교육학연구
Korean Journal of Educational Research
제61권 제4호, 2023, pp. 299-332
http://dx.doi.org/10.30916/KERA.61.4.299

Exploring the Possibility of Science-Inquiry Competence Assessment by ChatGPT-4: Comparisons with Human Evaluators*

Park, So-Young† Lee, ByungYoon‡ Lee, You-kyung
(Sookmyung Women's University)
Ham, Eun Hye
(Kongju National University)
Lee, Sunghye
(KAIST)

< Abstract >

This study aimed to explore the potential applications of ChatGPT in the field of education by comparing and analyzing the assessment results of students' descriptive responses to a science-inquiry task by human evaluators and ChatGPT-4. A total of 155 fifth-grade students' reports were evaluated by both human evaluators and ChatGPT-4 using 22 scoring criteria. The agreement between the two evaluators was assessed using the weighted kappa coefficient and correlation coefficient. As a result, the correlation coefficient between the human evaluators' total score and ChatGPT-4's score appeared to be .74, indicating a high level of correlation. However, varying levels of agreement between the two evaluators was found across the scoring criteria (weighted kappa coefficient = .02−.58 and correlation coefficient = .14−.58). Upon analyzing the agreement levels for each scoring criterion, it was found that the evaluation agreement was at a moderate level for the criteria assessing the content of the students' experiments. However, ChatGPT-4 appeared to apply more lenient scoring standards than the human evaluators when evaluating whether students utilized additional materials or prior knowledge, compared their responses to specific criteria (e.g., their hypotheses), and whether they reflected on their experiences while completing the tasks. The agreement was significantly lower for these scoring criteria. This study was able to explore the potential for automating the evaluation of descriptive responses using ChatGPT-4 and identifying scoring criteria that could be evaluated at a level similar to that of human evaluators.

Key words: ChatGPT, automatic scoring, scientific-inquiry competence, assessment of descriptive responses, agreement analysis

† First author: Park, So-Young (100, Cheongpa-ro 47-gil, Yongsan-gu, Seoul, Korea ; syngprk@sookmyung.ac.kr)
‡ Corresponding author: Lee, ByungYoon (100, Cheongpa-ro 47-gil, Yongsan-gu, Seoul, Korea ; lee.byungyoon12@gmail.com)

ChatGPT-4의 과학적 탐구 역량 평가 가능성 탐색: 인간평가자와의 비교를 중심으로*

박소영(숙명여자대학교, 교수)†
이병윤(숙명여자대학교 교육연구소, 전임연구원)‡
함은혜(공주대학교, 부교수)
이유경(숙명여자대학교, 조교수)
이성혜(한국과학기술원 과학영재교육연구원, 연구부교수)

< 요 약 >

본 연구는 교육학에서의 ChatGPT 활용 방안을 탐색하고자, 과학적 탐구 역량 과제 보고서에 대한 학생의 서술형 응답을 인간평가자와 ChatGPT-4에게 평가하게 한 후, 그 결과를 비교·분석하였다. 인간평가자와 ChatGPT-4가 각각 초등학교 5학년 학생 155명의 탐구 보고서를 22개의 채점항목으로 평가하였다. 두 평가자가 평정한 결과에 대해, 이차가중 카파계수와 상관계수를 가지고 평가일치도를 확인하였다. 연구결과, 인간평가자와 ChatGPT-4의 평가 총점 간 상관계수는 .74로 나타나, 높은 수준의 상관관계를 보이는 것으로 나타났다. 그러나 채점항목마다 두 평가자의 일치도는 다르게 나타났다(이차가중 카파계수 = .02~.58; 상관계수 = .14~.58). 또한 일치도에 따른 각 채점항목을 분석한 결과, 학생들이 수행한 실험 내용 자체에 대한 평가에서는 인간평가자와 ChatGPT-4 간의 평가일치도가 중간 수준 이상을 보였다. 그러나, ChatGPT-4는 학생들이 추가자료나 사전지식을 활용하였는지, 어떤 특정 기준(예: 자신이 세운 가설)과 비교하며 응답하였는지, 학생들이 과제를 수행하며 느낀 점 등을 반추하며 응답하였는지를 평가할 때, ChatGPT4가 인간평가자에 비해 관대한 채점기준을 적용하는 것으로 확인되었고, 관련한 채점항목에서는 인간평가자와 ChatGPT-4 간의 평가일치도가 상당히 낮게 나타났다. 본 연구에서는 단답형 응답뿐만 아니라 서술형 응답의 평가 자동화 가능성에 대해 ChatGPT-4를 활용하여 탐색하고, 인간평가자와 유사한 수준으로 평가할 수 있는 채점항목 등에 대해 확인하였다.

주제어: ChatGPT, 평가 자동화, 과학적 탐구 역량, 서술형 응답 평가, 일치도 분석

† 제1저자: 박소영(서울시 용산구 청파로47길 100 숙명여자대학교 교육연구소, syngprk@sookmyung.ac.kr)
‡ 교신저자: 이병윤(서울시 용산구 청파로47길 100 숙명여자대학교 교육연구소, lee.byungyoon12@gmail.com)

Ⅰ. 서론

고도화된 지능정보사회로의 변화에 따라 사회에서 요구하는 인재상이 변하고 있다. 이에 따라 우리나라에서도 2015 개정 교육과정을 통해 역량 중심의 교육과정을 실행하고 있다. 이런 교육과정의 변화에서 기대하는 고등사고 능력의 향상을 위해서는 평가방식의 변경이 수반되어야 한다는 주장이 지속되었다. 즉, 선다형 평가 중심의 평가 환경하에서는 고등사고 능력을 향상시킬 수도, 측정할 수도 없다는 비판이 제기되었다(박혜영 외, 2018, 2019). 이에 정부에서는 한국교육과정평가원을 중심으로 국가 수준에서 서·논술형 평가도구를 개발하여 보급하였고, 시도교육청별로도 서·논술형 평가도구를 개발, 배부하는 한편, 교사들의 서·논술형 평가 역량을 기르기 위한 교원연수를 실시하고 있다(충남일보, 2022; 대한경제, 2022). 더 나아가 정부는 이와 같은 평가방식을 수능에 반영하여 논술형 수능의 도입 가능성까지도 검토하고 있다(교육부, 한국대학교육협의회, 2023).

그러나 현실적으로 서·논술형 평가방식을 학교 현장에서 실행하는 데 어려움을 겪고 있는 것으로 나타났다. 서·논술형 평가 선도학교 및 연구 학교, 일반 학교 교사, 장학사, 전문가, 학생, 학부모 등을 대상으로 진행한 연구에 따르면, 현장에서 제기한 문제 중 하나가 바로 채점에서의 신뢰도가 확보하기 어렵다는 것이다(박혜영 외, 2019). 채점에서의 신뢰도는 채점자 간 신뢰도와 채점자 내 신뢰도를 모두 포함하는데, 채점자 간 신뢰도는 고부담 평가에서, 채점자 내 신뢰도는 수업 내 평가에서 문제제기되는 경우가 많았다. 특히 시험 결과가 공개되는 경우 채점자 간 교차평가가 이루어졌는데, 이 때 채점자 간 신뢰도를 확보하는 문제가 대두되었다. 채점자 내 신뢰도와 관련하여, 교사 단독으로 채점하는 경우에도 학생에 대한 사전경험이나 피로도 등에 따라 채점의 비일관성이 나타나, 서·논술형 평가에서 극복해야 할 문제로 제기되었다(이용상 외, 2013; 박혜영 외, 2019).

한편, 서·논술형 평가의 채점 공정성 논란을 해소하기 위해 인공지능 기술을 활용할 수 있을 것이라는 전망이 제시되고 있다(동아일보, 2023). 일부 연구에서는 인공지능 기술이나 기계를 활용한 자동채점을 통해 채점자 간 신뢰도와 채점자 내 신뢰도의 문제를 일부 해결할 수 있다고 밝혔다(Ercikan & McCaffrey, 2022). 이와 함께 자동채점기술에 대한 관심도 높아지고 있다. 현재 국외에서 이루어지는 자동채점 알고리즘은 좀 더 다양한 방식의 글쓰기에 대해 평가할 수 있는 단계이다. 물론, 국외의 자동채점 알고리즘이 모든 형태의 글쓰기에 완벽하게 적용될 수 있다고 보기는 어렵지만 한국어 자동채점 상황에 비해 나은 편으로, 국내 일부 연구는 학생의 한국어 답안을 영어로 번역하여 연구를 진행하기도 하였다(하민수, 2016). 한국어 자동채점 시스템과 관련하여 한국교육과정평가원에서는 2012년 이후 한국어 자동채점 시스템에 대한 연구를 지속적으로 실시하고 실행에 돌입하여 주로 단답형을 중심으로 연구와 채점을 수행해왔다. 이 연구에서 국외의

자동채점 알고리즘을 한국어에 적용하는 것은 한계가 있다는 점을 발견하고, 단답형태의 한국어 서답형 문항에 대한 채점을 수행하였다. 이 연구에서 더 나아가 2022년 연구 이후로 한 문단 정도의 논술형 자동채점 연구를 시도하고 있다(박종임, 2022: 43).

그동안 수행된 연구의 흐름으로 보았을 때, ChatGPT를 평가에 활용하고자 하는 시도는 자동채점 수행의 일환이라고 할 수 있다. 본 연구에서 시도하고자 하는 자동채점은 그동안 연구되었던 단답형이나 한 문단 정도의 일반적인 논술형 자동채점과는 다르게, 과학이라는 특정 분야에서 자주 활용되고 있는 과학 탐구 보고서 채점이라는 점에서 차별성이 있다. 과학 분야에서 과학 탐구 보고서와 실험 보고서 등은 학교 현장에서 흔히 쓰이는 수행평가 방식이지만 이에 대한 채점기준을 명확하게 하고 이를 자동화하는 시도에는 아직 이르지 못하였다. 본 연구에서는 학생들의 과학 탐구 보고서를 과학적 탐구 역량이라는 역량 기준에 맞추어 자동채점 할 수 있을 것인가에 대해 탐색하였다. 다만, 이 연구에서는 ChatGPT를 활용하여 학생 수행에 대한 개별적인 판단행위를 자동화하는 데 초점을 둔 반면, 대규모 답안에 대한 반복적인 채점 과정을 자동화한 것은 아니기 때문에 부분적 자동채점의 의미로 '기계채점'이라는 용어를 사용하였다.

ChatGPT를 활용한 과학 탐구 보고서 자동채점 가능성을 탐색하기 위해 인간 평가와 ChatGPT 평가를 비교하는 방식으로 연구를 진행하였다. 본 연구에서는 다음의 연구 문제를 설정하였다. 첫째, 한국의 비구조화된 과학 탐구 수행 과제에서 인간채점[1]과 기계(ChatGPT) 채점은 어느 정도 일치 수준을 보이는가? 둘째, 인간채점과 기계(ChatGPT) 채점 간 일치도가 높은 항목과 낮은 항목은 어떤 특징이 있는가? 본 연구결과를 통해 향후 과학적 탐구 보고서에 나타난 과학적 탐구 역량이 기계채점으로 가능할 것인가에 대해 논의하고, 기계채점의 현장 적용 가능성을 높이는 방안을 찾을 수 있을 것이라고 기대하였다.

Ⅱ. 이론적 배경

1. 서·논술형 평가의 자동채점과 ChatGPT의 등장

가. 서·논술형 평가의 자동채점과 발전 현황

선행연구에서 자동채점(automated scoring)은 인간의 채점과정 중 일부 혹은 전체를 자동화하는 광범위한 실천영역을 포함하며(Foltz et al., 2020, 기계채점(machine scoring)과 엄밀하게 구분하기보다 혼용하고 있다. 채점하고자 하는 문항의 유형(선다형 혹은 구성형)이나 복잡성(단답형

1) 여기에서 쓰인 인간채점의 '인간'이란 교육학 관련 분야에 종사하고 있는 전문가를 의미한다.

혹은 논술형)에 따라 채점 자동화의 의미가 달라질 수 있지만, 여기에서는 특히 서·논술형 문항에서의 자동채점 활용 관련 선행연구에 초점을 맞추었으며, 대부분 '자동채점'이라는 용어를 사용하였기 때문에 그대로 활용하여 정리하였다.

서술형 및 논술형 문항 채점을 자동화하기 위한 연구와 시도들은 지난 20여 년 동안 영미권에서 활발하게 진행되었다. 다수의 연구들이 학생들의 영어 작문 능력을 평가하기 위하여 작문 규칙(문법), 어휘(정교성 및 사용의 적절성), 담화 구조(조직 및 전개) 등의 평가 요소와 에세이에서 관찰되는 특질들의 관계를 규명하고, 그에 기반하여 점수를 예측하는 모형을 다양하게 탐색해왔다(Chodorow & Burstein, 2004; Lee et al., 2010; Rudner et al., 2006). 그리고 이러한 자동채점 모형들은 학생들의 글의 수준을 평가하는 데 인간채점자와 유사한 정도의 수행을 보일 수 있다는 것을 보여주었다(Cahill & Evanini, 2020).

국내에서는 한국교육과정평가원에서 수행된 진경애 등(2006)의 연구가 한국어 서답형 문항의 자동채점 도입의 가능성을 탐색한 최초의 연구이다. 이 연구에서는 기존 영미권에서 개발된 자동채점 시스템을 검토하고, 실제 중학교 3학년 수준의 영작문 자동채점 프로그램을 자체 개발하였으며, 그 결과 인간채점 결과와의 상관이 최소 .65에서 최대 .88까지 분포하는 것으로 나타났다. 한편, 사회과 학업성취도 평가의 서답형 문항 응답을 분석하고 향후 한국어기반 자동채점을 위해 필요한 과제나 단계를 구체화하기도 하였다. 그러나 당시 한국어 자연어처리 기술에 제한이 많아 자동채점 프로그램의 개발이 완료되지는 못하였다. 그 이후, 노은희 등(2012, 2013, 2014, 2015)의 연구에서 한국어로 된 서답형 문항을 자동으로 채점하는 프로그램을 단어와 구 수준에서 문장 수준으로 순차적으로 개발하였으며, 한국어 응답에 대한 자동채점 성능이 실질적으로 검토되었다. 초기 단어와 구 수준의 응답에 대한 자동채점과 인간채점 결과와의 상관은 대부분 .85 이상으로 높았으나, 문항에 따라 편차가 상당히 크고, 특히 응답의 자유도가 높은 문항의 채점 정확도가 낮았다(노은희 외, 2012). 이후 문장 단위 서답형 문항에 대한 자동채점에서는 채점의 정확성이 크게 향상되었으며(노은희 외, 2015; 송미영 외, 2016), 인간채점과 비교하여 약 60% 이상의 비용 절감 효과가 있음이 확인되었다(이상하 외, 2015).

최근에는 자연어처리 기술과 인공심층신경망 분석 모형에 대한 접근성이 확대되면서, 한국어 서술형 및 논술형 자동채점 연구도 증가하는 추세이다. 자동채점 연구가 수행되는 맥락을 고려할 때, 크게 교과기반 서술형 문항 평가(노은희 외, 2012; 박세진, 하민수, 2020)와 한국어 작문 평가(김승주, 2019; 박강윤, 이용상, 2022; 조희련 외, 2021)의 맥락으로 구분할 수 있다. 앞서 살펴본 초기 연구들과 박종임 등(2022)의 연구에서는 이 둘을 구분하지 않고 통합적으로 자동채점 적용 가능성을 검토해왔으나, 교과기반 서술형 평가와 한국어 작문 평가에서는 피험자 응답의 특성 즉, 응답의 길이와 내용이 뚜렷하게 구분되고 채점기준도 다르기 때문에 이 두 평가 맥락을 구분하는 것이 유용하다.

먼저, 교과기반 서술형 문항 평가의 맥락에서는 박세진과 하민수(2020)가 초등학생의 과학 서술

형 응답을 평가하기 위해 자동채점 시스템을 개발하였다. 구체적으로 4개 서술형 문항(온도와 열, 별과 행성, 태양과 광합성, 곰팡이의 역할)에 대한 학생 462명의 응답 자료에 순환신경망 모형을 적용하여 분류를 자동화하였다. 그 결과, 인간채점자와의 평정의 일관성(카파계수 기준)이 최소 .89에서 최대 .99까지로 상당히 높았다. 또한, 자동채점 시스템을 경험한 학생들은 자신의 응답에 대한 채점 결과가 정확하고 적시에 제공되어 도움이 되었다고 하였다. 노은희 등(2015)의 연구에서는 중학교 국어, 사회, 과학 및 고등학교 국어에서의 교과기반 서술형 문항에 자동채점을 적용하였으며, 인간채점자와의 채점의 일관성(카파계수 기준)이 최소 .76부터 최대 .99까지 높게 나타났다. 이와 같은 교과기반 서술형 평가에서 자동채점의 대상은 주로 1~2개 문장이고, 내용적으로 모범답안과 얼마나 유사한가가 채점기준이기 때문에, 피험자 응답에서의 채점자질 추출이 상대적으로 용이하였다.

반면, 한국어 작문 평가의 맥락에서는 자동채점을 위하여 채점자질(feature) 즉, 점수 산출의 근거가 되는 텍스트 표층의 언어학적 특성들을 규명하는 연구가 다수 이루어졌다(이현준, 박영민, 2019; 김승주 2019). 이 경우, 자동채점의 대상은 에세이 혹은 논술형 답안에 해당하는데, 문단을 기본단위로 하기 때문에 교과기반 서술형 문항에 비해 길이가 길고, 채점기준으로서의 모범답안을 구성하기도 어렵기 때문이다. 한국어 작문평가를 위한 자동채점모형을 구축하고 성능을 검토한 연구들은, 인간채점 결과와의 일관성을 평균 50% 혹은 그 이하로 보고하고 있다. 구체적으로 조희련 등(2021)은 유학생 한국어 쓰기 답안지 304편을 대상으로 여러 언어모델과 기계학습 분류기를 활용하여 점수 구간을 예측하였는데, 그 결과, 인간채점 결과와 비교한 분류 정확도가 평균 53% 내외였다. 또한, 박강윤과 이용상(2022)의 연구에서 심층신경망을 활용하여 대학생 530명의 에세이(답안 평균 길이 436자)에 대한 자동채점을 실시한 결과, 정확도가 21~34%로 상당히 낮았다. 적용된 모형과 채점기준(예: 내용, 표현, 구조, 문법)에 따라 차이가 있었으나, 향후 한국어 에세이나 논술형 답안에 대한 자동채점의 정확도를 개선하기 위한 방안이 다양하게 탐색될 필요가 있음을 시사하였다.

나. ChatGPT의 등장과 서술형 평가 자동채점의 가능성

본 연구에서는 서술형 평가 자동화의 전개 과정에서 새로운 시도의 하나로 ChatGPT를 활용하여 서술형 응답을 평가할 수 있는지 확인하였다. 전술하였듯이, 본 연구에서는 부분적 자동채점의 의미로 기계채점을 명명하였다. 주로 컴퓨터를 이용해 정답 템플릿에 채점기준을 입력하고 이를 바탕으로 채점하는 방법과 다양한 기계학습 기법(machine learning method)을 활용하여 채점하는 방법 등이 포함된다(조희련 외, 2021; 하민수 외, 2019). ChatGPT는 미국의 OpenAI에서 개발된 GPT(Generative Pretrained Transformer) 모델을 바탕으로 한 자연어처리 인공지능 서비스이다(성욱준, 2023). ChatGPT는 학습을 기반으로 하여 막대한 양의 데이터를 통해 인간과 같은 대화

를 만들어내는 딥러닝 기반 인공지능이다(공정식, 2023). 인공지능 기술도 하나의 기계에 속하는데, ChatGPT를 활용한 자동채점은 인공지능을 기반으로 동작하며, 자연어처리 알고리즘을 활용해 텍스트를 이해하고 분석한다. 웹상의 방대한 양의 텍스트를 사전학습하여 문장 구조, 문맥, 개념 등을 학습하고, 이를 통해 입력된 답변이 채점기준에 부합하는지를 판단한다. 그리고 이 결과를 사용자가 요청한 형태(점수, 등급, 피드백 등)로 제공한다(Lo, 2023). 특히 ChatGPT는 텍스트 분류에서 탁월한 성능을 보여(Chan, 2022) 이미 해외에서는 학생들의 과제 평가를 위해 ChatGPT를 활용한 연구들이 소수이지만 등장하고 있다(Bhat et al., 2022; Moore et al., 2022). 또한 이러한 연구들은 인간평가자와의 비교를 통해 그 효과를 검증하고 있다.

Bhat 등(2022)은 미국 대학생들이 생성한 단답형 질문(short-answer questions)이 교육적으로 유용한지 판단하기 위해 해당 질문들을 전문가와 GPT-3에게 평가하도록 하는 연구를 수행하였다. 학생들이 만든 데이터사이언스 관련 질문들에 대해, 5년 이상 경력의 해당 분야 전문가 2명과 GPT-3의 평가를 비교하였다. 연구 결과, 총 203개의 응답 중 135개(약 66.5%)에서 인간평가자와 GPT-3 간의 평가가 일치하는 것으로 나타났다. 일치하지 않은 68개의 질문에서는 대부분 GPT-3가 인간평가자보다 더 관대하게 채점한 것으로 확인되었다. 연구자들은 학생들이 질문을 생성할 때 구체적인 개념을 사용하였는지에 따라 인간평가자와 GPT-3 간의 평가 차이가 나타났다고 밝혔다. 인간평가자는 학생들이 작성한 질문에 구체화된 개념이 포함되어 있을 경우 그 질문이 유용하다고 판단하였다. 그러나, GPT-3는 데이터사이언스 분야에서 사용되는 전문용어나 기준을 완전히 파악하지 못해 해당 개념이 구체적인지 아닌지를 정확하게 판단할 수 없었고, 이에 따라 학생들이 생성한 질문의 교육적 유용성에 대한 평가가 상대적으로 유연했을 것으로 분석되었다.

더 나아가 Moore 등(2022)은 학생들의 단답형 질문에 대한 질적 평가뿐만 아니라, 각 질문의 인지적인 수준을 평가하기 위해 블룸의 분류 체계에 따라 질문을 분류할 수 있는지를 인간평가자와 GPT-3에게 평가하도록 하였다. 먼저, 학생들이 생성한 143개의 질문 중 57개(약 40%)에서 두 평가자의 평가가 일치하는 것으로 나타났다. Bhat 등(2022)의 연구와 비슷한 결과로, 일치하지 않은 86개 질문 대부분에서 GPT-3가 인간평가자보다 관대한 채점을 한 것으로 확인되었다. 또한, 학생들의 질문을 블룸의 분류 체계에 따라 분류한 결과, 총 6개의 분류 체계 중 2개(Evaluate, Create)에서는 일치율이 0%였고, 나머지 4개(Remember, Understand, Apply, Analyze)는 4%에서 48% 범위의 일치율을 보였다. 이러한 결과에 대해 연구자들은 GPT-3가 학생들의 결과물을 과대평가하고 블룸의 분류 체계에 따른 구분을 잘하지 못한다고 해석하였다. 그럼에도 불구하고 연구자들은 GPT-3가 인간평가자의 평가 전, 학생들의 결과물에 대해 일차적으로 합격 여부를 판단하는 데에는 충분한 역할을 수행할 수 있다고 주장하였다.

이로써 ChatGPT가 학생 응답에 대해 평가할 수 있는 가능성을 확인할 수 있었다. 위의 두 선행연구는 학생들의 단답형 질문에 대한 GPT-3 모델의 평가를 인간평가자와 비교 분석하였다. ChatGPT를 활용하여 자동채점을 구현한 선행연구와의 차별점은, 본 연구에서는 최근 교육현장에

서 활발하게 논의되고 있는 서술형 답안에 대한 자동채점의 가능성을 분석하였다는 것이다. 한국어 자동채점 시스템을 활용한 기존의 연구는 주로 단답형 문항에 대한 채점을 중심으로 이루어졌고, 최근에는 논술형 채점에 대한 시도가 이루어지고 있지만 한 문단 정도의 길이로 제한되어 이루어져 왔다(박종임, 2022). 그러나 ChatGPT-4라는 최신 버전이 공개되면서 서술형 답안의 자동채점 활용 가능성을 보다 체계적으로 탐색해볼 필요가 있다. 서술형 답안에 대한 자동채점의 가능성을 살펴보기 위해 본 연구에서는 특히 초등학생들이 과학 과목에서 많이 수행하는 과제인 과학탐구 보고서 답안을 수집하였다. 초등학교 과학 과목에서는 특히 지식과 개념을 확인하는 단답형 과제보다는 학생 스스로 이론과 가설을 적용하여 실험을 수행해나가는 서술형 탐구과제가 상대적으로 자주 이루어지고 있기 때문에 이러한 특성의 과목과 과제를 기반으로 자동채점의 타당성과 활용가능성을 검토해보는 것은 중요한 교육적 함의점을 제공해줄 수 있을 것이다.

2. 과학적 탐구 역량 평가

과학적 탐구 역량이 무엇인지를 알기 위해서는 과학적 탐구활동을 위해 필요한 요소가 무엇인지를 확인해볼 필요가 있다. 문헌들에 따르면 과학적 탐구활동은 자연 현상에서 파생된 인과적 질문을 과학적 ‘지식’과 ‘방법’을 통해 논리적으로 해결해나가는 과정이다(박인숙, 강순희, 2012). 따라서 본 연구에서는 과학적 지식과 방법, 이에 덧붙여 탐구과제 수행에 대한 태도까지 포함하여 과학적 탐구 역량을 평가하고자 하였다. 첫째, 과학적 지식은 자연 현상 및 사실에 대한 관찰, 실험, 분석에 의해 검증된 체계적 정보이다. 과학적 지식은 가설을 연역적으로 도출할 때뿐만 아니라, 가설 검증 후 결과를 해석하고 종합하는 귀납적인 문제 해결 과정에서도 중요한 역할을 한다(이정은, 정은영, 2013). 둘째, 과학적 방법에 해당하는 논리적 사고 과정은 과학적 지식이 적절하게 기능하도록 하는 역량의 필수적인 요인이다(박인숙, 강순희, 2012). 논리분석적 사고는 가설 생성 단계에서 자연 현상에 대한 원인을 예측하는 과정, 가설 검증 방법을 계획하고 수행하는 과정, 결과를 해석하고 결론을 도출해내는 과정 등 과학적 탐구활동의 전반적인 단계에 모두 적용된다(이정은, 정은영, 2013; 손정우, 2006). 셋째, 과학적 태도는 학습자 중심의 능동적인 활동이 요구되는 과학적 탐구활동 수행에서 빠질 수 없는 요소이다. 호기심에 의해 스스로 탐구활동에 필요한 자료를 탐색하고 수행 전, 중, 후에 걸쳐 자신의 탐구 수행에 대해 성찰하는 태도는 최근 과학적 탐구 역량을 구성하는 중요한 요인으로 평가되었다(함은혜 외, 2022). 추가적으로, 세 가지 역량 요소 외에 본 연구에서는 의사소통 능력을 평가하였다. 본 연구에서 평가한 과학적 탐구 역량은 학습자가 과제를 수행하면서 자유롭게 작성한 탐구 보고서를 기반으로 한다는 점에서 탐구 수행 과정 및 결과를 보고하는 의사소통 능력이 평가에 중요하게 작용할 수 있기 때문이다. 실제로

2015 개정 교육과정을 비롯하여 이전 연구에서는 과학적 의사소통 능력을 과학 탐구능력의 주요 요소 중 하나로 제시한 바 있다(백종호 외, 2020).

과학적 탐구과제의 수행 수준을 보다 타당하게 평가하기 위해서는 궁극적으로 이러한 과학적 탐구 역량의 요소들이 잘 반영된 평가기준을 마련할 필요가 있다. 학습자의 과학 탐구과제 수행을 평가하는 도구들은 꾸준히 개발되어 왔다. 과학 탐구과제의 개별적 특성에 따라 평가기준이 조금씩 다르게 제시되어 왔는데, 탐구문제를 선정하고 가설을 검증하는 일련의 실험 과정을 포함하는 통합 탐구 과정 평가도구들은 공통적으로 가설 생성, 실험 설계 및 수행, 자료 분석 및 해석, 그리고 결론 도출 및 평가의 요소들을 포함하여 왔다(김유향, 김영수, 2012; 최경애 외, 2017; 황진석 외, 2010). 구체적으로 가설 생성이란 독립변인과 종속변인을 판별해내고 이를 바탕으로 검증 가능한 가설을 도출해내는 것을 의미한다. 실험 설계 및 수행이란 가설에 따라 독립변인, 종속변인, 통제변인을 포함하여 실험을 디자인하고 실험 조건을 적절히 조정해 나가면서 실제로 실험을 수행하는 것을 의미한다. 자료 분석 및 해석이란 실험결과 자료를 보고 가설과 대조하면서 결과를 해석하는 과정으로 필요 시 그래프와 같은 시각적 자료로 변환하기도 한다. 마지막으로 결과 도출 및 평가는 실험결과를 종합하여 어떠한 원리를 찾아내는 등의 결론을 도출하거나 실험 과정에 대해 평가하고 성찰하는 과정을 포함한다.

본 연구에서는 이전에 개발된 과학적 탐구과제 평가도구들을 바탕으로 가설 생성, 실험 설계 및 수행, 자료 분석 및 해석, 결론 도출 및 평가라는 네 단계의 큰 과정 요소를 포함하고, 과학 탐구과정에 전반적으로 적용될 수 있는 요소를 하나 더 추가하였다. 구체적으로, 가설 생성 요소에서는 종속변인 규명 및 독립변인 탐색 그리고 가설 정당화를, 실험 설계 및 수행 요소에서는 독립변인 규명, 실험 수행, 혼재변인 통제를, 자료 분석 및 해석 요소에서는 가설 검증과 자료 해석을, 결론 도출 및 평가 요소에서는 평가 및 학습과정 성찰을, 마지막으로 전반에 대한 평가 요소에서는 계량적 접근이 관찰됐는지, 문서 기반 의사소통 수준이 어느 정도인지를 세부 요소로 포함하였다(<표 1> 참고).

<표 1> 과학 탐구과제의 과정 요소별 채점항목과 역량 구분

과학 탐구 과정 요소		번호	채점항목	역량 요소
가설 생성	종속변인 규명	C1	해당 차시에서 학습한 과학적 개념을 활용하여 치즈의 특징을 분석하는가?	과학적 지식
		C2	치즈의 특징을 분석하기 위하여 기타 사전지식(해당 차시 학습 이외)을 활용하는가?	과학적 지식
	독립변인 탐색	C3	치즈의 특징을 변화시킬 수 있을 것으로 제시되는 조건을 2개 이상 명료하게 제시하는가?	논리분석적 사고
		C4	각 조건의 변화가 치즈의 어떤 특징을 어떻게 변화시킬지를 예상(예측)하여 진술하는가?	논리분석적 사고
		C5	기타 자료를 참고하거나 스스로의 사고를 통해 고려할 만한 조건을 추가적으로 탐색하였는가?	탐구적 태도
	가설 정당화	C6	각 조건의 변화가 왜 혹은 어떻게 치즈의 특징을 변화시키게 될지를 사전에 학습한 과학적 개념이나 원리를 활용하여 설명하는가?	논리분석적 사고
		C7	각 조건의 변화가 왜 혹은 어떻게 치즈의 특징을 변화시키게 될지를 기타 과학적 개념이나 원리를 활용하여 설명하는가?	탐구적 태도
실험 설계 및 수행	독립변인 규명	C8	제시된(선정된) 조건이 앞에서(사전학습자료 혹은 개별 추가학습자료) 학습한 과학적 개념이나 원리와 관련이 있는가?	과학적 지식
		C9	실험을 위해 변화시킨 조건이 명확한가?	논리분석적 사고
	실험 수행	C10	선택한 조건을 변화시킬 수 있는 방법으로 치즈 제작 과정의 조건을 변화시켰는가?	논리분석적 사고
	혼재변인 통제	C11	자신이 선택한 조건 이외의 조건을 고려하여 통제하는가?	논리분석적 사고
자료 분석 및 해석	가설 검증	C12	가설을 바탕으로 결과를 분석하는가?: 조건 변화에 따른 결과물의 특성을 비교하여 기술하는가?	논리분석적 사고
		C13	가설을 바탕으로 결과를 분석하는가?: 분석 결과를 본인의 가설(예상했던 결과)과 비교하여 기술하는가?	논리분석적 사고
	자료 해석	C14	과학적 원리와 개념에 대한 사전지식을 활용하여 결과를 해석하는가?	과학적 지식
		C15	분석 결과를 이해(해석)하기 위해 추가자료를 탐색하는가?	탐구적 태도
결론 도출 및 평가	평가	C16	분석 결과에 근거하여 특정 조건의 적절성이나 유용성을 평가하는가?	탐구적 태도
	학습과정 성찰	C17	실험을 통해 자신이 무엇을 배웠는지 기술하는가?	탐구적 태도
		C18	본인이 수행한 실험의 강점이나 보완할 점 등을 기술하거나 향후 탐구과제를 제시하는가?	탐구적 태도
전반	계량적 접근	C19	조건을 변화시키는 과정 등에서 계량적 접근이 관찰되는가?	의사소통
	문서기반 의사소통	C20	치즈의 특징(관찰 결과)에 대한 언어적 기술이 풍부한가?	의사소통
		C21	참고한 자료의 출처를 제시하는가?	의사소통
		C22	진술이 명료하고, 문단을 구조화하는 등 독자의 가독성을 고려하는가?	의사소통

출처: 함은혜 외(2022). 초등학생 과학 탐구과제 수행 특성 분석 및 채점기준 개발. 한국과학교육학회지, 42(2), 244.

Ⅲ. 연구 방법

1. 연구참여자 및 연구 대상 과학 탐구활동 과제

본 연구에서 K대학 온라인 교육프로그램에 참여한 초등학교 5학년 155명(남학생 85명, 여학생 70명)이 제출한 탐구 보고서를 채점에 활용하였다. K대학 온라인 교육프로그램은 수학과 과학에 흥미와 재능이 있는 초·중학생들에게 과학적 탐구 기회를 제공하고 수·과학적 사고, 탐구 및 문제 해결 역량을 개발할 수 있도록 제공된다. 학생들은 별도의 선발 절차 없이 본인이 원하는 프로그램을 고르고, 해당 교육에 참여할 수 있다.

온라인 교육 콘텐츠는 '문제탐색-개념학습-문제해결' 단계로 구성된 e-book 형태로 제공된다. 문제탐색에서는 해당 주제에 대해 흥미와 학습 동기를 가질 수 있도록 실생활과 관련된 문제를 제시한다. 이를 통해 해당 주제에서 학습할 내용이 실제 생활에 어떻게 연결되는지 이해하도록 하며, 해당 주제와 관련하여 해결해야 할 문제를 제시하여 학생이 스스로 문제해결 계획을 생각해 볼 수 있도록 한다. 개념학습에서 학습자는 국가 교육과정에 포함된 수학, 과학 개념을 바탕으로 문제를 해결하기 위해 필요한 주요 개념들을 스스로 학습한다. 학생들은 학교에서 이미 학습한 개념을 스스로 정리하고 필요한 자료를 찾아보면서 학습할 수 있다. 문제해결 단계에서는 개념학습 단계에서 습득한 개념 및 원리를 활용하여 확장하고 심화시키면서 새로운 현상에 대해 고민하거나 새로운 아이디어를 제안하여 문제를 해결하는 과제를 수행한다. 특히 탐구과제는 "가설 생성, 실험 설계, 자료 분석 및 해석, 결과 도출 및 개선방안 제안" 등 일반적인 탐구 수행 절차에 따라 탐구를 수행하고 기록한 탐구 보고서를 제출하도록 한다(함은혜 외, 2022). 탐구과제는 3개의 하위 미션으로 제시된다. 온라인 교육프로그램에 참여하기 위해서 학생들에게는 회원가입이 요구되었으며, 회원가입 시 학생 본인과 법정 대리인은 개인 정보 수집 및 보호, 그리고 연구 목적으로 데이터를 활용하는 것에 대해 동의하였다.

2. 과학적 탐구활동 과제와 채점

본 연구에서 활용한 탐구과제 주제는 '치즈는 왜 맛이 다를까?'로 2015 교육과정 초등학교 고학년 과학 과목에서 다루는 균류와 세균의 특성, 기술·가정 과목의 식생활 및 식품과 영양 등의 개념을 활용하여 치즈의 생성과 맛, 특징에 어떻게 적용되는지 탐구해 볼 수 있는 기회를 제공하였다. 개념학습에서는 응고, 발효, 효소 등의 개념뿐만 아니라 치즈의 역사, 다양한 종류, 원료와 분

류, 치즈에 들어있는 영양소, 그리고 치즈로 만드는 요리 등과 관련된 내용이 제시되었다. 탐구과제에서는 개념학습에서 학습한 내용을 바탕으로 치즈의 맛과 특징에 영향을 미치는 변인을 조작하여 실험을 설계하고, 치즈를 직접 만들어보면서 조건에 따라 결과가 어떻게 변하는지, 치즈에서 발견되는 현상과 치즈의 특성은 어떤 것인지를 분석하고 정리하여 보고서로 제출하도록 하였다. 본 탐구과제를 통해 학생들은 관찰, 측정, 예상, 추리와 같은 기초 탐구과정과 가설 수립, (혼재)변인 통제, 자료 분석, 결론 도출 등의 통합 탐구과정을 수행하는 것이 기대되었다. 탐구과제는 <표 2>와 같이 구성되었으며, 채점에 활용된 학생의 탐구 보고서 실제 양식은 함은혜 등(2022)의 부록을 참고하기를 바란다.

<표 2> 과학적 탐구과제 내용

과제 단계	내용
과제	지금까지 우리는 치즈가 무엇인지, 치즈는 어떻게 만들어지는지, 다양한 치즈의 종류와 함유된 영양소가 무엇인지 배웠습니다. 앞서 배운 내용을 바탕으로 본격적으로 치즈를 만들어보면서 미션을 해결해봅시다. 이번 미션에서는 기본 리코타 치즈의 레시피에서 조작 변인을 설정하여 분석하고, 여러분만의 특별한 치즈를 만들어 봅시다
미션 1. 치즈의 특징 관찰	앞서 나온 레시피를 따라 '기본 리코타 치즈'를 만들어 보고 특징을 적어봅시다. 여러분은 기본 리코타 치즈 레시피를 충실히 따랐나요? 이 기본 레시피는 어떻게 만들어진 것일까요? 끓이는 시간, 불의 세기, 우유/생크림의 비율, 레몬즙의 양, 물기를 짜주는 정도 등의 조건들은 치즈의 특성에 어떠한 영향을 미칠까요?
미션 2. 치즈의 특성을 변화시키는 조건 탐색	리코타 치즈 레시피에서 치즈의 특성을 다르게 변화시킬 수 있는 조건을 적어도 2개 이상 생각해 봅시다. 변화시킨 조건에 따라 치즈에 어떤 결과가 기대되는지 역시 설명해 주세요. (예: 끓이는 시간)
미션 3. 실험 수행 과정 및 결과 보고	미션 2에서 제시했던 조건 중 하나를 선정하여, 그 조건에 변화를 주며(예: 물을 덜 넣는다/더 넣는다) 치즈를 직접 만들어 보고 결과를 분석해 봅시다.

주. 본 연구에서는 함은혜 외(2022)와 동일한 과학적 탐구 역량 평가도구와 과학 탐구과제를 사용했기 때문에 해당 논문에서 제시한 과학 탐구과제 내용 표를 제시함.

본 연구는 초등학생의 과학적 탐구 역량을 평가하기 위해 과학적 탐구활동 과제의 세 가지 하위 미션을 각기 다른 채점항목을 통해 평가하였다. 세 미션의 개별적 특성이 서로 다르기 때문에 각 미션에서 평가되는 역량 요소 또한 차이가 있다. 예를 들어 미션 1은 치즈의 특징을 관찰하는 것으로, 기존 레시피를 바탕으로 치즈를 직접 만든 후 그 특징을 기록하는 것이다. 미션 2에서는 치즈의 특성을 변화시킬 수 있는 조건을 탐색하게 하였으며, 학생들은 2개 이상의 조건을 제시하고 해당 조건들이 치즈에 어떤 변화를 가져올지 예상한 것을 작성하였다. 미션 3에서는 미션 2에서 제시한 조건 중 하나를 선정해서 실제로 실험을 진행한 후 결과를 분석하게 하였다.

본 연구에서 사용한 채점항목은 함은혜 등(2022)에서 개발한 채점항목을 그대로 적용하였다. 채

점항목은 미션 1에 2개, 미션 2에 5개, 그리고 미션 3에 11개가 있다. 과학적 탐구 역량 요소 중 의사소통 능력을 추가로 측정하기 위해 세 개의 미션 전반에 대한 평가 요소를 도입하였다. 이를 통해 학생들이 미션 1~3에서 작성한 응답 전반에 걸친 문서기반 의사소통 수준 4개 항목을 평가하였다. 결과적으로 본 연구에서는 총 22개의 채점항목을 통해 학생들의 서술형 응답을 평가하였으며, 미션별로 사용한 항목 번호는 <표 1>의 항목 번호를 사용하였다(<표 1>, <표 3> 참고). 이를 인간평가자와 ChatGPT가 각각 수행하였다.

<표 3> 과학적 탐구 역량의 하위 미션별 채점항목과 채점 내용

측정된 하위 미션	채점항목 번호	주요 채점 내용
미션 1. 치즈의 특징 관찰	C1~C2, 총 2개	치즈의 특징을 관찰하고 이를 적절히 분석하였는지에 대한 채점
미션 2. 치즈의 특성을 변화시키는 조건 탐색	C3~C7, 총 5개	치즈의 특징을 변화시킬 수 있는 조건을 제시하고, 그 조건에 따른 변화에 대해 적절히 기술하였는지에 대한 채점
미션 3. 실험 수행 과정 및 결과 보고	C8~C18, 총 11개	미션 2에서 제시한 조건을 토대로 실험을 적절히 설계하고 수행했는지, 그리고 실험 결과에 대해 적절히 기술하였는지에 대한 채점
문서기반 의사소통	C19~C22, 총 4개	언어적 기술, 가독성, 출처 등 응답의 언어적 표현에 대한 내용

주. 채점항목 번호는 <표 1>의 번호를 사용함.

3. 인간평가 과정

인간평가자는 총 11인으로 교육학 분야 연구 경력 10년 이상의 박사학위 소지자 3명과 교육학, 과학교육, 아동학 분야 박사수료 및 박사과정생 8명으로 구성되었다. 채점자별 최종학력과 전공분야, 초·중등 교육 및 연구 경험은 <표 4>에 제시되었다. 초·중등 교육 및 연구 경험은 초·중등 학생 대상으로 교육을 실시하는 영재교육원 등 학교 교육 이외에 교육을 수행한 경력이나 초·중등 학생 또는 교사 대상의 연구 수행 경험이 있는 경우 '있음'으로 표기하였으며, 초등학생 평가 경험은 학교 및 학교 외 교육기관에서 초등학생 대상 채점이나 평가활동을 수행한 경험이 있는 경우 '있음'으로 표기하였다.

<표 4> 채점자별 세부 정보

채점자 구분	최종학력	전공	교사자격증 소지 여부	초·중등 교육 및 연구 경험	초등학생 평가 경험
A	박사	교육학	중등	있음	없음
B	박사	교육학	중등	있음	있음
C	박사	교육학	중등	있음	없음
D	박사수료	교육학	없음	있음	있음
E	박사수료	아동학	없음	있음	있음
F	박사수료	과학교육	중등	있음	있음
G	박사수료	교육학	유아	있음	없음
H	박사수료	교육학	없음	없음	없음
I	박사과정	교육학	중등	있음	없음
J	박사과정	교육학	없음	있음	없음
K	박사과정	교육학	없음	있음	없음

보고서별로 2인의 교차채점자가 배정되었으며, 각 채점자별로 <표 1>에 제시된 채점항목(체크리스트)에 따라 독립적으로 평가를 실시하였다. 일차적으로 2인의 인간평가자는 이분 변수(0 혹은 1)로 채점하였으나, 채점자 간 평가의 불일치가 높을 경우에 대해서 2차 채점을 진행하였다. 최종 평가 점수는 교차채점자 2인의 합산 점수로 산출하였고, 채점항목에 따라 0, 1, 2점으로 최종 점수가 산출되었다. 이를 본 연구에서는 각각 1, 2, 3점으로 재코딩하여 ChatGPT-4가 평가한 점수와 비교하였다. 채점 결과 자료는 채점자, 채점항목, 피험자를 국면으로 하는 다국면 Rasch 모형에 적합하였고, 채점자 적합도와 채점항목 적합도도 모두 양호하였다. 구체적으로 채점자별 내적적합도 지수는 최소 .78에서 최대 1.21까지 분포하였으며, 채점항목별 내적적합도 지수는 최소 .79에서 최대 1.31까지 분포하였다. 또한 두 채점자 간의 판단의 일관성(spearman rank correlation)은 총점을 기준으로 최소 .63부터 최대 .95까지 분포하였다. 그 외 신뢰도 및 타당도 등의 세부적인 정보는 함은혜 등(2022)의 연구에 제시되었다.

4. ChatGPT-4 평가 과정

본 연구에서는 OpenAI의 ChatGPT-4를 활용하여 학생들의 응답을 평가하였다. 평가를 위해 과학적 탐구과제의 미션별로 학생들의 응답과 이와 관련한 평가기준(각 채점항목)을 대화창에 입력하였다. 대화창에 입력된 자세한 내용은 다음과 같다. 우선 ChatGPT-4가 평가 시 학생들이 적은 응답이 무슨 내용인지 알 수 있도록 실제 활동지에 제시되었던 각 미션별 과제에 대해 간략하게

요약하였다. 그 다음, 평가기준을 제시하며 학생들의 응답이 평가기준을 만족할 경우 3점, 보통일 경우 2점, 만족하지 않을 경우 1점으로 평가하도록 지시하였다. 이어서 학생들의 응답을 대화창에 함께 입력하였다([부록 1] 참고). ChatGPT-4의 대화창은 한 번에 입력할 수 있는 텍스트 양에 제한이 있어, 많은 양의 텍스트를 입력할 경우 처리할 수 없다는 경고 메시지가 출력되었다. 이를 고려하여 한 번에 입력하는 텍스트 양을 학생 5명의 응답으로 제한하였다. 단, 전체 학생들이 세 개의 미션에서 응답한 내용을 종합적으로 평가해야 하는 경우(예: 진술의 명료성 및 문단 구조화 등 독자의 가독성 고려 여부)에는 한 번에 처리할 수 있는 양을 고려하여 학생 3명의 응답만을 입력하여 평가하게 하였다.

5. 평가일치도 분석

본 연구는 인간평가자와 ChatGPT-4에 의한 채점 결과(점수)의 일치도를 검증하기 위해 Cohen 카파계수, 가중 카파계수(weighted kappa coefficient)와 Pearson 상관계수 분석을 실시하였다. 가중 카파계수는 두 점수 간 불일치 정도에 따라 가중치를 적용한 카파계수로서, 일반 카파계수와 달리 완전한 일치가 이루어지지 않는 경우를 모두 불일치로 간주하지 않고, 일치하는 정도에 따라 1부터 0 사이의 가중치가 부여된다(이상하 외, 2015). 본 연구에서 사용한 채점 점수 범위는 1점부터 3점이었다. 점수 차이가 1점인 경우(예: 인간은 3점 부여, ChatGPT-4는 2점 부여)와 2점 차이인 경우(예: 인간은 1점 부여, ChatGPT-4는 3점 부여)에 따라 가중치를 다르게 부여하는 것이 필요하다고 판단하였다. 따라서, Cohen 카파계수 외에도 가중 카파계수와 상관계수를 활용하여 일치도를 분석하였다.

가중 카파계수는 일차가중 카파계수와 이차가중 카파계수로 분류된다. 일차가중 카파계수는 점수 차이에 비례한 가중치를 사용하고, 이차가중 카파계수는 점수 차이의 제곱에 비례한 가중치를 사용한다(이상하 외, 2015). 본 연구에서는 인간과 ChatGPT-4의 점수 차이가 커짐에 따라 일치하는 정도를 이차가중 카파계수를 사용하여 분석하였다.

상관계수는 두 값의 선형관계를 표현하는 척도로, 두 가지 평가방식에 따른 점수가 완벽하게 일치하지 않아도 유사한 패턴으로 변화할 경우 높은 값을 나타낼 수 있다(이상하 외, 2015). 따라서, 본 연구에서는 상관계수도 함께 분석하여 인간평가자와 ChatGPT-4에 의한 채점 결과를 비교하였다.

Ⅳ. 연구결과

본 연구는 초등학생들이 작성한 과학적 탐구 활동 과제의 서술형 응답을 인간평가자와 ChatGPT가 얼마나 일치하게 평가하는지를 분석하였다. 먼저, 본 연구에서 사용한 22개의 채점항목에 대한 인간평가자와 ChatGPT-4 간의 평가일치도를 카파계수, 이차가중 카파계수, 상관계수로 확인하였다. 이후, 각각 일치도가 높은 채점항목과 낮은 채점항목이 어떤 평가 요소를 대상으로 하며, 각 채점항목이 어떤 특성을 가지고 있는지를 추가적으로 분석·제시하였다.

1. 채점항목별 일치도

연구 결과, 인간평가자와 ChatGPT-4의 평가 총점 간 상관계수는 .74로 나타나 높은 수준의 상관관계를 보이는 것으로 나타났다. 채점항목별 일치도 분석에 앞서, 155명의 학생 응답을 총 22개의 항목으로 채점한 총점 분포는 다음과 같다. 인간평가의 경우 총점이 22점에서 61점 사이에 분포하였고, ChatGPT-4의 경우 23점에서 65점 사이에 분포하는 것으로 나타나 ChatGPT-4가 근소한 차이로 더 높게 평가한 것을 확인하였다. 이후 일반 카파계수, 이차가중 카파계수 분석, 그리고 상관계수 분석을 실시하였다. 각 미션과 채점항목별로 인간평가자와 ChatGPT-4 평가 점수 간의 카파계수, 이차가중 카파계수와 상관계수를 정리하면 <표 5>와 같다.

본 연구에서 채점한 것처럼 채점 점수 범위가 3개 이상일 때에는 인접 범주 분류에 대해서 가중치를 부여하는 것이 실제 채점 상황을 반영한 정확도 추정치이기 때문에, 본 연구에서는 가중 카파계수의 값으로 주요 결과를 해석하였다. 전반적인 이차가중 카파계수의 범위는 .02에서 .58로, 그 범위가 넓은 것으로 나타났다. 이는 채점항목별 일반 카파계수의 값과 유사한 패턴으로 확인되었다(범위: .01～.38). 특히, 미션 2(리코타 치즈의 특성을 변화시킬 수 있는 조건을 적고, 기대되는 결과 작성)의 응답을 평가한 항목들이 미션 1과 3을 채점한 항목보다 상대적으로 높은 일치도를 보였다(범위: .12～.58). 이 중에서도 가장 높은 일치도를 보인 채점항목은 C4로, 각 조건의 변화가 치즈의 어떤 특징을 어떻게 변화시킬지를 예상하여 진술하는지를 평가한 부분이다. 이 채점항목과 관련하여, 인간평가자와 ChatGPT-4가 각각 부여한 점수 범주(1점～3점) 분포를 확인할 수 있는 교차표를 검토하였다. 인간평가자와 ChatGPT-4가 동일하게 점수를 부여한 비율은 C4 항목 전체 점수 분포의 61.9%로 나타났고, 특히 두 평가 주체가 동일하게 3점으로 평가한 비율이 37.4%로, 가장 높았다. 즉, 인간평자가자와 ChatGPT-4의 이차가중 카파계수가 높은 채점항목에서는 두 평가 주체 모두 학생의 응답에 대해 만족하는 비율이 높은 것으로 해석된다.

반면, 미션 3(리코타 치즈의 특성을 변화시킬 수 있는 조건 중 하나를 선정하여 그 조건에 변화

를 주어 실험 후 결과 분석)에 대한 학생들의 응답을 평가한 항목들은 상대적으로 낮은 일치도를 보였다(범위: .02~.47). 미션 3의 채점항목 중 일치도가 가장 낮았던 항목은 분석 결과를 이해하기 위해 추가자료를 탐색했는지에 관한 것이었다(C15). 추가적인 분석을 위해 C15 항목의 점수 범주별 인간평가자와 ChatGPT-4의 교차표를 확인한 결과, 인간평가자는 1점으로 평가했지만, ChatGPT-4는 2점이나 3점으로 평가한 경우가 더 많았다(인간평가자가 1점을 부여한 147개 응답의 약 85%). 이러한 결과는 ChatGPT-4가 해당 채점항목에 대해 인간평가자보다 관대하게 평가한 것으로 해석할 수 있다. 이 채점항목과 관련하여, ChatGPT-4가 학생 응답에 평가한 직후, 그 평가 결과에 대한 이유를 물어보았다. 이를 통해 인간의 평가와 왜 다르게 나타났는지를 확인하고자 하였다. 예를 들어, 치즈가 응고되기 위해서는 끓이는 시간이 충분해야 한다는 것을 알게 된 학생의 응답에 대해, ChatGPT-4는 가열의 영향이 치즈의 맛, 질감, 강도 등에 미친다는 발견을 하나의 추가자료로 인식하였다. 또한, 레몬 한 개의 분량을 넣어 치즈를 만들어서 변화를 관찰한 학생의 응답에 대해서도, 레몬 양의 차이를 추가적인 정보로 간주하여 인간평가자보다 높은 점수를 부여한 것으로 나타났다. 그러나, 인간평가자는 학생이 수행한 실험 내용 외에 추가적으로 찾은 자료를 토대로 작성했는지 여부를 평가하였기 때문에 ChatGPT-4의 평가와 큰 차이를 보였다.

각 채점항목별로 인간평가자와 ChatGPT-4의 평가 점수 간의 상관계수를 살펴본 결과, 전반적인 상관계수는 .14에서 .58로 나타났다(<표 5>). 이는 전술한 이차가중 카파계수의 패턴과 유사하다. 즉, 이차가중 카파계수가 작을수록 상관계수가 작았고, 이차가중 카파계수가 클수록 상관계수도 큰 것으로 나타났다. 뿐만 아니라, 가장 낮은 상관계수를 보인 항목과 가장 높은 상관계수를 보인 항목이 이차가중 카파계수 분석 결과에서 확인한 항목들과 일치하였다. 또한, 인간평가자와 ChatGPT-4의 평가 점수 간 상관관계가 미션별로 다르게 나타났다는 점도 동일하였다. 미션 2(리코타 치즈의 특성을 변화시킬 수 있는 조건을 적고, 기대되는 결과 작성)를 평가한 항목들에서 상대적으로 높았으며(범위: .29~.58), 미션 3을 채점한 항목들에서는 상관계수가 낮게 나타났다(범위: .14~.48).

일반 카파계수, 이차가중 카파계수 분석과 상관계수 분석을 종합해보면, 각 미션마다 평가되는 역량 요소와 평가해야 할 내용(항목)이 달라 채점항목별로 일치도가 매우 다양한 것으로 보인다. 낮은 일치도를 보인 채점항목들의 경우 ChatGPT-4의 평가가 인간의 평가보다 상대적으로 관대한 것으로 나타났다. 이에 본 연구에서는 일치도가 높은 항목과 낮은 채점항목이 각각 어떠한 평가 요소에 대한 것이며, 어떤 특성을 가지고 있는지를 추가적으로 분석하였다.

<표 5> 채점항목별 인간평가자 vs ChatGPT-4 평가일치도(카파계수, 이차가중 카파계수, 상관계수)

과학 탐구 과정 요소		번호	채점항목	카파계수	이차가중 카파계수	상관계수
가설생성	종속변인 규명	C1	해당 차시에서 학습한 과학적 개념을 활용하여 치즈의 특징을 분석하는가?	.11*	.19***	.31**
		C2	치즈의 특징을 분석하기 위하여 기타 사전지식(해당 차시 학습 이외)을 활용하는가?	.04	.12**	.28**
	독립변인 탐색	C3	치즈의 특징을 변화시킬 수 있을 것으로 제시되는 조건을 2개 이상 명료하게 제시하는가?	.29***	.37***	.41**
		C4	각 조건의 변화가 치즈의 어떤 특징을 어떻게 변화시킬지를 예상(예측)하여 진술하는가?	.38***	.58***	.58**
		C5	기타 자료를 참고하거나 스스로의 사고를 통해 고려할 만한 조건을 추가적으로 탐색하였는가?	.09**	.12***	.29**
	가설 정당화	C6	각 조건의 변화가 왜 혹은 어떻게 치즈의 특징을 변화시키게 될지를 사전에 학습한 과학적 개념이나 원리를 활용하여 설명하는가?	.34***	.54***	.56**
		C7	각 조건의 변화가 왜 혹은 어떻게 치즈의 특징을 변화시키게 될지를 기타 과학적 개념이나 원리를 활용하여 설명하는가?	.14**	.53***	.55**
실험설계 및 수행	독립변인 규명	C8	제시된(선정된) 조건이 앞에서(사전학습자료 혹은 개별 추가학습자료) 학습한 과학적 개념이나 원리와 관련이 있는가?	.05***	.27***	.26**
		C9	실험을 위해 변화시킨 조건이 명확한가?	.23***	.43***	.45**
	실험 수행	C10	선택한 조건을 변화시킬 수 있는 방법으로 치즈 제작 과정의 조건을 변화시켰는가?	.28***	.47***	.48**
	혼재변인 통제	C11	자신이 선택한 조건 이외의 조건을 고려하여 통제하는가?	.02	.05*	.19*
자료분석 및 해석	가설 검증	C12	가설을 바탕으로 결과를 분석하는가?: 조건 변화에 따른 결과물의 특성을 비교하여 기술하는가?	.09	.17*	.21*
		C13	가설을 바탕으로 결과를 분석하는가?: 분석 결과를 본인의 가설(예상했던 결과)과 비교하여 기술하는가?	.02	.06**	.23*
	자료 해석	C14	과학적 원리와 개념에 대한 사전지식을 활용하여 결과를 해석하는가?	.07	.21***	.29*
		C15	분석 결과를 이해(해석)하기 위해 추가자료를 탐색하는가?	.01	.02	.14
결론도출 및 평가	평가	C16	분석 결과에 근거하여 특정 조건의 적절성이나 유용성을 평가하는가?	.07*	.10**	.22**
	학습 과정 성찰	C17	실험을 통해 자신이 무엇을 배웠는지 기술하는가?	.02	.06*	.21**
		C18	본인이 수행한 실험의 강점이나 보완할 점 등을 기술하거나 향후 탐구과제를 제시하는가?	.01	.04*	.18*
전반	계량적 접근	C19	조건을 변화시키는 과정 등에서 계량적 접근이 관찰되는가?	.21***	.43***	.48**
	문서기반 의사소통	C20	치즈의 특징(관찰 결과)에 대한 언어적 기술이 풍부한가?	.29***	.47***	.53**
		C21	참고한 자료의 출처를 제시하는가?	.17***	.30***	.42**
		C22	진술이 명료하고, 문단을 구조화하는 등 독자의 가독성을 고려하는가?	.21***	.32***	.35**

* p < .05, ** p < .01, *** p < .001

2. 일치도에 따른 채점항목 특성 분석

본 연구를 통해 과학적 탐구활동 과제의 서술형 응답에 대한 인간평가자와 ChatGPT-4 간 평가 점수의 일치도를 확인하였다. 일치도에 따라 각 채점항목을 분석하면 다음과 같다. 우선, 일치도가 높은 채점항목의 대부분은 학생들이 제안한 치즈의 특성을 변화시키는 요소와 그 결과에 관한 사항들이었다(예: C4, C10). 특히 리코타 치즈 레시피에서 치즈의 특성을 변화시킬 수 있는 조건으로 실험을 진행할 때 치즈의 특징이 어떻게 변화하는지에 대해 진술하였는지에 대한 채점항목(C4)의 이차가중 카파계수는 .58로, 본 연구에서 가장 높은 일치도로 나타났다. 본 연구를 통해, 학생들이 수행한 실험 내용과 관련된 평가 측면에서는 인간평가자와 ChatGPT-4의 채점 결과가 중간 수준의 유사성을 보였음을 확인하였다.

반면, 낮은 일치도를 보이는 채점항목은 다음과 같은 특성을 가지는 것으로 파악되었다. 첫째, 기타 자료나 사전정보를 요구하는 경우다(예: C2, C15). 전체 22개 채점항목 중 사전지식이나 추가정보 등 기타 자료 등을 평가하는 항목은 총 6개였다. 그중 4개 항목의 이차가중 카파계수는 .00에서 .20 사이로 나타나 일치 정도가 매우 낮음으로 해석된다. 또한, 이 4개 항목 중 C15(분석 결과를 이해/해석하기 위해 추가자료를 탐색하는가?)의 이차가중 카파계수와 상관계수는 22개 항목 중 가장 낮은 것으로 나타났다. 특히 이러한 경우, 인간평가자보다 ChatGPT-4가 비교적 관대하게 평가한 것으로 보이는데, 이는 ChatGPT-4가 학생들의 응답에서 기타 자료나 사전정보를 정확히 구분해내지 못하는 것으로 해석될 수 있다.

둘째, 결과를 분석할 때 다른 기준을 토대로 비교하여 기술해야 하는 경우다. 주로 과학적 탐구 역량의 하위 요소 중 자료를 분석하고 기술할 때 필요한 논리분석적 사고(과학적 방법에 해당)를 평가한 항목들이다(예: C12, C13). 논리분석적 사고는 가설을 생성하고 검증하기 위한 과정, 그리고 결과 도출 및 해석에 대한 일련의 과정에 모두 적용된다. 특히 본 연구에서는 학생들이 본인의 가설과 비교하였는지 등을 평가할 때, 인간평가자와 ChatGPT-4 평가 점수의 이차가중 카파계수는 .06으로, 매우 낮은 것으로 나타났다. 이러한 수치는 논리분석적 사고를 평가한 다른 항목(예: 치즈의 특징을 어떻게 변화시킬지 예상하여 기술하였는지, 변화된 치즈의 특징을 과학적 개념을 활용하여 기술하였는지)의 이차가중 카파계수가 .50 이상인 것에 비해 상대적으로 매우 낮은 수치이다. 해당 채점항목들에서도 인간평가자는 1점(만족하지 않음)을 부여한 응답에 ChatGPT-4는 2점(보통)이나 3점(만족함)으로 채점한 것으로 나타나, ChatGPT-4의 평가가 더 관대한 것으로 확인되었다.

셋째, 학생이 과제에 임하면서 반추 및 자기평가를 하는 경우다. 학생이 실험을 통해 무엇을 배웠는지, 자신이 수행한 실험의 강점이나 보완할 점 및 향후 탐구과제에 대해 기술하는지를 평가하는 항목들이 포함된다. 이들은 과학적 탐구 역량의 하위 요소 중 탐구적 태도를 평가하는 항목이

다(예: C17, C18). 특히, 총 22개의 채점항목 중 탐구적 태도를 평가하는 5개 항목의 평균 이차가중 카파계수는 .15로, 이는 다른 역량 요소의 이차가중 카파계수 평균이 각각 .20(과학적 지식), .33(논리분석적 사고), .38(의사소통)이라는 점에서 상대적으로 낮은 수치이다. 뿐만 아니라 탐구적 태도를 평가하는 6개 채점항목 모두에서 ChatGPT-4가 인간평가자보다 관대하게 평가하는 것으로 나타나, 이를 통해서도 탐구적 태도 요소에서 다른 역량 요소에 비해 두 평가자의 채점 결과 차이가 큰 것을 확인하였다(<표 6> 참고).

<표 6> 역량요소별 인간평가자 vs ChatGPT-4 평가일치도의 평균, 표준편차

역량요소	카파계수		이차가중 카파계수		상관계수	
	평균	표준편차	평균	표준편차	평균	표준편차
과학적 지식	.07	.03	.20	.06	.29	.02
논리분석적 사고	.21	.14	.33	.21	.39	.16
탐구적 태도	.06	.05	.15	.19	.27	.15
의사소통	.22	.05	.38	.08	.45	.08

V. 논의 및 제언

1. 논의

본 연구는 서술형 응답 평가의 자동화 가능성을 확인하기 위한 첫 걸음으로, 과학적 탐구활동 과제 보고서의 서술형 응답에 대해 인간평가자와 ChatGPT가 채점한 결과를 비교 분석하였다. 이차가중 카파계수와 Pearson 상관계수 분석을 통해, 인간평가자와 ChatGPT-4 간의 평가일치도가 채점항목별로 다르게 나타나는 것을 확인하였다.

학생들이 치즈의 변화를 위해 설정한 실험 조건의 명확성, 그 조건을 토대로 실험할 경우 어떠한 변화가 예측되는지에 대한 진술, 그리고 치즈 제작 과정의 조건 변화 등 실험의 내용이나 결과에 대한 평가에서 인간평가자와 ChatGPT-4 간의 일치도는 각각 .43, .58, .47로, 이는 카파계수 해석에 따르면 중간(moderate) 수준으로 나타났다(박일남 외, 2013). 이와 같은 채점항목들은 학생들이 작성한 실험 내용(과정 및 결과) 자체에 대한 평가이다. 과학적 탐구활동 과제에서 학생들이 제안한 실험의 세부사항에 대해 채점하는 항목과 관련하여 ChatGPT-4는 인간평가자와 유사하게 평가하는 것을 본 연구를 통해 확인할 수 있었다. 학생들의 응답에서 '실험 조건', '변화 예측', '제작 과정의 조건 변화' 등과 관련된 내용이나 단어들이 쉽게 발견될 수 있었다는 점에서,

ChatGPT-4가 다른 채점항목들에 비해 상대적으로 용이하게 평가했을 것으로 판단된다.

반면, 학생들이 추가자료나 사전지식을 활용하였는지를 평가할 때 인간평가자와 ChatGPT-4 간의 일치도는 비교적 낮았다. 인간평가자는 본 과제에 대한 정보를 ChatGPT-4보다 더 많이 가지고 있었기 때문에, 본 과제에서 학습한 개념 외에 학생들이 가지고 있는 사전지식이나 학생들이 추가적으로 찾아본 기타 자료의 포함 여부를 더 엄격하게 판단하였을 것이다. 그러나 ChatGPT-4의 경우 해당 과제에 대한 학습 없이 대화창에 입력된 학생들의 응답만을 가지고 채점하였기 때문에, 학생들의 응답이 해당 과제를 통해 학습된 것인지, 학생이 실제로 찾아본 기타 자료인지에 대한 판단이 상대적으로 어려웠을 수 있다. 이런 경우 ChatGPT-4에게 사전에 해당 과제에 대한 정보를 미리 학습하게 하는 것을 통해 평가일치도를 향상시킬 수 있을 것으로 예상된다.

또한, ChatGPT-4는 어떤 특정 기준(예: 자신이 세운 가설이나 결과물의 특성 등)과 비교하며 응답하였는지를 평가할 때, 인간평가자에 비해 관대한 채점기준을 적용하는 것으로 확인되었다. 특히, 학생들이 결과를 기술할 때 스스로 세운 가설과 비교하였는지를 평가한 항목에서 인간평가자가 3점으로 평정한 학생은 155명 중 4명에 불과하였지만, ChatGPT-4는 총 86명 학생에게 3점으로 평가한 것으로 나타났다. 반대로 보면, 인간평가자가 해당 평가 항목에 대해 1점으로 평정한 학생은 135명이었던 반면, ChatGPT-4는 22명에 불과했다. 인간평가자가 3점을 준 4명 학생의 응답을 살펴보면, 실험 설계 시 예상했던 결과와 실제 결과가 얼마나 비슷했는지 혹은 얼마나 달랐는지를 명확하게 기술한 것을 확인할 수 있었다. 그러나 ChatGPT-4는 평가 시 가설과의 일치성이나 차이에 초점을 맞춘 것이 아니라, 학생이 제안한 변화 조건으로 결과를 기술했다면 이를 학생이 세운 가설과의 비교로 간주하여 3점으로 평가하였다. 이로 인해 ChatGPT-4는 훨씬 더 많은 학생에게 3점을 부여한 것으로 보인다. 이러한 두 평가자의 차이로 인해 이 채점항목에 있어 매우 낮은 일치도를 나타낸 것으로 판단된다.

마지막으로 학생들이 과제 수행 과정에서 경험한 느낌이나 생각들을 고찰하여 응답하는 부분에 대한 평가에서 ChatGPT-4는 채점항목의 맥락을 완전히 이해하지 못하였고, 이로 인해 인간평가자와의 평가일치도가 상대적으로 낮게 나타났다. 실험을 통해 무엇을 배웠는지를 평가하는 항목은 실험 결과 자체보다, 학생이 스스로 하나의 실험을 설계하고 실행하면서 얻은 경험의 가치 등의 의미를 담고 있다. ChatGPT-4는 학생들이 실험 결과에 대해 상세한 설명을 제공한 경우 이를 학생이 배운 것으로 간주하여 높은 점수를 부여한 것으로 나타났다. 뿐만 아니라 실험의 강점 기술 여부를 평가할 때, ChatGPT-4는 학생들이 치즈 레시피에 어떠한 변화를 주었는지에 대해 명확하게 서술했다면 그것을 모두 강점으로 판단하여 3점으로 평가하였다. 그러나 인간평가자의 경우 다른 학생들의 실험 내용과 비교하여 두드러진 변화를 제시한 경우에만 강점으로 인식한 것으로 보인다. 바꿔 말하면, ChatGPT-4는 본 채점기준의 맥락적 의미를 충분히 이해하지 못한 것으로 판단된다. 뿐만 아니라 두 평가자 사이에서 나타난 낮은 채점 일관성은, 두 평가자가 동일한 채점항목을 사용하였지만, 인간평가자는 채점훈련 과정을 거치면서 인간평가자끼리 암묵적인 평가

지식이 발생했기 때문일 수 있다. 반면 본 연구에서는 ChatGPT-4에게 별도의 사전채점훈련을 제공하지 않았기 때문에, 두 평가자 모두 동일한 채점항목을 사용했음에도 불구하고 낮은 평가일치도로 이어졌을 가능성이 있다. 이를 보완하기 위해 ChatGPT-4가 평가하기 전 각 점수별 응답 예시를 제시하여 학습(채점훈련)시키는 방법을 고안해볼 수 있다.

본 연구의 결과는 단답형 응답에 대해 인간평가자와 기계(ChatGPT)의 평가 간 일치도를 살핀 해외 선행연구결과와 유사한 패턴을 보여준다. 미국 대학생들이 생성한 질문의 학업적 유용성을 전문가와 GPT-3 모델이 평가한 결과, 한 연구에서는 전체 응답 중 약 40%에서 일치하는 판단이 이루어졌고(Moore et al., 2022), 다른 유사한 연구에서는 전문가와 GPT-3 모델 간의 평가일치율이 약 66.5%로 나타났다(Bhat et al., 2022). 두 선행연구에서 모두 GPT-3 모델이 유용하다고 평가한 응답이 인간평가자의 응답보다 상대적으로 많은 것으로 나타나, 본 연구와 마찬가지로 기계(GPT-3)의 평가가 비교적 관대함을 확인할 수 있었다.

그러나 본 연구는 위의 두 선행연구와 몇 가지 차별점을 가지고 있다. 첫째, 본 연구에서는 채점항목별 이차가중 카파계수와 상관계수를 사용하여 인간평가자와 기계 간의 평가일치도를 검증하였다. 앞선 두 연구(Bhat et al., 2022; Moore et al., 2022)에서는 학생 전체 응답에서 두 평가자가 일치하게 평가한 응답의 비율을 계산하는 완전일치도만을 제시하였다. 각 학생의 응답이 학업적으로 유용했는지 여부를 0점 또는 1점으로 평가한 뒤 일치된 횟수를 계산하였다. 반면 본 연구에서는 학생들의 응답을 채점할 때 평가 항목에 충분히 만족한 경우 3점, 보통인 경우 2점, 만족하지 않은 경우 1점으로 평가하였으므로, 두 채점자(인간과 기계)의 점수가 2점 차이나 1점 차이로 불일치할 수 있다. 본 연구에서는 불일치한 정도를 고려하여 이차가중 카파계수를 활용해 일치도를 계산함으로써 선행연구의 일치율 제시 방식보다 방법론적인 부분을 강화하였다.

둘째, 본 연구는 OpenAI의 최신 버전인 ChatGPT-4를 활용했다는 점에서 차별성을 가진다. 이는 단순히 최신의 기술을 사용했다는 사실보다, 본 연구를 통해서 ChatGPT-4의 한글처리 능력을 확인할 수 있다는 점에서 중요하다. 지금까지 국내에서 진행된 다양한 자동채점 프로그램 연구는 대부분 영어 작문 채점에 국한될 정도로(하민수 외, 2019), 한글로 된 문항의 자동채점은 단어나구, 혹은 한 문장 정도의 수준이거나 아직 시작 단계에 불과하다(박종임 외, 2022). 그러나 본 연구를 통해 한글로 된 서술형 응답도 ChatGPT-4를 통해 평가할 수 있음을 확인하였고, 일부 채점항목에서 인간평가자와의 일치도가 중간 수준 이상을 보임으로써 최신 버전인 ChatGPT-4의 한글 자연어처리 기술의 발전도 확인할 수 있었다.

셋째, 본 연구는 선행연구와 달리, 단일 문장 형태가 아닌 과학 탐구형 과제에 작성한 서술형 응답을 평가한 결과로 비교 분석하였다. 본 연구에서 평가한 데이터는 학생들이 참여한 과학 탐구형 과제에서 실제로 수행한 실험 내용을 기반으로 하였다. 학생들은 세 가지 미션을 수행하며 치즈의 조건 변화를 실험하고 그 결과를 보고서로 작성하였다. 본 연구에서는 ChatGPT-4에게 학생들의 응답에 대해 각 미션별로 총 22개의 다양한 채점항목을 적용하여 평가하게 함으로써 과학적

탐구 역량을 측정할 수 있음을 확인하였다. 무엇보다도 어떠한 채점항목에서 인간평가자와 일치했는지 혹은 차이가 있었는지를 추가적으로 분석했다는 점에서 선행연구와 큰 차이를 보인다. 일치도가 높은 채점항목은 주로 학생들의 실험 세부 내용(예: 치즈 특징 변화 기술, 치즈 제작 과정 조건 변경 기술 등)과 관련이 있었다. 반면 기존 자료와의 비교, 가설과의 비교, 또는 과제 수행 과정에서 느낀 점 등을 평가한 항목들에서는 일치도가 다소 낮았다. 본 연구는 이러한 차별점을 토대로 ChatGPT-4가 계속해서 발전해 나감에 따라, 가까운 미래에는 특정 과목 외에도 작문이나 보고서 평가 등 여러 유형의 서·논술형 응답도 자동화할 수 있음을 시사한다. 더불어 어떠한 평가 방식을 적용해야 ChatGPT-4가 인간평가자와 유사한 수준으로 평가할 수 있는지에 대한 통찰을 제공하였다.

2. 연구의 한계와 후속연구 제언

본 연구는 새로운 기술의 도입과 함께 실험적으로 시도한 연구로서 다음과 같은 제한점을 가진다. 우선, 본 연구에서는 ChatGPT-4의 자동채점과 관련한 학습 성능(zero-shot one-shot, few-shot) 중에, 특별한 사전훈련 없이 자동채점을 수행한 zero-shot 학습 성능만을 활용했다는 점에서 제한점을 가진다. ChatGPT-4와 같은 언어 AI 모델은 메타 학습자(meta-learners)로서, 제공된 데이터의 맥락 내에서 학습하여 결과를 추론한다(송민채, 신경식, 2022). 이에 따라 zero-shot보다는 few-shot(예: 예시 채점항목을 가지고 사전에 평가 연습을 한 후 평가) 성능이 더 우수하다고 알려져 있다(Brown et al., 2020). 다만, 본 연구는 이와 같은 미세조정(fine-tuning)의 유무나 세부 방식에 따라 ChatGPT-4의 채점 결과가 달라지는지에 초점을 두진 않았다. 그러나 이러한 사항은 ChatGPT-4를 사용한 자동채점에 대해 논의할 때 핵심적으로 다뤄져야 하는 주제이다. 따라서 후속연구의 한 형태로, ChatGPT-4의 미세조정 및 학습 성능에 따라 인간평가자와의 평가일치도가 달라지는지를 확인할 필요가 있다.

비슷한 맥락에서 GPT의 평가 신뢰도에 대한 제고 또한 하나의 제한점으로 볼 수 있다. 이는 특히 서·논술형 형태의 응답에 대한 채점에 있어 간과되어서는 안 되는 사항이다. 인간평가자의 경우, 채점을 훈련시키는 과정에서부터 신뢰도가 일정 수준에 도달할 때까지 다시 훈련을 받고 반복적으로 연습하는 과정을 거친다(이상하 외, 2015). 그만큼 평가자 간 신뢰도를 확보하는 과정은 매우 중요한 단계이다. 본 연구에서는 소수의 채점항목(특히 일치도가 매우 낮았던 항목)에 대해 ChatGPT-4에게 재채점하게 하였는데, 종종 첫 평가와 다른 점수를 부여하는 경향성을 보이기도 하였다. 물론 1점(채점항목에 만족하지 않은 경우)에서 3점(채점항목에 만족하는 경우)으로 바뀌는 등 극적인 점수 변화는 아니었으나, ChatGPT-4의 평가 결과가 다르게 나타나는 경우가 있었

기 때문에 신뢰도를 어떻게 해결할 것인지에 대한 추가 연구가 필요하다.

또한 본 연구의 주요 목적은 ChatGPT-4와 인간평가자와의 평가 결과를 비교·분석하는 것이었으며, 이 과정에서 좀더 체계적인 질적 분석의 필요성이 부각되었다. 본 연구는 ChatGPT-4의 서·논술형 평가 가능성을 탐색하는 초기 단계의 연구로, 두 평가자 간의 평가일치도에 있어 높고 낮은 차이를 보이는 채점항목의 특성을 규명하고자 하였다. 주로 채점항목의 특징(예: 각 역량요소의 차이), ChatGPT의 평가 한계(예: 채점항목의 맥락 이해 어려움), 두 평가자의 근본적 차이(예: 사전채점훈련 유무) 등에 초점을 두고 평가일치도를 분석하였다. 그러나 보다 깊이 있는 일치도 분석을 위해 추가적인 질적 분석 방법이나 절차를 도입하여 후속연구를 이어갈 필요가 있다.

이밖에도 본 연구를 통해 ChatGPT-4를 통한 채점 자동화에 있어 다음과 같은 측면에 대한 후속연구를 제안하고자 한다. 첫째, 인공지능 발전 속도를 고려할 때, 더욱 상세한 평가가 가능하며 특히 한글에 특화된 평가 기술에 대한 지속적인 연구적 관심이 필요할 것이다. 현재 서술형 평가의 자동채점 가능성 탐색이 그 어느 때보다도 활발한 시기이다. 최근 한국교육과정평가원에서 진행한 한국어 서·논술형 평가를 위한 자동채점 방안 설계 연구(박종임 외, 2022)는 기술적 측면(예: 한글 자연어처리 기술)뿐만 아니라 우리나라 교육 분야에서 한글의 고유한 언어적 특성을 반영한 말뭉치 구축의 필요성을 강조하였다. 이러한 결과는 자동채점 과정에서 텍스트의 표면적 평가를 넘어서 텍스트의 맥락과 숨겨진 의미까지도 평가되어야 함을 방증한다. 따라서, 후속연구에서는 인공지능을 활용한 평가에서 기술적인 측면과 의미적인 측면을 모두 고려한 방안에 대해 집중적으로 살펴봐야 할 것이다.

둘째, 컴퓨터나 GPT 같은 기계채점에 완전히 의존하기보다는 기계와 인간평가자가 적절히 시너지를 발휘할 수 있는 방안에 대한 탐색도 필요하다. 최근 국립국어원에서는 대규모 글쓰기 평가에 있어 “인공지능으로 80%, 사람 손으로 20%”의 채점 비율이 필요하다는 주장을 제기하였다(연합뉴스, 2022). 또한 최근 OpenAI의 공동 창업자인 그렉 브로크만은 ChatGPT와 같은 생성형 인공지능의 성공적인 기능 수행을 위해서는 인간의 피드백이 필수적이라고 강조하였다(국민일보, 2023). 인간이 ChatGPT 답변의 사실 여부를 확인하고 답변 생성 과정을 상세히 검토함으로써 ChatGPT의 학습을 보다 효과적으로 이루어질 수 있게 한다는 것이다. 즉, ChatGPT는 스스로 성장하기보다는 인간과 협력하며 발전할 수 있다는 것이다. 특히 평가 분야에서 전문가(인간)가 가지는 독특한 전문성은 더 중요하다. 예컨대 본 연구에서 발견된 바와 같이, 인공지능이 놓칠 수 있는 언어의 미묘한 뉘앙스는 인간의 언어적 감각을 통해 더욱 정확하게 파악될 수 있다. 이에 따라, ChatGPT와 같은 인공지능(혹은 기계)을 활용한 평가에 앞서 세심한 학습과정이 이루어져야 하며, 추후연구에서는 인간과 기계가 협력하여 평가를 수행하는 방안에 대한 논의도 고려해야 할 것이다. 이를 통해 인공지능 시스템의 성능을 더욱 향상시키고 평가의 정확도를 높일 수 있을 것으로 기대된다.

더 나은 미래 교육을 위해서는 서술형 평가가 활성화될 필요가 있다. 그러나 서술형 평가의 교

육적 중요성에도 불구하고, 높은 비용과 예산, 채점자 간 신뢰도 및 채점 공정성 문제로 활성화되지 못하였다. 이러한 문제에 대한 해결책으로, 본 연구는 ChatGPT-4라는 접근성이 편리한 하나의 인공지능 도구를 활용하여 그 가능성을 탐색하였다. 본 연구에서는 초등학교 5학년만을 대상으로 단일 과제에 대한 자료 수집이 이루어졌지만, 향후 다양한 학교급의 자료를 활용하고 과제 성격에 따른 차이를 분석하는 등의 연구가 추가적으로 이루어져야 할 것이다. 본 연구에서 수행한 것과 같이 평가 자동화의 가능성을 조명한 연구를 시작으로, 다양한 인공지능 기술을 적용한 연구 발전을 추구하기 위하여 일회성의 연구가 아닌 장기적인 계획과 여러 분야 전문가들의 협력을 통한 자동채점 시스템 구축이 필요할 것이다.

참고문헌

공정식(2023). 인공지능 ChatGPT와의 대화에서 엿본 미래의 희망. **대한토목학회지, 71**(3), 12-15.

(Translated in English) Kong, J. S. (2023). A glimpse of the hopeful future from a chat with ChatGPT. *The Magazine of the Korean Society of Civil Engineers, 71*(3), 12-15.

교육부, 한국대학교육협의회(2023). **제4차 2028 대입개편 전문가 포럼 자료집**. https://www.moe.go.kr/ 에서 2023. 4. 5 인출.

(Translated in English) Ministry of Education, Korean Council for University Education. (2023). The 4th 2028 college admission reform expert forum proceedings. Retrieved April 5, 2023, from https://www.moe.go.kr/

국민일보(2023. 4. 27). 챗GPT는 역사적 시기 진입 중… 인간의 피드백이 중요. https://m.kmib.co.kr/ 에서 2023. 4. 28 인출.

(Translated in English) ChatGPT is entering a historical period... Human feedback is important (April 27, 2023). Kukmin Ilbo. Retrieved April 28, 2023, from https://m.kmib.co.kr/

김승주(2019). 채점 자질 설계를 통한 지도 학습 기반 작문 자동채점의 타당도 확보 방안 탐색. **청람어문교육**, **69**, 265-295.

(Translated in English) Kim, S. (2019). Exploring the way to obtain validity of supervised-learning based automated writing scoring by feature engineering. *Journal of CheongRam Korean Language Education, 69*, 265-295.

김유향, 김영수(2012). 과학 탐구 사고력 측정을 위한 서술형 평가 도구 개발. **생물교육, 40**(1), 167-177.

(Translated in English) Kim Y., & Kim Y. (2012). The development of a free-response test for the assessment of science process skill. *Biology Education, 40*(1), 167-177.

노은희, 심재호, 김명화, 김재훈(2012). 대규모 평가를 위한 서답형 문항 자동채점 방안 연구. 서울: 한국교육과정평가원.

(Translated in English) Noh, E., Shim, J., Kim, M., & Kim, J. (2012). Research on automatic scoring methods for short-answer questions in large-scale assessments. Seoul: KICE.

노은희, 김명화, 성경희, 김학수(2013). 대규모 평가를 위한 서답형 문항 자동채점 프로그램 정교화 및 시범 적용. 서울: 한국교육과정평가원.

(Translated in English) Noh, E., Kim, M., Sung, K., & Kim, H. (2013). Refinement and pilot

implementation of an automatic scoring program for short-answer questions in large-scale assessments. Seoul: KICE.

노은희, 이상하, 임은영, 성경희, 박소영(2014). 한국어 서답형 문항 자동채점 프로그램 개발 및 실용성 검증. 서울: 한국교육과정평가원.

(Translated in English) Noh, E., Lee, S., Lim, E., Sung, K., & Park, S. (2014). Development and validation of a Korean short-answer question automatic scoring program. Seoul: KICE.

노은희, 송미영, 성경희, 박소영(2015). 한국어 문장 수준 서답형 문항 자동채점 프로그램 개발 및 적용. 서울: 한국교육과정평가원.

(Translated in English) Noh, E., Song, M., Sung, K., & Park, S. (2015). Development and implementation of a Korean sentence-level short-answer question automatic scoring program. Seoul: KICE.

대한경제(2022. 10. 13). AI 활용, 서·논술형 평가 실천 역량 강화 워크숍 개최. https://www.dnews.co.kr/ 에서 2023. 4. 15 인출.

(Translated in English) AI utilization, workshop on strengthening practical capabilities for descriptive and argumentative evaluations held (October 13, 2022). Daehan Economy. Retrieved April 15, 2023, from https://www.dnews.co.kr/

동아일보(2023. 3. 14). 논술형 수능 논의 시작... 사고력 측정 취지 좋지만 공정성 우려. https://www.donga.com/ 에서 2023. 4. 10 인출.

(Translated in English) Discussion on essay-type college entrance exam begins... Good intentions to measure critical thinking, but concerns about fairness (March 14, 2023). Donga Ilbo. Retrieved April 10, 2023, from https://www.donga.com/

박강윤, 이용상(2022). 한국어 에세이 문항 자동채점을 위한 딥러닝 알고리듬 탐색. **교육평가연구, 35**(3), 465-488.

(Translated in English) Park. K. Y., & Lee, Y. S. (2022). Deep learning algorithm exploration for automated Korean essay scoring. *Journal of Educational Evaluation, 35*(3), 465-488.

박세진, 하민수(2020). 순환신경망을 적용한 초등학교 5학년 과학 서술형 평가 자동채점시스템 개발 및 활용 방안 모색. **교육평가연구, 33**(2), 297-321.

(Translated in English) Park, S., & Ha, M. S. (2020). The development and application of automated scoring system for constructed-response assessment of 5th grade science in elementary schools using recurrent neural network. *Journal of Educational Evaluation, 33*(2), 297-321.

박인숙, 강순희(2012). 중학생의 과학 창의적 문제 해결 능력을 측정하기 위한 도구 개발. **한국과학교육학회지, 32**(2), 210-235.

(Translated in English) Park I., & Kang S. (2012). The development of assessment tools to measure scientific creative problem solving ability for middle school students. *Journal of The Korean Association for Science Education, 32*(2), 210-235.

박일남, 강승식, 노은희, 김명화, 성태제(2013). 정답 템플릿 작성 방식에 의한 한국어 서답형 문항 자동채점 시스템. **정보과학논문지: 컴퓨터의 실제 및 레터, 19**(12), 630-636.

(Translated in English) Park I., Kang S., Noh, E., Kim, M., & Seong, T. J. (2013). Automatic scoring of Korean short answers by answer template description. *KIISE Transactions on Computing Practices, KTCP, 19*(12), 630-636.

박종임, 이상하, 송민호, 이문복, 이민정, 최숙기(2022). 컴퓨터 기반 서·논술형 평가를 위한 자동채점 방안 설계(I). 진천: 한국교육과정평가원.

(Translated in English) Park, J., Lee, S., Song, M., Lee, M. B., Lee, M. J., & Choi, S. (2022). Design of automatic scoring solutions for computer-based descriptive and essay-type assessments. Jincheon: KICE.

박혜영, 이명애, 이명진, 김부연, 임해미, 이현숙, 이동엽(2018). 미래사회 대비 교육과정, 교수학습, 교육평가 비전 연구(Ⅲ): 초·중등학교의 교육평가 방향을 중심으로. 진천: 한국교육과정평가원.

(Translated in Englih) Park, H., Lee, M., Lee, M., Kim, B., Rim, H., Yi, H., & Lee, D. (2018). Education vision for the future curriculum, instruction, and evaluation in South Korea(Ⅲ). Jincheon: KICE

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

(Translated in English) Park, H., Kim, S., Kim, K., Lee, M., Kim, K., & Kim, J. (2019). Substanitializing methods of restricted and extended response essay assessment through enforcing the instruction-assessment alignment. Jincheon: KICE.

백종호, 변태진, 이동원, 심현표(2020). 2015 개정 교육과정 '과학탐구실험'평가 도구 및 평가 현황 탐색. **한국과학교육학회지, 40**(5), 515-529.

(Translated in English) Baek J., Byun T., Lee D., & Shim H. (2020). An investigation on the assessment tool and status of assessment in the 'Scientific Inquiry Experiment' of the 2015 revised curriculum. *Journal of The Korean Association For Science Education, 40*(5), 515-529.

성욱준(2023). [정책제안] AI 고도화에 따른 행정 영역에의 활용 가능성과 투명성 확보. **월간**

공공정책, 209, 59-61.

Sung, W. (2023). [Policy Proposal] Possibility of utilizing advanced AI in administrative areas and ensuring transparency. *Monthly Public Policy, 209*, 59-61.

손정우(2006). 과학논술능력 향상을 위한 과학적 사고력에 근거한 과학글쓰기 교수법. **교육과정평가 연구, 9**(2), 333-355.

(Translated in English) Son, J. (2006). A science writing teaching method based on scientific thinking for improving scientific essay writing ability. *The Journal of Curriculum and Evaluation, 9*(2), 333-355.

송미영, 노은희, 성경희(2016). 대규모 평가 서답형 문항 채점을 위한 문장 수준 자동채점 프로그램의 정확성 분석. **교육과정평가연구, 19**(1), 255-274.

(Translated in English) Song, M., Noh, E., & Sung, K. (2016). Analysis on the accuracy of automated scoring for Korean large-scale assessments. *The Journal of Curriculum and Evaluation, 19*(1), 255-274.

송민채, 신경식(2022). 한국어 자연어생성에 적합한 사전훈련 언어모델 특성 연구. **지능정보연구, 28**(4), 309-328.

(Translated in English) Song, M, & Shin, K. (2022). A study of pre-trained language models for Korean language generation. *Journal of Intelligence and Information Systems, 28*(4), 309-328.

연합뉴스(2022. 1. 18). 장소원 국립국어원장 "인공지능 활용 국어능력 진단체계 개발". https://www.yna.co.kr/ 에서 2023. 4. 22 인출.

(Translated in English) Yonhap News. (2022). Jang, the head of the National Institute of Korean Language "Development of Korean language proficiency assessment system utilizing artificial intelligence" Retrieved April 22, 2023, from https://www.yna.co.kr/

이상하, 노은희, 성경희(2015). 국가수준 학업성취도 평가 서답형 문항에 대한 자동채점의 실용성 분석. **교육과정평가연구, 18**(1), 185-208.

(Translated in English) Lee, S., Noh, E., & Sung, K. (2015). Contrasting automated and human scoring for short-answer NAEA questions. *The Journal of Curriculum and Evaluation, 18*(1), 185-208.

이용상, 구슬기, 이문복(2013). 채점 신뢰도 및 타당도 현황과 향상 방안. 서울: 한국교육과정평가원.

(Translated in English) Lee, Y., Gu, S., Lee, M. (2013). Currents and suggestions of scoring reliability and validity. Seoul: KICE

이정은, 정은영(2013). 과학 글쓰기를 활용한 과학적 사고력 평가 도구의 개발. **교사교육연구,**

52(3), 575-588.

(Translated in English) Lee J., & Jeong, E. (2013). Development of an evaluation tool for assessing scientific thinking ability using science writing. *Teacher Education Research, 52*(3), 575-588.

이현준, 박영민(2019). 자연어처리를 활용한 텍스트 연구 분야의 비교를 통한 자동채점 변인 탐색. **작문연구, 41**, 255-287.

(Translated in English) Lee H., & Park, Y. (2013). A study on the search for automatic scoring variables by comparison of text studies using natural language processing. *Writing Research, 41*, 255-287.

조희련, 임현열, 이유미, 차준우(2021). 한국어 학습 모델별 한국어 쓰기 답안지점수 구간 예측 성능 비교. **정보처리학회논문지. 소프트웨어 및 데이터 공학, 11**(3), 133-140.

(Translated in English) Cho, H., Im, H., Yi, Y.., & Cha, J. W. (2013). Comparison of Korean classification models' Korean essay score range prediction performance. *KIPS KTSDE, 11(3)*, 133-140.

진경애, 남명호, 김명화, 오상철, 김민정, 주형미(2006). 서답형 문항 자동채점 프로그램 도입 방안 연구(I). 서울: 한국교육과정평가원.

(Translated in English) Jin, K., Nam, M., Kim, M., Oh, S., Kim, M. J., Joo, H. (2006). Study on the introduction of short answer question automatic scoring program. Seoul: KICE.

최경애, 이성혜, 채유정(2017). 온라인 문제기반 과학 탐구과제 평가준거 개발. **한국과학교육학회지, 37**(5), 879-889.

(Translated in English) Choi K., Lee S., & Chae Y. (2017). Development of evaluation criteria for online problem-based science learning. *Journal of the Korean Association for Science Education, 37*(5), 879-889.

충남일보(2022. 9. 12). 대전교육과학연구원, '서·논술형 평가도구' 개발·배부. http://www.chungnamilbo.co.kr/ 에서 2023. 4. 15 인출.

(Translated in English) Chungnam Ilbo. (2022). Daejeon Educational Science Research Institute develops and distributes 'short answer and essay-type assessment tools' Retrieved April 15, 2023, from http://www.chungnamilbo.co.kr/

하민수, 이경건, 신세인, 이준기, 최성철, 주재걸, 김남형, 이현주, 이종호, 이주림, 조용장, 강경필, 박지선. (2019). 학습 지원 도구로서의 서술형 평가 그리고 인공지능의 활용: WA3I 프로젝트 사례. **현장과학교육, 13**(3), 271-282.

(Translated in English) Ha, M., Lee, G., Shin, S., Lee, J., Choi, S., Choo, J., Kim, N., Lee, H., Lee, J. H., Lee, J., Jo, Y., Kang, K., & Park, J. (2019). Assessment as a

learning-support tool and utilization of artificial intelligence: WA3I project case. *School Science Journal, 13*(3), 271-282.

하민수(2016). 영어기반 컴퓨터자동채점모델과 기계번역을 활용한 서술형 한국어 응답 채점-자연선택개념평가 사례. **한국과학교육학회지, 36**(3), 389-397.

(Translated in English) Ha, M. (2016). Scoring Korean written responses using English-based automated computer scoring models and machine translation: A case of natural selection concept test. *Journal of the Korean Association for Science Education, 36*(3). 389-397.

함은혜, 이유경, 박소영, 박혜진, 이성혜(2022). 초등학생 과학 탐구과제 수행 특성 분석 및 채점 기준 개발. **한국과학교육학회지, 42**(2), 239-252.

(Translated in English) Ham, E., Lee, Y., Park, S., Park, H., & Lee, S. (2022). Analysis on the characteristics and criteria development in performing science inquiry tasks for elementary school students. *Journal of the Korean Association for Science Education, 42*(2), 239-252.

황진석, 이현주, 곽대오(2010). 고등학생들의 생물교과 관련 과학 과정기능 평가. **교과교육학연구, 14**(1), 67-84.

(Translated in English) Hwang J., Lee h., & Kwak D. (2010). An assessment of high school students' performance on science process skills in biology. *Journal of Research in Curriculum Instruction, 14*(1), 67-84.

Bhat, S., Nguyen, H., Moore, S., Stamper, J., Sakr, M., & Nyberg, E. (2022). *Towards automated generation and evaluation of questions in educational domains*. Paper presented at the 15th International Conference on Educational Data Mining, Durham.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877-1901.

Cahill, A., & Evanini, K. (2020). Natural language processing for writing and speaking. In D. Yan, A. A. Rupp, & P. W. Foltz (Eds.). *Handbook of automated scoring: Theory and practice* (pp. 69-92). Boca Raton, FL: CRC Press.

Chan, A. (2022). GPT-3 and InstructGPT: Technological dystopianism, utopianism, and "contextual" perspectives in AI ethics and industry. *AI and Ethics, 3*, 53-64.

Chodorow, M., & Burstein, J. (2004). Beyond essay length: Evaluating e rater®'s performance on toefl® essays. *ETS Research Report Series, 2004*(1), i-38.

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

Foltz, P. W., Yan, D., & Rupp, A. A. (2020). Natural language processing for writing and speaking. In D. Yan, A. A. Rupp, & P. W. Foltz (Eds.). *Handbook of automated scoring: theory and practice* (pp. 69-92). Boca Raton, FL: CRC Press.

Lee, Y. W., Gentile, C., & Kantor, R. (2010). Toward automated multi-trait scoring of essays: Investigating links among holistic, analytic, and text feature scores. *Applied Linguistics, 31*(3), 391-417.

Lo, C. K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences, 13*(4), 410-425.

Moore, S., Nguyen, H. A., Bier, N., Domadia, T., & Stamper, J. (2022). *Assessing the quality of student-generated short answer questions using GPT-3*. Paper presented at the 17th European Conference on Technology Enhanced Learning, Toulouse.

Rudner, L. M., Garcia, V., & Welch, C. (2006). An evaluation of the IntelliMetric essay scoring system. *Journal of Technology, Learning, and Assessment, 4*(4), 3-21.

• 논문 접수 2023년 04월 30일 / 수정본 접수 6월 6일 / 게재 승인 6월 8일

• 박소영 : 서울대학교 교육학과에서 학사, 동 대학원에서 교육행정 전공으로 석사학위 취득. University of Wisconsin, Madison에서 교육행정 전공 박사학위 취득. 현재 숙명여자대학교 교육학부 교수로 재직 중임. 관심분야는 학교조직, 교육제도 및 정책, 학교성과, 교원정책, AI기반 평가 등임.

• 이병윤 : University of Minnesota, Twin Cities에서 심리학과 졸업. 서울대학교에서 교육학(교육심리)로 석사 및 박사학위 취득. 현재 숙명여자대학교 교육연구소에서 전임연구원으로 재직하고 있으며 청소년의 친사회성 및 교실공정성 등에 관심이 있음.

• 함은혜 : 서울대학교 교육학과에서 학사 및 석사 졸업. Michigan State University에서 교육측정 및 양적연구방법 전공으로 박사학위를 취득함. 현재 공주대학교 교육학과 부교수로 재직 중임. 관심 분야는 역량 진단, 기술향상평가, 문항반응이론과 잠재변인모형, 평가의 형성적 기능 등임.

• 이유경 : 연세대학교에서 교육학 및 국어국문학 학사, 서울대학교에서 교육학 석사, Michigan State University에서 교육심리 박사학위를 취득함. 현재 숙명여자대학교 교육학부 조교수로 재직 중임. 관심 분야는 학업동기의 발달과 영향요인, 협동학습, 학습동료 간 갈등조절, 학생 역량 등임.

• 이성혜 : 충북대학교 컴퓨터교육과 졸업. 서울대와 시라큐스대학교에서 교육공학으로 석사 및 박사학위 취득. 현재 KAIST 과학영재교육연구원에서 연구교수로 재직하고 있으며 에듀테크 및 AI 활용 교육, 학습 분석, 온라인 교육에 관심을 가지고 있음.

[부록 1] 실제 ChatGPT-4 대화창

Model: GPT-4

학생들에게 리코타 치즈 레시피에서 치즈의 특성을 다르게 변화시킬 수 있는 조건 중 하나를 선정해서 그 조건에 변화를 주며 치즈를 직접 만들어보고 결과를 분석하라고 했어. 아래 평가기준으로, 조건에 대해 해당 기준이 만족되면 3, 보통이면 2, 만족되지 않으면 1로 학생들을 상대평가 해줘.
평가 기준: 분석 결과를 이해(해석)하기 위해 추가 자료를 탐색하는가?

51: 차갑게 하기 vs 따뜻하게 하기실온보다 냉장온도 @ 2도씨 일 때 더 빠르게 단단해져 쫀득해졌다. 완성된 리코타치즈를 보온 @ 70도씨 에 넣어두었더니 유청이 더 생기고 푸딩처럼 변하였다.온도가 낮으면 단단해 진다.온도가 높으면 쫀득함이 약하다.치즈를 냉장고에 보관하면 보관기간을 늘릴 수 있고, 쫀득한 치즈를 먹을 수 있다. 보관하는 방법에 따라 식감과 모양이 달라진다.
52: 우유와 생크림의 비율 @ 우유 500mL, 생크림 500mL # 부피의 비 @ 1대1치즈와 생크림을 끓일 때 알갱이가 미세하며 걸죽하지 않다 완성된 치즈는 물렁하며, 표면이 상대적으로 울퉁불퉁 하며, 양이 적다.부드러우며, 표면이 거침신만, 짠맛, 고소한 맛이 높음알갱이가 작으며 양이 적다.부드러워 빵이나 샐러드랑 같이 먹기 좋다.우유와 생크림의 비율이 신맛에 영향을 준다는 것을 알았다. 그 이유는 우유와 생크림의 결합이 적어 비교적 적은 양의 치즈만 만들어 졌고, 나머지는 미세하여 치즈로 만들어지지 않아 상대적으로 소금의 양이 많아진 효과를 예상할 수 있다.
53: 물기를 적게 뺀다.부드러움덜 고소하다.부드럽다.우유맛이 많이 난다.맛이 연하다.치즈에 물기가 많아 고소함과 짭짤함이 단단한 치즈보다 덜 느껴졌지만 식감은 좋았다. 치즈 향이 진하지 않아 치즈를 싫어하는 사람도 먹을 수 있다.
54: 우유와 생크림의 양을 늘리기느끼한 맛과 물렁한 알갱이가 만들어졌다.부드럽다부드럽다 기본보다 느끼한 맛물렁한 알갱이가 짐치즈 맛이 진하게 나며 느끼한 맛이 있다. 물렁한 촉감과 부드러운 촉감이 섞여있다.
55: 레몬즙을 레몬 한 개 분량을 넣는다.치즈의 응고가 A에 비해 빨리 일어나고 덩어리가 크게 만들어진다.맛의 종류 @ 신맛이 많이 난다.맛의 세기 @ A보다 진하다.강도 @ 조금 단단하다.촉감 @ 알갱이가 있고 쫀득하다.레몬을 많이 넣으니 응고가 많이 되어 부피가 크고 조금 단단하고 맛은 신맛이 많이 난다.