

RESEARCH ARTICLE

The Impact of an AI-based Feedback System on the Improvement of Elementary Students' Statistical Inquiry Question Posing: The Moderating Effects of AI Perception and Feedback Self-efficacy

Juyeong Lee^{1*}, Hunkoog Jho²

¹Gwangju Dopyeong Elementary School

²Dankook University

AI 기반 피드백 시스템이 초등학생의 통계적 탐구 질문 설정 향상에 미치는 영향 : AI 인식과 피드백 자기 효율성의 조절 효과

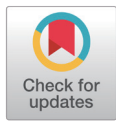
이주영^{1*}, 조현국²

¹광주도평초등학교, ²단국대학교

*Corresponding Author: Juyeong Lee, jy9307@dankook.ac.kr

ABSTRACT

With the advent of the digital age, the amount of data has exponentially increased making statistical analysis essential for decision-making and problem-solving. This research examines the impact of an AI-based feedback system (FS) using the GPT-4 model on the ability of sixth-grade students to formulate statistical inquiry questions. Conducted at an elementary school in Gyeonggi-do, the research involved 95 students divided into experimental and control groups, who participated in an eight-session program. The experimental group received feedback from the AI-based FS, while the control group received teacher feedback. Pre- and post-tests measured the improvement in students' statistical inquiry question levels. Additionally, the study analyzed how students' self-efficacy regarding feedback and their perception of AI moderated the effectiveness of the FS. Results showed that the AI-based FS significantly improved the students' ability to pose statistical inquiry questions compared to the control group. The study also found that a positive perception of AI enhanced the effectiveness of the FS, while self-efficacy regarding feedback did not show a significant impact. These findings suggest that AI-based FS can be an effective educational tool, particularly when students have a positive attitude toward AI. Future research should focus on developing fine-tuned AI-based FS capable of providing detailed feedback throughout all stages of statistical inquiry and investigate its effects on students with cognitive challenges in accepting feedback.



OPEN ACCESS

Brain, Digital, & Learning
2024, Vol. 14, No. 3, 459-473

<https://doi.org/10.31216/BDL.20240026>

Received: July 30, 2024

Revised: August 06, 2024

Accepted: September 07, 2024

© 2023. Institute of Brain based Education,
Korea National University of Education



This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Key words: AI-based feedback system, GPT-4o, statistical inquiry, statistical inquiry question, educational technology, feedback self-efficacy, perception of AI

Introduction

디지털 시대의 도래와 함께 데이터의 양은 기하급수적으로 증가하고 있다. 데이터는 의사결정과 문제 해결의 핵심 자원이기 때문에 기업, 정부, 학계 등 사회 각 분야에서 데이터의 중요성은 전례없이 강조되고 있다. 통계는 데이터를 수학적으로 분석하고 해석하는 데 필수적인 도구다. 범람하는 데이터 사이에서 적절한 데이터를 선택하고 분석하여 자신에게 필요한 정보를 얻기 위해, 미래 사회의 시민은 적절한 수준의 통계적 사고를 함양할 필요가 있다(Moore et al., 2017).

Wild & Pfannkuch (1999)에 따르면, 통계적 사고는 통계적 지식과 맥락적 지식, 그리고 데이터에 대한 정보가 올바르게 조합될 때 이루어지기에 맥락이 배제된 통계 교육은 학생들로 하여금 통계적 사고를 경험하지 못하게 한다. 따라서 실제 세계의 맥락을 담은 데이터를 통계적으로 접근하는 탐구를 통한 통계 교육에 대한 연구가 전 세계적으로 다양하게 이루어져 왔다(Franklin, 2007; Graham, 2006; Marriott et al., 2009; Wild & Pfannkuch, 1999). 통계적 탐구의 세부적인 과정과 용어는 연구마다 다르게 나타나지만, 대체로 통계적 탐구는 문제 설정, 데이터 수집, 데이터 분석, 결론 도출의 네 단계를 따르며 각 단계는 유연하게 순환적으로 이루어진다. 이 중 초기 단계인 문제 설정 단계의 중요성이 여러 연구들에서 강조되었다. Marriot et al. (2009)는 문제 설정 단계가 학생들로 하여금 통계적 문제에 대해 자세히 논의하고, 데이터를 활용해 문제를 해결하는 방법을 모색하도록 이끈다고 이야기했으며, Shaughnessy (2007)은 실제 탐구에서 통계적 문제를 형성하는 데 있어 필요한 능력들을 갖추기 위해서는 문제 설정 단계에 대한 교육이 강조되어야 한다고 이야기했다. 이에 따라 학생들의 탐구 질문 설정 능력을 향상시키기 위한 방안에 대한 연구가 이루어진 바 있다(Arnold, 2013; Arnold & Franklin, 2022; Frischemeier & Leavy, 2020; Lee & Yoo, 2020).

문제 설정 단계에 대한 교육이 잘 이루어지기 위해서는 교사의 적절한 피드백이 요구된다. 교사의 피드백을 통해 학생들은 자신이 설정한 통계적 문제가 어떤 형태의 문제인지 인식할 수 있으며, 문제가 가지고 있는 오류를 파악하고 이를 수정할 수 있다(Arnold & Franklin, 2022). 그러나 Kim & Han (2021)에서 드러나는 것처럼 현재 많은 교사들도 통계적 탐구 문제 설정에 있어 개념적 혼란을 겪고 있기에 통계적 탐구 문제 설정에 대한 명료한 피드백을 제공하는 것에 어려움이 있을 수 있다. 또한 피드백의 효과는 학생의 인지적 특성, 성별 또는 신념에 따라 크게 달라지기에(Banaruee et al., 2017; Han, 2017; Narciss et al., 2014; Winstone et al., 2021) 명료한 개념을 지닌 교사라 해도 이를 모두 고려한 피드백을 제공하는 것은 매우 어려운 일이다.

LLM(Large Language Model)을 활용한 AI(Artificial Intelligence) 기반의 FS(Feedback System)는 이러한 한계를 극복하고 통계능력을 향상시킬 수 있는 위한 효과적인 학습도구로 주목받고 있다(Ellis & Slade, 2023). AI 기반의 FS는 인간 교사와 같은 방식으로 학생에게 피드백을 제공할 수 있으며, 프롬프트 엔지니어링을 통해 개별 맞춤형 피드백을 제공하는 데 유리하다(Li & Xing, 2021; Wang et al., 2023; Shin et al., 2024; Xu et al., 2023). Dai et al.(2023)은 데이터 과학 수업에서 ChatGPT가 인간 교사보다 유창하고 일관성 있는 피드백을 제공할 수 있음을 보이기도 했으며 여러 연구들이 AI 기반의 FS가 학생의 성취에 있어 긍정적인 효과를 보임을 증명하는 바 있다(Jia et al., 2022; Shin et al. 2024; Zheng et al., 2021).

선행연구를 분석해본 결과 통계적 문제 설정에 대한 연구(Arnold, 2013; Arnold & Franklin, 2022; Frischmeier & Leavy, 2020; Lee & Yoo, 2020) 및 AI 기반 FS의 교육적 활용에 대한 많은 연구(Ellis & Slade, 2023; Jia et al., 2022; Shin et al., 2024; Zheng et al., 2021)가 이루어지고 있었다. 하지만 AI기반 FS를 활용한 통계적 탐구의 문제 설정 교육에 대한 연구는 찾아보기 어려웠다. 또한 학생의 특성이나 인식이 인간 교사의 피드백 효과에 미치는 영향에 대한 연구들이 다수 이루어지고 있었던 반면(Banaruee et al., 2017; Han, 2017; Narciss et al., 2014; Winstone et al., 2021), 학생의 특성이나 인식이 AI가 제공하는 피드백의 효과에 미치는 영향을 다룬 연구는 찾아보기 어려웠다.

따라서 본 연구에서는 대표적인 LLM중 하나인 GPT4o에 프롬프트 엔지니어링을 실시하여 이를 통계적 탐구 문제에 대한 피드백을 줄 수 있는 AI기반 FS로 활용한다. 이를 통해 AI기반 FS가 통계적 탐구 문제 설정 단계에서 학생이 설정한 통계 문제 수준에 미치는 영향을 확인하고자 한다. 또한 AI에 대해 긍정적인 태도를 지닌 학생은 AI를 사용할 의도가 높고, 피드백에 대한 자기 효용성(Self-efficacy)이 피드백의 사용에 유의미한 매개 작용을 한다는 연구 결과에 착안하여, AI에 대한 긍정적 인식과 피드백에 대한 자기 효용성이 AI 기반 FS의 효과에 유의미한 조절작용을 하는지 확인하고자 한다(Gado et al., 2021; Otto et al., 2024; Winstone et al., 2021).

연구문제는 다음과 같다. 첫째, AI 기반 FS의 활용이 학생들의 통계적 탐구 문제 수준 향상에 유의미한 영향을 미칠 것인가? 둘째, AI에 대한 긍정적 인식과 피드백에 대한 자기 효용성이 AI 기반 FS를 통한 통계적 탐구 문제 수준 향상에 영향을 줄 것인가?

Materials and Methods

Participants

본 연구는 경기도 G시 소재의 D초등학교 6학년 다섯 학급의 학생 115명 중, 연구 참여에 동의한 95명을 대상으로 하였다. 해당 학생들은 2015 개정 교육과정을 통해 평균, 최대, 최소와 같은 기본 통계량과 막대, 띠 그래프를 그리는 방법에 대해 학습하였으나 통계적 관점에서 탐구 질문을 만드는 과정에 대해서는 교육을 받지 않았다. 2024년도 6월 한 달간 총 다섯 학급의 학생들을 총 39명의 실험집단과 56명의 통제집단으로 나누어 8차시의 통계 수업을 실시하였다. 모든 학생들은 수업 전후로 사전, 사후 검사에 참여하였다.

Development of Program

학생들로 하여금 통계적 탐구 질문이 무엇인지 이해하게 하고, 탐구 질문을 기반으로 한 통계적 탐구를 실행해볼 수 있도록 하기 위해 Arnold (2013)의 연구에 따라 PPDAC 순환단계에 따른 8차시 수업을 기획했다. 또한 학생들에게는 이미 형성된 데이터를 기반으로 통계적 탐구 질문을 도출하는 경험이 필요하다(Lee et al., 2021). 이를 위해 형성된 데이터를 활용하여 탐구 질문을 도출하는 연습을 진행한 후, 실제 탐구를 위한 통계적 탐구 질문을 만들어볼 수 있도록 설계하였다. Arnold (2013)에 따르면 통계적 탐구 질문은 요약질문, 비교 질문 및 관계 질문으로 구분 가능하다. 요약 탐구 질문은 데이터의 묘사가 필요할 때 활용되며, 보통 단일 데이터 세트를 다룬다. 예를 들어, “반 학생들의 평균 키는 얼마인가?”가 요약질문에 속한다. 비교 질문은 두 개 이상의 데이터 하위 집합을 비교할 때 활용된다. 예를 들어, “남학생과 여학생 중 평균 키가 더 큰 그룹은 어

느 쪽인가?”라는 질문이 이에 해당한다. 관계 질문은 두 변수 간의 상호 관계를 조사할 때 활용된다. 예를 들어, “공부 시간과 시험 점수 사이에 관계가 있는가?”라는 질문이 이에 해당한다. 이 중 관계 질문은 초등 교육 과정에서 벗어나는 상관관계나 산점도 등의 개념을 필요로 하기에 본 연구에서는 활용하지 않았다.

6학년 학생들은 교육과정 내에서 통계적 관점에서 탐구 질문을 만드는 과정에 대해서는 학습하지 못했다. 따라서 해당 프로그램은 통계적 탐구 질문이 무엇인지 이해하고, 이를 직접 작성해볼 수 있도록 하는 것에 초점을 맞추었다. 또한 학생들은 평균, 최대, 최소와 같은 기본 통계량과 막대, 띠 그래프를 그리는 방법에 대해 학습하였으나 아직 데이터의 분포와 특성 사이의 상관관계 개념이 무엇인지 학습하지 못했다. 따라서 프로그램은 평균, 최소, 최대 통계량만을 활용하였으며, 데이터의 특징을 파악하기 위한 요약 질문과 데이터를 구분하여 비교하기 위한 비교 질문만을 활용하였다.

Table 1. Program for improving the level of statistical inquiry question posing

Lesson number	Lesson Contents	Use of AI-based FS
1	▶ Understanding Statistical Inquiry Questions - Understanding the Concept of Statistical Inquiry Questions	
2	▶ Understanding Summary Questions - Create summary questions based on the data	O
3	▶ Understanding Comparison Questions - Create comparison questions based on the data	O
4	▶ Posing Statistical Inquiry Questions - Write one summary question and one comparison question based on the data	O
5	▶ Preparing for Inquiry - Select a research topic from the activity sheet and write a statistical inquiry question in groups	O
6	▶ Interpreting Inquiry Results - Interpret the collected data	
7-8	▶ Sharing Inquiry Results - Create presentation materials based on your team's inquiry results and present them	

2~5차시에서는 실험집단의 학생들을 대상으로 AI기반의 FS를 활용하였다. 이를 위해 실험집단의 학생들은 2차시에서 GPT-4 모델에 질문을 입력하고 피드백을 받는 방법을 교육받았으며 실제로 모델에 접속하여 질문을 입력하고 응답을 받는 과정을 직접 경험하였다. 이후 학생들은 예제를 통해 GPT-4 모델에 간단한 질문을 입력하고 피드백을 확인하는 실습을 진행하였다. 학생들은 질문이 양호하다는 피드백이 나타날 때까지 피드백을 바탕으로 질문을 수정하고 다시 모델에 입력하여 새로운 피드백을 받는 과정을 반복하였다. 비교집단에 해당하는 세 학급의 학생들은 교사로부터 피드백을 받도록 하였다. 학생들은 질문을 작성하고 교사로부터 검토를 받았으며 질문이 양호하다는 피드백을 받을 때까지 수정과 검토를 반복하였다.

실험집단의 수업을 담당한 교사들은 각각 경력이 5년과 10년인 일반 초등교사였으며, 통제집단의 수업을 담당한 교사들은 경력이 각각 3년, 6년, 10년인 일반 초등교사였다. 이들은 연구에 참여하기 전에 동일한 교육 프로그램에 대한 사전 교육을 받고, 프로그램의 목적과 방법에 대해 충분히 이해한 상태에서 수업을 진행하였다. 특히, 실험집단의 수업을 담당한 교사들은 AI 도구 사용에 대한 추가적인 훈련을 받았다.

Construction of an AI-based Feedback System(FS)

본 연구에서는 OpenAI의 GPT-4o 모델을 활용하여 학생의 학습을 지원하는 FS를 구축하였다(Fig. 1). 시스템의 작동 과정은 다음과 같다. 학생이 웹 애플리케이션 인터페이스를 통해 질문을 입력하면, 이 입력은 Nginx 웹 서버로 전송된다. Nginx는 HTTP 요청을 처리하는 웹 서버로서, 학생의 질문 요청을 수신하고 이를 Gunicorn WSGI 서버로 전달한다. Gunicorn은 Nginx로부터 전달받은 요청을 Flask 애플리케이션에 중계한다. Flask는 학생의 질문이 도착하면 GPT-4o API를 호출하여 질문에 대한 적절한 답변을 생성한다. Flask 애플리케이션이 API를 통해 질문을 전달하면, GPT-4o는 자연어 처리 알고리즘을 활용하여 가장 적절한 답변을 생성하고 이를 반환한다. 생성된 답변은 Flask 애플리케이션을 통해 학생에게 반환되며, 모든 질문과 답변은 SQLAlchemy ORM을 통해 데이터베이스에 저장한다. 모든 구성 요소는 24시간 운영되는 Amazon EC2 클라우드 서버에서 호스팅된다.

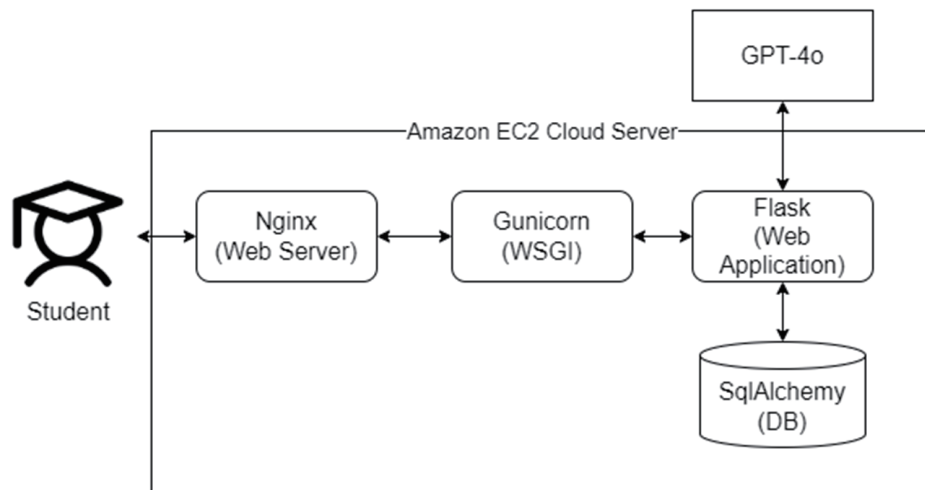


Fig. 1. Diagram of AI-based FS using GPT-4o

Flask를 통해 구축된 웹 애플리케이션은 사전에 설정한 학생 정보에 따라 아이디를 학생 이름, 비밀번호를 반, 번호의 조합으로 하는 세션을 통해 로그인을 하도록 설정하였다. 이를 통해 웹 애플리케이션을 활용하는 사용자의 정보를 데이터베이스에 저장할 수 있도록 하였다. 또한 요약질문 페이지와 비교질문 페이지를 구분하여 학생이 피드백을 받고 싶은 질문 유형에 따라 다른 피드백을 받을 수 있도록 설정하였다(Fig. 2).

AI 질문 설정 도구미 요약 질문 비교 질문 로그 아웃

요약 질문

[사용 안내]

1. 질문을 입력한 뒤 엔터 키를 누르거나 입력 버튼을 누르면 메시지가 서버로 전송됩니다.
2. 대체로 10초 이내에 응답하지만 통행의 종류에 따라 오래 걸리기도 합니다.
3. 응답이 없거나 오래 걸리는 경우에는 로그아웃 이후 다시 연결하시기 바랍니다.

동아료! "학생들의 국어 점수 평균은?"이라는 질문은 아주 잘했어요! 평균을 구하는 것은 학생들의 전반적인 성취도를 이해하는 데 도움이 되니까요. 이 질문을 통해 국어 점수의 중요 경향을 파악할 수 있습니다. 아주 잘했어요!

혹시 다른 과목에 대해서도 비슷한 질문을 만들 수 있을까요? 예를 들어, "학생들의 수학 점수 평균은?" 또는 "학생들의 사회 점수 평균은?" 같은 질문도 좋은 요약 질문이 될 수 있어요.

자신의 설정한 질문을 입력하세요

엔터 버튼

(A)

AI 질문 설정 도구미 요약 질문 비교 질문 로그 아웃

비교 질문

[사용 안내]

1. 질문을 입력한 뒤 엔터 키를 누르거나 입력 버튼을 누르면 메시지가 서버로 전송됩니다.
2. 대체로 10초 이내에 응답하지만 통행의 종류에 따라 오래 걸리기도 합니다.
3. 응답이 없거나 오래 걸리는 경우에는 로그아웃 이후 다시 연결하시기 바랍니다.

이두 좋은 질문이야! "남자들은 국어와 수학 중 평균적으로 어떤 시험을 더 잘 보았는가?"라는 질문은 두 과목의 평균 점수를 비교하는 것이기 때문에 매우 적절한 통계적 탐구 질문이야.

이 질문을 통해 남학생들의 국어와 수학 점수의 평균을 구하고, 이를 비교함으로써 어떤 과목에서 더 높은 성취를 보이는지 알 수 있을 거야.

이 질문은 명확하고 구체적이어서 분석하기에 딱 좋아. 잘했어!

자신의 설정한 질문을 입력하세요

엔터 버튼

(B)

Fig. 2. AI-based FS using GPT-4o. (A) AI-based FS for summary questions. (B) AI-based FS for comparison questions.

이를 위해 웹 애플리케이션과 연결된 GPT4o 모델은 예상 질문 유형과 통계적 탐구 문제 설정의 기준 및 데이터에 대한 정보를 바탕으로 프롬프트 엔지니어링을 실시하여 이에 따라 피드백을 제공할 수 있도록 하였다. 통계적 탐구 문제 설정의 기준은 Arnold (2013)가 제시한 기준 중, 학생들이 통계적 탐구 질문을 처음으로 작성해보는 상황이라는 점을 고려하여 '질문 대상 집단이 명확한가?', '의도가 명확한가?', '데이터를 활용하여 답변 가능한 질문인가?'의 3가지 기준만을 적용하였다. 또한 학생들이 활용할 데이터가 'A초등학교 6학년 학생들의 시험성적' 데이터이며 특성으로는 '반', '성별', '국어성적', '수학성적', '사회성적'을 가지고 있는 데이터셋이라는 정보를 입력하였다.

본 연구에서는 FS의 사용 중 발생할 수 있는 문제를 확인하기 위해 사전에 2명의 6학년 학생을 대상으로 시험 구동을 실시하였다. 이를 통해 발견된 문제점과 이에 대한 개선 방안은 다음과 같다.

첫째, FS가 원하지 않는 형식으로 답변하거나 뉘앙스가 다른 답변을 생성하는 경우가 있었다. 이러한 문제를 해결하기 위해 GPT-4o 모델의 temperature 설정을 0.1로 낮추어 모델의 출력을 더 일관되게 조정하였다. 또한, 일관된 답변을 제공하기 위해 본 연구에서 지향하는 모범답안을 예시 프롬프트(example prompt)로 추가하여 모델이 특정한 형식과 내용으로 답변을 생성하도록 유도하였다. 이를 통해 FS는 보다 안정적이고 일관된 답변을 제공할 수 있게 되었다.

둘째, 학생들이 FS의 답변을 기반으로 자신의 질문을 어떻게 수정해야 할지 알지 못하는 경우가 있었다. 이를 해결하기 위해 FS가 학생의 질문을 바탕으로 개선할 수 있는 방법을 예시로 제공하도록 프롬프트를 추가하였다. 예를 들어, 학생의 질문이 불명확하거나 개선이 필요한 경우, FS는 어떻게 질문을 수정해야 하는지에 대한 구체적인 예시를 함께 제공함으로써 학생들이 보다 명확하고 효과적인 질문을 작성할 수 있도록 하였다.

Table 2. An Example of Questions and Answers from the GPT-4 Model Conducted Through Prompt Engineering

Types of inquiry questions handled by the AI model	examples of query	examples of answer provided by the AI model
Summary question	" "	"The question has not been provided yet. Please let me know the student's question, and I will provide feedback."
	"What is the highest test score in this table?"	"It's a good start, but the question needs to be more specific. For example, you should clarify which subject's highest test score you are referring to. You can improve your question by changing it to: 'What is the highest test score among Korean, Math, and Social Studies in this table?'" This way, you can find the highest score in each subject."
	"What is the average Social Studies score among female students?"	"Great job! This question asks for the average Social Studies score for a specific gender (female), making it statistically very useful. Keep creating various summary questions like this!"
Comparison question	"What is the average math score of boys?"	"The question 'What is the average math score of boys?' is useful for summarizing the math performance of a specific group, but it is not a 'comparative question.' To make a comparative question, you can ask something like, 'Which gender has a higher average math score, boys or girls?'"
	"Between boys and girls, which gender has the higher average score on the math test?"	"That's a great question! 'Which gender has a higher mean score in the math exam, boys or girls?' is very suitable for comparing the average math scores between genders. By analyzing data with such a specific and clear question, you can easily understand the differences in math performance between genders. Well done!"
	"Which gender has the highest maximum score in math, boys or girls?"	"The question is very well crafted! 'Which gender has the highest maximum score in math, boys or girls?' is an excellent summary question for comparing the top math scores between the two genders. Comparing the maximum values is a very useful method for analyzing data. Great job on creating a good question!"

Table 2는 프롬프트 엔지니어링을 실시한 모델이 학생들의 질문에 따라 실제로 반환했던 답변의 예시를 나타낸다. 해당 모델은 아무것도 없는 질문이나, 받을 것으로 기대되는 질문 유형과 다른 질문을 받았을 때 질문이 적절하지 않다는 답을 반환하며 참고할 수 있는 적절한 예시를 제공해준다. 또한 적절한 질문을 받았을 때에는 학생에게 잘한 점에 대한 칭찬을 제공하며 현재 질문을 더 발전시킬 수 있는 방안을 제공해준다.

Measurements

Level of Statistical Inquiry Problem Posing

학생들의 통계적 질문 설정 수준의 변화를 확인하기 위해 사전, 사후 검사를 실시한다. Fig. 3은 검사에서 활용된 데이터셋을 나타낸다. 검사에서는 공공데이터 포털에서 제공하는 '2022년도 전국 학생 건강검사'의 데이터 중 일부를 '시도', '학년', '성별', '태어난 달', '키', '몸무게', '시력' 특성에 한해 제공한다. 사전검사는 학생들에게 주어진 데이터를 활용하여 데이터를 보고 답할 수 있는 통계적 탐구 질문을 두 개 만들 것을 요구한다. 사후검사는 학생들이 주어진 데이터를 활용하여 답할 수 있는 요약질문과 비교질문을 하나씩 만들 것을 요구한다. 학생들의 통계적 질문 설정 수준은 학생이 만든 질문들의 평균을 사용하였다.

시도	학년	성별	태어난 달	키_cm	몸무게_kg	시력
세종	2	남	1월	125	25	1.1
인천	5	남	11월	150	53	1.1
서울	4	여	9월	148	56	1.4
인천	3	남	11월	135	36	0.9
경기	3	남	3월	131	30	1.1
서울	6	남	2월	168	56	0.9
서울	6	남	9월	158	50	0.8
경기	3	여	1월	133	41	1
세종	6	남	4월	167	82	1.2
세종	2	여	1월	132	25	0.8
인천	4	여	1월	148	57	0.9
서울	1	남	5월	122	25	1
인천	1	여	3월	120	22	0.9
경기	3	남	1월	133	31	1
세종	3	여	3월	129	25	1.2
인천	4	여	7월	132	25	0.4
경기	5	여	12월	136	34	1.2
서울	4	남	3월	149	37	0.7

Fig. 3. Dataset used in pre-and post-tests

학생들이 데이터를 보고 설정한 통계적 질문의 수준을 정량화하였다. 통계적 탐구 질문의 수준은 Arnold (2013)의 연구를 참고하여 비통계적 질문, 전-요약질문/전-비교질문, 요약질문/비교질문의 3단계로 구분하였고 각 수준에 대한 판단 준거를 설정하였다. 각 통계적 문제 수준에 대한 판단 기준은 Table 3과 같다. 예를 들어, 학생이 만든 질문이 비 통계적 질문에 해당한다면, 해당 질문은 1의 점수를 부여하였으며 학생이 만든 질문이 통계적 질문에 해당한다면 해당 질문은 3의 점수를 부여하였다.

Table 3. Criteria for the level of statistical inquiry question posing

level	Criteria for level of statistical inquiry problem posing	
1	Non-statistical questions	- Nonsense, not related or not a summary question
2	Pre-summary questions	- A question that asks how many of a particular category or within a specified range.
		- A question that asks for the most popular or most common.
	Pre-comparison questions	- A question that hints at comparison.
		- A question that has all of one group bigger/smaller than all of another group or compares individuals.
3	Summary questions	- A question that asks about a summary statistic; for example, the mean.
		- A question that asks about the overall distribution of the data or what are typical (more than a single value, more than the typical or the mean).
	Comparison questions	- A question that compares categorical data
		- A question that compares a summary statistic.
		- A question that assumes the idea of tendency. This includes questions that ask how much bigger or if there is a difference. Uses a comparing word
		- A question that includes the idea of tendency; for example, on average, generally, tends.

사전검사와 사후검사 사이의 통계적 문제 수준 향상을 평가하기 위해, 사전검사 점수 대비 사후검사 점수의 비율을 향상도(Improvement)로 사용하였다.

Perception of AI

AI에 대한 학생들의 긍정적 인식 수준을 조사하기 위해 Jeffrey (2020)의 설문과 Gado et al. (2022)의 설문 중 AI에 대한 태도(Attitudes towards AI)를 참고하여 AI에 대한 인식(Perception of AI)을 다섯 문항으로 조사하였다. 각 문항은 1(매우 그렇지 않다)부터 5(매우 그렇다)의 리커트 척도(Likert scale)를 활용하였다. 예를 들어, 설문에서 학생들은 ‘인공지능이 우리 사회에 좋은 영향을 줄 것이다.’에 1~5의 점수를 매긴다.

Self-efficacy about Feedback

피드백에 대한 자기 효율성(Self-efficacy)을 조사하기 위해 Linderbaum & Levy (2010)을 참고하여 자기 효율성을 다섯 문항으로 조사하였다. 문항은 학교에서의 상황에 맞게 설문을 조정하였다. 특히 ‘조언을 처리한다(Deal with feedback)’라는 표현을 학생들이 어려워할 것이라 판단하여 ‘조언을 처리한다’는 것은 조언을 받았을 때 나에게 필요한 조언이라고 생각할 경우 이를 받아들이고, 나에게 맞지 않는 조언이라고 느꼈을 때 이를 받아들이지 않는 것을 뜻한다.’라는 설명을 추가하였다. 각 문항은 1(매우 그렇지 않다)부터 5(매우 그렇다)의 리커트 척도를 활용하였다. 예를 들어, 설문에서 학생들은 ‘다른 친구들과 비교했을 때, 나는 조언을 더 잘 처리하는 편이다.’에 1~5의 점수를 매긴다.

Data Collection and Analysis

본 연구는 사전, 사후 검사를 통해 실험집단 39명과 비교집단 56명의 자료를 수집하여 이를 두 가지 관점에서 분석하였다. 첫번째로 학생들의 통계적 탐구 질문 수준이 AI기반 FS 활용 여부에 따라 유의미하게 다른지 확인하기 위해 독립 표본 t-검정을 실시하였다. 두번째로 학생들의 통계적 탐구 질문 수준의 향상 정도가 학생들의 피드백에 대한 자기 효율성 또는 AI에 대한 긍정적 인식 수준에 따라 조절되는지 확인하기 위해 Hayes (2022)의 Process macro model 2를 활용하여 조절 분석(Mediation analysis)을 실시하였다.

Results

Impact of AI Feedback on Improvement in Levels of Students' Statistical Inquiry Questions

프로그램이 적용되기 전 시행된 사전조사 결과에 대한 Levene's test결과, $F = 1.351$, $p = 0.248$ 로, 유의수준 95%에서 두 집단의 분산이 동질적임을 확인하였다. 이후 독립표본 t-검정 결과, $t = -0.752$, $p = 0.454$ 로 유의수준 95%에서 귀무가설이 채택되었다. 따라서 프로그램 적용 전, 실험집단과 통제집단의 탐구 문제 설정 수준에는 유의미한 차이가 없다는 것을 확인할 수 있었다.

프로그램을 실시한 이후에 진행된 사후조사의 결과에 대한 Levene's test결과, $F = 0.392$, $p = 0.533$ 로, 유의수준 95%에서 두 집단의 분산이 동질적임을 확인하였다. 이후 독립 표본 t-검정을 실시하였을 때, $t = -3.156$, $p = 0.002$ 로, 유의확률 95%에서 귀무가설을 기각하였다. 이는 실험집단과 통제집단 간의 탐구 문제 설정 수준

에 유의미한 차이가 있음을 나타낸다.

AI기반 FS 적용 후, 실험집단은 탐구 문제 설정 수준에서 통제집단보다 유의미하게 높은 성취를 보였다. 이러한 결과는 AI기반 FS의 활용이 학생들의 통계적 탐구 능력을 향상시키는 데 효과적일 수 있음을 시사한다.

Table 4. Results of the independent samples t-test for pre-test and post-test

Type of test	Group	N	M(SD)	t	df	p
Pre	Experimental group	56	1.366(0.584)	-0.752	93	0.454
	Control group	39	1.462(0.643)			
Post	Experimental group	56	1.857(0.679)	-3.156	93	0.002*
	Control group	39	2.321(0.739)			

Note. N= population, M = Mean, SD = Standard deviation, df = degree of freedom, p = significance

*p<.01

Impact of AI Perceptions and Feedback Beliefs on Improvement in Levels of Statistical Inquiry Questions

Fig. 4는 두 번째 연구문제를 개념적 도표로 나타낸 것이다. AI기반 FS가 통계적 탐구 질문 설정 향상에 영향을 미칠 때 피드백에 대한 자기 효율성과 AI에 대한 인식 수준이 조절 작용을 하는지 확인하기 위해 SPSS에서 PROCESS v4의 model 2를 활용하여 다중 조절 변수에 대한 조절 분석을 실시하였다(Hayes, 2022). AI 기반 FS의 사용여부를 독립변수(X)로, 통계적 탐구 문제 설정 향상도를 종속 변수(Y)로 사용했으며 피드백에 대한 자기 효율성과 AI에 대한 긍정적 인식 수준을 각각 조절변수(W, Z)로 활용하였다. 분석 과정에서는 상호작용 효과를 살펴보기 위해 연속형 변수들에 대해 평균 중심화(mean centering)를 실시했다. 상호작용 효과의 유의성은 0.05 수준에서 판단했다.

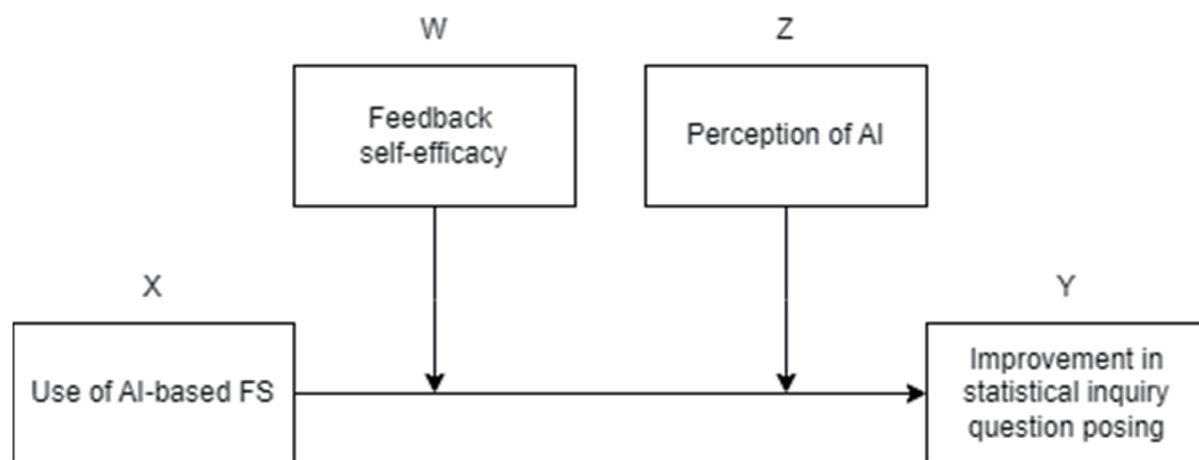


Fig. 4. Conceptual diagram of moderation analysis

피드백에 대한 자기 효율성과 AI에 대한 긍정적 인식 수준을 측정하기 위해 사용된 설문지는 교육학 교수와 인공지능 융합교육 석사 출신 교사에게 검토받았으며 이를 바탕으로 수정하였다. 설문 실시 후, 수집된 데이터를 바탕으로 신뢰도를 검증하였다. 피드백에 대한 자기 효율성 설문지의 신뢰도를 확인한 결과, Cronbach's α 계수는 0.742로 나타나 적절한 신뢰도를 보였다. AI에 대한 긍정적 인식 수준 설문지의 신뢰도는 Cronbach's α 계수가 0.815로 나타나 높은 신뢰도를 보였다.

Table 5에 따르면 조절 분석을 통해 얻은 예측 모델의 결정계수 R^2 는 0.139로 나타났으며, 이는 모델이 종속 변수의 변동성의 약 16.8%를 설명함을 의미한다. F 값은 2.872 ($p = 0.019$)으로 모델이 통계적으로 유의미함을 나타냈다. 상호작용 항목에 대한 주요 발견은 다음과 같다. 'AI 기반 FS의 활용'과 '피드백에 대한 자기 효율성' 간의 상호작용 항(Use of AI-based FS*Feedback self-efficacy)의 계수는 0.012 ($p = 0.958$)으로 유의미하지 않게 나타났다. 'AI 기반 FS의 활용'과 'AI에 대한 긍정적 인식' 간의 상호작용 항(Use of AI-based FS*perception of AI)의 계수는 0.495 ($p = 0.004$)로 유의미하게 나타났다.

Table 5. Results of moderation analysis examining the effects of AI-based FS, feedback self-efficacy, and perception of AI on statistical inquiry question improvement

Variables	B	SE	t	p
Con	1.497	0.100	15.026	0.000***
Use of AI-based FS	0.249	0.156	1.592	0.115
Feedback self-efficacy	0.142	0.141	1.003	0.319
Perception of AI	-0.269	0.114	-2.349	0.021*
Use of AI-based FS*Feedback self-efficacy	0.012	0.226	0.053	0.958
Use of AI-based FS*Perception of AI	0.495	0.169	2.930	0.004**
F				2.872(0.019*)
R^2				0.139
ΔR^2				0.083(F=4.298, p=0.017*)

Note. SE = Standard error, df = degree of freedom, p = significance

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

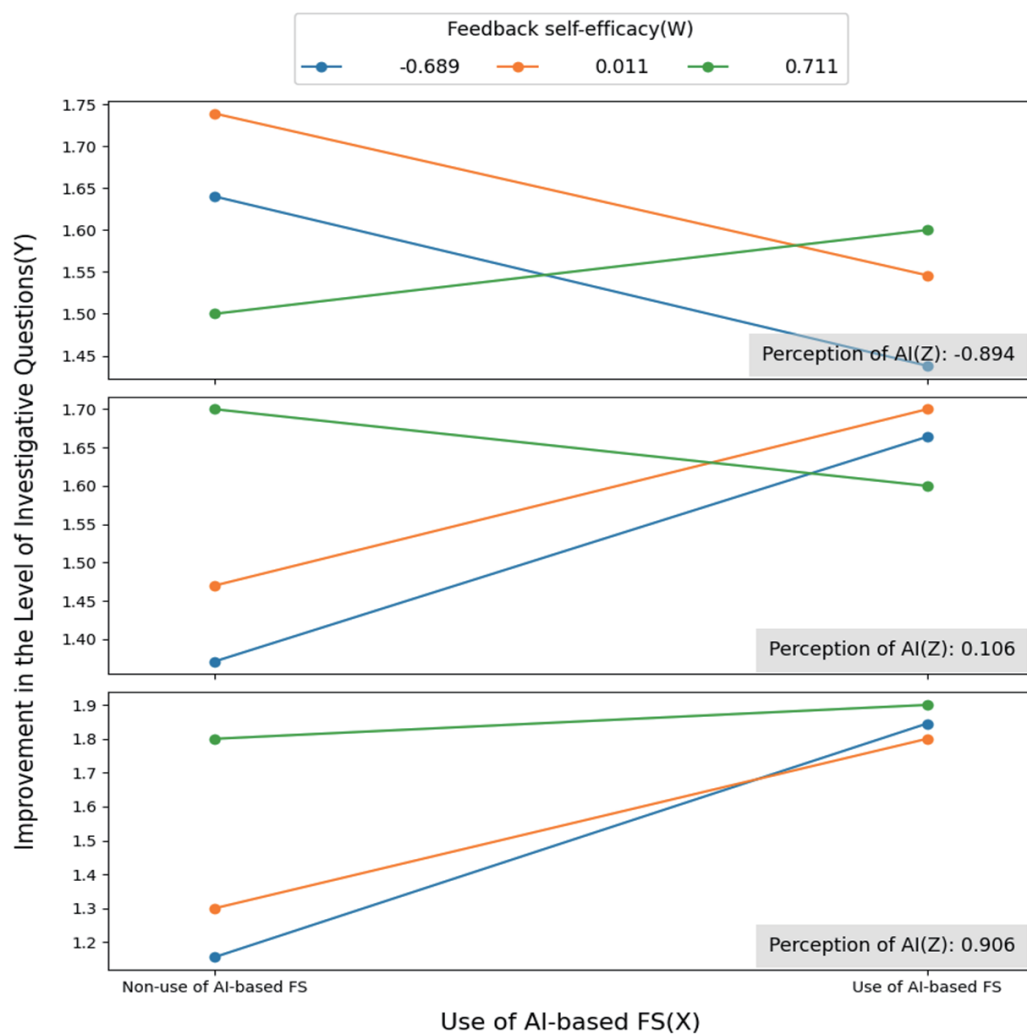
Table 6에 따르면 조건부 효과 분석 결과, AI에 대한 긍정적 인식이 높을 경우, 피드백에 대한 자기 효율성이 낮을 때(Feedback self-efficacy = -0.689, $p = 0.011$)와 피드백에 대한 자기 효율성이 평균 수준일 때(Feedback self-efficacy = 0.011, $p = 0.002$), 그리고 피드백에 대한 자기 효율성이 높은 수준일 때(Feedback self-efficacy = 0.711, $p = 0.010$) 모두 통계적으로 유의미한 영향을 미치는 것으로 나타났다. 반대로 AI에 대한 긍정적 인식이 평균 수준이거나 낮은 경우에는 피드백에 대한 자기 효율성의 효과가 통계적으로 유의미하지 않았다. AI에 대한 긍정적 인식이 평균 수준일 때(Perception of AI = 0.106), 피드백에 대한 자기 효율성이 낮을 때(Feedback self-efficacy = -0.689, $p = 0.184$), 피드백에 대한 자기 효율성이 평균 수준일 때(Feedback self-efficacy = 0.011, $p = 0.057$), 그리고 피드백에 대한 자기 효율성이 높은 수준일 때(Feedback self-efficacy = 0.711, $p = 0.175$) 모두 통계적으로 유의미하지 않았다. 또한 AI에 대한 긍정적 인식이 낮은 경우(Perception of AI = -0.894)에도 피드백에 대한 자기 효율성의 효과는 낮을 때(Feedback self-efficacy = -0.689, $p = 0.452$), 평균 수준일 때(Feedback self-efficacy = 0.011, $p = 0.386$), 그리고 높은 수준일 때(Feedback self-efficacy = 0.711, $p = 0.508$) 모두 통계적으로 유의미하지 않았다. 이 결과는 AI에 대한 긍정적 인식이 높을 때 피드백 자기 효율성의 효과가 증대되며, AI에 대한 인식이 낮거나 평균일 때는 유의미한 효과가 나타나지 않음을 시사한다.

Table 6. Conditional effects of feedback self-efficacy and perception of AI on statistical inquiry question improvement

Feedback self-efficacy	Perception of AI	Effect	SE	t	p
-0.689	-0.894	-0.202	0.267	-0.755	0.452
-0.689	0.106	0.293	0.219	1.340	0.184
-0.689	0.906	0.689	0.265	2.604	0.011*
0.011	-0.894	-0.193	0.222	-0.871	0.386
0.011	0.106	0.301	0.157	1.925	0.057
0.011	0.906	0.697	0.214	3.259	0.002**
0.711	-0.894	-0.185	0.278	-0.664	0.508
0.711	0.106	0.310	0.227	1.366	0.175
0.711	0.906	0.705	0.268	2.634	0.010*

Note. SE = Standard error, df = degree of freedom, p = significance

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

**Fig. 5.** Conditional effects of use of ai-based FS, feedback self-efficacy, and perception of AI on improvement in the level of investigative questions

Discussions

본 연구의 결과는 AI 기반 FS의 활용이 학생들의 통계적 탐구 질문 수준을 유의미하게 향상시킨다는 것을 보여준다. 프로그램 적용 전에는 실험집단과 통제집단 모두 통계적 질문을 형성하는 능력에 유의미한 차이가 없었으나, 프로그램 적용 후 AI 기반 FS를 활용한 실험집단은 통제집단에 비해 통계적 탐구 질문 수준에서 유의미하게 높은 성취를 보였다. 이러한 결과는 AI 기반 FS가 학습의 성취에 효과적인 영향을 미쳤다는 기존 연구들의 결론과도 일치한다(Dai et al., 2023; Shin et al., 2024). 하지만 본 연구에서 활용한 AI 기반 FS는 통계적 탐구 단계 중 문제 설정 단계에 국한되어 사용이 가능하다는 점에서 통계적 탐구를 전문적으로 지원할 수 있는 AI 기반 FS의 개발에 대한 연구가 추가적으로 이루어질 필요가 있다.

또한, 조절 분석 결과는 학생들의 AI에 대한 인식 및 피드백에 대한 자기 효율성과 AI 기반 FS의 효과 사이의 상호작용 여부에 대한 통찰을 제공하였다. 첫째로, 'AI 기반 FS의 활용'과 '피드백에 대한 자기 효율성' 간의 상호작용 항은 유의미하지 않았다. 기존 연구에 따르면, 자기 효율성은 피드백의 활용에 영향을 미친다(Winstone et al., 2021). 이는 피드백을 받아들이는 데 자신감이 없는 학생이 피드백을 사용하지 않는다는 Winstone et al. (2017)의 연구와도 일맥상통한다. 그러나 본 연구에서는 AI 기반 FS를 사용할 때 자기 효율성이 학습 효과에 유의미한 영향을 미치지 않았다. 이는 AI 기반 FS가 자기 효율성에 관계없이 학생들이 피드백을 수용할 수 있도록 도울 수 있음을 시사한다. 하지만 본 연구만으로는 이러한 상호작용을 완전히 설명하기 어려우므로, 자기 효율성이 학습 효과에 미치는 영향에 대해 AI 기반 FS가 어떤 작용을 하는지 심층적으로 조사하는 추가 연구가 필요하다.

둘째로, 'AI 기반 FS의 활용'과 'AI에 대한 긍정적 인식' 간의 상호작용 항은 유의미한 것으로 나타났다. 이는 AI에 대한 긍정적 인식이 높은 학생들이 AI 기반 피드백 시스템으로부터 더 큰 혜택을 받는다는 것을 시사한다. 특히, AI에 대한 긍정적 인식이 높은 학생들(평균보다 1 표준편차 높은 경우)은 AI 기반 FS를 활용할 때 통계적 탐구 능력에서 큰 향상을 보였다. 이러한 결과는 AI에 대한 학생들의 긍정적인 인식 수준이 AI를 더 적극적으로 활용하게 한다는 이전 연구 결과와 일치한다(Gado et al., 2021; Otto et al., 2024). 이는 AI에 대한 인식을 향상시킬 수 있는 교육프로그램이 함께 적용되었을 경우 AI기반 FS가 더 효과적으로 쓰일 수 있음을 시사한다. 따라서 학생들이 AI에 대해 가지는 인식에 대한 세부적인 연구와 AI에 대한 인식을 향상시킬 수 있는 프로그램에 대한 연구가 지속적으로 이루어질 필요가 있다(Gherhes & Obrad, 2018; Jang et al, 2022; Oh, 2023).

Conclusions

본 연구는 경기도 G시 소재 초등학교 6학년 학생들을 대상으로 AI 기반 자동 피드백 시스템(FS)이 학생들의 통계적 탐구 질문 설정 능력에 미치는 영향을 조사하였다. 연구 참여자는 총 92명으로, 실험집단과 통제집단으로 나누어 8차시의 통계 수업을 진행하였다. 실험집단은 GPT4 모델을 활용한 AI 기반 FS를 통해 피드백을 받았고, 통제집단은 교사로부터 피드백을 받았다. 연구는 사전·사후 검사를 통해 학생들의 통계적 탐구 질문 수준 변화를 측정하였으며, 피드백 자기 효율성과 AI에 대한 긍정적 인식이 FS의 효과에 미치는 영향을 분석하였다. 연구 결과는 다음과 같다.

첫째, AI 기반 FS의 활용이 학생들의 통계적 탐구 질문 수준을 유의미하게 향상시키는 것으로 나타났다. 통계적 탐구 질문 설정은 통계적 탐구에서 가장 중요한 단계로 여겨지며, AI 기반 FS를 통해 학생들은 반복적

인 피드백을 받으며 자신들의 질문을 지속적으로 개선할 수 있었다. 이는 통계적 탐구 질문 설정 능력의 향상으로 이어졌다. 둘째, AI 기반 FS의 효과는 학생들의 AI에 대한 인식에 따라 달라졌다. AI에 대한 긍정적 인식이 높은 학생들은 AI 기반 피드백 시스템을 통해 통계적 탐구 능력이 더 크게 향상되었다. 반면, 피드백에 대한 자기 효능성은 본 연구에서 유의미한 영향을 미치지 않았다.

본 연구의 결과를 바탕으로 한 제언은 다음과 같다.

첫째, 통계적 탐구의 전 과정에서 자세하고 정확한 피드백을 제공할 수 있도록 Fine-tuning된 AI 기반 FS의 개발과 그 효과를 연구할 필요가 있다. 둘째, 피드백 수용에 인지적 어려움을 겪는 학생들을 대상으로 AI 기반 FS의 긍정적 효과를 심층적으로 조사하는 연구가 필요하다.

References

- Arnold, P. M. (2013). *Statistical Investigative Questions: An Enquiry into Posing and Answering Investigative Questions From Existing Data* (Unpublished Doctoral Dissertation). The University of Auckland, United States.
- Banaruee, H., Khoshsima, H., & Askari, A. (2017). Corrective feedback and personality type: A case study of Iranian L2 learners. *Global Journal of Educational Studies*, 3, 14-21.
- Dai, W., Lin, J., Jin, H., Li, T., Tsai, Y. S., Gašević, D., & Chen, G. (2023). Can large language models provide feedback to students? A case study on ChatGPT. 2023 IEEE International Conference on Advanced Learning Technologies, Orem, Utah, United States.
- Ellis, A. R., & Slade, E. (2023). A new era of learning: Considerations for ChatGPT as a tool to enhance statistics and data science education. *Journal of Statistics and Data Science Education*, 31, 128-133.
- Frischemeier, D., & Leavy, A. (2020). Improving the quality of statistical questions posed for group comparison situations. *Teaching Statistics*, 42, 58-65.
- Gado, S., Kempen, R., Lingelbach, K., & Bipp, T. (2022). Artificial intelligence in psychology: How can we enable psychology students to accept and use artificial intelligence? *Psychology Learning & Teaching*, 21, 37-56.
- Graham, A. (2006). *Developing Thinking in Statistics*. Paul Chapman Publishing.
- Gherhes, V., & Obrad, C. (2018). Technical and humanities students' perspectives on the development and sustainability of Artificial Intelligence (AI). *Sustainability*, 10, 1-16.
- Han, Y. (2017). Mediating and being mediated: Learner beliefs and learner engagement with written corrective feedback. *System*, 69, 133-142.
- Hayes, A. F. (2022). *Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach* (3rd ed.). The Guilford Press.
- Jeffrey, T. (2020). Understanding college student perceptions of artificial intelligence. *Systemics, Cybernetics and Informatics*, 18, 8-13.
- Jang, J., Jeon, J., & Jung, S. (2022). Development of ST. EM-based AI education program for sustainable improvement of elementary learners. *Sustainability*, 14, 1-16.
- Jia, Q., Young, M., Xiao, Y., Cui, J., Liu, C., Rashid, P., & Gehringer, E. (2022). Insta-reviewer: A Data-driven Approach for Generating Instant Feedback on Students' Project Reports. 15th International Conference on Educational Data Mining, Durham, UK.
- Lee, J., & Yoo, Y. (2020). Analysis of problem statements in poster presentations of statistical investigation. *School Mathematics*, 22, 967-993.

- Lee, K., Yoo, Y., & Tak, B. (2021). Towards data-driven statistics education: An exploration of restructuring the mathematics curriculum. *School Mathematics*, 23, 361-386.
- Li, C., & Xing, W. (2021). Natural language generation using deep learning to support MOOC learners. *International Journal of Artificial Intelligence in Education*, 3, 186-214.
- Linderbaum, B. A., & Levy, P. E. (2010). The development and validation of the Feedback Orientation Scale (FOS). *Journal of Management*, 36, 1372-1405.
- Marriott, J., Davies, N., & Gibson, L. (2009). Teaching, learning and assessing statistical problem solving. *Journal of Statistics Education*, 17, 5.
- Moore, D. S., McCabe, G. P., & Craig, B. A. (2017). *Introduction to the Practice of Statistics* (9th ed.). W. H. Freeman and Company.
- Oh, S. (2023). A study on the case of using ChatGPT & learners' perceptions in college liberal arts writing. *Korean Journal of General Education*, 17, 11-23.
- Wang M., Wang M., Xu X., Yang L., Cai D., & Yin M. (2023) Unleashing ChatGPT's power: A case study on optimizing information retrieval in flipped classrooms via prompt engineering. *IEEE Transactions on Learning Technologies*, 17, 629-641.
- Narciss, S., Sosnovsky, S., Schnaubert, L., Andrés, E., Eichelmann, A., Goguadze, G., & Melis, E. (2014). Exploring feedback and student characteristics relevant for personalizing feedback strategies. *Computers & Education*, 71, 56-76.
- Otto, D., Assenmacher, V., Bente, A., Gellner, C., Waage, M., Deckert, R., Siepermann, M., Kammer, F., & Kuche, J. (2024). Student Acceptance of Ai-Based Feedback Systems: An Analysis Based on the Technology Acceptance Model (TAM). 18th International Technology, Education and Development Conference, Valencia, Spain.
- Shaughnessy, J. (2007). Research on statistics learning and reasoning. In F. K. Lester (Ed.), *Second Handbook of Research on Mathematics Teaching and Learning* (pp. 957-1010). Information Age Publishing.
- Shin, B., Lee, J., & Yoo, Y. (2024) Exploring automatic scoring of mathematical descriptive assessment using prompt engineering with the GPT-4 model: Focused on permutations and combinations. *The Mathematical Education*, 63, 187-207.
- Winstone, N. E., Nash, R. A., Rowntree, J., & Parker, M. (2017). It'd be useful, but I wouldn't use it': Barriers to university students' feedback seeking and recipience. *Studies in Higher Education*, 42, 2026-2041.
- Winstone, N. E., Hepper, E. G., & Nash, R. A. (2021). Individual differences in self-reported use of assessment feedback: The mediating role of feedback beliefs. *Educational Psychology*, 41, 844-862.
- Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *Journal of Statistics and Data Science Education*, 67, 223-265.
- Xu, W., Meng, J., Suba Raja, S. K., Priya, M. P., & Devi, M. K. (2023). Artificial intelligence in constructing personalized and accurate feedback systems for students. *International Journal of Modeling, Simulation, and Scientific Computing*, 14, 1-21.
- Zheng, L., Niu, J., Zhong, L., & Gyasi, J. F. (2021). The effectiveness of artificial intelligence on learning achievement and learning perception: A meta-analysis. *Interactive Learning Environments*, 31, 5650-5664.

Authors' Information

Lee, Juyeong: Gwangju Dopyeong Elementary School, Teacher, 1st Author. <https://orcid.org/0009-0007-0374-360X>

Jho, Hunkoog: Dankook University, Associate Professor, Co-Author. <https://orcid.org/0000-0002-8740-6550>