

AI 활용 교육의 학업성취 효과 : 메타분석

곽성희 · 허 군[†]

국립부경대학교(학생 · [†]교수)

Effects of AI-Supported Education on Academic Achievement : A Meta-Analysis

Sung-Hee KWAK · Gyun HEO[†]

Pukyong National University(student · [†]professor)

Abstract

This meta-analysis examined the effect of AI-supported education on academic achievement in Korea. A total of 24 experimental and quasi-experimental studies published in KCI-indexed journals from 2017 to February 2026 were analyzed. Hedges' g was calculated from post-test means, standard deviations, and sample sizes, and effects were pooled using a random-effects model. The overall effect was positive and significant, Hedges' $g = 0.89$, 95% CI [0.56, 1.22], $p < .001$. AI intervention type showed the clearest moderator pattern, with adaptive tutoring showing the largest point estimate. Education level, subject area, study design, and publication year were not significant moderators. The findings suggest a positive effect of AI-supported education, but should be interpreted cautiously because of high heterogeneity, small subgroup sizes, and possible publication bias.

Key words : AI-supported education, Academic achievement, Meta-analysis, Adaptive tutoring, Conversational tutoring, AI-integrated learning environment

I. 서론

최근 인공지능(AI) 기술이 보편화되면서 교육 현장에서도 AI를 학습 지원 도구로 활용하는 사례가 증가하고 있다. AI 기술은 학습자의 학습 과정과 수행 결과를 분석하고, 이에 근거하여 피드백이나 개별화된 학습 지원을 제공하는 도구로 활용되고 있다. 특히 지능형 튜터링 시스템은 학습자의 응답을 토대로 교수적 처치를 조정할 수 있다는 점에서 AI 활용 교육의 대표적인 형태로 논의되어 왔다(VanLehn, 2011). AI는 단순 보조 도구가 아닌 교수·학습 과정에서 학습자를 직접

지원하는 교수적 기능을 수행하는 요소로 주목받고 있다(Holmes et al., 2019). 이러한 변화는 교육 정책에서도 확인된다. 국내에서는 디지털 전환과 연계하여 ‘모두를 위한 맞춤형교육’을 제시하고, AI 디지털 교과서 도입과 데이터 기반 수업설계를 통한 교사의 맞춤형 학습 지원 역할을 강조하였다(Ministry of Education, 2023). 국제적으로도 AI를 교육의 디지털 전환을 견인하는 핵심 요소로 규정하고, 개인화 학습, 교수 반응성, 학습자 지원 강화와 함께 포용성, 형평성에 대한 필요성을 함께 강조하고 있다(UNESCO, 2021; Miao and Holmes, 2023). 디지털 교육 전환과 맞춤형 학습

[†] Corresponding author : 051-629-5970, E-mail: gyunheo@pknu.ac.kr

* This article was derived from the first author's doctoral research.

지원의 필요성이 확대되면서 AI 활용 교육에 대한 정책적 관심도 증가하고 있다. 교육분야에서 AI 활용 교육은 즉각적 피드백을 제공하는 시스템, 대화형 튜터링을 통한 상호작용을 지원하는 시스템, 학습자의 수행 수준에 따라 과제의 순서·난이도·지원을 조정하는 적응형 시스템, 이러한 기능을 복합적으로 통합한 학습환경 등 다양하게 적용되고 있다. 피드백·상호작용·적응적 지원이 교수과정 안에서 어떻게 조절되는가에 따라 학업성취에 미치는 영향도 달라질 수 있다(Hattie and Timperley, 2007; VanLehn, 2011).

본 연구에서 AI 활용 교육은 AI 기술 또는 AI 기반 시스템이 교수·학습 과정에 활용되어 학습자의 학습을 지원하는 교육 중재로 정의하였다. 학업성취는 특정 교과 또는 학습영역에서 나타나는 인지적 학습성으로 정의하였으며, 시험점수·수행평가·표준화 평가 등 정량적 지표를 통해 수치로 측정된 학습 결과를 포함하였다.

AI 활용 교육의 선행 메타분석 연구들은 학업성취에 대해 다음과 같이 보고하고 있다. Steenbergen-Hu and Cooper(2013)는 K-12 수학 학습에서 ITS 효과가 소폭의 긍정적 효과에 그쳤다고 보고하였고, Ma et al.(2014)은 107개 효과크기를 종합하여 ITS가 교사 주도 대집단 수업 및 비 ITS 컴퓨터 기반 수업보다 유의하게 우수한 효과를 나타냄을 보고하였다. 챗봇 기반 학습에 대한 선행 메타분석에서도 AI 챗봇이 학습성취에 긍정적인 효과를 보일 수 있음이 보고되었다(Wu and Yu, 2024). 생성형 AI 관련 메타분석에서도 K-12 및 고등교육 학습자의 학습성취에 긍정적 효과가 보고되었다(Liu et al., 2025).

최근 교육 현장에서는 챗봇, 지능형 튜터링 시스템, 생성형 AI 등 다양한 형태의 AI 활용 교육 유형이 확대되고 있다. 최근 연구를 중심으로 전체 효과뿐 아니라 중재 유형에 따른 효과 차이를 검토해볼 필요가 있다. 이에 본 연구는 국내 AI 활용 교육 연구를 대상으로 메타분석을 실시하여 AI 활용 교수·학습 중재 유형들이 학업성취에 미

치는 효과를 종합하고자 한다. 분석 대상은 초등학교부터 대학교까지의 일반 학습자를 대상으로 한 국내 실험 및 준실험 연구로 한정하였다. 특히 실험집단과 통제집단 모두 포함한 비교 설계 연구만을 포함하였다. 이는 단일집단 사전·사후 연구에서 발생할 수 있는 내적 타당도 위협요인을 최소화하고 효과크기 해석의 일관성을 확보하기 위한 것이다. AI 활용 교육을 단일 범주로 취급하지 않고 중재 유형, 교육수준, 교과영역에 따른 효과크기 양상을 탐색함으로써 국내 AI 활용 교육 연구의 실증적 근거를 제공하고자 한다.

이를 위한 연구문제는 다음과 같다.

첫째, AI 활용 교육 연구의 일반적인 특징은 무엇인가?

둘째, AI 활용 교육이 학업성취에 미치는 전체 효과성은 어떠한가?

셋째, AI 활용 교육의 학업성취 효과는 주요 하위변인에 따라 어떠한 차이를 보이는가?

II. 연구 방법

1. 연구대상 논문의 선정

본 연구는 교육 분야에서 AI 활용 교육이 학습자의 학업성취에 미치는 영향을 종합적으로 분석하기 위해 KCI 학술지에 게재된 실험 및 준실험 연구를 대상으로 메타분석을 실시하였다. 분석대상 논문은 국내 학술 데이터베이스인 RISS, KISS, DBpia를 활용하여 검색하였다. 검색어는 AI 활용 교육과 학업성취를 나타내는 용어를 중심으로 설정하였다. 최종 분석에는 실험집단과 통제집단이 명확히 구분되고, 효과크기 산출에 필요한 학업성취 결과의 표본수, 평균, 표준편차가 보고되어 있는 연구만을 포함하였다. 이는 본 연구의 AI 활용 교육이라는 중재변인과 학업성취라는 결과변인에 부합하는 연구를 선별하고 동일한 기준으로 효과크기를 산출할 수 있기 때문이다. 문헌 검색에 앞서 연구대상 문헌 선정기준을

명확히 하기 위해 PICOS 기준(Littell, Corcoran and Pillai, 2008)을 적용하였다(<Table 1>).

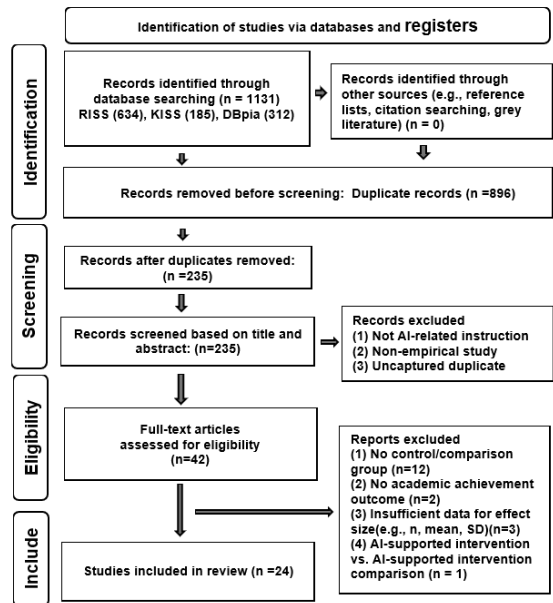
<Table 1> PICOS-Based Study Selection Criteria

Selection Criteria	Details
Population (P)	General learners from elementary school to university.
Intervention (I)	AI-supported education in which AI technologies or AI-based systems were used to support learners' academic learning processes.
Comparison (C)	Non-AI comparison instruction, including traditional instruction and non-AI digital learning.
Outcomes (O)	Academic achievement measured by test scores, performance scores, or other cognitive learning outcomes.
Study Design (S)	Experimental or quasi-experimental studies with both an experimental group and a control group, reporting mean, standard deviation, and sample size.

본 연구에서는 2017년부터 2026년 2월까지 국내 KCI 학술지에 발표된 연구를 분석대상으로 설정하여 현재 AI 활용 교육의 특성을 반영하고, 이질성을 줄이며, 실험·준실험 연구를 충분히 확보하고자 하였다. 학습자의 학업성취를 지원하기 위해 AI 기술 또는 AI 기반 시스템을 활용한 교육 증재를 포함하였다. 포함된 연구는 교수적 기능에 따라 AI 기반 피드백, 대화형 튜터링, 적응형 또는 지능형 튜터링, AI 통합 학습 환경으로 분류하였다.

문헌 선정 과정은 PRISMA 2020 흐름도에 따라 [Fig. 1]과 같이 정리하였다. AI 관련 검색어에는 “인공지능, AI, GPT, Chatbot, 생성형 AI, 적응형 학습, 지능형 튜터링 시스템”을 포함하였으며, 학업성취 관련 검색어에는 “학업성취, 학습성과, 학습효과, 성취도”를 포함하였다. 검색은 AI 관련 용어와 학업성취 관련 용어를 조합하여 실시하였으며 그 결과 1,131편의 문헌이 확인되었다. 중복 문헌을 제외하고 제목, 초록, 핵심어를 추가적으

로 검토하여 최종적으로 효과크기를 산출할 수 있는 24편의 연구가 메타분석에 포함되었다. 문헌 검색과 선정, 제목 및 초록 검토, 전문 검토, 자료 코딩, 비뮌 위험 평가는 두 명의 연구자가 독립적으로 수행하였다. 의견이 일치하지 않는 경우, 원문을 재검토한 후 협의를 통해 최종 선정 여부를 결정하였다.



[Fig. 1] PRISMA flow Diagram.

2. 코딩 및 분석 방법

효과크기 산출을 위한 자료 입력은 논문제목, 저자명, 발표연도, 연구대상(학교급), 교과목, 표본수, 실험집단과 통제집단의 구성정보를 포함하였다. 효과크기 산출을 위해서는 실험집단과 통제집단의 학업성취 사후검사 평균, 표준편차, 표본수를 코딩하였다. 여러 학업성취 지표가 제시된 경우에는 연구자가 주된 학습성가로 제시한 성취도 점수 또는 총점을 우선적으로 사용하였다. 효과크기의 방향을 통일하기 위하여 실험집단을 AI 활용교육 집단으로 비교집단을 비AI 비교수업 집단으로 설정하였다. 효과크기가 양수이

면 AI 활용교육이 비AI 비교수업보다 학업성취 향상에 더 효과적인 것으로 해석하였다.

AI 활용 교육 유형은 교수·학습 과정에서 AI가 수행한 기능을 기준으로 분류하였다. 비교집단은 전통수업과 비AI 비교수업(COMP)으로 코딩하였다. AI 대화형 튜터링(CT), AI 적응형 튜터링 시스템(AT), AI 통합 학습환경(IE)으로 구분하였다. 후보문헌 단계에서는 AI 피드백 시스템(FB) 연구가 확인되었으나, 해당 연구는 AI 활용교육 간 비교 연구였으므로 주 분석에서 제외하였다.

교육 대상은 초등(L1), 중등(L2), 고등(L3), 대학(L4)으로 구분하였고, 교과영역은 Language(S1), STEM(S2), Medical/Health(S3), Social science(S4)으로 구분하였다. Language(S1)에는 영어 관련 과목과 국어 글쓰기를 포함하였고, STEM(S2)에는 수학, 과학, 프로그래밍을 포함하였다. Medical/Health(S3)에는 치위생 과목을 포함하였고, Social science(S4)에는 사회, 디지털 윤리 및 언어습관 관련 과목을 포함하였다.

Medical/Health(S3)과 Social science(S4)는 포함 연구 수가 적어 교과영역별 하위집단 분석 결과는 탐색적으로 해석하였다. 수업기간, 회기 수, 1회기 시간 등 중재 기간 관련 변인은 조절변수에서 제외하였다. 그 이유는 포함연구 간 보고 방식이 통일되어 있지 않거나, 제시되지 않아 확인하기 어려워 시간 및 기간 변인을 정량적 조절변수로 투입하지 않았다.

본 연구의 종속변수는 학업성취이며, 효과크기는 실험집단과 통제집단의 사후검사 점수를 기준으로 산출하였다. 포함 연구들은 서로 다른 교과, 검사도구, 점수체계를 사용하고 있으므로 단순 평균차가 아니라 표준화 평균차를 사용하였다. 실험집단과 통제집단의 사후검사 평균, 표준편차 및 표본수를 이용하여 표준화 평균차를 산출하였고, 표본수가 작은 연구에서 효과크기가 과대추정될 가능성을 보정하기 위해 Hedges' g를 사용하였다. 포함 연구 간 보고 방식의 일관성을 확보하기 위해 사후검사 평균과 표준편차를 기준으

로 효과크기를 산출하였다. 그러나 다수의 연구가 준 실험 설계로 수행되었기 때문에, 사전 성취 수준의 차이가 완전히 통제되지 않았을 가능성이 있다. 본 연구의 효과크기는 AI 활용교육의 순수 인과효과라기보다 비AI 비교수업 대비 사후 학업성취 차이를 종합한 결과로 해석할 필요가 있다.

자료 분석은 R 프로그램을 활용하여 실시하였다. 효과크기 산출, 전체 효과 분석, 하위집단 분석, 메타회귀 분석, 출판편향 검토 및 민감도 분석을 수행하였다. 분석에는 주로 meta, metafor 패키지를 활용하였다.

전체 효과크기는 랜덤효과모형으로 산출하였다. 이는 본 연구에 포함된 24편의 연구가 교육대상, 교과영역, AI 활용교육 유형, 연구설계 등에서 서로 차이를 보이기 때문이다. 랜덤효과모형은 개별 연구의 효과크기가 서로 다른 참효과의 분포에서 추출되었다고 가정하며, 연구 간 분산을 모형에 포함하여 전체 평균 효과를 추정한다(Borenstein et al., 2009; Higgins et al., 2022).

연구 간 이질성은 Q , I^2 , τ^2 를 산출하여 확인하였다. 또한 AI 활용교육 유형, 교육대상, 교과영역에 따른 효과 차이를 확인하기 위해 하위집단 분석을 실시하였다. 출판편향에 따른 효과 변화는 메타회귀분석으로 검토하였다. 분석 결과의 안정성을 확인하기 위해 leave-one-out 민감도 분석을 실시하였다. 출판 편향 가능성은 funnel plot, Egger 회귀검정 및 trim-and-fill 분석 등을 통해 검토하였다.

III. 연구 결과

1. AI 활용 교육 연구의 일반적인 특성 및 비뚤림 위험 평가

본 연구의 최종 분석 대상은 총 24편의 AI 활용 교수·학습 중재 연구이다. 전체 표본수는 1,546명이었으며, 이 중 실험집단은 777명, 통제집단은 769명이다. 연구당 평균 표본 수는 64.4

명, 중앙값은 50명, 최소 표본수는 30명, 최대 표본수는 138명으로 확인되었다.

<Table 2>와 <Table 4>를 보면 교육수준별로는 초등학교 대상 연구가 12편(50.0%)으로 가장 많았다. 다음으로 대학교 대상 연구가 10편(41.7%)이었다. 중학교와 고등학교 대상 연구는 각 1편(4.2%)으로 주로 초등학교와 대학교에 집중되어 있음을 알 수 있다. 출판연도별로는 2023년 연구가 6편(25.0%)으로 가장 많았고, 2025년 연구가 5편(20.8%), 2024년 연구가 3편(12.5%)이었다. 2023년 이후 연구가 16편(66.7%)으로 AI 활용 교육 연구는 최근 들어 증가하고 있음을 보여준다. AI 활용 교육 유형별로는 AI 대화형 튜터링 (CT)이 12편(50.0%)으로 가장 많았다. 다음으로는 AI 통합 학습환경 유형(IE)이 7편(29.2%), AI 적응형 튜터링 시스템 유형(AT)이 5편(20.8%)이었다. 이는 국내 AI 활용 교육 연구에서 챗봇, AI 말하기 앱 등 학습자와 AI 간 상호작용을 기반으로 한 수업이 많이 활용 되었음을 나타낸다. 교과영역별로는 언어영역이 15편(62.5%)으로 가장 높았고, STEM영역이 6편(25.0%), 사회영역이 2편(8.3%), 보건·의료 영역이 1편(4.2%)이었다. 분석 대상 연구는 영어말하기, 발음, 어휘, 문법, 읽기, 쓰기 등 언어교육 분야에 집중되었다. 연구설계별로는 준실험 연구가 21편(87.5%)으로 대부분을 차지하였고 실험연구는 3편(12.5%)이었다. AI 활용 교육 연구는 학교에서 기존 학급 또는 분반을 활용하여 수행된 교육현장 기반 연구가 많았음을 알 수 있다.

Risk-of-Bias 평가를 실시한 결과는 <Table 3>과 같다. 무작위 배정 연구에는 RoB 2를 적용하였고 (Sterne et al., 2019), 비무작위 또는 준실험 연구에는 ROBINS-I를 적용하였다(Sterne et al., 2016). 판정 범주와 평가 기준이 서로 다르므로, 두 도구의 결과를 하나의 점수로 합산하지 않고 도구별 전체 판정 수준을 요약하였다. RoB 2를 적용한 3편은 모두 some concerns로 판정되었고, ROBINS-I를 적용한 21편 중 20편은 moderate

risk, 1편은 serious risk로 판정되었다. 이는 포함 연구의 상당수가 준실험 설계로 수행되었고, 집단 배정 및 사전 동질성 확보 측면에서 편향 가능성이 있음을 보여준다.

2. AI 활용 교육 전체 효과크기와 민감도 분석

AI 활용 교육이 학업성취에 미치는 전체 효과크기를 분석한 결과는 <Table 5>와 같다. AI 활용 교육의 전체효과크기는 Hedges' $g = 0.89$, 95% CI[0.56, 1.22]로 통계적으로 유의하였다 ($z=5.298$, $p < .001$). <Table 6>의 연구간 이질성은 $Q(23)=140.98$, $p < .001$, $I^2=89.47\%$, $\tau^2=0.60$, $\tau = 0.77$ 로 매우 높은 수준이었다. 개별 연구의 효과크기는 대체로 양의 방향으로 나타났으나 일부 연구에서는 매우 큰 효과크기 또는 0을 포함한 신뢰구간이 확인되었다.

<Table 2> General Characteristics of the Included Studies

Category	Sub-group	k	%	Exp. n	Ctrl. n	Total N	Mean N
Publication year	2018	2	8.3	48	45	93	46.5
	2019	2	8.3	51	49	100	50
	2021	2	8.3	55	62	117	58.5
	2022	2	8.3	57	58	115	57.5
	2023	6	25	227	236	463	77.2
	2024	3	12.5	90	84	174	58
	2025	5	20.8	202	190	392	78.4
	2026	2	8.3	47	45	92	46
Grade level	L1	12	50	323	323	646	53.8
	L2	1	4.2	46	46	92	92
	L3	1	4.2	27	32	59	59
	L4	10	41.7	381	368	749	74.9
Subject area	S1	15	62.5	493	483	976	65.1
	S2	6	25	172	176	348	58
	S3	1	4.2	46	44	90	90
	S4	2	8.3	66	66	132	66
AI intervention type	AT	5	20.8	129	134	263	52.6
	CT	12	50	433	429	862	71.8
	IE	7	29.2	215	206	421	60.1
Research design	Exp.	3	12.5	77	73	150	50.0
	Quasi-exp.	21	87.5	700	696	1,396	66.5

Note. Exp. = experimental study; Quasi-exp. = quasi-experimental study; Exp. n = experimental group sample size; Ctrl. n = control group sample size; k = number of studies.

<Table 3> Summary of Risk-of-Bias Assessment

Tool	Overall judgment	k	%
RoB 2	Low risk	0	0.0
	Some concerns	3	100.0
	High risk	0	0.0
ROBINS-I	Low risk	0	0
	Moderate risk	20	95.2
	Serious risk	1	4.8
	Critical risk	0	0

Note. RoB 2 was applied to randomized studies, and ROBINS-I was applied to non-randomized or quasi-experimental studies.

<Table 4> Characteristics of Included Studies (N = 24)

ID	Author(year)	Level	Sub- ject	AI intervention Type	N
1	Jeong & Ma (2023)	L1	S4	CT	40
2	Jeong & Shin (2023)	L1	S2	CT	86
3	Jin (2022)	L4	S1	IE	50
4	Jin (2025)	L4	S1	IE	50
5	Jo & Lim (2025)	L4	S3	IE	90
6	Kang & Kim (2024)	L1	S1	IE	45
7	Kang & Pho (2025)	L4	S1	CT	130
8	Kang & Kim (2026)	L4	S2	IE	48
9	Kim & Choi (2021)	L1	S2	IE	58
10	Kim (2018a)	L4	S1	CT	47
11	Kim (2018b)	L4	S1	CT	46
12	Kim (2019)	L4	S1	CT	70
13	Kim (2023)	L4	S1	CT	138
14	Lee (2019)	L1	S1	CT	30
15	Ko & Han (2023)	L1	S2	AT	48
16	Oh & Choe (2025)	L2	S4	CT	92
17	Oh & Kim (2021)	L3	S2	AT	59
18	Park & Lee (2024)	L4	S1	IE	80
19	Seo & Kim (2023)	L1	S1	CT	109
20	Seo & Kim (2025)	L1	S1	CT	30
21	Seong (2022)	L1	S1	AT	65
22	Whang & Lee (2024)	L1	S2	AT	49
23	Woo & Kim (2023)	L1	S1	AT	42
24	Yang (2026)	L1	S1	CT	44

<Table 5> Overall Effect Size of AI-Supported Education on Academic Achievement

Model	k	Hedges' g	SE	95% CI	z	p
Random -effects model	24	0.89	0.17	[0.56, 1.22]	5.29	< .001

Note. k = number of studies; SE = standard error; CI = confidence interval.

<Table 6> Heterogeneity Statistics for the Overall Effect Size

Model	k	Q	df	p	I ² (%)	τ^2	τ
Random -effects model	24	140.98	23	< .001	89.47	0.60	0.77

AI 활용 교육의 긍정적인 효과와 함께 연구 간 이질성을 고려한 신중한 해석이 필요하다.

일부 연구에서는 효과크기가 매우 크게 나타났고, 0을 포함하여 통계적으로 유의하지 않은 결과는 연구간 높은 이질성과 관련될 수 있어 이후 하위집단 분석이나 메타회귀 분석을 통하여 효과 차이를 분석해 보았다. 민감도 분석을 통해 전체 효과크기의 안정성을 확인하였다. <Table 7>의 Leave-one-out 분석 결과, 각 연구를 하나씩 제외 하더라도 전체 효과크기는 Hedges' $g = 0.780 \sim 0.929$ 범위에서 유지되었고, 모든 경우에서 통계적으로 유의하였다($p < .001$). 이는 특정 연구가 전체 결과를 결정적으로 좌우하지 않았음을 의미한다. 또한 이상치 및 영향력 있는 연구 2편을 제외한 후 재분석한 결과, 전체 효과크기는 Hedges' $g = 0.67$, 95% CI [0.49, 0.85], $p < .001$ 로 나타나 원분석보다 다소 감소하였으나 유의한 정적 효과는 유지되었다. 이와 함께 이질성은 $I^2 = 89.47\%$ 에서 61.82%로, τ^2 는 0.599에서 0.108로 감소하였다. 따라서 일부 연구가 전체 효과크기와 이질성에 영향을 미쳤을 가능성은 있으나, AI 활용 교육이 학업성취에 긍정적 효과를 가진다는 본 연구의 결론은 민감도 분석 후에도 유지되는 것으로 해석할 수 있다.

<Table 7> Sensitivity Analysis of the Overall Effect Size

Analysis	k	Excluded studies	Hedges' g	95% CI	p	I ² (%)	τ^2	Interpretation
Original random-effects model	24	—	0.89	[0.56, 1.22]	< .001	89.47	0.599	Significant positive effect
Leave-one-out analysis	23	23	0.780 ~0.929	All CIs excluded 0	< .001	82.77 ~90.14	0.331 ~0.638	Effect remained significant
Reanalysis after excluding outliers/influential cases	22	2	0.67	[0.49, 0.85]	< .001	61.82	0.108	Effect remained significant

3. 하위변인 분석 및 메타회귀

<Table 8>은 하위변인 메타분석 결과이다. AI 활용교육의 효과크기는 AI 중재 유형에 따라 유의한 차이를 보였다($Q_B(2) = 8.056, p = .018$). AI 적응형 튜터링의 효과크기 가 가장 크게 나타났으며(Hedges' $g = 1.818$, 95% CI [0.571, 3.065]), AI 통합 학습환경(Hedges' $g = 0.669$, 95% CI [0.283, 1.056]), AI 대화형 튜터링(Hedges' $g = 0.641$, 95% CI [0.400, 0.882]) 순으로 나타났다. AI 적응형 튜터링의 경우 신뢰구간이 넓고, 이질성도 $I^2 = 94.46\%$, $\tau^2 = 1.900$ 으로 매우 높게 나타났다. 이 결과는 AI 적응형 튜터링의 효과가 가장 우수하다는 단정적 근거라기보다, 제한된 연구 수와 높은 연구 간 편차 속에서 관찰된 추정치로 해석할 필요가 있다. 하위집단 간 차이는 교육수준($Q_B(3) = 2.201, p = .532$), 교과영역($Q_B(3) = 0.953, p = .813$), 연구설계($Q_B(1) = 0.076, p =$

.782)에서 통계적으로 유의하지 않았다. 다만 교육수준의 중학교(L2)와 고등학교(L3)는 각각 $k = 1$ 이었고, 교과영역의 보건·의료(S3) 영역도 $k = 1$ 에 불과하였다. 따라서 이들 하위집단의 효과크기는 통계적으로 안정적인 하위집단 효과로 해석하지 않았으며, 분석 대상 연구의 분포를 보여주는 기술적 정보로만 제시하였다. AI 적응형 튜터링에서 비교적 큰 효과크기 추정치가 관찰된 것은 학습자 수행 분석, 피드백 제공, 상호작용 지원, 적응적 학습경로 조정 등 다양한 교수적 기능을 강조한 선행연구와 맥락을 같이한다(Ouyang and Jiao, 2021; Holmes et al., 2019). 또한 지능형 튜터링 시스템 관련 선행 메타분석에서도 학습자의 수행 상태를 모델링하고 개별화된 지원을 제공하는 시스템이 학업성취에 효과적일 수 있음을 보고하였다(Ma et al., 2014; VanLehn, 2011). 그러나 높은 이질성은 향후 추가 연구를 통해 재검증

<Table 8> Subgroup Meta-analysis Results by Moderator

Moderator	Subgroup	k	Hedges' g	95% CI	p	I ² (%)	τ^2	$Q_B(df)$	p_B
AI intervention type	CT	12	0.641	[0.400, 0.882]	< .001	64.21	0.109	8.056(2)	.018
	AT	5	1.818	[0.571, 3.065]	.004	94.46	1.900		
	IE	7	0.669	[0.283, 1.056]	< .001	72.97	0.198		
Education level	L1	12	1.177	[0.531, 1.822]	< .001	92.91	1.189	2.201(3)	.532
	L2	1	0.300	[-0.111, 0.711]	.153	—	—		
	L3	1	0.692	[0.165, 1.219]	.010	—	—		
	L4	10	0.697	[0.415, 0.980]	< .001	70.68	0.143		
Subject area	S1	15	0.842	[0.444, 1.240]	< .001	88.36	0.530	0.953(3)	.813
	S2	6	0.999	[0.068, 1.930]	.035	93.98	1.260		
	S3	1	1.628	[1.151, 2.105]	< .001	—	—		
	S4	2	0.639	[-0.110, 1.388]	.095	73.29	0.217		
Research design	Exp.	3	1.055	[-0.178, 2.289]	.093	91.40	1.068	0.076(1)	.782
	Quasi-exp.	21	0.876	[0.524, 1.228]	< .001	89.68	0.594		

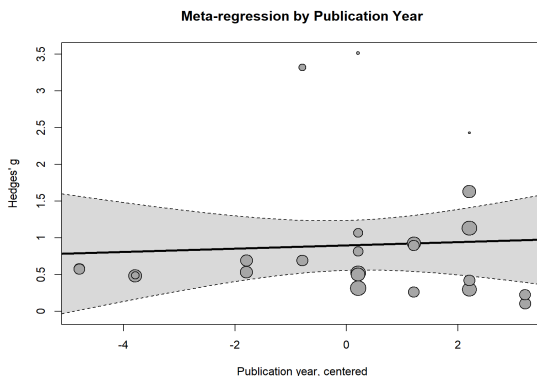
<Table 9> Meta-regression of Effect Size by Publication Year

Moderator	k	β	SE	95% CI	QM(df)	p	τ^2	I ² (%)	Interpretation
Publication year	24	0.022	0.073	[-0.121, 0.165]	0.091(1)	.763	0.634	90.00	ns

Note. β indicates the estimated change in Hedges' g per one-year increase in publication year.

CI = confidence interval; QM = moderator test statistic.

될 필요가 있다. 출판연도 메타회귀 분석 결과는 <Table 9>와 [Fig. 2]에 제시된 바와 같다. 출판연도는 AI 활용교육의 학업성취 효과크기를 유의하게 설명하지 못하였다($\beta = 0.022$, $QM(1) = 0.091$, $p = .763$). 회귀선은 완만한 정적 방향을 보였으나 통계적으로 유의하지 않았으며, 잔여이질성도 여전히 높게 나타났다($\tau^2 = 0.634$, $I^2 = 90.00\%$). 이는 본 연구에 포함된 국내 KCI 연구들에서 출판연도와 효과크기 간 유의한 선형 관계가 확인되지 않았음을 의미한다.



[Fig. 2] Meta-regression of Effect Size by Publication Year.

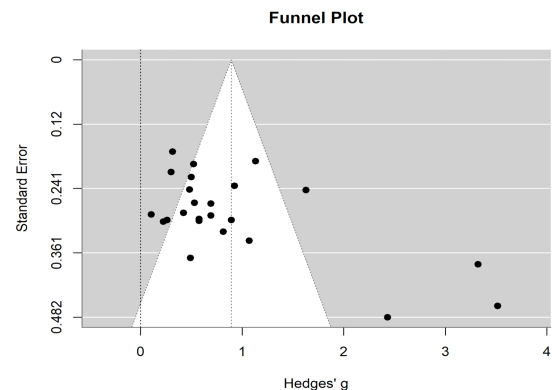
4. 출판편향 분석

Funnel plot, Egger 회귀검정, Begg 순위상관 검정 및 trim-and-fill 분석을 통해 출판편향 가능성을 검토하였다. [Fig. 3]의 funnel plot에서는 일부 연구의 효과크기가 큰 방향으로 분포하는 비대칭적 양상이 확인되었다. Egger 회귀검정은 3.814, $p < .001$, Begg 순위상관 검정은 Kendall's $\tau = 0.341$, $p = .020$ 으로 모두 통계적으로 유의하여 funnel plot 비대칭성이 확인되었다. 이는 효과가

작거나 유의하지 않은 소규모 연구가 누락되었거나, 소규모 연구에서 효과가 크게 나타나는 경향이 있었을 가능성을 시사한다. <Table 11>의 trim-and-fill 분석에서는 보정되어 추가된 연구 수가 0편이었고, 보정 전후의 전체 효과크기는 모두 Hedges' $g = 0.894$, 95% CI [0.563, 1.225], $p < .001$ 로 동일하였다. 그러나 보정 연구가 추가되지 않았다는 결과를 출판편향이 없다는 의미로 해석하기는 어렵다. Egger 검정과 Begg 검정에서 유의한 비대칭성이 확인되었고 연구 간 이질성도 높게 나타났으므로, 본 연구의 전체 효과크기는 다소 과대추정되었을 가능성을 배제할 수 없다. 따라서 AI 활용교육의 긍정적 효과는 확인되었으나, 효과크기의 일반화에는 주의가 필요하다.

<Table 10> Publication Bias Tests

Test	Statistic	p	Interpretation
Egger's regression test	3.814	< .001	Asymmetry detected
Begg's rank correlation test	Kendall's $\tau = 0.341$	.020	Asymmetry detected



[Fig. 3] Funnel Plot.

<Table 11> Trim-and-Fill Analysis

Model	Observed studies	Imputed studies	Total studies	Hedges' g	SE	95% CI	p
Original random-effects model	24	0	24	0.894	0.169	[0.563, 1.225]	< .001
Trim-and-fill adjusted model	24	0	24	0.894	0.169	[0.563, 1.225]	< .001

Note. The trim-and-fill analysis imputed no missing studies

IV. 결론

본 연구는 국내 KCI 학술지에 게재된 AI 활용 교육 연구를 대상으로 학업성취에 미치는 효과를 메타분석하였다. 이를 위해 2017년부터 2026년 2월까지 발표된 실험 및 준실험 연구 중 실험집단과 통제집단이 구분되고 효과크기 산출에 필요한 통계치를 제시한 24편을 최종 분석 대상으로 선정하였다. 첫째, AI 활용 교육 연구는 2023년 이후에 집중되어 있었다. 교육수준은 초등학교와 대학교, 교과영역은 언어영역, AI 활용 유형은 AI 대화형 튜터링이 가장 많았다. 연구설계는 준실험 연구가 대부분이었다. 이는 국내 AI 활용 교육 연구가 실제 학교 및 대학 수업 맥락에서 수행되며, 챗봇, AI 말하기 앱, AI 기반 학습환경 등 학습자와 AI 간 상호작용을 활용한 교수·학습 방식이 활발히 시도되고 있음을 보여준다. 둘째, AI 활용 교육은 학업성취 향상에 유의한 효과를 보였다. 랜덤효과모형의 전체 효과크기는 Hedges' $g = 0.89$ 로 나타나 비AI 비교수업보다 학업성취 향상에 긍정적이었다. 그러나 연구 간 이질성이 매우 높았으므로, 전체 효과크기는 단일한 평균 효과로 해석하기보다 교육수준, 교과영역, AI 활용 방식, 연구설계의 차이를 고려하여 신중하게 해석해야 한다. 셋째, AI 활용 유형별 효과크기는 점추정치 기준으로 AI 적응형 튜터링, AI 통합 학습환경, AI 대화형 튜터링 순으로 나타났다. 다만 AI 적응형 튜터링은 신뢰구간이 넓고 이질성이 높아 가장 우수한 유형으로 단정하기 어렵다. 이는 AI 활용 교육의 효과가 교수·학습 과정에서 AI가 수행하는 기능에 따라 달라질 수 있음을 시

사하지만, 유형별 효과 차이는 추가 연구를 통해 재검증될 필요가 있다. 넷째, 교육수준, 교과영역, 연구설계에 따른 하위집단 간 효과 차이는 통계적으로 유의하지 않았다. 일부 하위집단은 포함 연구 수가 적었으므로, 교육수준 및 교과영역별 효과 차이는 더 많은 연구가 축적된 후 재검토할 필요가 있다. 다섯째, 출판연도는 효과크기를 유의하게 설명하지 못하였다. 최근 AI 활용 교육 연구가 증가했지만, 출판연도의 증가가 효과크기의 증가 또는 감소와 직접적으로 관련되지 않았음을 의미한다. 여섯째, 출판편향 검정에서는 funnel plot 비대칭성이 확인되었고, Egger 회귀검정과 Begg 순위상관 검정에서도 유의한 비대칭성이 나타났다. Trim-and-fill 분석에서는 보정 연구가 추가되지 않았고 보정 후 효과크기에도 변화가 없었으나, 이를 출판편향이 없다는 의미로 해석하기는 어렵다. 효과가 작거나 유의하지 않은 소규모 연구가 누락되었을 가능성이 있으므로, 전체 효과크기는 다소 과대추정되었을 수 있다.

본 연구의 제한점은 다음과 같다. 첫째, 분석대상 연구가 24편으로 제한되어 일부 하위집단 결과를 일반화하는 데 한계가 있다. 특히 중학교, 고등학교, 보건·의료 영역은 각각 1편에 불과하므로 해당 효과크기는 기술적 정보로 해석해야 한다. 둘째, AI 적응형 튜터링은 큰 효과크기 추정치를 보였으나 신뢰구간이 넓고 이질성이 높아 향후 추가 연구를 통한 재검증이 필요하다. 셋째, 포함 연구의 수업기간, 회기 수, 1회기 시간, 총 학습시간 등이 일관되게 보고되지 않아 중재 기간과 강도를 조절변인으로 분석하지 못하였다. 향후 연구에서는 중재 주수, 회기 수, 회기당 시

간, 총 중재 시간, 학습자 노출 빈도 등을 표준화하여 보고할 필요가 있다. 넷째, 포함 연구의 상당수가 준실험 연구였고, RoB 2 적용 연구 3편은 모두 some concerns, ROBINS-I 적용 연구 21편 중 20편은 moderate risk, 1편은 serious risk로 판정되었다. 따라서 본 연구의 효과크기는 AI 활용 교육의 강한 인과적 효과라기보다 포함 연구들에서 관찰된 사후 학업성취 차이의 종합 결과로 해석해야 한다. 다만 leave-one-out 민감도 분석에서 각 연구를 하나씩 제외하더라도 전체 효과크기는 Hedges' $g = 0.780 \sim 0.929$ 범위에서 모두 유의하게 유지되어, 전체 효과가 특정 단일 연구에 의해 결정된 것은 아니었다. 다섯째, 출판편향 검정에서 유의한 비대칭성이 확인되었고 연구 간 이질성도 높았으므로 전체 효과크기와 하위집단 효과크기는 연구 맥락의 차이를 고려하여 신중하게 해석해야 한다. 이상의 결과를 바탕으로 본 연구는 국내 KCI 학술지 연구를 대상으로 AI 활용 교육의 학업성취 효과를 정량적으로 종합하고, AI 활용 유형별 효과 양상을 제시했다는 점에서 의의가 있다.

References

- Borenstein M, Hedges LV, Higgins JPT and Rothstein HR(2009). Introduction to meta-analysis. Wiley.
- Hattie J and Timperley H(2007). The power of feedback. Review of Educational Research, 77(1), 81~112.
<https://doi.org/10.3102/003465430298487>
- Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ and Welch VA(Eds.)(2022). Cochrane handbook for systematic reviews of interventions (Version 6.3). Cochrane.
<https://training.cochrane.org/handbook>
- Holmes W, Bialik M and Fadel C(2019). Artificial intelligence in education: Promises and implications for teaching and learning. Center for Curriculum Redesign.
- Littell JH, Corcoran J and Pillai V(2008). Systematic reviews and meta-analysis. Oxford University Press.
- Liu X, Guo B, He W and Hu X(2025). Effects of generative artificial intelligence on K - 12 and higher education students' learning outcomes: A meta-analysis. Journal of Educational Computing Research, 63(5), 1249~1291.
<https://doi.org/10.1177/07356331251329185>
- Ma W, Adesope OO, Nesbit JC and Liu Q(2014). Intelligent tutoring systems and learning outcomes: A meta-analysis. Journal of Educational Psychology, 106(4), 901~918.
<https://doi.org/10.1037/a0037123>
- Ministry of Education(2023). Digital-driven education reform plan announced: Unlocking opportunities for personalized learning in education.
<https://english.moe.go.kr>
- Ouyang F and Jiao P(2021). Artificial intelligence in education: The three paradigms. Computers and Education: Artificial Intelligence, 2, Article 100020.
<https://doi.org/10.1016/j.caeai.2021.100020>
- Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, Shamseer L, Tetzlaff JM, Akl EA, Brennan SE, Chou R, Glanville J, Grimshaw JM, Hróbjartsson A, Lalu MM, Li T, Loder EW, Wilson EM, McDonald S, McGuinness LA, Stewart LA, Thomas J, Tricco AC, Welch VA, Whiting P and Moher D(2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372, 71.
<https://doi.org/10.1136/bmj.n71>
- Steenbergen-Hu S and Cooper H(2013). A meta-analysis of the effectiveness of intelligent tutoring systems on K - 12 students' mathematical learning. Journal of Educational Psychology, 105(4), 970~987.
<https://doi.org/10.1037/a0032447>
- Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, Cates CJ, Cheng HY, Corbett MS, Eldridge SM, Emberson JR, Hernán MA, Hopewell S, Hróbjartsson A, Junqueira DR, Jüni P, Kirkham JJ, Lasserson T, Li T, McAleenan A, Reeves BC, Shepperd S, Shrier I, Stewart LA, Tilling K, White IR, Whiting PF and Higgins JPT(2019). RoB 2: A revised tool for assessing risk of bias in randomised trials. BMJ, 366, l4898.
<https://doi.org/10.1136/bmj.l4898>
- Sterne JAC, Hernán MA, Reeves BC, Savović J,

- Berkman ND, Viswanathan M, Henry D, Altman DG, Ansari MT, Boutron I, Carpenter JR, Chan AW, Churchill R, Deeks JJ, Hróbjartsson A, Kirkham J, Jüni P, Loke YK, Pigott TD, Ramsay CR, Regidor D, Rothstein HR, Sandhu L, Santaguida PL, Schünemann HJ, Shea B, Shrier I, Tugwell P, Turner L, Valentine JC, Waddington H, Waters E, Wells GA, Whiting PF and Higgins JPT(2016). ROBINS-I: A tool for assessing risk of bias in non-randomised studies of interventions. *BMJ*, 355, i4919.
<https://doi.org/10.1136/bmj.i4919>
- UNESCO(2021). AI and education: Guidance for policy-makers. UNESCO Publishing.
<https://doi.org/10.54675/PCSP7350>
- Miao F and Holmes W(2023). Guidance for generative AI in education and research. UNESCO.
<https://doi.org/10.54675/EWZM9535>
- VanLehn K(2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, 46(4), 197~221.
<https://doi.org/10.1080/00461520.2011.611369>
- Wu R and Yu Z(2024). Do AI chatbots improve students' learning outcomes? Evidence from a meta-analysis. *British Journal of Educational Technology*, 55(1), 10~33.
<https://doi.org/10.1111/bjet.13334>
-
- Received : 04 May, 2026
 - Revised : 27 May, 2026
 - Accepted : 18 June, 2026