설계자이자 AI 결과의 비판적 검토자로 확장되어야 함을 인식하고 있었다. 특히, 연구 참여자 B는 루브릭을 정밀하게 설계하는 역량과 AI 결과를 교육적으로 재해석하는 역량이 교사에게 필요한 전문성임을 직접적으로 언급하고 있었다.

루브릭을 좀 디테일하게 하고 해야 AI가 채점도 디테일하고 하겠구나가 이제 좀 생각이 들었고, 루브릭을 잘 해서 쿼리를 잘 던질 수 있게끔 하는 것까지도 우리의 전문성으로 같이 향상이 돼야 되지 않을까 생각을 했어요.

이는 AI가 교사의 역할을 축소하기보다는 오히려 더 높은 수준의 전문성을 요구하는 방향으로 변화시키고 있음을 보여준다. 특히 루브릭 설계 역량은 2절에서 확인된 바와 같이 AI 기반 평가의 질을 결정하는 핵심 변수로, 교사가 평가 의도를 명확하게 진술하고 측정 가능한 항목으로 루브릭을 구성할수록 AI는 교사의 의도에 부합하는 결과를 산출할 수 있었다. 이 점에서 AI 기반 평가 시스템의 도입은 기술 수용 이상의 의미를 지닌다. 교사는 AI가 평가하기 어려운 고차원적 사고와 정의적 영역을 자신의 고유한 관찰과 판단으로 담당하고, AI 결과를 비판적으로 검토하여 교육적으로 재해석하며, 나아

가 AI와의 협력적 평가 설계를 통해 시스템의 정확도를 높여가는 방향으로 자신의 전문성을 재구성하게 된다.

### 다. 학생-교사 관계의 가치 재확인

마지막으로 연구 참여자들은 AI 기반 평가 시스템이 교육에 깊숙이 자리잡게 될 경우 상실될 수 있는 가치에 대해서도 우려를 표명하였다. 연구 참여자 D는 학교가 지식만 습득하는 곳이 아니라 교사와 학생 간의 상호작용 속에서 삶의 태도나 눈에 보이지 않는 가치를 배우는 곳임을 강조하였다. 연구 참여자 B는 학생들이 AI 평가를 지나치게 신뢰하게 되면 교사의 평가를 수용하지 않게 될 수 있으며, 이는 교사와 학생 간 신뢰 관계의 약화로 이어질 수 있다고 지적하였다.

이러한 우려는 AI 기반 평가 시스템의 활용이 효율성 추구에만 치우쳐서는 안 되며, 교사와 학생 간의 교육적 관계를 유지하고 강화하는 방향에서 이루어져야 함을 보여준다. 궁극적으로 교사는 AI 채점 결과를 비판적으로 검토하고, AI가 평가하기 어려운 영역에 대해서는 자신의 전문적 판단을 적극적으로 발휘하며, AI와의 적절한 역할 분담을 통해 평가의 질적 수준과 교육적 관계의 가치를 함께 높여나가는 방향으로 나아가야 할 것이다.

## V. 결론 및 제언

본 연구는 G영재학교에서 AI 기반 과정 중심 평가 시스템을 개발하고 1, 2차 실행연구를 통해 그 교육적 의미를 탐색하였다. 연구 결과, 이 시스템은 가능성과 한계를 동시에 드러내면서 교육 현장에서의 활용 방향과 교사 역할의 재정립에 대한 실천적 지식을 생성하였다.

첫째, 본 연구에서 채택한 실행연구 방법론은 계획-실행-성찰의 순환 구조를 통해 단순한 기술 도입의 효과 검증을 넘어 AI 기반 평가 시스템이 실제 교수·학습 현장에서 어떤 의미를 갖는지를 심층적으로 탐색하였다. 1차 실행의 문제 경험이 2차 실행의 설계 변화로 이어지는 순환적 과정은 학교 현장에 AI 기반 평가 시스템이 적용될 때, 교사의 지속적인 성찰을 전제로 해야 함을 보여준다.

둘째, AI 기반 과정 중심 평가 시스템은 “학습을 위한 평가(AfL)” 실현에 기여할 수 있는 가능성을 확인해 주었다. AI 시스템은 즉각적이고 세분화된 피드백을 제공함으로써 학생이 자신의 사고 과정을 점검하고 다음 학습 단계로 나아갈 수 있도록 지원하였

다. 그러나 AfL의 실현은 피드백의 양적 확대만으로 이루어지지 않으며, AI가 제공하는 구조화된 피드백과 교사가 제공하는 관계적 피드백이 상호 보완적으로 기능할 때 가장 가깝게 실현될 수 있음도 확인되었다.

셋째, AI 기반 평가 시스템의 도입은 교사의 전문성을 약화시키기보다 오히려 더 정교하고 고차원적인 전문성을 요구했다. 루브릭을 정밀하게 설계하는 역량, AI 채점 결과를 비판적으로 검토하고 교육적으로 재해석하는 역량, AI가 평가하기 어려운 고차원적 사고와 정의적 영역을 교사 고유의 판단으로 담당하는 역량이 새로운 전문성 영역으로 요청된다. 이는 단순히 개별 교사의 적응 문제가 아니라, AI 기반 평가 시스템을 도입하는 모든 교육 현장에서 교사 전문성의 방향 자체가 재정의되어야 함을 시사한다.

본 연구의 결과는 AI 기반 평가 시스템의 도입과 관련한 교육 정책 수립과 후속 연구에 다음과 같은 시사점을 제공한다.

정책적 측면에서 AI 기반 평가 시스템의 성공적 현장 안착을 위해서는 기술 보급에 앞서 루브릭 설계 역량, AI 결과의 비판적 검토 역량, AI와의 협력적 평가 설계 역량을 포함하는 교원 연수 프로그램 개발이 선행되어야 한다. 또한 AI 기반 평가 시스템이 공교육 평가 체계와 연계될 경우, AI 채점 결과에 대한 근거 제시, 채점 결과에 대한 이의 제기 절차, AI 평가의 적용 범위에 대한 명확한 지침 수립 등의 제도적 장치가 마련되어야 한다.

후속 연구 측면에서, 본 연구는 단일 학교의 특수한 맥락에서 수행된 실행연구로 결과의 일반화에 한계가 있다. G영재학교는 학점제와 절대평가를 운영하는 특수한 학교로, 일반 학교를 포함한 다양한 학교 맥락에서의 적용 사례 탐색이 요청된다. 또한 루브릭의 구체성 수준에 따른 AI 채점의 일관성을 측정하는 연구나, AI 기반 평가 의존도 증가가 교사–학생 간 교육적 관계에 미치는 영향에 대한 심층적 탐색도 필요하다.

본 연구는 AI 기반 과정 중심 평가 시스템을 교육 현장에 도입하고 그 교육적 의미를 실행연구의 순환적 과정을 통해 탐색한 연구이다. AI와 인간 교사가 공존하는 평가 생태계를 구성하는 방향에 관한 실천적 지식을 생성하였다는 점에서 의의를 갖는다. 궁극적으로 AI 기반 평가 시스템의 도입은 평가의 효율성을 높이는 기술적 전환이 아니라, 교사의 전문성이 재구성되고 학생의 성장을 위한 평가 문화가 심화되는 교육적 전환의 계기로 이해되어야 할 것이다.

## 참고문헌


강지영, 소경희(2011). 국내 교육관련 실행연구(action research) 동향 분석. **아시아교육연구, 12**(3), 197-224.

교육부(2015). **2015 개정 교육과정 초·중등학교 교육과정 총론**. 세종: 교육부.

김성훈(2008). 교육평가는 교육을 교육답게 하는가?. **교육원리연구, 12,** 73-91.

김신영(2015). 교실 내 학생평가의 운영실태와 개선 방향. **교육과정평가연구, 18**(3), 257-282.

김욱태, 김영식(2024). 수학과 '확률과 통계' 영역에서 ChatGPT를 활용한 서답형 평가 피드백이 학생들의 피드백 리터러시에 미치는 영향. **컴퓨터교육학회 논문지, 27**(3), 19-30.

김유정, 이경건, 홍훈기(2019). 교사의 과정 중심 평가 역량에 관한 사례연구: 중고등학교 과학교사 사례를 중심으로. **한국과학교육학회지, 39**(6), 695-706.

박정(2017). 수업에서 학생평가 의미 탐색. **교육평가연구, 30**(3), 397-413.

박정(2019). 과정 중심 평가를 위한 수업과 학생 자기평가 의미 탐색. **교육평가연구, 32**(3), 421-440.

박혜영, 이명애, 이명진(2019). 우리나라 미래 초·중등학교 교육평가 방향 탐색. **교육과정평가연구, 22**(3), 147-171.

반재천, 김선, 박정, 김희경(2018). 교사별 과정 중심 평가에 대한 교사의 인식. **교육과정평가연구, 21**(3), 105-130.

성열관(2006). 교육과정 실행연구의 성장과 주요 특징에 대한 이론적 고찰. **교육과정연구, 24**(2), 87-109.

송민호, 김동영, 김진구, 박상욱, 박종임, 정수진(2024). **교과 서논술형 평가 자동채점을 위한 인공지능 모델 적용 방안 연구(Ⅰ)**. 한국교육과정평가원, 연구보고 RRE 2024-9.

신병철, 이준수, 유연주(2024). 프롬프트 엔지니어링을 통한 GPT-4 모델의 수학 서술형 평가 자동 채점 탐색: 순열과 조합을 중심으로. **수학교육, 63**(2), 187-207.

신세인, 이준기, 하민수, 박지선(2022). 초·중등 과학교사들의 인공지능 기반 서술형 자동채점 시스템 수용의도에 대한 연구. **현장과학교육, 16**(2), 195-205.

신혜진, 안소연, 김유원(2017). 과정 중심 평가 활용의 정책적 분석. **교육과정평가연구, 20**(2), 135-162.

이서교, 신민철, 황경빈, 박인우(2024). 인공지능 융합교육 필요성과 활성화 방안에 관한 질적 연구: 인공지능 교육 업무 담당 교육 전문직의 인식을 중심으로. **교육공학연구, 40**(2), 571-611.

이현주, 백종민, 곽영순(2020). 2015 개정 교육과정에 따른 중학교 과학교사들의 과정중심평가 실행 및 어려움에 대한 인식 탐색. **과학교육연구지, 44**(2), 133-144.

이형빈(2023). 실행연구. 한국교육과정학회 (편저), **교육과정 연구방법론** (pp. 385-418). 경기도: 교육과학사.

조수영(2017). 고등학교 현장에서의 역량기반 교육과정 수업 평가 연계의 현실화 방안. **중등교육연구, 65**(1), 255-281.

최훈원, 최윤정(2024). 키워드 네트워크 분석을 활용한 과정 중심 평가의 연구 동향 분석. **교육과정평가연구, 27**(2), 251-277.

Agarwal, A., Agarwal, H., & Agarwal, N. (2023). Fairness score and process standardization: Framework for fairness certification in artificial intelligence systems. *AI and Ethics, 3*(2), 267-279.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V.2. *Journal of Technology, Learning, and Assessment, 4*(3), 1-30.

Baker, R. S. (2021). *Getting past the current trade-off between privacy and equity in educational technology.* In The economics of equity in K-12 education: Connecting financial investments with effective programming (pp. 97-107). New York, NY: Routledge.

Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*(2), 139-144.

Boehm, B. W. (1986). A spiral model of software development and enhancement. *ACM SIGSOFT Software Engineering Notes, 11*(4), 14-24.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77-101.

Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. *Assessment & Evaluation in Higher Education, 43*(8), 1315-1325.

Dikli, S. (2006). An overview of automated scoring of essays. *Journal of Technology, Learning, and Assessment, 5*(1), 1-35.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81-112.

Holmes, W., Bialik, M., & Fadel, C. (2019). *Artificial intelligence in education: Promises and implications for teaching and learning*. Boston, MA: Center for Curriculum Redesign.

Kemmis, S., & McTaggart, R. (2000). *Participatory action research*. In N. K. Denzin & Y. S. Lincoln (Eds.), Handbook of qualitative research (2nd ed., pp. 567-605). Thousand Oaks, CA: Sage Publications.

Luckin, R., Holmes, W., Griffiths, M., & Forcier, L. B. (2017). Towards artificial intelligence-based assessment systems. *Nature Human Behaviour, 1*(3), 0028.

Malouff, J. M., & Thorsteinsson, E. B. (2016). Bias in grading: A meta-analysis of experimental research findings. *Australian Journal of Education, 60*(3), 245-256.

Mehta, K., & Sharma, R. (2023). *Prioritizing the critical success factors of e-learning systems by using DEMATEL*. In S. Daniel (Ed.), Redefining virtual teaching learning pedagogy (pp. 401-420). Singapore: Springer.

Reason, P., & Bradbury, H. (2001). *Handbook of action research: Participative inquiry and practice*. Thousand Oaks, CA: Sage.

Winstone, N. E., Nash, R. A., Parker, M., & Rowntree, J. (2017). Supporting learners' agentic engagement with feedback: A systematic review and a taxonomy of recipience processes. *Educational Psychologist, 52*(1), 17-37.

Wiliam, D. (2011). *Embedded formative assessment*. Bloomington, IN: Solution Tree Press.

논문접수 : 2026.4.2. / 수정본접수 : 2026.4.28. / 게재승인 : 2026.5.11.

ABSTRACT

# Exploring the Educational Implications and Improvement Processes of an AI-Based Process-Centered Assessment System: An Action Research Approach

**Yoonri Jeong, Yujung Kim, Jeongin Kwon, Sangmi Chae, Byoungchul Shin**
Teacher, Gyeonggi Science High School for the Gifted

This study explored the educational implications and improvement processes involved in applying an AI-based process-centered assessment system at G Science High School for the Gifted. Drawing on action research conducted by four teacher-researchers, the study collected and analyzed self-report journals, semi-structured interviews, and student and teacher surveys. Four key findings emerged. First, the AI-based assessment system demonstrated meaningful educational potential by providing personalized feedback and reducing repetitive scoring workload, thereby enabling teachers to devote greater attention to individual student interaction. Second, notable limitations were identified, including inconsistent AI scoring, difficulties in recognizing handwritten responses, and inadequate evaluation of higher-order thinking and affective domains. Third, through reflective cycles across two implementation phases, participants inductively identified three conditions for effective field application: selectively determining the scope of AI use based on subject and task characteristics, designing precise rubrics, and establishing teacher-led verification systems for AI-generated results. Fourth, teachers reconstructed their professional identities by positioning AI as an auxiliary assessment tool, expanding their roles as rubric designers and critical reviewers of AI outputs, and reaffirming the value of the teacher-student educational relationship. These findings suggest that AI-based assessment systems can meaningfully contribute to realizing the aims of process-centered assessment, provided that teachers' active engagement and professional judgment remain central to the evaluation process.

*Key Words: process fortified assessment, AI-based assessment system, assessment for learning, action research*

## [부록 1] 학생 설문 응답 결과 요약 (1, 2차 비교)

| 하위영역 | 문항 | 1차 M | 1차 긍정(%) | 2차 M | 2차 긍정(%) | 영역 평균 |
|---|---|---|---|---|---|---|
| | (n=) | n=103 | n=103 | n=71 | n=71 | |
| AI 피드백 활용 역량 | 잘 몰랐던 부분·틀린 부분 파악 | 3.29 | 42.7 | 3.55 | 60.6 | 1차 3.23<br>2차 3.41 |
| | 잘하고 있는 것 파악 | 3.30 | 44.7 | 3.44 | 53.5 | |
| | 더 노력할 부분 파악 | 3.25 | 45.6 | 3.52 | 57.7 | |
| | 학습 전략 수립 | 3.09 | 37.9 | 3.20 | 38.0 | |
| | 학습 방법 수정 | 3.15 | 41.7 | 3.25 | 39.4 | |
| | 부족한 학습 내용 수정 | 3.27 | 47.6 | 3.49 | 54.9 | |
| AI 피드백 인식 | 피드백 이해·학습 활용 기대 | 3.25 | 46.6 | 3.31 | 45.1 | 1차 3.33<br>2차 3.46 |
| | 학습 장단점 파악에 중요 | 3.16 | 39.8 | 3.37 | 43.7 | |
| | 다양한 방법으로 제시 | 3.76 | 67.0 | 3.93 | 74.6 | |
| | 학습 방법·전략 등 다양한 정보 제공 | 3.15 | 39.8 | 3.25 | 40.8 | |
| 학습자 주도성 | 피드백 과정 핵심 역할은 학생 자신 | 4.10 | 80.6 | 3.96 | 77.5 | 1차 4.10<br>2차 3.96 |
| 학습 효과 기대 | 피드백 후 학습 향상 기대 | 3.01 | 31.1 | 3.24 | 40.8 | 1차 3.14<br>2차 3.27 |
| | 새 내용 학습의 즐거움 | 3.27 | 39.8 | 3.30 | 43.7 | |
| 정서 조절 | 부정 피드백 수용·회피 않음 | 4.07 | 77.7 | 3.94 | 77.5 | 1차 4.10<br>2차 4.00 |
| | 피드백 후 감정 진정(들뜸) | 4.15 | 80.6 | 4.01 | 77.5 | |
| | 피드백 수용적 태도(분노) | 4.09 | 81.6 | 4.06 | 82.8 | |
| | 피드백 후 감정 진정(창피) | 4.10 | 80.6 | 3.99 | 80.3 | |
| 도움 추구·지속 의향 | 이해 안 될 때 도움 추구 | 3.74 | 64.1 | 3.45 | 56.3 | 1차 3.56<br>2차 3.51 |
| | 부정 피드백 후에도 도움 추구 | 3.60 | 59.2 | 3.68 | 62.0 | |
| | 향후 AI 피드백 활용 의향 | 3.33 | 45.6 | 3.41 | 49.3 | |

주: 1) 5점 리커트 척도(1=매우 그렇지 않다 ~ 5=매우 그렇다). 긍정 응답(%)은 4·5점 응답 비율.

2) 1차: 2025년 1학기 말(n=103), 2차: 2025년 2학기 말(n=71). 설문 도구는 김욱태·김영식(2024)의 피드백 리터러시 척도를 수정·적용함.

3) 본 설문은 삼각검증 보조자료로 활용되었으며, 질적 분석 결과를 보완하는 참고 자료임.

## [부록 2] 교사 설문 응답 결과 요약 (1, 2차 비교)

| 하위영역 | 문항 | 1차 M | 1차 긍정(%) | 2차 M | 2차 긍정(%) | 영역 평균 (1차→2차) |
|---|---|---|---|---|---|---|
| | (n=) | n=4 | n=4 | n=4 | n=4 | |
| AI 친숙도 | AI 관련 기술·제품 사용 경험 | 4.40 | 80.0 | 5.00 | 100.0 | 4.30 ↓ 4.75 |
| | AI 정보 미디어 접촉 | 4.60 | 100.0 | 5.00 | 100.0 | |
| | AI 기반 제품 사용 익숙함 | 4.20 | 80.0 | 4.60 | 100.0 | |
| | 일상에서 AI 관련 기술 접촉 빈도 | 4.00 | 80.0 | 4.80 | 100.0 | |
| | AI 기반 인터넷 기능 활용 익숙함 | 4.60 | 100.0 | 4.60 | 100.0 | |
| 감정 반응 (긍정) | 흥미 유발 | 4.25 | 100.0 | 4.75 | 100.0 | 4.19 ↓ 4.53 |
| | 기분 좋은 느낌 유발 | 4.00 | 75.0 | 4.50 | 100.0 | |
| | 호기심 유발 | 4.25 | 100.0 | 4.75 | 100.0 | |
| | 기대감 유발 | 4.25 | 100.0 | 4.00 | 75.0 | |
| 감정 반응 (부정) | 불편한 감정 유발 | 2.75 | 25.0 | 2.50 | 25.0 | 2.60 ↓ 2.55 |
| | 긴장감 유발 | 2.75 | 25.0 | 2.50 | 25.0 | |
| | 불안감 유발 | 2.75 | 25.0 | 2.50 | 25.0 | |
| | 두려움 유발 | 2.25 | 25.0 | 2.25 | 25.0 | |
| | 걱정 유발 | 2.50 | 25.0 | 2.50 | 25.0 | |
| 교육적 유용성 | 평가 소요 시간 절약 | 4.50 | 100.0 | 4.50 | 75.0 | 4.22 ↓ 4.08 |
| | 평가 결과 신뢰도 증가 | 4.25 | 100.0 | 3.25 | 50.0 | |
| | 정보기술 역량 증가에 도움 | 4.25 | 75.0 | 3.75 | 75.0 | |
| | 업무 부담 감소에 효과적 | 4.75 | 100.0 | 4.75 | 100.0 | |
| | 교사 전문성 신장에 도움* | 3.33 | 66.7 | 4.25 | 75.0 | |
| | 교사의 평가 불안 감소* | 4.00 | 75.0 | 4.00 | 75.0 | |
| 부정 결과 인식 | 교사 권위 상실 | 2.25 | 25.0 | 1.75 | 0.0 | 2.42 ↓ 2.50 |
| | 정보기술 비익숙 교사 소외 | 3.00 | 50.0 | 2.25 | 0.0 | |
| | 부가 연수·업무 요구 증가 | 4.50 | 100.0 | 4.25 | 75.0 | |
| | 교사 전문성 퇴보 | 1.50 | 0.0 | 2.00 | 0.0 | |
| | 교사 무기력화 | 2.00 | 0.0 | 2.25 | 25.0 | |
| | 비인간적 문화 조성 | 1.25 | 0.0 | 2.50 | 0.0 | |
| 긍정 효과 인식 | 개인별 맞춤형 교육 도움* | 4.50 | 100.0 | 4.25 | 100.0 | 4.25 ↓ 3.83 |
| | 일관적 평가 가능 | 4.00 | 50.0 | 3.00 | 25.0 | |
| | 채점 객관성 확보 도움 | 4.25 | 75.0 | 3.75 | 75.0 | |
| | 공정한 평가 문화 정착 | 3.50 | 25.0 | 3.50 | 75.0 | |
| | 즉각적 피드백 제공 도움* | 4.75 | 100.0 | 4.00 | 75.0 | |
| | 서술형 평가 확산에 기여 | 4.50 | 100.0 | 4.50 | 100.0 | |
| 기술 한계 인식 | 학생 서술 의도 평가 미흡 | 3.00 | 50.0 | 2.25 | 0.0 | 2.87 ↓ |
| | 일관적 평가 불가 | 3.50 | 25.0 | 3.50 | 50.0 | |

| 하위영역 | 문항 | 1차 M | 1차 긍정(%) | 2차 M | 2차 긍정(%) | 영역 평균 (1차→2차) |
|---|---|---|---|---|---|---|
| | 학습자 수준 측정 부정확 | 3.50 | 50.0 | 3.25 | 25.0 | 2.79 |
| | 평가 공정성 훼손 | 2.33 | 0.0 | 2.25 | 0.0 | |
| | 적절한 피드백 제공 불가 | 2.50 | 25.0 | 2.50 | 0.0 | |
| | 적절한 학습행동 유도 불가 | 2.25 | 25.0 | 3.00 | 25.0 | |
| 수용 의도 | 기회 시 활용 의향 | 4.50 | 100.0 | 4.25 | 75.0 | 4.45 ↓ 4.17 |
| | 교육현장 수용 필요 | 4.50 | 100.0 | 4.25 | 75.0 | |
| | 활용 의사 | 4.67 | 100.0 | 4.50 | 100.0 | |
| | 타 교사 추천 의사 | 4.25 | 100.0 | 4.00 | 75.0 | |
| | 모든 서술형 응답 AI 채점 의사 | 4.33 | 100.0 | 3.75 | 75.0 | |
| | 일부 서술형 AI 채점 의사 | 4.50 | 100.0 | 4.25 | 100.0 | |

주: 1) 5점 리커트 척도(1=매우 그렇지 않다 ~ 5=매우 그렇다). 긍정 응답(%)은 4·5점 응답 비율.

2) 설문 도구는 신세인 외(2022)의 AI 기반 서술형 자동채점 수용의도 측정 도구를 바탕으로 구성함.

3) * 표시 문항은 논문 본문에서 직접 인용된 핵심 문항.

4) n=4 소규모 응답으로 통계적 일반화에는 한계가 있으며, 삼각검증 보조자료로서의 의미를 지님.

5) 일부 문항에 무응답이 있어 유효 응답 수가 n=3인 경우가 있음(1차: 전문성 신장, 평가 공정성 훼손, 활용 의사 문항 각 1건; 2차: 기대감 유발 문항 1건). 무응답은 분석에서 제외하였음.