

교육과정평가연구  
The Journal of Curriculum and Evaluation  
2025, Vol. 28, No. 3, pp.319~359  
DOI: <https://doi.org/10.29221/jce.2025.28.3.319>

## GAI-HITL 기반 독서 문항 자동 생성(AIG)의 심리측정학적 타당성 분석 연구

박고운 (한국교원대학교 국어교육전공 박사과정)\*  
최숙기 (한국교원대학교 국어교육과 교수)\*\*

### 요약

본 연구는 생성형 인공지능과 교사 전문성의 구조적 협업을 통한 GAI-HITL(Generative AI-Human-in-the-Loop) 기반 자동 문항 생성 방식이 고등학교 국어과 독서 평가에서 심리측정학적 타당성을 확보할 수 있는지를 실증적으로 검증하였다. ChatGPT-4o와 현장 교사의 협업을 통해 8단계 GAI-HITL 프로토콜을 설계하고, 2015 개정 교육과정 독서 성취기준에 기반하여 20개의 선다형 문항을 개발하였다. 고등학교 3학년 53명을 대상으로 실시한 평가 결과를 고전검사이론(CTT)과 문항반응이론(IRT)을 통해 분석하였다. 연구 결과, GAI-HITL 기반 문항 세트는 내적 일관성 신뢰도(Cronbach's  $\alpha = .7897$ )가 교육평가 기준을 상회하였으며, 평균 정답률 66.8%로 적절한 난이도를 보였다. CTT를 주 분석틀로 하고 IRT를 보조적으로 활용한 결과, 교사 예측 난이도와의 정합성은 55%, 변별도는 점이연상관 기준 55%의 문항이 양호 이상 수준을 나타냈다. 특히 문항별 품질 차이가 개발 과정에서의 교사-AI 협업 양상과 밀접한 관련이 있음을 확인하였다. 높은 변별도를 보인 문항은 교사의 구체적인 오답 구성 피드백을 통해 완성된 반면, 낮은 변별도 문항은 AI의 자기 점검과 교사 개입이 미흡한 사례였다. 본 연구는 GAI-HITL 방식이 현재 기술 수준에서도 교육 현장에 적용 가능한 신뢰도 높은 평가 도구를 생산할 수 있음을 실증하였으며, AI 시대 교육평가에서 인간 전문가와 인공지능 협업의 시너지 효과를 보여준다는 점에서 의의가 있다.

주제어: 자동 문항 생성(AIG), 생성형 인공지능(GAI), Human-in-the-Loop(HITL), 고전검사이론(CTT), 문항반응이론(IRT), 독서 평가

\* 제1저자, [koniholic@gmail.com](mailto:koniholic@gmail.com)

\*\* 교신저자, [agreement@knue.ac.kr](mailto:agreement@knue.ac.kr)

## I. 서론

교육평가는 학습자의 학업 성취를 진단하고 교육목표 달성 여부를 규명하는 필수적 학술 활동으로, 교수·학습 과정의 개선뿐 아니라 교육 정책 수립의 근거를 제공한다(박도순, 2009; 성태제, 2014). 특히 고등학교 국어과 독서 영역 평가는 단순한 정보 이해를 넘어 비판적·창의적 사고, 정보 통합 등 고차원적 사고 역량을 심층적으로 측정해야 한다(권태현, 이정찬, 김승현, 2017). 이러한 복합적 평가 요구는 사실적 이해, 추론, 비판, 창의 사고를 위계적으로 진단할 수 있는 정교한 문항 설계를 전제하며, 출제자의 고도화된 전문성을 필수 조건으로 설정한다(김혜정, 2008; 정혜승, 2008).

그러나 현재 단위학교 문항 개발은 교사의 직관과 경험에 크게 의존하고 있으며, 수작업 중심의 개발 체계는 평가 도구의 일관성과 타당성 확보에 구조적 한계를 노출한다(Gierl & Haladyna, 2013). 실제로 남민우 외(2022)는 고등학교 국어과 지필 평가 문항을 분석한 결과, 성취기준과 문항 간의 불일치, 선택지 논리 오류, 사고 수준 불일치 등의 문제가 광범위하게 확인되었으며, 이는 문항 개발 역량의 차이가 문항의 질적 불균형으로 이어질 수 있음을 시사한다. 정혜승(2008) 역시 이러한 문제를 지적하며 교사의 평가 역량의 격차가 문항의 타당성과 신뢰도에 직접적인 영향을 미친다고 보고하였다. 이러한 문제는 평가 주체의 전문성 차이를 반영하며, 문항의 질적 불균형을 초래할 수 있는 실질적 요인으로 작용한다.

이러한 문제의식 속에서 자동 문항 생성(Automatic Item Generation, AIG)에 대한 관심이 지속적으로 확대되고 있다. AIG는 초기에 템플릿 기반의 정형적 접근에서 출발하여, 이후 대규모 언어 모델(Large Language Models, LLMs)과 자연어처리(Natural Language Processing, NLP)의 발전에 힘입어 생성형 인공지능(Generative AI, GAI)을 활용한 방식으로 진화하고 있다(Gierl, Hollis & Tanygin, 2021; Kasneci et al., 2023). GAI 기반 AIG는 생산성과 다양성을 극대화할 수 있다는 점에서 주목받고 있으나, 교육적 타당성, 논리적 정합성, 교육과정과의 연계성 확보에 한계가 있다는 비판이 지속적으로 제기되어 왔다. Cong-Lem, Soyoof & Tsering(2025)은 ChatGPT와 같은 대규모 언어 모델이 정확성과 신뢰성, 비판적 사고와 문제 해결 능력, 기술적 제약 등에서 다면적인 한계를 지니고 있어 교육평가 영역에서는 신중한 접근이 필요하다고 지적하였다.

이러한 한계를 보완하기 위한 대안으로 Human-in-the-Loop(HITL) 접근이 대두되고 있다. HITL은 AI 시스템의 의사결정 과정에 인간 전문가의 판단과 개입을 통합하여 결과물의 품질을 향상시키는 방법론으로, 교육 분야에서는 AI가 생성한 문항을 교사가 검토하고 개선하는 협업 구조를 의미한다(Memarian & Doleck, 2024; U.S. Department

of Education, 2023). 실제 최근 생성형 AI를 활용한 자동 문항 생성의 품질 검증을 위해 심리측정학적으로 분석하고 타당화를 실시하는 연구가 확산하는 추세이다. 이러한 국제적 연구 동향과 맥을 같이하여, 국내에서도 AIG 과정에서 교사와 AI의 협업 필요성을 강조하는 연구들이 나타나고 있다. 이하늘, 이용상(2024)은 GPT-4.0을 활용한 한국어 읽기 문항 자동 생성에서 교사 피드백과 조정이 포함된 협업 구조의 필요성을 제안하였으며, 박고운, 최숙기(2025)는 사고 연쇄(Chain-of-Thought, CoT) 프롬프팅 기반 국어과 AIG 모델 연구에서 교사의 피드백이 반복적으로 개입하는 HITL 구조가 필요하다고 제안하였다.

그러나 현존하는 HITL 기반 AIG 연구는 대부분 개념적 모델 제시나 전문가의 직관적 평가에 국한되어 있으며, 생성 문항을 실제 교육 현장에 투입하고 학습자 응답 데이터를 바탕으로 심리측정학적 특성을 정량적으로 검증한 실증 연구는 여전히 제한적이다 (Bejar et al., 2003; Memarian & Doleck, 2024). Song, Du & Zheng(2025)의 체계적 문헌 고찰에 따르면, 최근 15년간 수행된 AIG 관련 연구 중 문항 난이도 및 변별도를 실측 데이터를 통해 평가한 사례는 매우 드문 것으로 나타났다. Young et al.(2025)은 ChatGPT-4가 생성한 다지선다형 문항을 문항반응이론(Item Response Theory, IRT) 과 전문가 평가를 통해 분석하여, 생성된 문항들이 논리적으로 타당하고 학습 목표에 부합하나 전반적으로 난이도가 낮고 변별력이 높아 총괄 평가보다는 형성 평가에 적합하다고 보고하였다. 이러한 선행연구의 흐름을 고려할 때, 국내에서도 AIG 결과에 대한 객관적이고 정량적인 심리측정 분석의 필요성이 강하게 제기된다.

본 연구에서는 이러한 선행연구의 흐름을 고려하여 생성형 AI와 교사 전문가의 협업을 통해 문항을 점진적으로 고도화하는 반복적이고 상호보완적인 협업 과정을 Generative AI-Human-in-the-Loop(이하 GAI-HITL)로 정의하였다. 이를 기반으로 한 AIG 프로토콜에 의해 생성된 문항 결과를 학습자 응답 데이터를 기반으로 문항의 타당성을 실증적으로 분석하여 관련 이론을 검증하고자 하는 것을 목적으로 한다.

이에 본 연구는 GAI-HITL 방식을 기반으로 한 고등학교 국어과 독서 AIG 결과의 심리측정학적 특성을 난이도 및 변별도 측면에서 체계적으로 분석함으로써, 해당 방식의 타당성과 실용 가능성을 실증적으로 평가하고자 한다. 이를 위해 본 연구는 다음과 같은 세 가지 연구 문제를 설정하였다.

첫째, GAI-HITL 방식으로 개발된 독서 문항의 난이도는 교사 예측, CTT, IRT 분석에서 어떤 특성과 상관성을 보이는가?

둘째, GAI-HITL 방식으로 개발된 독서 문항의 변별도는 CTT 및 IRT 기준에서 어떤 특성과 관계를 나타내는가?

셋째, 난이도와 변별도에 대한 분석 결과를 종합할 때, GAI-HITL 방식의 문항 개발이 평가 도구로서 갖는 심리측정학적 품질은 어떠한가?

본 연구는 다음과 같은 학술적 및 실천적 기여를 지닌다. 첫째, 이론적 기여 측면에서 본 연구는 GAI-HITL을 통해 생성된 문항의 난이도와 변별도를 정량적으로 분석하고, 교사 예측과 실제 반응 간의 상관성을 검증함으로써 실증적 근거를 제공한다(Leslie & Gierl, 2023; Song, Du & Zheng, 2025). 둘째, 실천적 기여 측면에서 본 연구는 GAI-HITL 기반 문항의 심리측정학적 품질을 검증함으로써, 교사들이 AI 생성 문항을 신뢰하고 교육 현장에 적용할 수 있는 실용적 근거를 제공한다(U.S. Department of Education, 2023).

## II. 이론적 배경

### 1. 2015 개정 교육과정 독서 역량과 평가 문항 개발의 과제

2015 개정 국어과 교육과정은 독서 역량을 학습자의 인지적 성장과 고차 사고 능력을 심화하는 핵심 축으로 설정하고 있으며, 독서 교육의 목표를 사실적 이해에서 시작하여 추론, 비판, 창의적 사고로 이어지는 위계적 구조 속에 체계화하였다(교육부, 2015). 이러한 위계는 단순한 정보 해독을 넘어, 텍스트의 구조와 논리를 분석하고, 사회문화적 맥락 속에서 의미를 재구성하며, 타당성과 관점을 비판적으로 평가하는 고차 사고 능력으로 연결된다(김혜정, 2011; 정혜승, 2008). 독서 교육은 따라서 학습자의 능동적 의미 구성과 인지적 전이를 이끄는 복합적 사고 활동으로 자리매김하며, 이는 평가 장면에서도 구조적 정합성과 타당성을 갖춘 문항 설계를 통해 구현되어야 한다(정민주 외, 2022). 특히 교육과정은 독서 능력을 단순한 기술적 독해가 아니라 지식의 구성과 문제 해결, 창의적 발상까지 아우르는 복합적 사고 능력으로 간주하며, 이에 따라 독서 평가는 단일 정답을 도출하기 어려운 복합 사고 과정의 구조적 반영을 요한다(남민우 외, 2022).

이러한 교육과정의 목표는 대학수학능력시험 국어 영역의 독서 문항 구성과 밀접하게 연결되어야 한다. 정혜승(2008)은 독서 평가 문항이 텍스트의 구조적 특성과 독자의 인지적 처리 과정을 종합적으로 고려해야 하며, 특히 정보 간의 관계 분석과 추론적 사고를 측정할 수 있는 설계가 필요하다고 강조하였다. 김혜정(2011)은 독서 능력 평가에서 사실적 이해, 추론적 이해, 비판적 이해의 위계적 구조가 문항 설계에 체계적으로 반영

되어야 하며, 각 수준별 사고 과정의 특성을 고려한 평가 도구 개발이 중요하다고 지적하였다. 정민주 외(2022)는 독서 평가가 단순한 정보 확인을 넘어서 텍스트 내 정보의 통합과 재구성, 비판적 판단을 요구하는 복합적 사고 과정을 측정해야 한다고 보았다. 이러한 관점에서 독서 문항은 학습자의 인지적 처리 과정과 사고의 깊이를 구조적으로 평가할 수 있는 설계 원리를 필요로 한다.

그러나 이러한 문항 설계는 고도의 평가 설계 역량을 요구하며, 실제 문항 개발 현장에서는 다양한 난제가 제기된다. 특히 독서 능력의 위계적 구조를 평가 문항에 충실히 반영하려 할 때, 각 사고 수준을 명확히 구분하면서도 하나의 문항 내에서 사고의 흐름을 논리적으로 구성하는 데 현실적 제약이 발생한다(남민우 외, 2022; 정민주 외, 2022). 문항이 평가하려는 사고 수준이 불명확하거나 선택지가 오답 유도 요인으로 작용하지 못할 경우, 문항의 변별력이 약화되고 평가의 타당성도 저해될 수 있다(김혜정, 2008). 이러한 상황은 특히 선다형 문항 형식의 구조적 한계와도 연결되며, 복합적 사고를 요구하는 평가 내용과 단일 선택지 형식 간의 불균형이 근본적인 긴장을 형성한다.

이러한 문제는 수능과 같은 국가 수준 평가뿐 아니라 학교 현장의 지필 평가에서도 유사하게 나타난다. 장성민(2024)은 고등학교 국어과 교사들이 독서 영역의 평가 문항을 출제하는 과정에서 위계적 사고에 기반한 문항 설계에 어려움을 겪고 있으며, 이는 문항 개발자의 평가 문식성과 독서 이론에 대한 이해 부족, 검토 체제의 부재, 그리고 실무적 제약 등과 복합적으로 연결된다고 지적하였다. 특히 교사들이 경험적으로 문항을 구성할 경우, 고차 사고를 측정하기 위한 평가 목적이 구체화 되지 않거나, 평가 도구로서의 논리적 완결성이 부족해질 가능성이 크다. 정혜승(2008) 또한 교사의 독해력, 글 구조 분석 능력, 질문 설계 능력의 격차가 문항의 타당성과 신뢰도에 직접적인 영향을 미친다고 보고하였다. 이는 평가 주체인 교사의 전문성 신장을 요구하는 것이며, 문항의 질적 불균형을 초래할 수 있다.

결국 2015 개정 교육과정이 지향하는 독서 능력의 위계적 발달과 고차 사고 중심의 평가 설계는 선다형 문항 체제에서도 충분히 반영되어야 하며, 이를 위해서는 문항의 사고 수준 설계, 지문과 문항 간 정합성, 선택지 구성의 정교함 등 다양한 측면에서 질적 기준이 요구된다. 이러한 기준은 단지 평가도구로서의 기능을 넘어서, 학습자의 사고 경로를 추적 가능하게 하고, 교수·학습 과정에서의 환류 자료로 기능하기 위해 필수적이다. 본 연구는 이러한 관점에서 수능에서 요구하는 수준의 독서 능력을 평가할 수 있는 선다형 문항을 생성형 인공지능 기반(AIG)으로 개발하고, 그 타당성을 실증적으로 분석하고자 한다. 특히 생성형 AI의 도입은 기존 문항 개발의 전문성·효율성 문제를 일정 부분 보완할 수 있는 기술적 가능성을 제시하며, 이는 평가의 질적 향상을 위한 새로운 전환점으로 기능할 수 있다.

## 2. 자동 문항 생성(AIG) 개념 및 연구 동향

자동 문항 생성(AIG)은 평가 설계 및 개발 과정을 디지털화하여 효율성과 일관성을 높이는 평가 공학(Assessment Engineering) 기반의 접근법이다(오규설, 2022). 초기 AIG는 미리 정의된 문항의 구조적 틀에 특정 변수 값을 대입하는 ‘템플릿 기반’ 방식으로 이루어졌다. 이 방식은 문항의 형식이 고정된 수학, 과학 등에서는 유용했으나, 다양한 맥락과 의미를 다루어야 하는 언어 평가 문항을 생성하는 데에는 명백한 한계를 가졌다.

이러한 한계는 GPT와 같은 대규모 언어 모델(LLM)을 활용한 AIG 연구를 통해 새로운 국면을 맞이하고 있다. 이 접근 방식은 주어진 텍스트를 바탕으로 지문은 물론이고 발문, 정답 및 선택지를 통합적으로 생성할 수 있어 문항 구성의 전 과정 및 핵심 요소를 도출하는데 AIG 기법을 적극적으로 활용하고 있다. von Davier(2018)는 신경망 기반의 AIG 과정을 ‘사전 준비-문항 생성-검토 및 수정-문항 최적화’의 4단계로 체계화하여, 이후 LLM 기반 AIG 연구의 이론적 토대를 마련하였다.

LLM을 활용한 문항 생성 연구는 모델이 발전함에 따라 그 품질 또한 향상되는 추세를 보인다. AI가 생성한 문항은 전문가 검토를 거칠 경우 교사의 업무 부담을 획기적으로 줄일 수 있는 초별 안으로서 잠재력을 보여준다(Biancini, Ferrato & Limongelli, 2024; 임상목 외, 2024). 그러나 AI가 생성한 문항에는 내용적 오류나 논리적 비약이 포함될 수 있으며, 특히 정답이 모호하거나 오답 선택지가 비현실적인 경우가 있어 최종 사용을 위해서는 전문가의 검토가 필수적이라는 점이 국내외 연구에서 공통적으로 지적된다(Grévisse, 2023; 박고운, 최숙기, 2025; 임상목 외, 2024; 이하늘, 이용상, 2024).

해외에서는 생성된 문항의 교육적 품질을 체계적으로 분석하려는 연구가 활발히 진행되고 있다. 특히, 생성 문항을 블룸의 분류법(Bloom’s Taxonomy)과 같은 인지적 위계에 맞춰 정렬하고 그 품질을 평가하는 연구가 주목받고 있다. 한 연구에서는 GPT-3.5가 블룸의 분류법 수준에 따라 문항을 생성하도록 지시했으나, 인간 전문가가 ‘고품질’이라고 판단한 문항은 57개 중 12개에 불과했다. 특히, 블룸의 분류법에서 ‘평가’나 ‘종합’과 같이 상위 수준으로 갈수록 GPT-3.5가 고품질 문항을 생성하는 데 더 큰 어려움을 겪는 경향이 관찰되었다. 이러한 연구는 AIG 기술이 기초적인 지식 확인에는 유용하지만, 고차원적인 사고를 측정하는 문항을 안정적으로 생성하기 위해서는 개선이 필요함을 시사한다(Hwang et al., 2024).

이러한 한계를 극복하기 위해 최근에는 생성 문항의 논리적 완결성을 높이는 사고 연쇄(CoT) 프롬프팅과 인간-AI 협력 모델이 주목받고 있다. CoT 프롬프트를 적용하면 단순히 예시를 제공하는 방식보다 문항의 완성도가 높아지지만, 여전히 고차원적인 비

판적 사고를 요구하는 문항 생성에는 한계를 보인다(박고운, 최숙기, 2025). 이와 함께 인간-인공지능 협력(HITL)은 생성-검증-재구성의 순환을 안정화하는 운영 원리로 자리매김하였다. 교육 분야 AI에 대한 체계적 문헌 고찰 연구는 인간 전문가의 개입이 목표 정렬, 책임성 확보, 그리고 맥락 적합성을 높이는 데 결정적인 역할을 함을 보여준다(Memarian & Doleck, 2024). 나아가 비판적 사고 평가 문항 생성을 목표로 한 STAIR-AIG 연구는, AI가 1차로 생성한 문항을 다단계 전문가 검토와 재생성 과정을 반복하는 모듈화된 루프를 통해 문항의 품질을 체계적으로 향상시킬 수 있음을 실증하였다(Kim et al., 2025).

요약하면, 최신 AIG 기술은 LLM의 발전을 통해 문항 생성의 자동화 수준을 빠르게 높여가고 있으나, 생성된 문항의 사고 수준을 정교화하고 교육과정과의 정합성을 확보하며 선택지의 타당성을 검증하는 등의 고차원적 판단은 여전히 인간 전문가의 개입이 필수적이다. 따라서 CoT 프롬프팅과 같은 고도화된 기술과 HITL 기반의 인간-AI 협업 모델을 결합하는 방식이 현실적인 대안으로 주목받고 있다(신동광, 2024). 이러한 연구들은 자동화 기술과 교사의 전문성을 융합하는 것이 미래 교육 평가의 질을 높이는 핵심 전략임을 보여준다.

### 3. 프롬프트 엔지니어링 전략과 연구 동향

대규모 언어 모델(LLM)의 성능이 발전함에 따라, 모델의 잠재력을 최대한 이끌어내기 위한 프롬프트 엔지니어링(Prompt Engineering)의 중요성이 부각되고 있다. 단순한 질의응답을 넘어 복잡한 과업을 수행하기 위해서는 모델의 추론 과정을 체계적으로 유도하고 제어하는 정교한 프롬프트 설계가 필수적이다. 특히 교육 평가 문항 생성과 같이 고도의 논리적 정확성과 교육적 타당성을 요구하는 분야에서는 최신 프롬프트 전략에 대한 이해가 중요하다.

최신 프롬프트 전략의 근간에는 인-컨텍스트 학습(In-Context Learning, ICL)이라는 LLM의 독특한 학습 방식이 자리 잡고 있다(Brown et al., 2020). ICL은 모델의 가중치를 직접 수정하는 파인튜닝(fine-tuning)과 달리, 프롬프트 내에 몇 가지 과업 예시(demonstrations)를 함께 제공하여 모델이 주어진 맥락(context) 안에서 과업의 패턴을 학습하고 해결하도록 유도하는 기법이다. 예를 들어, 프랑스 → 파리, 일본 → 도쿄, 한국 → ? 와 같이 예시를 제시하면, 모델은 ‘국가-수도’라는 패턴을 맥락 안에서 학습하여 ‘서울’이라는 답을 도출한다. 이처럼 예시를 제공하는 방식을 ‘퓨샷(Few-Shot) 프롬프팅’이라 하며, 본 연구에서 AI에게 기출 문항 예시를 제공한 것 역시 이 원리를 활용한 것이다.

ICL을 기반으로 한 획기적 전환은 사고 연쇄(CoT) 전략의 등장으로 이루어졌다. Wei et al.(2022)은 복잡한 추론 문제가 주어졌을 때, 최종 답변만 요구하는 대신 중간 추론 단계를 명시적으로 포함하도록 유도하면 LLM의 성능이 비약적으로 향상됨을 실증했다. 이 방식은 모델이 단계별로 논리적 사고를 수행하게 함으로써 문제 해결의 정확성과 일관성을 높였으며, 본 연구의 8단계 프로토콜 설계에 핵심적인 이론적 기반을 제공했다. CoT의 등장은 모델의 ‘생각’ 과정을 외부로 드러내어 통제하려는 후속 연구들로 이어졌다. Huang, Zhang & Liu(2022)는 여기서 더 나아가 LLM이 스스로의 추론 과정을 서술하는 ‘내적 독백(Inner Monologue)’ 기법을 제안했다. 이는 AI가 현재 사고하는 지점과 문제점을 스스로 밝히게 함으로써, 인간 전문가가 어느 지점에 개입하여 맥락을 재구성해야 할지를 명확히 알려주는 역할을 한다.

최근에는 한 단계 더 발전하여, AI가 자신의 초기 답변을 스스로 비판하고 수정하는 ‘자기 수정(Self-Correction)’ 또는 ‘자기 개선(Self-Refinement)’ 전략이 활발히 연구되고 있다(Madaan et al., 2023; Shinn et al., 2023). 이는 AI의 메타인지 능력을 활용하여 결과물의 질을 내부적으로 향상시키는 방식으로, 본 연구 프로토콜의 ‘6단계: 자기 점검 및 오류 탐지’ 과정에 직접적으로 반영되었다. 또한, 단일한 사고 경로를 넘어 여러 추론 경로를 탐색하고 최적의 답을 선택하는 ‘사고의 나무(Tree-of-Thoughts, ToT)’ 와 같은 더 발전된 프롬프팅 기법도 제안되며(Yao et al., 2023), 프롬프트 전략은 더욱 정교화되는 추세이다.

이러한 프롬프트 전략들은 복잡한 과업 수행 중 발생할 수 있는 맥락 상실(context loss)과 논리적 비일관성 문제를 보완하는 데 효과적이다(Maity, Derooy & Sarkar, 2024). 특히 고품질의 평가 문항을 생성하기 위해서는 결과물의 최종 검토뿐만 아니라 생성 과정 전반에 걸친 체계적인 품질 관리가 필수적이다. Shah(2024)가 제안한 ‘다단계 검증(multi-stage verification)’ 개념은 이러한 필요성을 충족하는 전략으로, AI가 생성한 결과물을 여러 단계에 걸쳐 체계적으로 검증함으로써 논리적 일관성과 타당성을 확보하고자 한다.

결론적으로, 본 연구의 GAI-HITL 프로토콜은 인-컨텍스트 학습(ICL)을 바탕으로 사고 연쇄(CoT)를 기본 골격으로 삼고, 내적 독백을 통해 AI의 사고 과정을 투명하게 드러내어 인간의 개입 지점을 확보하며, 자기 수정 개념을 활용해 AI의 자체적인 품질 관리 능력을 유도하는 최신 프롬프트 전략들을 종합적으로 반영한 결과물이다.

### III. 연구 방법

#### 1. 피험자 정보

본 연구는 GAI-HITL 기반 자동 문항 생성 프로토콜을 통해 개발된 평가 문항을 실제 교육 현장에 적용하고, 학습자의 반응 자료를 수집함으로써 문항의 심리측정학적 타당성과 실용성을 검증하고자 하였다. 문항 적용은 정규 수업의 자연스러운 맥락 내에서 이루어졌으며, 이를 통해 생성형 AI 기반 문항이 실제 학습자에게 유의미한 반응을 유도하고 문항으로서의 기능을 충실히 수행하는지를 실증적으로 확인하고자 하였다.

연구 참여자는 전라남도 소재 일반계 A고등학교 3학년 재학생 55명으로 구성되었으며, 이들은 모두 동일한 국어 교사의 <독서> 과목 정규 수업을 수강 중인 집단이었다. 평가 문항은 2025년 5월 1주차 정규 수업 시간에 일괄적으로 실시되었으며, 각 학생은 개별적으로 제공된 문항지에 응답하였다. 자료 수집 직후 응답지를 검토한 결과, 전 문항에 무응답하거나 전 문항에 정답만을 표시한 2명의 응답은 비정상적 반응 사례로 간주되어 최종 분석에서 제외되었다.

응답 자료는 각 문항별로 학습자가 선택한 선지와 정답 여부(정답: 1, 오답: 0)를 기준으로 코딩되었으며, 이후 문항별 정답률, 선지 선택 빈도, 학습자별 총점 분포 등 기초 통계량을 산출할 수 있는 형태로 전처리되었다. 사전 정제 과정에서는 결측 응답 확인, 이상 반응 제거, 정오 처리 일관성 검토 등이 포함되었으며, 이는 후속 통계 분석(고전 검사이론 및 문항반응이론 기반)의 정확성과 해석 가능성을 확보하기 위한 필수 절차로 수행되었다.

#### 2. 자동 문항 생성 절차

##### 가. 성취기준 및 독해 지문 선정 과정

본 연구에서는 고등학교 독서 교육에서 요구되는 평가 타당성을 확보하기 위해, 문항 생성을 위한 기준으로 『2015 개정 국어과 교육과정』의 독서 영역 성취기준을 적용하였다. 적용된 성취기준은 크게 두 영역으로 구분되며, 첫째, 독서의 방법 영역에서는 글에 드러난 정보를 바탕으로 중심 내용, 주제, 구조 등을 파악하는 사실적 이해(12독서02-01), 드러나지 않은 정보를 추론하는 능력(12독서02-02), 글의 관점이나 표현, 사회적 이념 등을 비판적으로 이해하는 능력(12독서02-03), 필자의 관점에 대한 대안 탐색

과 같은 창의적 읽기(12독서02-05)를 포함한다. 둘째, 독서의 분야 영역에서는 인문·예술(12독서03-01), 사회·문화(12독서03-02), 과학·기술(12독서03-03) 분야별 특성과 관련된 비판적 이해 능력을 요구하고 있다(〈표 1〉 참고).

〈표 1〉 문항 생성 대상 독서 성취 기준(2015 개정 국어과 교육과정)

구분	성취기준	핵심 내용
독서의 방법	12독서02-01	글에 드러난 정보를 바탕으로 중심 내용, 주제, 구조 등 사실적 내용 파악
	12독서02-02	드러나지 않은 정보를 예측하여 의도, 목적, 숨겨진 주제 등 추론
	12독서02-03	글의 관점·표현·사회적 이념 등을 비판적으로 이해
	12독서02-05	사회 및 필자 생각에 대한 대안 탐색 등 창의적 읽기
독서의 분야	12독서03-01	인문·예술: 인문학적 세계관, 예술·삶의 태도, 인간 성찰 비판적 이해
	12독서03-02	사회·문화: 사회적 요구·신념, 현상 특성, 역사적 맥락 비판적 이해
	12독서03-03	과학·기술: 지식·정보의 객관성, 논거 입증, 과학 원리의 응용과 한계 비판적 이해

이러한 성취기준을 바탕으로, 문항 생성을 위한 지문은 『2026학년도 수능특강 국어 영역 독서』(한국교육방송공사, 2025) 교재에서 선정하였다. 해당 교재는 고등학교 3학년 수업에서 실제로 활용되는 대표적인 연계 교재로, 교육부와 한국교육과정평가원의 감수를 받은 공신력 있는 자료이다. 본 연구에서는 2025년 3~4월 중 실제 수업 시간에 다루어진 총 10편의 지문<sup>1)</sup>을 1차 후보군으로 설정한 뒤, 후보 지문에 대한 세 명의 교사 협의가 이루어졌다. 이때, 이 과정에 참여한 교사 3인의 평균 교육 경력은 10년이며, 고등학생을 대상으로 다년간 대학수학능력시험 대비 독서 교육 및 평가 문항 출제 경험을 보유하고 있었기에 논의의 전문성을 확보하였다. 특히 문항 생성 프로토콜에 직접 참여한 교사 1인(이하 ‘개발 교사’)은 GAI-HITL 프로토콜에 대한 사전 이해를 바탕으로 AI와의 협업을 주도하였으며, 최종 문항 검토에 참여한 교사 2인(이하 ‘검토 교사’)은 동료 검토자로서 문항의 객관성과 교육적 적합성을 검증하는 역할을 수행하였다. 이들의 반복 협업을 통해 최종적으로 6편의 지문이 선정 되었고, 이때의 선정 기준은 ①교육과정 성취기준과의 정합성, ② 다양한 독해 기능과 수준별 문항 도출 가능성, ③평가 목적과 수업 맥락과의 부합성 등을 중심으로 설정하였다.

1) 본 연구에서 활용한 지문은 당해 연도 EBS 수능특강 교재에서 선정하였으므로, 피험자인 고3 학생들이 해당 지문에 사전 노출되었을 가능성이 있다는 점은 연구의 잠재적 제한점으로 고려될 수 있다. 다만 본 연구는 이러한 사전 학습 효과를 통제 변인으로 설정하기보다, 연구의 생태학적 타당성(ecological validity)을 확보하는 방향으로 설계되었다. 즉, 문항 생성에 활용된 지문은 평가 시행 직전인 2025년 3~4월 중 해당 학급의 정규 <독서> 수업 시간에 실제로 다루어졌던 내용으로 한정하였다. 따라서 본 평가가 측정하는 것은 ‘처음 보는 글에 대한 독해력’이 아니라, ‘학습한 내용을 바탕으로 한 이해 및 적용 능력’에 가깝다. 이는 대학수학능력시험의 EBS 연계 방식이나 일반적인 학교 내신 평가 상황을 반영한 것이기도 하다.

최종 선정된 지문은 기술, 사회, 인문 등 세 가지 분야에 고르게 분포하였으며, 각 지문은 학습자의 사고력을 촉진할 수 있는 주제를 담고 있다. 예를 들어, ‘관세의 기능과 종류’와 같은 사회 영역 지문은 비판적 사고를 유도할 수 있는 맥락을 포함하고 있으며, ‘지각에 대한 김창협 의 주장’과 같은 인문 지문은 상반된 관점을 제시함으로써 고차 사고를 자극하는 데 적절하다. 각 지문에 대한 구체적인 제재 영역, 주제, 페이지, 문항 수는 <표 2>에 정리하였다. 이후 단계에서는 이러한 지문을 바탕으로 ChatGPT와 교사 협업을 통해 사고 유형별 문항이 생성되었다.

<표 2> 문항 생성에 활용된 지문 개요

	지문 제재 영역	세부 주제	페이지	문항수
1	기술	자석에서 자기장이 발생하는 이유	21~23쪽	2문항
2	사회	관세의 기능과 종류	139~141쪽	4문항
3	인문	지각에 대한 김창협 의 주장	25~26쪽	4문항
4	기술	형광등과 LED 전등	17~19쪽	2문항
5	사회	채권의 수익률	113~116쪽	4문항
6	사회	임차인 보호를 위한 임대차 계약의 갱신	131~134쪽	4문항

최종 선정된 6개 지문은 <표 2>와 같으며, 이후 문항 개발 및 검토 과정을 거쳐 각 지문별로 2~4개의 문항이 최종 확정되었다.

#### 나. 자동 문항 생성 절차 및 프로토콜

본 연구는 고등학교 독서 교육과정의 성취기준에 부합하는 평가 문항을 자동으로 생성하고, 생성형 인공지능(ChatGPT-4o)과 현장 국어 교사의 구조적 협업을 통해 정교화함으로써 문항의 교육적 타당성과 심리측정학적 적절성을 동시에 확보하고자 하였다. 이를 위해 박고운, 최숙기(2025)가 제안한 CoT 프롬프트 설계 방식과 Huang, Zhang & Liu(2022)의 내적 독백(inner monologue) 기반 대형 언어 모델(LLM) 구성 원리를 이론적 배경으로 삼아, AI의 생성 과정을 외화하고 교사의 개입 가능성을 구조화한 GAI-HITL 기반 자동 문항 생성 프로토콜을 설계하였다.

이 프로토콜은 총 8단계로 구성되며, AI의 사고 흐름을 명시적으로 구조화하고 교사가 각 단계마다 실시간으로 협업하여 문항의 구성과 논리를 정교화할 수 있도록 설계되었다.

① 페르소나 설정 및 조건 제공: 개발 교사가 AI에게 ‘수능 국어 독서 문항 출제 전문가’와 같은 구체적인 역할을 부여하고, 5지 선다형 형식, 측정하려는 사고 유형(사실, 추론, 비판, 창의), 교육과정 성취 기준 등 문항 생성의 기본 조건을 명확히 설정한다. 이

단계에서 교사는 초기 출제 맥락이 평가 목적에 부합하는지 점검한다.

② 교육과정 분석 및 평가 요소 도출: AI는 제시된 성취기준을 분석하여 문항이 측정해야 할 교육 목표와 평가 요소를 구조화한다. 교사는 AI가 성취기준의 핵심을 정확하게 이해하고 반영했는지 검토하고 감독한다.

③ 지문 구조 분석 및 개념 정리: AI는 주어진 독서 지문을 읽고 핵심 개념, 논지 전개 방식, 문단 간의 관계 등 텍스트를 구조화하고 그 내용을 요약한다. 교사는 AI의 지문 해석에 논리적 오류나 왜곡이 없는지 검토한다.

④ 예시문항 풀이 및 근거 도출: 개발 교사는 AI에게 해당 지문과 관련된 기출 문항(예: EBS 수록 문항)을 제시하고, 그 풀이 과정과 정답 및 오답의 근거를 지문에서 찾아 설명하도록 요구한다. 이를 통해 교사는 AI가 실제 평가 문항의 출제 패턴과 사고의 흐름을 정확히 학습했는지 확인하고, 잘못된 추론 과정을 보일 경우 즉시 개입하여 수정한다.

⑤ 인지 영역별 문항 설계: 앞선 분석을 바탕으로 AI는 설정된 인지 영역(사실, 추론 등)의 목표에 맞춰 새로운 문항의 발문과 선택지를 구체적으로 설계한다. 교사는 이 단계에서 문항의 구성 논리가 타당한지, 의도한 인지 수준과 일치하는지를 집중적으로 검토한다.

⑥ 자기 점검 및 오류 탐지: AI는 스스로 생성한 문항 초안에 대해 논리적 비약, 발문과 선택지 간의 불일치, 매력적이지 않은 오답 등 잠재적 오류를 메타인지적으로 점검하고 그 결과를 보고한다. 이는 교사가 AI의 사고 과정을 파악하고 이후 정교한 피드백을 제공하는 중요한 기반이 된다.

⑦ 문항 초고 및 해설 작성: 자기 점검을 마친 AI는 발문, 정답, 오답 선택지, 그리고 상세한 해설을 포함한 완전한 형태의 문항 초고를 생성한다. 교사는 이 초고의 전체적인 완성도와 명료성을 검토한다.

⑧ 교사-AI 반복 협업 정교화: 개발 교사는 최종 초고를 바탕으로 “이 선택지는 매력도가 떨어진다” 또는 “발문의 표현이 모호하다”와 같이 구체적인 피드백을 제공한다. AI는 이 피드백을 반영하여 문항을 수정하며, 이 피드백과 수정의 순환 과정은 문항이 최종적인 완성도를 갖출 때까지 반복된다. <표 3>은 이러한 프로토콜 절차에 대한 개요이다.

<표 3> GAI-HITL 기반 자동 문항 생성 프로토콜 개요

단계	주요 절차	AI CoT 기능	교사(HITL) 개입 시점	회귀조건 및 회귀대상 단계
①	AI에게 '출제자' 역할 부여 형식, 사고 유형, 성취기준 등 조건 명시	역할 기반 사고를 정립 조건 기반 의도 설정	조건 일치 여부 판단 출제 맥락 타당성 검토	조건 부적절 시 → ① 반복 또는 성취기준(②) 재검토
②	성취기준 구조화 평가 요소 명시	평가 목표 중심의 구조화 측정 요소 명확화	교육과정 해석 정확성 점검	해석 오류 시 → ② 반복 기준 불일치 시 → ① 회귀

단계	주요 절차	AI CoT 기능	교사(HITL) 개입 시점	회귀 조건 및 회귀대상 단계
③	지문 핵심 논지, 문단 구조 분석 주요 개념 도출	지문 구조화 텍스트 간 논리 흐름 정리	정보 누락/왜곡 여부 점검 용어/관계 오류 수정	지문 해석 오류 시 → ③ 반복 지문 부적합 시 → ② 회귀
④	기출 문항 풀이 근거 위치 명시 정오답 판단 과정 서술	문제 해결 사고 흐름 재현 정답오답 판단 근거화	사고 경로 타당성 점검 출제 패턴 반영 여부 확인	추론 오류 시 → ④ 반복 지문 해석 문제 시 → ③ 회귀
⑤	설정된 인지 수준에 따른 문항 초안 작성 발문 및 선지 설계	인지 목표 반영 문항 구조 논리 구성	발문 표현·선지 구성 논리 검토 인지 적합성 점검	발문 불명확 시 → ⑤ 반복 구성 오류 시 → ④ 회귀
⑥	문항 내 오류 자동 탐색 정답 단서, 선지 유사성 등 점검	메타인지 기반 자기 평가 오류 유형·위치 명시	AI 자기 진단 결과 해석 오류 보정 방향 제시	오류 명확 시 → ⑤ 회귀 구조 전반 문제 시 → ④ 회귀
⑦	문항 요소(발문, 선지) 정리 상세 해설(정답 근거, 오답 설명) 포함	문항 전체 구조 적합성 점검 해설의 논리적 구성	해설 누락/부적절성 난이도 조절 필요 여부 판단	해설 부적절 시 → ⑦ 반복 문항 문제 시 → ⑤ 회귀
⑧	교사의 피드백 반영 AI 재구성 및 수정 루프 수행	피드백 기반 재설계 자기 수정 논리 CoT 표현	발문 명료성, 선지 매력도, 정답 유도성 등 반복 점검	전체 품질 미달 시 → ⑦ 반복 필요 시 ⑤·④까지 회귀 가능

특히 6단계 이후부터는 AI가 메타인지 기반 자기 진단을 수행하고, 교사는 이를 기반으로 반복적인 피드백과 수정을 통해 최종 문항을 완성한다(Shah, 2024). 이 구조는 AI가 사고 흐름을 내적으로 구성한 상태에서 교사가 지속적으로 개입함으로써 문항의 설계 오류와 사고 왜곡을 사전에 방지할 수 있도록 한 점에서 기존의 단순 생성·검토 방식과 본질적으로 구별된다(Huang, Zhang & Liu, 2022).

〈표 4〉는 GAI-HITL 프로토콜의 6단계와 8단계가 실제로 적용된 과정을 보여준다. 개발 교사는 먼저 6단계의 일환으로, AI에게 기출 문제를 참고하여 “개선안을 제시”하라는 메타 인지적 자기 점검 프롬프트를 입력했다. 이 결과를 바탕으로, 8단계 ‘반복 협업 정교화’에서 개발 교사는 “4번 선지가 너무 명확하게 할당관세인 게 보인다”며 난이도 상향을 요구하는 1차 피드백을 제공했다. AI는 이 지적을 수용하며 “정답 선지임을 단서로 추론하기 쉬운 표현들이 많다”고 스스로 분석하고, 수정 방향을 제시했다. 이어서 개발 교사는 “발문에서 ‘두 정책 목표를 모두 고려’하라고 했는데 1번 선지는 하나만 언급해서 오답 판별이 너무 쉽다”는 2차 피드백을 추가로 제공했다. AI는 이 피드백을 반영하여 모든 선지가 두 정책 목표의 긴장 관계를 포함하도록 문항 전체를 재구성하였다. 이처럼 교사의 전문적 판단과 AI의 분석적 제안이 결합되는 구체적이고 반복적인 상호작용을 통해 문항의 완성도를 높여나갔다.

<표 4> 6단계와 8단계 HITL 적용 예시

단계	HITL 지점의 교사 프롬프트	생성 결과 예시															
6 단 계	<p>이전 출제 조건들을 고려하여 각 문항을 비판적으로 검토할 것. 특히, 139쪽 지문의 기출 문제 4번을 적극적으로 활용하여 개선안을 마련할 것.</p>	<p><b>변형 방향 제안</b></p> <p>기존 문항의 구조를 바탕으로 하되, 다음 중 하나 이상의 축을 변형 요소로 삼으면 좋습니다:</p> <table border="1"> <thead> <tr> <th>변형 요소</th> <th>설명</th> <th>예시</th> </tr> </thead> <tbody> <tr> <td>① 수요공급 수치 변경</td> <td>할당 수량 도출 방식 변화</td> <td>수요량 1만2천, 공급량 4천 등</td> </tr> <tr> <td>② 산업 보호 대상 변경</td> <td>수입품이 아닌 원료 사용자 산업 보호 강조</td> <td>반대 논리 구성 가능</td> </tr> <tr> <td>③ 제시된 관세 방식 중 틀린 방식 삽입</td> <td>실제 조건에 맞지 않는 관세 전략 포함</td> <td>슬라이드 관세 부적절하게 삽입</td> </tr> <tr> <td>④ 보기 항목 수를 축소 or 구조 단순화</td> <td>편단 부담을 줄이고 핵심 사고에 집중</td> <td>예: 선택지 4개, 숫자 정보 명확화</td> </tr> </tbody> </table> <p>네, 알고 있습니다. 139쪽 지문(「관세의 기능과 종류」)에 딸린 기출 4번 문항은 다음과 같은 유형이었습니다:</p> <p><b>기출 4번 문항</b></p> <p>문항 유형: 사례 적용 / 관세 종류 판단 문항 구조: &lt;보기&gt; 상황 제시 + 5인의 제안 → 가장 적절한 해결책 고르기</p> <p>&lt;보기&gt; 요약</p> <ul style="list-style-type: none"> <li>어떤 원료의 국내 총수요량: 1만 톤, 국내 총공급량: 5천 톤</li> <li>수입 원료 5천 톤 필요</li> <li>국내 산업 보호를 위해 20% 관세를 필요</li> <li>BUT 원료 전체에 20% 관세 적용 시 원료를 사용하는 산업에 부담이 큼</li> </ul> <p>▶ 핵심 쟁점: ① 국내 산업 보호 vs ② 원료 사용자 산업의 부담 경감 → 이 두 가지를 균형 있게 조정할 수 있는 관세 방식을 고르는 문제</p>	변형 요소	설명	예시	① 수요공급 수치 변경	할당 수량 도출 방식 변화	수요량 1만2천, 공급량 4천 등	② 산업 보호 대상 변경	수입품이 아닌 원료 사용자 산업 보호 강조	반대 논리 구성 가능	③ 제시된 관세 방식 중 틀린 방식 삽입	실제 조건에 맞지 않는 관세 전략 포함	슬라이드 관세 부적절하게 삽입	④ 보기 항목 수를 축소 or 구조 단순화	편단 부담을 줄이고 핵심 사고에 집중	예: 선택지 4개, 숫자 정보 명확화
	변형 요소	설명	예시														
① 수요공급 수치 변경	할당 수량 도출 방식 변화	수요량 1만2천, 공급량 4천 등															
② 산업 보호 대상 변경	수입품이 아닌 원료 사용자 산업 보호 강조	반대 논리 구성 가능															
③ 제시된 관세 방식 중 틀린 방식 삽입	실제 조건에 맞지 않는 관세 전략 포함	슬라이드 관세 부적절하게 삽입															
④ 보기 항목 수를 축소 or 구조 단순화	편단 부담을 줄이고 핵심 사고에 집중	예: 선택지 4개, 숫자 정보 명확화															
<p>선지에 구체적인 수치를 반영해 주고, 특히 4번 선지가 너무 명확하게 할당관세인 게 보이는 데 좀 더 난이도를 높일 수 있는 선지로 바꿀 수 있을까?</p>	<p>네, 아주 날카로운 지적입니다.</p> <p><b>현재 ④번 선지는 다음과 같은 이유로 정답이 너무 눈에 띄는 구조를 가지고 있습니다:</b></p> <ul style="list-style-type: none"> <li>"8,000톤까지 0%", "초과분엔 25%"라는 구체 수치</li> <li>"균형 있게 달성할 수 있다"는 긍정적 결론 문장</li> <li>구조적으로 할당 관세를 그대로 설명하는 문장</li> </ul> <p><b>즉, 정답 선지임을 단서로 추론하기 쉬운 표현들이 많습니다.</b></p>																
8 단 계	<p>발문에서 '두 정책 목표를 모두 고려해야한다'고 했는데 1번 선지의 경우엔 '국내 생산 업체를 보호한다'라고만 나와 있으니 오답 판별이 바로 될 것 같아. 각 선지들을 이런 사항들을 고려해서 수정해 줘.</p>	<p><b>④번 선지 리라이팅 전략</b></p> <table border="1"> <thead> <tr> <th>항목</th> <th>수정 방향</th> </tr> </thead> <tbody> <tr> <td>수치</td> <td>유지 (사고력 유도 위해 필수)</td> </tr> <tr> <td>어휘</td> <td>"균형 달성" → 중립적·조건적 표현으로 변경</td> </tr> <tr> <td>구조</td> <td>정답임을 암시하는 구성 회피 (ex. "~이 가장 이상적이다" 지양)</td> </tr> <tr> <td>조점</td> <td>"균형"보다는 "한계 내 현실적 절충" 같은 표현 활용</td> </tr> </tbody> </table> <p>정확하고 핵심적인 지적입니다.</p> <p><b>발문에서 "두 정책 목표를 모두 고려해야 한다"라고 명시했기 때문에, 선지들이 어느 한쪽(①의 경우 '국내 생산 업체 보호')만 언급할 경우 학생들이 곧바로 오답으로 간주할 수 있는 단서가 노출됩니다.</b></p> <p>따라서 모든 선지는 반드시 **두 목표(① 국내 생산 업체 보호, ② 중소기업 수입 부담 완화)**를 모두 언급하거나, 혹은 그 둘의 균형이나 긴장 관계를 자연스럽게 녹여 줍니다.</p> <p>이를 통해 <b>표면 단서 없이 오직 사고를 통해 판단해야 하도록 하겠습니다.</b></p>	항목	수정 방향	수치	유지 (사고력 유도 위해 필수)	어휘	"균형 달성" → 중립적·조건적 표현으로 변경	구조	정답임을 암시하는 구성 회피 (ex. "~이 가장 이상적이다" 지양)	조점	"균형"보다는 "한계 내 현실적 절충" 같은 표현 활용					
	항목	수정 방향															
수치	유지 (사고력 유도 위해 필수)																
어휘	"균형 달성" → 중립적·조건적 표현으로 변경																
구조	정답임을 암시하는 구성 회피 (ex. "~이 가장 이상적이다" 지양)																
조점	"균형"보다는 "한계 내 현실적 절충" 같은 표현 활용																
		<p>(수정 결과↓)</p>															

**【발문】**

어느 금속 원료의 국내 중수요량은 12,000톤, 국내 중공급량은 4,000톤이다.  
 이 원료는 국내 중소 제조기업의 생산에 필수적인 자재로,  
 국내 생산 업체 보호와 중소기업의 수입 부담 완화를 위해 "두 정책 목표를 동시에 고려해야 한다."  
 이 상황에서 다음 중 관세 정책으로 가장 적절한 것은?

**【선지】**

- ㉠ 수입량 전체인 8,000톤에 25%의 고정 세율을 적용하여, 산업 보호 효과를 확보하면서도 일정 수준의 수입 수요를 유지할 수 있도록 한다.
- ㉡ 수입 원료의 가격이 국내 생산가보다 낮을 때에만 8,000톤 전량에 30% 세율을 적용하여, 가격 경쟁을 억제하고 수입 부담을 선택적으로 줄이도록 한다.
- ㉢ 수입 가격이 낮을수록 30%, 높을수록 10% 세율을 적용하여, 수입 가격 변동에 대응하면서도 과도한 저가 수입으로부터 국내 산업을 보호한다.
- ㉣ 8,000톤까지는 세율을 면제하고, 초과분에만 25% 세율을 부과하여, 수입 원료에 대한 안정적 접근을 보장하면서도 일정 수준의 산업 보호를 실현한다.
- ㉤ 수입량 중 절반인 4,000톤에는 15%, 나머지 4,000톤에는 35%의 세율을 자동 적용하여, 관세 평균을 조정하면서 산업 보호와 수입 수요를 병행한다.

프롬프트 상에서는 AI가 학습자의 인지적 복잡성을 고려하여 각 지문과 성취 기준에 따라 ① 사실적 독해, ② 추론적 독해, ③ 비판적 독해, ④ 창의적 독해의 네 가지 사고 유형을 균등하게 반영하도록 지시되었으며, 총 6개 지문에서 유형별로 4문항씩, 총 24 문항이 생성되었다. 프롬프트는 문항 생성자의 역할 설정, 성취기준 연계, 사고 유형 정의 및 예시 제공, 5지선다형 형식 고정 및 해설 포함 등의 조건을 포함하였으며, 교사는 프롬프트 설계와 AI 출력물에 대한 실시간 피드백을 통해 표현 오류나 논리적 결함이 없는 고품질 문항이 생성되도록 지속적으로 개입하였다. 즉, 문항 생성 초기부터 GAI-HITL 프로토콜 8단계 전체에 걸쳐 교사 1인이 개발자로서 함께 참여하였으며, AI의 사고 흐름을 보조하고 문항 구조와 표현을 교육적 기준에 맞게 정교화하였다.

이와 같이 생성된 24개의 초고 문항은 GAI-HITL 프로토콜을 거쳐 구조화되었고, 이후 교사 중심의 2단계 검토 절차를 통해 교육적 완성도를 최종적으로 확보하였다. 이 검토는 동일 학교 소속 교사 3인이 참여한 협의 평가 형태로 이루어졌으며, 다음의 두 가지 하위 과정으로 구성되었다. 먼저 문항 품질 검토에서는 ① 평가 목적 부합 여부, ② 정보 제시의 명료성, ③ 선지 간 변별력, ④ 언어 표현의 적절성 등을 중심으로 오류와 논리적 결함을 점검하고 수정하였으며, 이어 교육적 적합성 검토에서는 ① 지문 특성에 따른 인지 유형의 적합성, ② 교육과정 연계성, ③ 학습자 수준에 맞는 난이도(인지적 부하)를 종합적으로 평가하여 최종 선별을 수행하였다.

이 과정에서 확인된 주요 결과 중 하나는, 생성형 AI가 특히 비판적 이해 영역에서 문항 구성의 타당도를 확보하는 데 어려움을 보인다는 점이었다. 이는 박고운, 최숙기(2025)의 분석과 일치하며, Maity, Deroy & Sarkar(2024), Alfertshofer 외(2024)가 지적한 바와 같이 생성형 AI 기반 문항에서 빈번히 발생하는 문맥 손실과 논리 비일관성 문제의 일환으로 해석할 수 있다. 예를 들어, AI가 생성한 비판적 이해 문항의 다수는 글의 핵심 논지에 대한 타당한 비판보다, 지엽적인 내용이나 표현을 문제 삼거나, 지문

에 근거하지 않은 외부 지식을 끌어와 논지를 왜곡하는 오류를 보였다. 이에 검토 교사 3인은 해당 문항들이 교육과정 성취기준(12독서02-03)에서 요구하는 비판적 사고 능력을 타당하게 측정하기 어렵다고 판단하였고, 논의를 통해 질적 완성도가 확보된 문항만을 선별하는 과정에서 자연스럽게 해당 유형의 문항 수가 조정되었다. 이에 따라 초기 계획되었던 사고 수준 간 균형은 지문별 특성과 문항 완성도를 고려하여 부분적으로 조정되었으며, 최종적으로 총 20개 문항이 채택되었다. 각 문항에는 교사의 사전 예측 난이도(상·중·하)가 함께 부여되어 문항 난이도의 분포도 파악할 수 있도록 하였다(〈표 5〉 참조).

<표 5> 최종 독해 선다형 자동 생성 문항 정보

문항	지문 제재 영역	세부 주제	사고 유형	교사 예측 난이도	비고
1	기술	자석에서 자기장이 발생하는 이유	사실적 이해	하	독해 수준별 문항 분포: 사실적 독해(9개) 추론적 독해(5개) 비판적 독해(1개) 창의적 독해(5개)  예상 난이도별 문항 분포 사실적 독해: 상(0) 중(2) 하(7) 추론적 독해: 상(0) 중(3) 하(3) 비판적 독해: 상(1) 중(0) 하(0) 창의적 독해: 상(2) 중(3) 하(0)
2			창의적 이해	중	
3	사회	관세의 기능과 종류	사실적 이해	하	
4			추론적 이해	중	
5			추론적 이해	하	
6			창의적 이해	중	
7	인문	지각에 대한 김창협 의 주장	사실적 이해	하	
8			사실적 이해	중	
9			비판적 이해	상	
10			사실적 이해	하	
11	기술	형광등과 LED 전등	사실적 이해	하	
12			추론적 이해	하	
13	사회	채권의 수익률	사실적 이해	하	
14			추론적 이해	중	
15			창의적 이해	상	
16			창의적 이해	상	
17	사회	임차인 보호를 위한 임대차 계약의 갱신	사실적 이해	하	
18			사실적 이해	중	
19			추론적 이해	중	
20			창의적 이해	중	

〈표 5〉는 최종 선정된 20개 문항의 제재 영역, 사고 유형, 교사 예측 난이도를 보여 준다. 독해 수준별로는 사실적 독해 문항이 9개로 가장 많았으며, 교사 예측 난이도 기준으로는 ‘상’으로 분류된 문항이 3개(비판적 이해 1, 창의적 이해 2)였다.

<표 6> 관세의 기능과 종류 지문 사회 영역 독서 문항 최종안

EBS 기술 문항(원문)	자동 문항 생성 결과
<p>1. 윗글을 통해 답할 수 있는 질문이 아닌 것은?                      ① 국가적 차원에서 관세는 어떠한 기능을 하는가?                      ② 긴급 관세를 발동하기 위해 정부는 어떠한 행정적 절차를 거쳐야 하는가?                      ③ 관세를 경제적 수단 이상의 대외적 정책 수단으로 보는 이유는 무엇인가?                      ④ 관세에 조세 법률주의의 원칙을 엄격하게 적용하지 않는 이유는 무엇인가?                      ⑤ 현대의 국가에서 모든 상품에 일률적인 관세를 부과하지 않는 이유는 무엇인가?</p>	<p>3. 다음 중 ㉠과 ㉡에 대한 설명으로 적절하지 않은 것은?                      ① ㉠은 국산품과 수입품의 가격 차이를 조정하여 국내 산업 보호 효과를 낼 수 있다.                      ② ㉡은 수입량이 일정 기준을 초과하면 2차 관세율이 적용되며, 할당 수량은 국내 수요량에서 국내 공급량을 뺀 수량으로 정해진다.                      ③ ㉠은 수입 가격에 따라 세율이 유동적으로 조정되는 반면, ㉡은 수입량을 기준으로 세율이 차등 적용된다.                      ④ ㉠은 수입품 가격이 낮아질수록 높은 세율이 적용되며, ㉡은 일정 수량을 초과한 물량에만 높은 세율이 적용된다.                      ⑤ ㉠은 국산품 보호를 위해 높은 세율을 유지하는 반면, ㉡은 수입품의 가격이 급변하는 상황에서 유동적으로 사용된다.</p>
<p>2. ㉠과 ㉡에 대한 설명으로 적절하지 않은 것은?                      ① ㉠에서는 동일한 상품이라도 수출국의 가격 수준에 따라 관세가 달라진다.                      ② ㉡에서는 상품의 품질에 대한 고려 없이 관세가 매겨진다.                      ③ 과세 가격이 올라가면 ㉠에서는 관세가 올라가지만 ㉡에서는 관세가 달라지지 않는다.                      ④ 과세 가격을 파악하기 어려운 물건에 대해서는 ㉠보다 ㉡을 적용하는 것이 효율적이다.                      ⑤ 고가품의 수입을 억제하기 위해서는 ㉠보다 ㉡을 통해 관세를 매기는 것이 더 효과적이다.</p>	<p>4. 다음 중 '긴급 관세'의 발동 사유로 가장 적절하지 않은 것은?                      ① 외국 정부의 지원으로 저가에 수입된 품목이 급증하고, 해당 산업의 국내 기업들이 가격 경쟁에서 밀리고 있다.                      ② 특정 품목의 수입 가격이 하락하면서 수입량이 급증하였고, 이로 인해 유사 품목을 생산하던 국내 기업들이 큰 타격을 입고 있다.                      ③ 수입 가격이 변동 없이 유지되고 있지만, 해외 공급량 증가로 인해 동일 품목의 수입량이 단기간 내 두 배 이상 늘어났다.                      ④ 특정 품목의 수입 가격이 상승하고 수입량은 줄어들었으며, 이로 인해 유사 상품을 생산하던 국내 업체들의 마진율이 악화되고 있다.                      ⑤ 특정 수입품과 유사 품목을 생산하는 국내 산업의 손해 우려가 존재하며, 이를 방지하기 위해 정부가 관세를 조정 시기를 논의하고 있다.</p>
<p>3. ㉠을 이해한 내용으로 적절하지 않은 것은?                      ① 복합 관세의 실효세율은 선택 관세의 실효세율보다 크겠군.                      ② 과세 가격이 일정액보다 커지면 종가세 액이 총량세 액을 초과하겠군.                      ③ 복합 관세를 총량세로 변경하면 관세에 의한 국고 수입이 줄어들겠군.                      ④ 복합 관세를 적용하면 과세 대상 가격 변화에 비례하여 실효세율이 변화하겠군.                      ⑤ 수입품의 수입을 최대한 억제하기 위해서는 선택 관세보다 복합 관세를 적용하는 것이 좋겠군.</p>	<p>5. 다음 중 '관세'와 '조세 법률주의'의 관계에 대한 설명으로 가장 적절한 것은?                      ① 관세는 조세 법률주의의 적용을 받지만, 국내외 경제 여건에 따라 시행령을 통해 세율을 유연하게 조정할 수 있게 되어 있다.                      ② 관세도 조세이므로 법률에 따라 세율이 정해지고, 경제 상황과 관계없이 국회에서 정한 세율만을 적용해야 한다.                      ③ 관세는 과세 대상의 특수성과 국제적 협정 문제로 인해 오히려 법률보다는 고시와 행정 지침을 중심으로 운영된다.                      ④ 관세는 외국과의 협정이 많은 분야이므로 조세 법률주의가 적용되지 않으며, 대부분 대통령 시행령에 따라 결정된다.                      ⑤ 조세 법률주의가 적용되는 경우에도 관세는 행정부에서 정한 조건과 기준에 따라 행정부가 자의적으로 세율을 결정할 수 있다.</p>
<p>4. 윗글을 참고하여 &lt;보기&gt;에 제시된 상황에 대한 해결책을 제안한 내용으로 가장 적절한 것은?                      &lt;보기&gt;                      어느 원료의 국내 총수요량은 1만 톤이며 이 원료의 국내 총공급량은 5천 톤이다. 국내 총공급량은 5천 톤을 유지하고 이 원료의 국내 생산 가격과 수입 가격이 현재 수준과 같다고 할 때, 이 원료를 생산하는 국내 산업의 경쟁력 확보를 위해서는 수입 원료에 대해 20%의 관세율을 적용할 필요가 있다. 그러나 원료 수입량 전체에 대해 20%의 관세율을 적용하면 이 원료를 사용하여 제품을 생산하는 산업의 부담이 가중된다.</p>	<p>6. 다음 &lt;보기&gt; 어느 금속 원료의 국내 수요·공급 상황과 정책 목표에 대한 설명이다. 정책 목표 달성을 위해 필요한 관세 정책으로 가장 적절한 것은?                      &lt;보기&gt;                      • 국내 총수요량: 12,000톤                      • 국내 총공급량: 4,000톤                      • 수입량 필요: 8,000톤                      • 이 원료는 중소 제조기업에 필수적인 자재이며, 정책 목표 (1) 국내 원료 생산 업체 보호 (2) 국내 중소 제조기업 보호</p>
<p>① 갑: 원료의 국내 생산 가격이 현재 수준을 유지한다면 수입 가격에 상관없이 수입량 전체에 대해 관세율을 20%로 고정하도록 합시다.                      ② 을: 국내 생산 가격보다 원료의 수입 가격이 낮으면 높은 세율을 적용하고 원료의 수입 가격이 높으면 낮은 세율을 적용하도록 합시다.                      ③ 병: 원료의 수입 가격이 국내의 원료 가격보다 낮으면 관세를 부과하지 말고 높은 경우에는 30%의 관세율을 적용하여 관세를 부과하도록 합시다.                      ④ 정: 국내 생산 가격과 수입 가격이 현재 수준을 유지한다면 원료 수입량 5천 톤에 대해서는 20%의 관세율을 적용하고 나머지 수입량에 대해서는 슬라이드 관세를 적용하도록 합시다.                      ⑤ 무: 국내 생산 가격과 수입 가격이 현재 수준을 유지한다면</p>	<p>① 수입량 전체인 8,000톤에 25%의 고정 세율을 적용하여, 산업 보호 효과를 확보하면서도 일정 수준의 수입 수요를 유지할 수 있도록 한다.                      ② 8,000톤까지는 세율을 면제하고, 초과분에만 25% 세율을 부과하여, 수입 원료에 대한 안정적 접근을 보장하면서도 일정 수준의 산업 보호를 실현한다.</p>

EBS 기출 문항(원문)	자동 문항 생성 결과
원료 수입량 5천 톤에 대해서는 관세율을 0%로 하고 5천 톤을 초과하는 수입량에 대해서는 20%의 관세율을 적용하도록 합시다.	③ 수입 가격이 낮을수록 30%, 높을수록 10% 세율을 적용하여, 수입 가격 변동에 대응하면서도 과도한 저가 수입으로부터 국내 산업을 보호한다. ④ 수입 원료의 가격이 국내 생산가보다 낮을 때에만 8,000톤 전량에 30% 세율을 적용하여, 가격 경쟁을 억제하고 수입 부담을 선택적으로 줄이도록 한다. ⑤ 수입량 중 절반인 4,000톤에는 15%, 나머지 4,000톤에는 35%의 세율을 차등 적용하여, 관세 평균을 조정하면서 산업 보호와 수입 수요를 병행한다.

### 3. 분석 방법

본 연구에서는 GAI-HITL 방식으로 생성된 고등학교 독서 문항의 심리측정학적 특성을 검토하기 위해 고전검사이론(CTT)과 문항반응이론(IRT)을 상호보완적으로 적용하였다. 이는 문항의 난이도와 변별도를 다층적으로 해석하여 생성형 AI 기반 문항 개발의 타당성과 실용성을 입체적으로 평가하기 위함이다(Cappelleri, Lundy & Hays, 2014; Hu, Liu & Zhang, 2021).

분석 전략 측면에서, 본 연구는 CTT를 주 분석틀로, IRT를 보조적 분석틀로 설정하였다. 이는 유효 응답자 53명이라는 표본 수의 한계를 고려한 현실적 판단이다. CTT는 문항 난이도와 변별도의 안정적 추정에 요구되는 최소 표본 크기( $N \approx 50$ )를 비교적 충족하여(Kline, 2005), 분석의 기본적인 신뢰성을 확보할 수 있다. 따라서 본 연구는 개념이 명확하고 적은 표본에서도 안정적인 통계치를 제공하는 CTT의 장점을 우선적으로 활용하였다. 반면, IRT는 문항과 피험자 특성을 분리하여 분석하는 장점이 있지만(Cappelleri, Lundy & Hays, 2014), 안정적인 모수 추정을 위해 본 연구의 표본보다 훨씬 큰 규모를 요구한다. 따라서 IRT 분석은 CTT만으로는 파악하기 어려운 문항 특성을 탐색적으로 살펴보는 데에만 제한적으로 활용되었다.

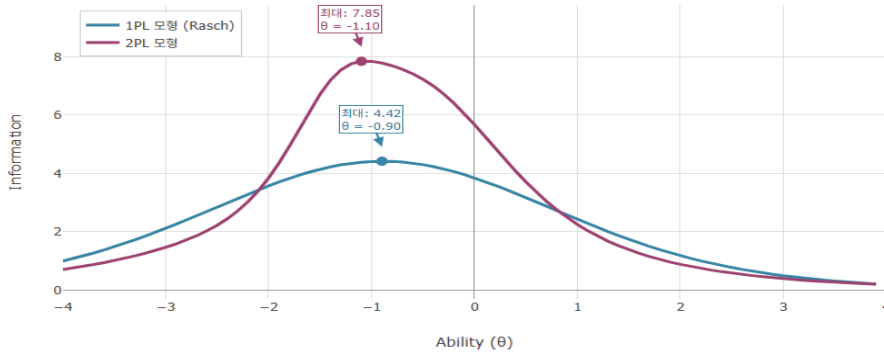
CTT는 개인의 관찰 점수(X)가 진점수(T)와 측정 오차(E)의 합으로 구성된다는 기본 가정( $X = T + E$ )에 근거하며(Crocker & Algina, 1986), 문항 난이도는 정답자 비율(p-value)로, 변별도는 점이연 상관계수(r) 및 상·하위 집단 정답률 차이(DI)로 산출된다. 이때 상·하위 집단은 검사 총점을 기준으로 상위 27%와 하위 27%의 학생들로 분류하였다(Kelley, 1939).

&lt;표 7&gt; 1PL 모형과 2PL 모형 간의 적합도 지표 및 우도비 검정 결과

모델	Log_Likelihood	AIC	BIC	우도비 검정 $\chi^2(df)$	p-값
1PL	-517.99	1075.98	1175.30	18.72(20)	0.540
2PL	-508.63	1097.27	1295.91		

<표 7>과 같이 IRT 분석에서는 모델 적합도 검정 결과, 1PL 모형이 데이터에 적절히 부합하는 것으로 나타났다( $\chi^2(20) = 18.72, p = .540$ ). 우도비 검정에서도 2PL 모형이 1PL 모형에 비해 통계적으로 유의미한 개선을 보이지 않아, 간명성의 원칙에 따라 1PL 모형을 채택하는 것이 타당하였다. 그럼에도 본 연구에서는 CTT의 변별도 지표와 비교·분석한다는 탐색적 목적을 위해 2PL 모형의 결과를 참고하였다.

그러나 본 연구의 표본 크기( $N = 53$ )는 2PL 모형의 안정적인 모수 추정을 위해 권장되는 최소 수준(일반적으로 약 400~500명)에 미치지 못하는 명백한 한계를 갖는다(Thorpe & Favia, 2012). 두 모형의 측정 정밀도를 비교하기 위해 피험자 능력 수준( $\theta$ )에 따른 측정표준오차(Standard Error of Measurement, SEM)를 분석하였다. 측정표준오차는 검사정보함수(Test Information Function, TIF)의 제곱근에 반비례하므로, 정보량이 많을수록 측정 오차는 작아진다. 본 연구에서 2PL 모형은  $\theta \approx -2.0 \sim +3.0$  구간에서 1PL 모형보다 일관되게 높은 검사정보량을 제공하였고, 이에 따라 이론적 기대에 부합하게 대부분의 구간에서 더 낮은 SEM 값을 나타냈다. 예를 들어, 평균적인 능력 수준인  $\theta = 0$ 에서는 1PL 모형의 SEM이 0.66, 2PL 모형은 0.57이었고,  $\theta = -1$  구간에서도 각각 0.71과 0.52로 2PL 모형의 정밀도가 더 높았다. 그러나 극단적 능력 구간(예:  $\theta = -4.0$ )에서는 2PL 모형의 SEM이 1PL보다 약 3배 이상 높게 나타나는 등 일부 구간에서는 이론적 기대와 상반된 결과가 확인되었다. 이는 소표본 상황에서 2PL 모형의 문항 변별도(a) 모수 추정이 불안정해짐에 따라 TIF 자체의 정밀도가 저하되고, 결과적으로 실제 측정오차가 증가했을 가능성을 시사한다(Reise & Waller, 2009; Finch & French, 2019). 따라서 이론적 SEM과 실측된 총 SEM 간의 구분이 필요하며, 특히 소규모 표본 기반 연구에서는 2PL 모형의 정밀도 해석에 신중을 기할 필요가 있다.



주:  $\theta$ 는 표준화된 능력 수준을 나타내며, 정보량이 높을수록 해당 능력 구간에서 더 정밀한 측정이 가능함

[그림 1] 1PL 모형과 2PL 모형의 검사정보함수(Test Information Function) 비교

이에 2PL 모형의 분석 결과는 문항 특성에 대한 확정적인 판단의 근거로 활용하기보다, CTT 분석 결과와의 경향성을 비교하는 탐색적 자료로만 제한적으로 참고하였다. 2PL 모형의 변별도( $a$ ) 값은 CTT의 변별도 지수와 경향성을 비교하는 참고 자료로만 활용하였으며, 모수 값 자체의 절대적 의미는 신중하게 해석하였다. 문항 해석 기준은 CTT 기반 변별도(DI)의 경우 높은 변별도( $D \geq .40$ )부터 매우 낮음( $D < .20$ )으로 분류하고(Wu, Tam & Jen, 2016), IRT 기반 변별도( $a$ )는 참고를 위해 매우 높음( $a \geq 2.0$ )부터 매우 낮음( $a < 0.5$ )으로 구분하였다(Magis, Yan & von Davier, 2017).

본 연구의 모든 통계 분석은 R 소프트웨어(Ver. 4.3.x)를 활용하였으며, CTT 분석에는 CTT 패키지를, IRT 분석에는 ltm 패키지를 사용하였다.

## IV. 연구 결과

### 1. 기초 응답 통계 분석

본 절에서는 GAI-HITL 기반 자동 문항 생성 프로토콜을 통해 개발된 총 20개 문항에 대해 고등학교 3학년 학습자들이 실제로 응답한 결과를 바탕으로 수행한 기초 통계 분석 결과를 제시한다. 이는 후속 심리측정 분석(CTT 및 IRT)에 앞서 문항군 전체의 난이도 분포, 정답률 경향, 내적 일관성 수준 등을 확인하고, GAI-HITL 기반 문항이 학습자 수준에 비추어 교육적으로 적절한 문항 특성을 지니고 있는지를 검토하기 위한 목적을 지닌다. 주요 기초 통계 분석 결과는 <표 8> 및 <표 9>와 같다.

&lt;표 8&gt; 자동 생성 문항의 피험자 기초 응답 통계 분석

문항	지문 제재 영역	세부 주제	사고 유형	교사 예측 난이도	선택지 반응 분포 (명)					정답률 (%)
					1	2	3	4	5	
1	기술	자석에서 자기장이 발생하는 이유	사실적 이해	하	44	2	1	5	1	81.82
2			창의적 이해	중	4	36	7	3	3	67.27
3	사회	관세의 기능과 종류	사실적 이해	하	1	0	5	1	46	85.45
4			추론적 이해	중	2	1	16	29	5	54.55
5			추론적 이해	하	41	2	3	4	3	76.36
6			창의적 이해	중	5	25	6	8	9	47.27
7	인문	지각에 대한 김창협 의 주장	사실적 이해	하	36	0	15	0	2	67.27
8			사실적 이해	중	5	4	3	41	0	76.36
9			비판적 이해	상	0	41	5	5	2	76.36
10			사실적 이해	하	3	5	37	1	7	69.09
11	기술	형광등과 LED 전등	사실적 이해	하	4	2	0	0	47	87.27
12			추론적 이해	하	2	4	39	6	2	72.73
13	사회	채권의 수익률	사실적 이해	하	4	7	38	1	3	70.91
14			추론적 이해	중	9	7	6	26	5	47.27
15			창의적 이해	상	28	7	5	7	6	52.73
16			창의적 이해	상	4	13	9	21	6	40
17	사회	임차인 보호를 위한 임대차 계약의 갱신	사실적 이해	하	2	38	6	6	1	70.91
18			사실적 이해	중	8	4	4	7	30	56.36
19			추론적 이해	중	6	2	30	9	6	56.36
20			창의적 이해	중	3	2	4	7	37	67.27

&lt;표 9&gt; 정답률 분포에 따른 기초 통계량(N=53)

최솟값	1/4값	중앙값	평균값	3/4 값	최댓값
5.00	11.00	13.00	13.36	17.00	19.00

GAI-HITL 기반 자동 문항 생성 프로토콜을 통해 개발된 총 20개 문항에 대해 고등학교 3학년 학습자들이 실제로 응답한 결과를 바탕으로 기초 통계 분석을 실시하였다. <표 8>과 <표 9>에 제시된 바와 같이, 전체 20문항에 대한 검사 총점의 평균은 13.36 점, 중앙값은 13.00점, 표준편차는 3.42점으로 나타났다. 사분위수 간 범위(IQR)는 6.00점이었으며, 최솟값은 5점, 최댓값은 19점으로 총 14점의 범위를 보였다. 문항군의 내적 일관성 신뢰도를 검토한 결과, Cronbach's  $\alpha$  계수는 0.7897로 산출되었다. 이는 Nunnally & Bernstein(1994)이 제시한 심리측정 도구의 신뢰도 기준인 0.70을 상회하는 수치로, GAI-HITL 기반으로 설계된 문항 세트가 평가 도구로서 안정적인 내적 일관성을 확보했음을 보여준다.

## 2. 자동 생성 문항 난이도 분석

본 절에서는 GAI-HITL 기반 자동 문항 생성 프로토콜을 통해 개발된 고등학교 독서 문항 20개에 대해 문항 난이도 분석을 실시하였다. 분석은 고전검사이론(CTT)과 문항 반응이론(IRT)을 병행 적용하여, 각각 정답률(p)과 난이도 모수(b)를 추정하고, 교사의 사전 예측 난이도와와의 정합성을 비교하는 방식으로 이루어졌다. <표 10>은 문항별 난이도 분석 결과를 정리한 것이다.

<표 10> 문항별 난이도 분석 결과

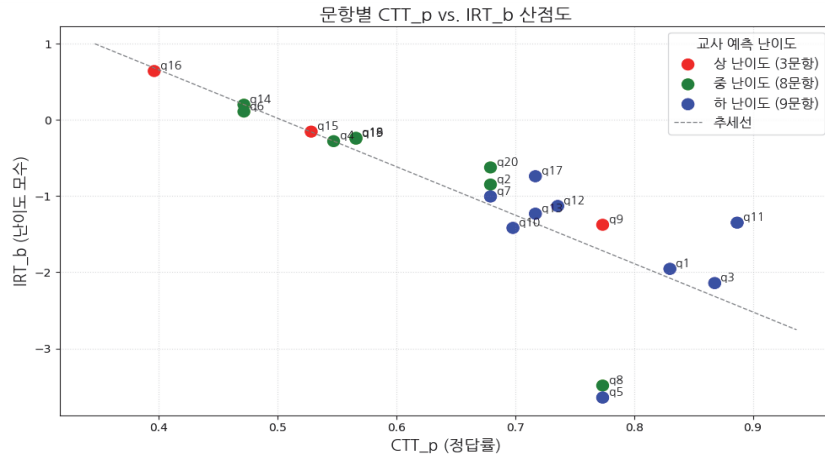
문항	교사 예측 난이도	CTT_p	IRT_b
q1	하	0.8302	-1.9555
q2	중	0.6792	-0.8488
q3	하	0.8679	-2.1421
q4	중	0.5472	-0.2795
q5	하	0.7736	-3.6451
q6	중	0.4717	0.1086
q7	하	0.6792	-1.0057
q8	중	0.7736	-3.4871
q9	상	0.7736	-1.377
q10	하	0.6981	-1.4181
q11	하	0.8868	-1.3496
q12	하	0.7358	-1.1304
q13	하	0.717	-1.2318
q14	중	0.4717	0.1993
q15	상	0.5283	-0.1547
q16	상	0.3962	0.641
q17	하	0.717	-0.7398
q18	중	0.566	-0.235
q19	중	0.566	-0.2478
q20	중	0.6792	-0.6245

<표 11> 난이도 지표별 기술통계 요약

난이도 지표	최소값	최대값	평균	표준편차
CTT_p	0.3962	0.8868	0.6709	0.1134
IRT_b	-3.6451	0.641	-1.0562	1.0393

<표 11>에 제시된 바와 같이, 전체 문항의 CTT 기반 정답률 평균은 0.6709(SD = 0.1134), IRT 기반 난이도 평균은 -1.0562(SD = 1.0393)로 나타났으며, 이는 전체 문항군이 비교적 쉬운 수준으로 구성되었음을 보여준다. 특히 정답률이 0.70 이상인 문항

이 절반 이상을 차지하고, IRT b값이 대부분 음수에 분포함에 따라, 학습자 평균 능력 수준보다 낮은 인지 요구 수준에서 문항이 구성되었음을 알 수 있다.



[그림 2] 문항별 CTT\_p vs. IRT\_b 산점도

<표 12> 난이도 해석 지표 간 등급 일치도

지표 쌍	일치 문항 수	일치율 (%)
교사 예측 vs CTT_p	11	55%
교사 예측 vs IRT_b	13	65%
CTT_p vs IRT_b	15	75%

<표 13> 난이도 상관분석 결과(Spearman / Pearson)

(Pearson Correlation)	스피어만 상관계수(Spearman Correlation)			
		교사예측난이도	CTT_p	IRT_b
	교사예측난이도		0.619	-0.598
	CTT_p	0.623		-0.911
	IRT_b	-0.478*	-0.786	

유의수준 : \*p < 0.05, p < 0.01

교사 예측 난이도와와의 정합성은 <표 12>에서 확인할 수 있다. CTT 기준 일치율은 55%, IRT 기준 일치율은 65%로, IRT가 교사의 판단과 더 높은 일치도를 보였다. CTT와 IRT 간의 정합성은 75%로 가장 높았으며, 이는 두 측정 이론이 서로 다른 계산 논리를 가지면서도 실질적으로는 유사한 문항 특성을 측정하고 있음을 보여준다. <표 13>의 상관분석 결과에서도 교사 예측 난이도와 CTT 간에는 중간 정도의 정적 상관관계

(Pearson  $r = .619$ ,  $p < .01$ ), 교사 예측과 IRT 간에는 유의미한 부적 상관관계(Pearson  $r = -.598$ ,  $p < .01$ ; Spearman  $\rho = -.478$ ,  $p < .05$ )가 나타났고, CTT와 IRT 간에는 매우 높은 부적 상관(Pearson  $r = -.911$ ,  $p < .01$ )이 나타났다. [그림 2]의 산점도에서도 이러한 높은 상관관계를 시각적으로 확인할 수 있다. 특히 교사 예측과 IRT 간의 부적 상관은 교사가 더 어렵다고 예측한 문항일수록 IRT에서는 상대적으로 낮은 b값(쉬운 문항)을 보인다는 것을 의미하며, 이는 교사의 직관적 판단과 IRT 모델의 능력 기반 추정 간 체계적 차이를 시사한다. <표 10>의 문항별 분석 결과와 <표 11>의 기술통계 요약을 통해 세부적인 난이도 특성을 살펴보면, CTT와 IRT의 난이도 해석이 일치한 문항은 q1, q5, q11, q16 등 총 15개 문항으로, 양 이론 모두 동일하게 쉬운 또는 어려운 문항으로 판단하였다. 예를 들어, q5는 CTT 정답률이 0.7736, IRT b값이  $-3.6451$ 로 매우 쉬운 문항으로 해석되었으며, q16은 정답률 0.3962, b값 0.641로 가장 어려운 문항으로 분석되어, 두 지표가 높은 정합성을 보였다.

반면 일부 문항은 CTT와 IRT 간 해석이 상이하였다. q2는 CTT 기준으로는 정답률 0.6792로 ‘중간’ 수준이지만, IRT에서는 b값이  $-0.8488$ 로 쉬운 문항으로 해석되었다. 이는 CTT의 난이도(p)가 단순히 정답자 비율에 의해 결정되는 것과 달리, IRT의 난이도(b)는 ‘50%의 정답 확률을 갖는 능력 수준’을 의미하기 때문에 나타나는 현상이다 (Baker, 2001). 즉, q2 문항의 경우 평균적인 정답률에도 불구하고 상대적으로 낮은 능력 수준의 학생들도 다수 정답을 맞았기에, IRT 모델은 해당 문항이 높은 능력을 요구하지 않는 ‘쉬운’ 문항이라고 판단한 것이다(Hambleton, Swaminathan & Rogers, 1991). 유사하게 q4도 CTT 정답률은 0.5472로 ‘중간’ 수준이나, IRT b값은  $-0.2795$ 로 평균(0)보다 약간 낮은 ‘쉬운 문항’으로 분석되었다. 이 두 지표 간의 불일치는 해당 문항이 가진 인지적 요구의 복잡성에서 기인한 것으로 해석된다. 예를 들어, q4 문항(‘채권의 수익률’ 지문)은 수치를 활용한 추론을 요구하는데, 선택지 중 하나가 매우 매력적인 오답으로 기능하여(1번 선택지 반응률 17.0%) 능력 수준이 중간 이하인 학생들이 오답을 선택할 확률을 높였다. 이처럼 선택지 구조의 미세한 차이가 CTT와 IRT 난이도 추정치 간의 불일치에 영향을 미친 것으로 보인다. 그러나 본 연구의 이러한 추정 은 향후 후속 연구를 통해 보다 심화된 질적 분석과 연계하여 논의할 필요가 있다.

&lt;표 14&gt; 문항 난이도 특이 문항 분석 사례

유형	대표 문항	교사 예측	실제 정답률(%)	내용 영역	사고 수준	선지 정답 집중률(%)	불일치 원인 및 특성
교사 과대추정	q8	중	76.36	인문	사실적 이해	77.40	지문 내 명시적 정보로 인한 용이성, 내용 추상성 과대평가
	q9	상	76.36	인문	비판적 이해	77.40	논리구조 명확성, 철학 영역 난이도 과대추정 경향
교사 과소추정	q6	중	47.27	사회	창의적 이해	47.20	선택지 간 고른 분산, 창의적 적용의 복잡성 과소평가
	q14	중	47.27	사회	추론적 이해	49.10	수치적 추론 복잡성, 매력적 오답(1번 17.0%) 존재
적절한 예측	q1	하	81.82	기술	사실적 이해	83.00	명시적 정보, 높은 정답 집중도로 예측과 일치
	q16	상	39.62	사회	창의적 이해	39.60	선택지 간 적절한 분산, 고차원적 사고 요구 적절 예측

<표 14>는 교사의 사전 예측과 실제 결과 간 불일치가 나타난 특이 문항의 분석 사례를 요약한 것이다. 교사가 실제보다 어렵다고 예측한 대표적인 사례는 q8과 q9이다. q8은 ‘중’ 난이도로 분류되었으나 정답률 0.7736, b값 -3.4871로 실제로는 쉬운 문항으로 기능하였으며, q9는 ‘상’ 난이도로 분류되었으나 동일한 정답률과 b값 -1.377로 쉬운 문항으로 나타났다. 두 문항 모두 인문(철학) 영역의 지문에 기반하며, 지문 내 정보가 명시적으로 제공되거나 논리 구조가 선명하여 학습자의 추론 과정이 단순화된 것이 주요 원인으로 보인다. 반대로 교사가 쉽게 예측한 문항 중 실제로는 어려웠던 사례로는 q6과 q14가 있다. q6은 창의적 이해 문항으로, 정답률 0.4717, b값 0.1086으로 어려운 문항으로 작용하였다. 선택지 간 반응 분산이 크고, 정답 응답이 확연히 집중되지 않아 학습자들이 정답을 명확히 구별하기 어려웠던 것으로 해석된다. q14는 추론적 이해를 요구하는 사회영역 문항으로, 수치적 개념 혼동이나 매력적인 오답의 존재가 학습자의 판단을 어렵게 만든 것으로 분석된다. 반면, 교사의 예측과 실제 결과가 일치한 대표 문항으로는 q1과 q16이 있으며, 이는 명시적 단서의 제공 여부, 고차적 사고 요구의 명확성 등과 연관되어 있음이 확인되었다.

### 3. 문항 변별도 분석

본 절에서는 GAI-HITL 기반 자동 문항 생성 프로토콜을 통해 개발된 고등학교 독서 문항 20개에 대해 변별도 분석을 실시하였다. 분석은 고전검사이론(CTT)의 점이연상관

계수(r) 및 상하위집단 변별도(DI), 문항반응이론(IRT)의 변별도 모수(a)를 중심으로 이루어졌으며, 각 지표 간의 기술통계적 특성과 정합성, 그리고 문항별 특이 사례에 대한 해석을 통해 자동 생성 문항의 판별력 수준을 종합적으로 평가하고자 하였다. 문항별 상세 변별도 지표는 <표 15>에 제시하였다.

<표 15> 문항별 변별도 지표값

문항	CTT		IRT_a
	점이연상관 (r)	상하위집단 변별도 (DI)	
q1	0.3188	0.2857	0.9419
q2	0.3522	0.3571	1.1029
q3	0.3192	0.2857	1.0382
q4	0.2788	0.5	0.7684
q5	0.1484	0.2143	0.3452
q6	0.4051	0.7143	1.2847
q7	0.3522	0.5714	0.8633
q8	0.16	0.3571	0.3617
q9	0.398	0.5	1.093
q10	0.303	0.3571	0.6433
q11	0.4734	0.4286	3.3111
q12	0.3515	0.4286	1.1308
q13	0.343	0.5	0.8724
q14	0.2715	0.5	0.6235
q15	0.3731	0.7143	0.8655
q16	0.2973	0.6429	0.7371
q17	0.4227	0.5714	2.2913
q18	0.5406	0.7857	2.2035
q19	0.4865	0.7143	1.7851
q20	0.4857	0.7143	2.0854
변별도 수준별 문항 분포 요약			
구분	양호	보통	미흡
CTT_r	11 (55%)	5 (25%)	4 (20%)
DI	13 (65%)	5 (25%)	2 (10%)
IRT a	10 (50%)	7 (35%)	3 (15%)

<표 15>를 바탕으로 문항별 특성을 살펴보면, 먼저 변별도가 우수한 문항들의 경우, q18(CTT\_r = 0.541, IRT\_a = 2.2035)과 q19(CTT\_r = 0.486, IRT\_a = 1.785), 그리고 q6(CTT\_r = 0.405, IRT\_a = 1.2847) 등이 대표적이다. <표 18>에서 확인할 수 있듯, 이들 문항은 정답률이 40~60% 구간에 분포하며 오답 선택지에 응답이 고르게 분산되는 이상적인 패턴을 보였다. 이는 정답률 중간 수준에서 변별력이 극대화된다는 Popham(2017)과 Haladyna(2004)의 이론을 실증적으로 뒷받침한다. 반면, q5(CTT\_r = 0.1484)와 q8(CTT\_r = 0.1600)은 매우 낮은 변별력을 나타냈다. 두 문항 모두 정답

선택지에 응답자가 과도하게 집중되었고(정답률 77.4%), 특히 q8은 하나의 오답 선택지에 아무도 응답하지 않아 오답의 가능성이 결여된 문제를 보였다(Haladyna & Rodriguez, 2013). 예외적으로 q11은 정답률이 88.7%로 매우 쉬웠음에도,  $CTT_r = 0.473$ ,  $IRT_a = 3.311$ 이라는 매우 높은 변별력을 기록하여, 매력적인 오답이 소수 학생들을 효과적으로 변별해 낸 특이 사례로 분석되었다. <표 15>의 결과를 종합한 전체 문항의 변별도 기술통계는 <표 16>과 같다. CTT 기반의 점이연 상관계수( $r$ )는 평균 0.3563, DI는 평균 0.5061, IRT의 변별도 모수( $a$ )는 평균 1.1992로, 대부분의 문항이 중간 이상 수준의 변별력을 갖추고 있음을 시사한다.

<표 16> 변별도 지표별 기술통계 요약

지표	최소값	최대값	평균	표준편차
CTT_r	0.1484	0.5406	0.3563	0.0914
DI	0.2143	0.7857	0.5061	0.173
IRT a	0.3452	3.3111	1.1992	0.7623

문항별로 살펴보면, q18은 정답률 56.6%,  $CTT_r = 0.541$ ,  $IRT_a = 2.2035$ 로 모든 지표에서 뛰어난 변별도를 보였다. 정답 외 선택지에도 고르게 분산된 응답 패턴을 보여 학습자의 능력 수준 차이를 민감하게 포착한 문항이다. 유사하게 q19는 정답률 56.6%,  $CTT_r = 0.486$ ,  $IRT_a = 1.785$ 로 높은 변별력을 보였으며, 선택지 간 분산이 이상적으로 이루어졌다. q6 또한 정답률 47.2%,  $CTT_r = 0.405$ ,  $IRT_a = 1.2847$ 로 우수한 변별력을 보였고, 창의적 이해 유형이라는 점에서 인지적 복잡성이 적절히 작용한 것으로 해석된다. 이들 문항은 정답률이 40~60% 구간에 분포하며, 선택지 간 반응 분산이 잘 이루어져 이상적인 문항 반응 패턴을 형성한 사례이다. 특히 q18과 q19는 선택지별 응답이 균형 있게 분산되면서 학습자의 능력에 따른 정답 도달 양상이 명확히 구분되었으며, 이는 정답률 중간 수준에서 변별력이 극대화된다는 Popham(2017)과 Haladyna(2004)의 이론을 실증적으로 뒷받침한다.

반면, q5는 정답률 77.4%,  $CTT_r = 0.1484$ ,  $IRT_a = 0.3452$ 로, q8은 정답률 77.4%,  $CTT_r = 0.1600$ ,  $IRT_a = 0.3617$ 로 매우 낮은 변별력을 나타냈다. 두 문항 모두 정답 선택지에 응답자가 과도하게 집중되었고, 특히 q8은 5번 선택지에 아무도 응답하지 않아 실질적으로 4지선다 문항으로 작동하였다. 이는 Haladyna & Rodriguez(2013)가 지적한 오답 선택지의 가능성 결여 문제를 그대로 드러낸 사례이며, 선택지 설계상의 미흡함이 문항의 구분 능력을 심각하게 저해함을 보여준다. Tarrant, Ware & Mohammed(2009)의 연구에서도 언급된 바와 같이, 선택지 간 응답의 불균형은 문항의 타당성

과 가능성을 심각하게 약화시키며, q5, q8의 사례는 이를 그대로 입증한다. 예외적으로 q11은 정답률 88.7%로 매우 높은 정답 집중도를 보였음에도 불구하고,  $CTT_r = 0.473$ ,  $IRT_a = 3.311$ 이라는 매우 높은 변별력을 기록하였다. 이는 대부분의 응답자가 정답을 맞췄지만, 일부 오답 선택지가 제한된 수의 학습자에게 혼란을 유발할 수 있는 매력적 오답으로 기능했을 가능성을 보여준다. 오답 선택지의 가능성이 완전히 상실되지 않았고, 정답을 선택한 학습자들의 사고 경로가 보다 일관되었다는 점에서, 문항의 논리 구조가 명확하면서도 교육적 설계가 성공적으로 구현된 사례로 해석할 수 있다. 그러나 이는 단일 문항의 특이 현상으로 일반화에는 주의가 필요하다.

<표 18>에 제시된 정답률 집중도별 분류에 따르면, 60~75% 구간에 위치한 문항들(q20, q17, q2, q12, q13 등)은  $CTT_r = 0.35048$ ,  $IRT_a = 1.0\sim 2.1$ 로 안정적인 변별력을 보였다. 예를 들어 q20은 정답률 67.3%,  $CTT_r = 0.4857$ ,  $IRT_a = 2.085$ 로 변별력이 우수하였으며, 선택지 반응 분산도 적절히 분포되었다. 반면 q1, q3, q4와 같은 문항들은 정답률은 높지만,  $CTT_r$ 이 0.30 내외로 상대적으로 낮은 변별도를 보였고, 선택지 분산도 불균형한 편이었다. 이들 문항은 매력적인 오답 구성의 결여로 인해 응답자의 실제 능력 차이를 충분히 드러내지 못한 사례로 해석되며, 난이도 조정이나 선택지 개선이 필요한 문항으로 분류된다.

변별도 지표 간 정합성을 분석한 결과(<표 17> 참조),  $CTT_r$ 과 DI 간 Pearson 상관계수는 0.883, Spearman 상관계수는 0.809로 매우 높은 수준의 일치도를 보였으며, 이는 두 CTT 기반 지표가 측정 대상의 동일성과 정합성을 입증한다. 반면  $CTT_r$ 과  $IRT_a$  간 상관은 Pearson  $r = 0.681$ , DI와  $IRT_a$  간 상관은  $r = 0.574$ 로 중간 수준의 수렴 타당성을 보여주었다. 이는 CTT가 전통적 응답 데이터를 기반으로 하고, IRT는 능력 수준에 따른 확률 모델에 기반한다는 이론적 차이에서 기인하며, 서로 다른 방식으로 변별력을 측정하면서도 문항의 기능성에 대한 판단은 일정 부분 일치하고 있음을 보여준다. 특히 Spearman 상관계수가 전반적으로 Pearson 상관계수보다 높게 나타난 점은 변별력 지표 간의 순위 기반 해석에서 높은 일관성을 의미하며, 문항 선별이나 개선 우선순위 설정 시 실질적 활용 가능성을 시사한다.

<표 17> 변별도 상관분석 결과(Spearman / Pearson)

Correlation (Pearson)	스피어만 상관계수(Spearman Correlation)		
	CTT_DI	CTT_r	IRT_a
CTT_DI		0.809	0.557*
CTT_r	0.883		0.574*
IRT_a	0.681	0.687	

유의수준 : \*p < 0.05, p < 0.01

이러한 분석 결과는 본 검사가 측정하고자 하는 구인, 즉 ‘독해 능력’에 대한 구인 타당도(construct validity)를 지지하는 하나의 증거로 고려될 수 있다. Messick(1989)에 따르면, 구인 타당도에 대한 주장은 다양한 증거를 통해 종합적으로 뒷받침되어야 하며, 그 증거의 원천 중 하나로 검사의 내적 구조(internal structure)가 언급된다. 본 검사는 선행 연구 및 교육과정 분석을 통해 ‘독해 능력’의 핵심 요소를 반영하여 개발되었으므로, 검사 총점은 학생의 독해 능력을 나타내는 타당한 대리 변인(proxy variable)으로 간주할 수 있다. 이러한 맥락에서 문항 변별도는 검사의 내적 구조 일관성을 살펴볼 수 있는 유용한 지표 중 하나이다. 특정 문항의 변별도가 높게 나타났다는 것은, 검사 총점으로 대변되는 잠재 능력 수준이 높은 학생과 낮은 학생을 해당 문항이 비교적 성공적으로 구별해 낼 가능성을 시사한다. 이는 개별 문항이 전체 검사와 유사한 구인, 즉 ‘독해 능력’을 일관되게 측정하고 있을 개연성을 뒷받침하는 결과로 해석될 수 있다(Allen & Yen, 1979).

이러한 관점에서 <표 18>의 결과를 살펴보면, 정답률 40~60% 구간에 위치한 q6, q18, q19, q15는 CTT 변별도 기준으로 0.37 이상을 기록하여 상대적으로 우수한 변별력을 보이는 것으로 나타났다. 또한, 정답 외 선택지들에도 응답이 고르게 분산된 경향은 효과적인 오답이 가능하고 있음을 보여주며(Haladyna, 2004), 해당 문항들이 능력 수준이 다른 학습자들을 변별하는 데 긍정적으로 기여하고 있음을 시사한다.

<표 18> 선택지 집중도 구간별 요약표

선택지 집중도 구간	문항	CTT_r (점이연상관)	정답률(%)	선택지 분산 특성 요약	핵심 시사점
40-60%	q18, q19, q6	0.405-0.541	47.2-56.6	균형 · 고른 분산	최상 · 우수 변별력, 이상적 분산
60-75%	q20, q17, q2, q12, q13	0.352-0.486	67.9-73.6	적절한 분산	안정적 변별력, 매력적 오답 구성
>75%	q11, q9, q15, q3, q1, q10, q16, q4, q14, q8, q5	0.148-0.473	47.2-88.7	과도 집중 · 극도 집중 · 혼합 분산	고집중 문항 위험 및 예외 사례(q11)

본 연구의 의의를 명확히 하기 위해, 변별도 분석 결과를 GAI-HITL 프로토콜과 연계하여 심층적으로 분석하면 다음과 같다.

첫째, 우수 변별도 문항(q18, q19)의 성공 요인은 ‘교사-AI 반복 협업 정교화(8단계)’ 과정에서 찾을 수 있다. 예를 들어, 가장 높은 변별도(CTT\_r=0.541, IRT\_a=2.2035)를 보인 q18(‘임대차 계약’ 지문)의 경우, AI가 생성한 초고의 오답 선택지들은 단순히 ‘틀린’ 내용으로 구성되어 매력도가 낮았다. 그러나 개발 교사가 8단계에서 “정답과 유사한

논리 구조를 갖지만 결정적인 오류를 포함하는 매력적인 오답으로 수정해달라” 구체적인 피드백을 제공했고, AI는 이 피드백을 반영하여 법률 용어와 조건을 미묘하게 변경한 오답들을 생성했다. 이처럼 교사의 전문적 판단에 기반한 정교한 피드백이 AI의 생성 능력을 특정 방향으로 유도함으로써, 결과적으로 학습자의 고차원적 사고 능력을 변별하는 데 성공적인 문항이 개발될 수 있었다.

둘째, 낮은 변별도 문항(q5, q8)의 원인은 ‘자기 점검 및 오류 탐지(6단계)’의 한계와 관련이 깊다. 변별도가 매우 낮았던 q5(‘관세’ 지문)의 경우, AI는 6단계 자기 점검에서 “논리적 오류 없음”로 자체 진단했다. 개발 교사 역시 8단계 검토 과정에서 해당 문항의 오답 선택지가 가진 문제점(지나치게 정답과 거리가 났음)을 간과하고 다른 문항에 피드백을 집중했다. 이는 AI의 메타인지적 자기 점검이 완벽하지 않으며, 이를 보완해야 할 교사의 개입(HITL)이 충분히 이루어지지 않았을 때 문항의 질적 저하가 발생할 수 있음을 명확히 보여주는 사례이다.

이처럼 GAI-HITL 기반 문항의 심리측정학적 특성은 단순히 문항 자체의 내적 속성만으로 결정되는 것이 아니라, 생성 프로토콜의 각 단계에서 이루어지는 교사와 AI의 상호작용 방식과 질에 직접적인 영향을 받는다. 따라서 본 연구의 난이도 및 변별도 분석 결과는 GAI-HITL이라는 개발 방식의 가능성과 한계를 동시에 보여주는 구체적인 증거라 할 수 있다.

## V. 결론 및 제언

본 연구는 생성형 인공지능과 교사 전문성의 협업을 구조화한 GAI-HITL(Generative AI-Human-in-the-Loop) 기반 자동 문항 생성 방식이 고등학교 국어과 독서 평가에서 심리측정학적 타당성과 교육적 실효성을 확보할 수 있는가를 실증적으로 검토하였다. 총 53명의 고등학교 3학년 학습자의 응답을 바탕으로 고전검사이론(CTT), 문항반응이론(IRT), 변별도 지표 등을 통합적으로 분석함으로써, 자동 생성된 독서 문항의 난이도, 변별도, 신뢰도 등 주요 심리측정 특성에 대한 다면적 검토를 수행하였다.

연구의 결과를 요약하면 다음과 같다.

첫째, GAI-HITL 기반 문항 세트는 평가 도구로서 요구되는 기본적인 양호도를 확보하였다. 구체적으로, 전체 문항의 평균 정답률은 66.8%로 학습자 집단에 적절한 난이도로 기능했으며, 점수 분포 역시 특정 수준에 편중되지 않고 고른 변별 가능성을 보였다. 또한, 내적 일관성 신뢰도(Cronbach's  $\alpha = .7897$ )는 교육평가에서 요구하는 기준을 상

회하여, 개발된 문항들이 측정하고자 하는 구인을 일관되게 측정하고 있음을 뒷받침했다.

둘째, 난이도 분석 결과, 문항 대부분이 학습자 평균 능력보다 약간 쉬운 수준에 분포하였으며, 교사 예측 난이도와 의 정합성은 CTT 기준 55%, IRT 기준 65%로 비교적 높은 수준을 보였다. 다만 문항별로는 사고 수준, 지문 구조, 선택지 유사성 등 다양한 요인이 실제 난이도에 영향을 주었으며, 단순한 제재 영역 분류나 직관적 판단만으로 문항 난이도를 예측하는 데 한계가 있음을 보여주었다.

셋째, 변별도 분석 결과, 다수의 문항이 양호한 변별력을 갖춘 것으로 나타났다. CTT 점이연상관( $r$ ) 기준 55%, 상하위집단 변별도(DI) 기준 65%, IRT 변별도 모수( $a$ ) 기준 50%의 문항이 양호 이상 수준으로 분석되었다. 특히 정답률이 40~60% 구간에 위치한 문항들( $q6$ ,  $q18$ ,  $q19$  등)은 오답 선택지 간 응답 분산이 고르게 이루어져 매우 우수한 변별력을 보였으며 이는 Haladyna(2004)와 Messick(1989)의 문항 설계 원칙을 실증적으로 뒷받침하는 결과이다.

넷째, 지표 간 정합성 분석에서는 CTT\_r과 DI 간 Pearson  $r = 0.883$ , CTT\_r과 IRT  $a$  간  $r = 0.681$ 로 높은 수렴 타당성이 확인되었고, 스피어만 상관계수가 더 높게 나타나 실용적 평가 상황에서 지표 간 순위 일치도가 높음을 시사하였다. 이는 서로 다른 분석 이론 간에도 자동 생성 문항의 판별력이 일관된 방향으로 평가됨을 의미한다.

다섯째, 가장 핵심적인 발견으로, 생성 문항의 심리측정학적 품질이 GAI-HITL 프로토콜 내 교사와 AI의 상호작용 방식 및 질과 직접적으로 연관됨을 확인했다. 우수한 변별도를 보인 문항( $q18$  등)은 '8단계: 반복 협업 정교화'에서 교사의 구체적인 피드백을 통해 오답의 매력도가 향상된 성공 사례였던 반면, 변별력이 낮았던 문항( $q5$  등)은 '6단계: 자기 점검'에서 AI와 교사 모두 문제점을 간과한 사례로 분석되었다. 이는 GAI-HITL 방식의 성공이 단순히 기술을 사용하는 것을 넘어, 인간 전문가의 시의적절하고 정교한 개입에 좌우됨을 시사한다.

이상의 결과는 GAI-HITL 방식의 자동 문항 생성이 현재의 기술 수준에서도 실용 가능한 평가 도구를 생산할 수 있는 가능성을 보여주는 실증적 근거로 해석할 수 있다. 특히 GIA의 출력물을 교사의 판단과 구조화된 검토 절차를 통해 정교화함으로써 교육 현장에 적용 가능한 신뢰도 높은 평가 도구를 개발할 수 있음을 확인하였다.

후속 연구의 제언 사항을 제시하면 다음과 같다. 본 연구의 탐색적 결과를 바탕으로 GAI-HITL 기반 국어과 독서 문항 개발의 실용화를 위한 단계적 발전 방안을 다음과 같이 제안한다.

첫째, 선택지 품질 관리 체계의 내재화가 최우선 과제이다. 본 연구에서 확인된 바와 같이, 오답 선택지의 기능성 결여는 문항 변별력 저하의 직접적 원인으로 작용하며, 이는 본 연구에서 분석한 바와 같이 교사의 개입이 미치지 못한 부분에서 두드러지게 나

타났다. Haladyna(2004)와 Messick(1989)이 강조한 바와 같이 선택지의 질적 수준이 문항의 구인 타당도를 결정하는 핵심 요소임을 고려할 때, 단기적으로는 정답 집중도와 선택지 분산 패턴을 자동 진단하는 품질 관리 도구의 개발이 필요하며, 중장기적으로는 선택지 매력도 예측 알고리즘의 도입을 통해 문항 생성 초기 단계부터 체계적 품질 관리가 이루어져야 한다(Rodriguez, 2005).

둘째, 국어 교사의 문항 개발 역량 강화를 위한 체계적 연수 프로그램이 핵심 전략이다. 소수 교사 참여로 인한 본 연구의 한계를 극복하기 위해, 다양한 배경의 교사들이 GAI-HITL 방식을 효과적으로 활용할 수 있도록 하는 전문성 강화가 필수적이다. 특히 사고 수준별 문항 구성, CoT 기반 프롬프팅 기법, 선택지 기능성 진단 등의 실습 중심 역량 훈련이 필요하며, 교사 학습공동체를 통한 사례 공유와 검토 워크숍이 병행되어야 한다(Kasneji et al., 2023; U.S. Department of Education, 2023). 이는 AI 시대 교육에서 인간 교사의 전문성이 기술과 협력하는 새로운 형태로 진화해야 한다는 관점과 일치한다.

셋째, 후속 연구에서는 보다 다양한 학교급(중학교, 초등학교 포함), 학년 수준, 제재 영역, 사고 수준, 문항 유형(객관식 외 단답형, 서술형 등), 평가 맥락(형성평가, 총괄평가, 진단평가 등)을 아우르는 확장적 설계가 필요하다. 특히 표집 수가 53명에 불과했던 본 연구의 한계를 극복하기 위해 대규모 학습자 데이터를 기반으로 한 CTT 및 IRT 모수 안정화, 문항군별 반응 패턴 군집 분석 등의 후속 실증 연구가 요구된다. 이러한 확장적 접근을 통해 GAI-HITL 방식의 일반화 가능성과 안정성을 보다 체계적으로 검증할 수 있을 것이다.

넷째, 국어과 독서 영역 내 제재별 차별화된 문항 개발 가이드라인 정립이 요구된다. 본 연구에서 사회 영역 독서 문항의 우수한 변별력과 인문 영역에서의 교사 난이도 과대평가 현상이 확인된 만큼, 제재별 특성을 반영한 정교한 접근이 필요하다. Baker(2001)가 제시한 문항 특성과 피험자 능력의 상호작용을 고려할 때, 제재별 문항 설계 원칙과 교사 예측 보정 방안을 포함한 실무 가이드라인을 개발하고 이를 교사 연수 프로그램에 체계적으로 반영해야 한다(Ebel & Frisbie, 1991; Popham, 2017).

다섯째, 장기적으로는 학습자 맞춤형 독서 평가 시스템으로의 발전을 지향하되, 현실적 접근이 필요하다. 현재의 제한적 연구 결과를 고려할 때, Deep Knowledge Tracing 과 같은 고도화된 기술의 도입은 충분한 후속 연구와 기술적 검증을 거친 신중한 접근이 요구된다. 우선은 현재 기술 수준에서 구현 가능한 문항 품질 분석 도구의 안정화에 집중하고, Nunnally & Bernstein(1994)이 강조한 점진적 타당화 과정에 따라 시스템을 단계적으로 고도화해 나가는 것이 현실적 대안이다(Hambleton & Swaminathan, 1985; Pandey & Karypis, 2019; Piech et al., 2015).

이와 같은 단계적 발전을 통해 GAI-HITL 방식은 국어과 독서 교육에서 공정하고 신

뢰도 높은 평가 도구로 정착될 수 있을 것이며, AI 시대의 교육 혁신과 학습자 중심 평가 체제 구축에 핵심적인 역할을 수행할 수 있을 것이다. 생성형 AI와 교사의 협업이 갖는 시너지 효과는 단순한 자동화의 효율성을 넘어, 교육 전문성과 기술 융합의 새로운 평가 생태계를 구현하는 기반이 될 수 있다. 본 연구는 그 실현 가능성을 실증적으로 제시했다는 점에서 교육 현장과 정책 차원 모두에 유의미한 시사점을 제공한다.

## 참고문헌

- 교육부(2015). 2015 개정 국어과 교육과정(교육과정 고시 제2015-74호). 교육부.
- 김혜정(2008). 고등학교 국어과 평가의 문제점과 개선 방안: 평가 사례를 중심으로. **국어교육학연구**, 32, 97-127.
- 김혜정(2011). 국가 수준 학업 성취도 평가 개선을 위한 읽기 평가의 비판적 검토. **국어교육**, 134, 61-86.
- 권태현, 이정찬, 김승현(2017). 수능 국어 영역 독서 평가 문항의 양호도 조사 연구. **독서연구**, 45, 131-159.
- 남민우, 이상일, 최숙기, 서수현, 남가영, 정민주(2022). 국어과 평가 문항의 양호도 분석틀 개발을 위한 기초 연구. **청람어문교육**, 86, 71-95.
- 박고운, 최숙기(2025). 국어과 읽기 영역 선다형 평가를 위한 자동 문항 생성 방안 연구. **교육과정평가연구**, 28(1), 215-246. <https://doi.org/10.29221/jce.2025.28.1.215>
- 박도순(2009). **교육평가**. 교육과학사.
- 성태제(2014). **현대교육평가**(4판). 학지사.
- 신동광(2024). LLM 기반 맞춤형 자동문항생성기의 활용가능성 탐색: Poe AI 챗봇빌더를 중심으로. **영어교과교육**, 23(2), 181-206.
- 오규설(2022). 자동문항생성기법을 활용한 읽기평가 개발에 대한 시론: 사실적 독해 문항 모형을 중심으로. **청람어문교육**, 87, 7-34.
- 이하늘, 이용상. (2024). 한국어 읽기 평가에서 자동 문항 생성의 가능성 탐색. **교육문화연구**, 30(3), 659-686.
- 임상묵, 조혜원, 이정우, 이현숙(2024). 사회과 자동문항생성을 위한 대규모 언어모델의 활용 가능성 탐색. **교육정보미디어연구**, 30(3), 1035-1060.
- 장성민(2024). 학교 지필평가 출제 과정에서 느끼는 국어 교사의 실천적 어려움 탐색: 고등학교의 독서 선다형 평가 상황을 중심으로. **리터러시연구**, 15(2), 89-110.
- 정민주, 서수현, 남민우, 최숙기, 이상일, 남가영(2022). 좋은 국어과 평가 문항 특성에 관한 질적 분석 연구: 국어과 평가 문항 양호도 분석틀 개발 연구(2). **청람어문교육**, 89, 43-78.
- 정혜승(2008). 교사의 읽기 평가 전문성 실태: 지필 평가 문항 분석을 중심으로. **독서연구**, 19, 307-346.
- 한국교육방송공사(2025). 2026학년도 수능특강: 국어영역 독서. 한국교육방송공사.

- Alfertshofer, M., Hoch, C. C., Funk, P. F., Hollmann, K., Wollenberg, B., Knoedler, S., & Knoedler, L. (2024). Sailing the seven seas: A multinational comparison of ChatGPT's performance on medical licensing examinations. *Annals of Biomedical Engineering*, 52, 1542–1545. <https://doi.org/10.1007/s10439-023-03338-3>
- Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Brooks/Cole.
- Baker, F. B. (2001). *The basics of item response theory*(2nd ed.). ERIC Clearinghouse on Assessment and Evaluation.
- Bejar, I. I., Lawless, R. R., Morley, M. E., Wagner, M. E., Bennett, R. E., & Revuelta, J. (2003). A feasibility study of on-the-fly item generation in adaptive testing. *Journal of Technology, Learning, and Assessment*, 2(3), 1–29.
- Biancini, G., Ferrato, A., & Limongelli, C. (2024). Multiple-choice question generation using large language models: Methodology and educator insights. In *Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization* (pp. 584–590). Association for Computing Machinery. <https://doi.org/10.1145/3631700.3665233>
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., & Amodei, D. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.
- Cappelleri, J. C., Lundy, J. J., & Hays, R. D. (2014). Overview of classical test theory and item response theory for the quantitative assessment of items in patient-reported outcome measures. *Clinical Therapeutics*, 36(5), 648–654. <https://doi.org/10.1016/j.clinthera.2014.04.006>
- Cong-Lem, N., Soyooof, A., & Tsering, D. (2025). A systematic review of the limitations and associated opportunities of ChatGPT. *International Journal of Human – Computer Interaction*, 41(7), 3851–3866. <https://doi.org/10.1080/10447318.2024.2344142>
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart and Winston.
- Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement*(5th ed.). Prentice Hall.
- Finch, W. H., & French, B. F. (2019). A comparison of estimation techniques for IRT

- models with small samples. *Applied Measurement in Education*, 32, 77-96. <https://doi.org/10.1080/08957347.2019.1577243>
- Gierl, M. J., & Haladyna, T. M. (2013). *Automatic item generation: Theory and practice*. Routledge.
- Gierl, M. J., Hollis, L. P., & Tanygin, V. (2021). *Advanced methods of automatic item generation*(1st ed.). Routledge. <https://doi.org/10.4324/9781003025634>
- Grévisse, C. (2023). Comparative quality analysis of GPT-based multiple choice question generation. In H. Florez, I. Galpin, & C. Grévisse (Eds.), *Communications in Computer and Information Science: Vol. 1881. Applied Informatics* (pp. 435–447). Springer, Cham. [https://doi.org/10.1007/978-3-031-46813-1\\_29](https://doi.org/10.1007/978-3-031-46813-1_29)
- Haladyna, T. M. (2004). *Developing and validating multiple-choice test items*(3rd ed.). Lawrence Erlbaum Associates.
- Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*(1st ed.). Routledge. <https://doi.org/10.4324/9780203850381>
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage Publications..
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Springer.
- Hu, Y., Liu, Y., & Zhang, Z. (2021). Automatic item generation: A review of recent developments and future directions. *Educational Measurement: Issues and Practice*, 40(2), 41–52. <https://doi.org/10.1111/emip.12380>
- Huang, Y., Zhang, Z., & Liu, T. (2022). Inner monologue for chain-of-thought reasoning in language models. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics(ACL 2022)*, 1235–1246. <https://doi.org/10.18653/v1/2022.acl-main.115>
- Hwang, K., Wang, K., Alomair, M., Choa, F.-S., & Chen, L. K. (2024). Towards automated multiple choice question generation and evaluation: Aligning with Bloom's taxonomy. In A. M. Olney, I. A. Chounta, Z. Liu, O. C. Santos, & I. I. Bittencourt (Eds.), *Artificial intelligence in education. AIED 2024*(Lecture Notes in Computer Science, Vol. 14830, pp. 389–396). Springer, Cham. [https://doi.org/10.1007/978-3-031-64299-9\\_35](https://doi.org/10.1007/978-3-031-64299-9_35)

- Kasneji, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., & Kasneji, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, Article 102274. <https://doi.org/10.1016/j.lindif.2023.102274>
- Kelley, T. L. (1939). The selection of upper and lower groups for the validation of test items. *Journal of Educational Psychology*, 30(1), 17–24. <https://doi.org/10.1037/h0057123>
- Kim, E., Li, S., Khalil, S., & Shin, H. J. (2025). STAIR-AIG: Optimizing the automated item generation process through human-AI collaboration for critical thinking assessment. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (pp. 920–930)*. Association for Computational Linguistics.
- Kline, T. J. B. (2005). *Psychological testing: A practical approach to design and evaluation*. Sage Publications.
- Leslie, T., & Gierl, M. J. (2023). Using automatic item generation to create multiple-choice questions for pharmacy assessment. *American Journal of Pharmaceutical Education*, 87(10), 100081. <https://doi.org/10.1016/j.ajpe.2023.100081>
- Madaan, A., Tandon, N., Gupta, P., Hall, K., Gao, L., Wiegrefe, S., Bansal, M., Clark, P., & Yazdanbakhsh, A. (2023). *Self-refine: Iterative refinement with self-feedback*. arXiv preprint arXiv:2303.17651.
- Magis, D., Yan, D., & von Davier, A. A. (2017). *Computerized adaptive and multistage testing with R: Using packages catR and mstR*. Springer.
- Maity, S., Deroy, A., & Sarkar, S. (2024). Harnessing the power of prompt based techniques for generating school level questions using large language models. In *Proceedings of the 15th Annual Meeting of the Forum for Information Retrieval Evaluation (pp. 30–39)*.
- Memarian, B., & Doleck, T. (2024). Human-in-the-loop in artificial intelligence in education: A review and entity-relationship (ER) analysis. *Computers in Human Behavior: Artificial Humans*, 2(1), 100053. <https://doi.org/10.1016/j.chbah.2024.100053>
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement*(3rd ed.,

- pp. 13–103). American Council on Education/Macmillan.
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*(3rd ed.). McGraw–Hill.
- Pandey, S., & Karypis, G. (2019). A self-attentive model for knowledge tracing. *Proceedings of the 12th International Conference on Educational Data Mining*, 384–389.
- Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L. & Sohl–Dickstein, J. (2015). Deep knowledge tracing. *Advances in Neural Information Processing Systems*, 28, 505–513.
- Popham, W. J. (2017). *Classroom assessment: What teachers need to know*(9th ed.). Pearson.
- Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology*, 5, 27–48. <https://doi.org/10.1146/annurev.clinpsy.032408.153553>
- Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice*, 24(2), 3–13.
- Shah, C. (2024). From prompt engineering to prompt science with human in the loop. *arXiv preprint arXiv:2401.04122*.
- Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., & Yao, S. (2023). *Reflexion: Language agents with verbal reinforcement learning*. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Advances in Neural Information Processing Systems* (Vol. 36, edge-proceedings of NeurIPS 2023). Neural Information Processing Systems Foundation
- Song, Y., Du, J., & Zheng, Q. (2025). Automatic item generation for educational assessments: A systematic literature review. *Interactive Learning Environments*. Advance online publication. <https://doi.org/10.1080/10494820.2025.2482588>
- Tarrant, M., Ware, J., & Mohammed, A. M. (2009). An assessment of functioning and non-functioning distractors in multiple-choice questions: A descriptive analysis. *BMC Medical Education*, 9, 40. <https://doi.org/10.1186/1472-6920-9-40>
- Thorpe, G. L., & Favia, A. (2012). *Data analysis using item response theory methodology: An introduction to selected programs and applications* [Technical Paper

- No. 20]. University of Maine. Retrieved from [https://digitalcommons.library.umaine.edu/psy\\_facpub/20](https://digitalcommons.library.umaine.edu/psy_facpub/20)
- U.S. Department of Education. (2023). *Artificial Intelligence and the Future of Teaching and Learning: Insights and Recommendations*. Office of Educational Technology. <https://www.ed.gov/sites/ed/files/documents/ai-report/ai-report.pdf>
- von Davier, M. (2018). Automated item generation with recurrent neural networks. *Psychometrika*, 83(4), 847–857. <https://doi.org/10.1007/s11336-018-9630-3>
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35, 24824–24837.
- Wu, M. L., Tam, H. P., & Jen, T. H. (2016). Classical Test Theory. In M. L. Wu, H. P. Tam, & T. H. Jen (Eds.), *Educational Measurement for Applied Researchers* (pp. 5–27). Springer.
- Yao, S., Yu, D., Zhao, J., Yu, I., Du, Y., Tsvetkov, Y., Cao, Y., & Chen, D. (2023). *Tree of thoughts: Deliberate problem solving with large language models*. arXiv preprint arXiv:2305.10601.
- Young, R., Courtney, E., Kah, A., Wilkerson, M., & Chen, Y.-H. (2025). Content and item response theory analysis of ChatGPT-4-generated multiple-choice items. *Teaching of Psychology*, 52(3), 305–313. <https://doi.org/10.1177/00986283241311220>

논문접수 : 2025.7.3. / 수정본접수 : 2025.7.26. / 게재승인 : 2025.8.7.

ABSTRACT

## Psychometric Validity Analysis of GAI-HITL-Based Automatic Item Generation (AIG) for Reading Comprehension Items

Goun Park\*

Doctoral candidate, Korean Education, Korea National University of Education

Sookki Choi\*\*

Professor, Korean Education, Korea National University of Education

This study empirically examined whether a GAI-HITL (Generative AI-Human-in-the-Loop) based automatic item generation approach, through structured collaboration between generative artificial intelligence and teacher expertise, can ensure psychometric validity in high school Korean reading comprehension assessment. An 8-stage GAI-HITL protocol was designed through collaboration between ChatGPT-4o and classroom teachers, and 20 multiple-choice items were developed based on the 2015 Revised National Curriculum reading achievement standards. Assessment results from 53 high school seniors were analyzed using Classical Test Theory (CTT) and Item Response Theory (IRT). Results showed that the GAI-HITL-generated item set achieved internal consistency reliability (Cronbach's  $\alpha = .7897$ ) above educational assessment standards and demonstrated appropriate difficulty with a mean correct response rate of 66.8%. Using CTT as the primary analytical framework with IRT as supplementary analysis, 55% of items showed acceptable agreement with teacher-predicted difficulty levels, and 55% of items demonstrated satisfactory discrimination based on point-biserial correlation criteria. Notably, item quality differences were closely related to the patterns of teacher-AI collaboration during development. Items with high discrimination were completed through teachers' specific feedback on attractive distractor construction, while items with low discrimination resulted from cases where AI self-checking and teacher intervention were insufficient. This study

---

\* First Author, koniholic@gmail.com

\*\* Corresponding Author, agrement@knue.ac.kr

demonstrates that the GAI-HITL approach can produce reliable assessment tools applicable to educational settings even at current technological levels, showing the synergistic effects of human-AI collaboration in educational assessment for the AI era.

*Key Words: Generative Artificial Intelligence, Human-in-the-Loop, Reading Assessment, Item Analysis, Psychometrics, Classical Test Theory, Item Response Theory*