Check for updates

JFMSE, 38(5), pp. 1154~1163, 2026.
수산해양교육연구, 제38권 제5호, 통권143호, 2026.
www.ksfme.or.kr
https://doi.org/10.13000/JFMSE.2026.10.38.5.1154

생성형 AI 교수통합경험 척도 예비개발: 독립 합성 페르소나 표본을 활용한 문항 구조 예비 점검

최현규 · 원효헌†
국립부경대학교(학생 · †교수)

Preliminary Development of the Generative AI Instructional Integration Experience Scale: A Synthetic Persona-Based Pilot Examination of Item Structure

Hyeon Kyoo CHOI · Hyo Heon WON†
Pukyong National University(student · †professor)

Abstract

This study preliminarily developed a scale for measuring teachers' generative AI instructional integration experience and conducted a pilot examination of its item structure using expert content validation and two independent synthetic persona-based simulated response datasets. Based on prior studies on TPACK, AI self-efficacy, teacher AI competence, AI acceptance, and ethical use of generative AI, 40 initial items were developed across four dimensions: AI lesson preparation, AI lesson implementation, AI assessment and feedback, and AI use reflection. Twelve experts reviewed item relevance, clarity, representativeness, redundancy, and factor fit, resulting in 28 preliminary items. For exploratory factor analysis, 549 valid responses were obtained after excluding 11 inattentive cases from 560 Nemotron-Personas-KOREA-based simulated responses. Maximum likelihood extraction and direct oblimin rotation retained 20 items. For cross-validation within synthetic persona-based data, a separate synthetic persona sample(N=600), generated independently from the EFA sample, was used for confirmatory factor analysis. The one-factor model was inadequate, whereas the correlated four-factor model showed acceptable fit and the second-order model also showed acceptable fit. The findings provide preliminary evidence for the item structure of a multidimensional construct and demonstrate a methodological example for early-stage item refinement in the AI era. Because the simulated responses were generated from synthetic personas rather than actual teachers, the results should be interpreted as a supplementary pilot examination of item structure, not as final validation evidence for the teacher population.

Key words : Generative AI, Nemotron-Personas-KOREA, Instructional integration, Scale development, Cross-validation

I. 서 론

생성형 AI는 교사의 수업 설계, 수업 운영, 평가 · 피드백, 전문적 성찰의 방식을 빠르게 변화시키고 있다. ChatGPT, Gemini, Claude, Copilot과 같은 도구는 수업자료 생성, 활동 아이디어 탐색, 학습자 수준별 설명 구성, 평가문항 개발, 피드백 문장 작성 등 교수 · 학습의 여러 국면에서 활용

† Corresponding author: 051-629-5972, wonhyo@pknu.ac.kr

될 수 있다. 그러나 이러한 활용은 단순한 기술 사용이 아니라 교육과정, 교과내용, 학습자 특성, 평가 준거, 윤리적 판단이 결합되는 교수 전문성의 문제로 이해되어야 한다.

UNESCO(2023)는 생성형 AI의 교육적 활용에서 인간 중심성, 교사의 전문적 판단, 개인정보 보호, 저작권, 평가 공정성 등을 함께 고려해야 한다고 보았다. 또한 UNESCO(2024)의 교사용 AI 역량 프레임워크는 AI 기초 이해뿐 아니라 AI 윤리, AI 교수법, 인간중심적 접근, 전문적 성장의 중요성을 강조한다. 이러한 논의는 생성형 AI 활용 연구가 단순한 사용 의도나 긍정적 인식을 넘어, 교사가 AI를 실제 교수활동에 어떻게 통합하고 성찰하는지를 측정해야 함을 시사한다.

기존 기술통합 연구의 대표적 이론인 TPACK은 기술지식, 교수지식, 내용지식이 수업 맥락에서 상호작용한다고 설명한다(Mishra and Koehler, 2006). 이 관점에서 생성형 AI는 교사를 대체하는 자동화 도구라기보다, 교사의 교수학적 판단을 보조하고 확장하는 기술적 매개체로 이해될 수 있다. 따라서 생성형 AI 교수통합경험은 AI 사용능력이나 호감이 아니라, AI가 수업목표 설정, 활동 설계, 수업 중 조정, 평가·피드백, 수업 후 성찰에 실제로 연결된 경험으로 개념화되어야 한다.

최근 AI 관련 척도 연구는 AI 자기효능감, AI 리터러시, 교사 AI 역량, AI 수용 및 인식 등을 중심으로 발전해 왔다. Wang and Chuang(2024)은 AI 자기효능감 척도를 통해 개인이 AI 기술을 사용할 수 있다고 지각하는 신념을 측정하였고, Chiu et al.(2025)은 교사의 AI 역량 자기효능감 척도를 개발하여 AI 지식, AI 교수법, AI 평가, AI 윤리, 인간중심교육, 전문적 참여의 차원을 제시하였다. Işık et al.(2024)은 교사의 AI 도구 활용 인식을 측정하는 도구를 개발하였다. 이러한 연구들은 AI 교육연구의 중요한 기반이지만, 주로 인식, 역량 신념, 수용 태도에 초점을 둔다는 점에서 실제 수업과정에서의 통합 경험을 직접 측정하는 데에는 한계가 있다.

생성형 AI의 교육적 활용은 특히 평가와 피드백 영역에서 새로운 쟁점을 낳는다. Perkins et al.(2023)은 생성형 AI를 평가에 통합할 때 허용 범위와 윤리적 기준을 명시해야 한다고 보았고, Kasneci et al.(2023)은 대규모 언어모델이 교육적 기회를 제공하는 동시에 오류, 편향, 과의존, 학습자 평가의 신뢰성 문제를 유발할 수 있다고 지적하였다. 이러한 논의는 생성형 AI 교수통합경험 척도에 평가·피드백 경험과 함께 오류·환각 가능성, 개인정보, 저작권, 표절 등 성찰적 문항이 포함되어야 하는 근거가 된다.

한편 AI 시대의 교육연구는 연구방법 측면에서도 새로운 가능성과 한계를 동시에 마주하고 있다. 실제 교사 표본을 수집하기 전 단계에서 AI 페르소나 기반 모의응답자료를 활용하면, 전문가 검토 이후 예비문항의 요인분리 가능성, 응답품질 점검 문항의 작동 여부, 문제 문항의 구조적 특성을 사전에 검토할 수 있다. 그러나 모의응답자료는 실제 인간 교사 자료를 대체할 수 없으며, 본 타당화의 최종 근거로 과도하게 해석되어서는 안 된다. 따라서 본 연구는 전문가 내용타당도, 탐색적 요인분석, 독립 합성 페르소나 표본 기반 교차타당화를 결합하여 AI 시대의 척도 예비개발 절차를 실험적으로 제시하고자 한다.

이에 본 연구는 생성형 AI 교수통합경험을 측정하기 위한 예비 척도를 개발하고, 전문가 내용타당도와 Nemotron-Personas-KOREA 기반 합성 페르소나 모의응답자료를 활용하여 문항 구조를 예비 점검하고자 한다. 본 연구는 실제 교사 표본을 활용한 최종 타당화 연구가 아니라, 실제 표본 수집 이전 단계에서 문항의 요인분리 가능성과 구조적 안정성을 사전에 검토하는 방법론적 파일럿 연구로 위치하게 된다. 연구문제는 다음과 같다. 첫째, 선행연구에 근거하여 구성한 문항은 전문가 내용타당도 검토를 통해 어떻게 정제되는가? 둘째, 합성 페르소나 기반 모의응답자료

에서 척도의 탐색적 요인구조는 어떠한가? 셋째, 탐색적 요인분석 표본과 독립적으로 생성된 별도의 합성 페르소나 기반 모의응답자료에서 20문항 4요인 구조는 예비적으로 재현되는가?

Ⅱ. 연구 방법

1. 척도 구인 및 문항 개발

본 연구에서는 <Table 1>과 같이, 생성형 AI 교수통합경험을 교사가 생성형 AI를 수업 준비, 수업 운영, 평가 · 피드백, 수업 성찰 과정에 실제로 활용하거나 그 가능성과 한계를 경험적으로 인식한 정도로 정의하였다. 이 정의는 생성형 AI 활용을 기술적 숙련도나 태도 차원이 아니라 교수활동 과정에서의 실제 통합 경험으로 파악한다는 점에 특징이 있다.

하위요인은 AI 수업준비 경험, AI 수업운영 경험, AI 평가 · 피드백 경험, AI 활용성찰 경험의 네 차원으로 설정하였다. 수업준비 경험은 수업 목표와 학습 흐름 구체화, 성취기준과 학습목표에 맞는 수업 내용 구성, 학생의 오개념 예상, 기존 수업자료 개선 경험을 포함한다. 수업운영 경험은 학생 반응에 따른 설명 · 활동 조정, 학생 질문 대응 설명 마련, 토의 · 탐구 · 문제해결 활동 운영, AI 결과 검토 경험을 포함한다. 평가 · 피드백 경험은 평가문항이나 수행과제 구성, 피드백 작성, 수행 내용 해석, 평가 결과 종합 판단 경험을 의미한다. 활용성찰 경험은 교사 역할 변화, AI 결과의 오류와 환각 가능성, 개인정보 · 저작권 · 표절 등 윤리적 쟁점, 동료 교사와의 공유를 성찰하는 경험을 포함한다.

<Table 1> Prior literature and implications for item development

Research stream	Implication for item development
TPACK	Conceptualized generative AI use as instructional practice embedded in lesson contexts rather than as simple technology use.
AI self-efficacy	Distinguished actual instructional experience from perceived ability or confidence in using AI.
Teacher AI competence	Provided a conceptual basis for organizing items around lesson preparation, implementation, assessment, ethics, and reflection.
AI acceptance	Clarified the need to develop experience-frequency items rather than attitude-, usefulness-, or intention-based items.
Generative AI ethics and assessment	Supported the inclusion of items on AI errors, hallucinations, privacy, copyright, plagiarism, and assessment judgment.

문항 개발은 TPACK, AI 자기효능감, 교사 AI 역량 자기효능감, AI 도구 활용 인식, 생성형 AI 평가 · 윤리 논의를 종합하여 이루어졌다. TPACK은 기술통합의 교수학적 관점을 제공하였고, AI 자기효능감 및 AI 역량 척도 연구는 기존 척도와 본 연구의 차별성을 설정하는 비교 기준이 되었다. 생성형 AI의 오류, 환각, 윤리, 평가 공정성 논의는 활용성찰 및 평가 · 피드백 문항 구성의 근거로 반영되었다. 초기 문항은 4요인 40문항으로 구성하였다.

문항은 생성형 AI에 대한 일반적 태도, 수용의도, 능력 신념이 아니라 최근 1년간의 교수활동 경험 빈도를 묻는 방식으로 작성하였다. 응답자는 각 문항에 대해 최근 1년 동안 해당 경험을 어느 정도 자주 해보았는지를 기준으로 응답하도록 안내받았다. 응답 범주는 1점 '전혀 경험한 적 없다', 2점 '거의 경험한 적 없다', 3점 '가끔 경험한 적 있다', 4점 '자주 경험한 적 있다', 5점 '매우 자주 경험한 적 있다'로 설정하였다. 역채점 문항은 포함하지 않았으며, 하위요인 점수는 각 하위요인에 속한 문항의 평균으로 산출하였다. 전체 점수는 최종 20문항의 평균으로 산출하였다.

2. 전문가 내용타당도 검토

초기 40문항은 전문가 12인의 검토를 거쳐 정

제하였다. 전문가 패널은 교육공학, AI 교육, 교육평가, 교사교육, 교육과정, 컴퓨터교육, AI 리터러시, 교과교육 관련 전문가로 구성하였다. 전문가들은 각 문항에 대해 적절성, 명확성, 대표성, 중복성을 4점 척도로 평정하고, 요인 적합성 및 수정 의견을 제시하였다. 내용타당도 검토에서는 I-CVI, 중복성 평정, 전문가 서술 의견을 종합하였으며, 이는 전문가 패널 검토와 CVI를 활용한 문항 수준 내용타당도 검토 절차에 근거한 것이다(Davis, 1992; Polit et al., 2007).

전문가 검토 결과, 초기 40문항 중 28문항이 예비조사용 문항으로 유지되었다. 유지된 28문항의 적절성 및 대표성 I-CVI는 모두 1.000이었고, 명확성 I-CVI 평균은 .935로 나타났다. 중복성 평균은 1.423이었다. 중복성은 1점 '전혀 중복되지 않음'에서 4점 '매우 중복됨'까지의 척도로 평정되었으므로, 1.423은 전반적으로 문항 간 중복성이 낮고 내용 변별성이 비교적 확보되었음을 의미한다. 다만 일부 문항에서 내용 중복과 요인 경계 모호성이 지적되어 12문항을 삭제하였다.

전문가의 서술 의견은 크게 세 방향으로 수렴하였다. 첫째, AI에 대한 일반적 호감, 필요성 인식, 활용 의도보다 실제 교수활동 과정에서의 경험 빈도를 측정해야 한다는 의견이 제시되었다. 둘째, 일부 문항은 기존 문항과 내용상 중복되거나 다른 요인으로 교차적재될 가능성이 있다는 의견이 제시되었다. 셋째, 응답방식이 경험의 정도가 아니라 최근 1년간의 경험 빈도를 묻는다는 점이 명확히 제시되어야 한다는 의견이 제시되었다. 이에 따라 태도·인식 중심 문항, 중복 문항, 요인 경계가 모호한 문항을 삭제 또는 재검토하고, 요인당 7문항씩 총 28문항을 예비조사용 문항으로 유지하였다. 이러한 절차는 문항의 통계적 적합성을 검토하기에 앞서, 각 문항이 '교수활동에 기반한 실제 경험'을 묻는지 확인하기 위한 내용타당도 판단 과정이었다.

3. Nemotron-Personas-KOREA 기반 모의응답자료

Nemotron-Personas-KOREA는 성별, 연령, 혼인상태, 교육수준, 직업, 지역 등 한국 사회의 인구통계학적 속성을 포함하는 합성 페르소나 데이터셋이다(Kim et al., 2026). 데이터셋 설명에 따르면 KOSIS, 대법원, 국민건강보험공단, 한국농촌경제연구원, NAVER Cloud 등 공식 또는 준공식 자료를 바탕으로 한국 인구의 사회·인구학적 다양성을 반영하도록 구축되었다. 다만 합성 페르소나의 주변분포 정렬이 실제 결합분포의 완전한 재현을 보장하는 것은 아니므로(Bae, 2026), 본 연구에서는 이를 실제 교사 표본의 대체자료가 아니라 예비문항의 응답구조를 점검하기 위한 모의응답자료로 제한하여 활용하였다.

합성 페르소나 기반 모의응답은 Python 스크립

<Table 2> Expert review issues and item refinement decisions in content validation

Expert review issue	Main expert opinion	Item refinement decision
Construct clarity	Items should measure instructional experience rather than general attitudes, perceived importance, or intention to use AI.	Attitude- or perception-oriented items were deleted.
Item redundancy	Some items overlapped substantially with other items within the same factor.	Redundant items were removed before the EFA stage.
Factor boundary	Some items were more closely related to ethical reflection or digital literacy than to lesson implementation or assessment.	Items with ambiguous factor membership were deleted or reconsidered.
Response format	Experts emphasized that the scale should ask about frequency of experience, not level of agreement.	A five-point experience-frequency response format was adopted.
Preliminary item pool	Experts supported retaining enough items for later statistical refinement.	Seven items per factor were retained, resulting in 28 preliminary items.

트를 활용하여 생성 · 정리하였다. 먼저 Nemotron-Personas-KOREA에서 제공하는 성별, 연령, 교육수준, 직업, 지역 등의 기본 인구통계학적 속성을 불러온 뒤, 본 연구 목적에 맞게 교직경력, 학교급, 교과, 디지털 활용수준, 생성형 AI 사용경험, AI 관련 연수경험, 학교의 AI 활용 분위기, 수업혁신 성향, 기술불안 등의 교육맥락 속성을 추가로 부여하였다. 이후 각 페르소나의 속성과 최근 1년간의 생성형 AI 활용 경험 수준을 반영한 응답 생성 프롬프트를 구성하고, 대규모 언어모델 기반 응답 생성 절차를 통해 문항별 1~5점 경험빈도 응답을 생성하였다. 생성된 응답은 Python을 통해 정리 · 점검한 뒤 CSV와 Excel 형식의 분석자료로 변환하였다. 응답 생성 과정에서는 하위요인명이나 기대되는 4요인 구조를 제시하지 않았으며, 문항은 무작위 순서로 제시하였다. 또한 EFA용 표본과 CFA용 표본은 서로 독립적으로 생성하여 동일자료 기반 재검토 가능성을 줄이고자 하였다. 구체적인 문항, 문항 매핑표, 모의응답자료, 코드북 및 분석 요약자료는 Zenodo에 공개하였다(Choi and Won, 2026).

탐색적 요인분석에는 전체 560건의 모의응답자료 중 응답품질 확인 문항에서 지시된 응답을 선택하지 않은 11건(2.0%)을 제외한 549건을 사용하였다. 확인적 요인분석에는 탐색적 요인분석 표본과 완전히 독립적으로 생성된 별도의 합성 페르소나 표본 600건을 사용하였다. 이는 동일자료에서 문항을 선별하고 동일자료로 다시 확인하는 한계를 줄이고, EFA에서 도출된 문항 구조가 독립 합성 페르소나 기반 모의응답자료에서도 예비적으로 재현되는지를 점검하기 위한 절차이다.

두 표본은 동일한 척도 문항을 사용하되, 분석 단계에서 서로 독립된 표본으로 구분하였다. EFA 표본은 문항 정제와 요인구조 탐색에 사용하였고, CFA 표본은 EFA에서 도출된 20문항 4요인 구조가 별도 자료에서도 유지되는지를 예비적으로 점검하는 데 사용하였다. 따라서 본 연구의 CFA 결과는 실제 교사 표본에 대한 최종 타당화 근거가 아니라, 독립 합성 페르소나 표본 내에서 문항 구조의 재현 가능성을 검토한 보조적 근거로 해석하였다.

4. 분석 방법

분석은 SPSS와 AMOS를 활용하였다. 먼저 전문가 검토를 거친 28문항에 대해 탐색적 요인분석을 실시하여 문항의 주적재량, 교차적재, 요인별 내용 대표성을 검토하였다. 생성형 AI 교수통합경험의 하위요인들은 수업 준비, 운영, 평가 · 피드백, 활용성찰이라는 교수활동의 연속적 과정에 위치하므로 서로 독립적이라고 보기 어렵다. 이에 본 연구에서는 요인 간 상관을 허용하는 사각회전 방식이 적절하다고 판단하고, 최대우도법(Maximum Likelihood)과 직접 오블리민(Direct Oblimin) 회전을 적용하였다.

탐색적 요인분석에서 문항 선별은 패턴계수의 크기, 교차적재 여부, 요인별 문항 수의 균형, 이론적 대표성을 함께 고려하였다. 이후 독립적으로 생성된 600건의 합성 페르소나 표본을 대상으로 단일요인 모형, 4요인 상관모형, 2차요인 모형의 확인적 요인분석을 실시하였다. 모형적합도는 χ^2(df), CMIN/df, CFI, TLI, RMSEA, SRMR을 중심으로 판단하였으며, 4요인 상관모형을 기준으로 합성신뢰도(CR), 평균분산추출(AVE), Fornell-Larcker 기준, HTMT를 보조적으로 검토하여 수렴타당도와 판별타당도를 점검하였다.

확인적 요인분석에서는 기본모형의 구조를 우선 검토하였으며, 적합도를 높이기 위한 사후적 오차공분산은 추가하지 않았다. 이는 모의응답자료에서 얻어진 결과가 연구자의 사후적 수정지수 선택에 의해 과도하게 개선되는 것을 피하기 위한 것이다. 또한 4요인 상관모형과 2차요인 모형을 함께 비교하여 하위요인의 구분 가능성과 상위 구성개념으로서의 해석 가능성을 검토하되, 본 연구의 결과는 실제 교사 표본에 대한 최종 타당화가 아니라 합성 페르소나 기반 모의응답자

료 내 문항 구조의 예비 점검으로 제한하여 해석하였다.

Ⅲ. 연구 결과

1. 탐색적 요인분석 및 신뢰도

전문가 내용타당도 검토를 통해 40문항 중 28문항이 예비조사용 문항으로 유지되었다. 28문항 예비 탐색 단계에서 KMO는 .951, Bartlett의 구형성 검정은 χ^2=9453.023(df=378, p<.001)로 나타나 요인분석에 적합한 자료구조가 확인되었다. 최종 문항 선별에서는 선행연구와 전문가 검토에 근거한 4요인 구조와의 이론적 일치성, 주적재량의 크기와 교차적재 가능성, 요인별 문항 수의 균형, 문항 내용의 대표성과 해석 명료성을 종합적으로 고려하였다. 특히 통계적으로 적재량이 양호하더라도 문항 내용이 태도, 효능감, 일반적 인식으로 해석될 가능성이 있거나 다른 요인과 의미상 경계가 불분명한 경우에는 제외하였다. 이에 따라 전문가 검토 후 유지된 28문항 중 요인당 5문항, 총 20문항을 최종 후보 문항으로 선정하였다.

<Table 3>과 같이, 최종 20문항에 대한 탐색적 요인분석에서도 KMO=.932, Bartlett의 구형성 검정은 χ^2=7588.678 (df=190, p<.001)로 나타났다. 직접 오블리민 회전 결과, 네 하위요인에 해당하는 문항들이 이론적으로 의도한 구조와 대체로 일치하였다. 전체 신뢰도는 Cronbach's α=.920으로 나타났고, 하위요인별 신뢰도는 .911~.917 범위로 안정적이었다. 이는 최종 문항이 통계적 간

<Table 3> Exploratory factor analysis and reliability of the retained items

Factor	Item	Abbreviated item content	Pattern coefficient				α
			1	2	3	4	
Preparation	P1	Lesson objectives and flow	0.033	0.023	0.780	0.006	.911
	P2	Curriculum-aligned lesson content	0.056	0.033	0.795	-0.037	
	P3	Anticipating misconceptions	0.043	0.022	0.799	-0.004	
	P4	AI prompts for lesson purposes	0.071	0.044	0.773	-0.004	
	P5	Revising instructional materials	0.032	0.040	0.796	-0.009	
Implementation	O1	Adjusting instruction to student responses	0.812	0.029	0.089	-0.030	.917
	O2	Additional explanations for questions	0.831	-0.015	-0.038	0.095	
	O3	Discussion, inquiry, and problem-solving activities	0.811	0.047	-0.027	0.023	
	O4	Reviewing AI outputs with students	0.845	0.004	-0.032	-0.004	
	O5	Supporting participation across levels	0.767	-0.006	0.078	0.064	
Assessment	F1	Assessment items and performance tasks	0.013	0.025	0.001	0.800	.912
	F2	Feedback statements by student level	0.019	0.003	-0.027	0.848	
	F3	Interpreting learning outcomes	0.005	-0.014	-0.007	0.807	
	F4	Reviewing alignment with lesson objectives	0.023	0.054	0.002	0.791	
	F5	Judging AI-related results with other evidence	-0.011	0.047	0.034	0.805	
Reflection	R1	Reflection on lesson preparation	-0.010	0.822	-0.065	0.086	.915
	R2	Teacher role as learning facilitator	-0.024	0.849	-0.002	0.016	
	R3	AI errors and hallucinations	-0.025	0.791	0.007	0.088	
	R4	Privacy, copyright, and plagiarism	0.038	0.823	-0.025	0.018	
	R5	Sharing AI-use experiences with colleagues	0.052	0.803	-0.017	0.001	

Note. Maximum likelihood extraction and direct oblimin rotation were applied. KMO=.932, Bartlett's χ^2=7588.678(df=190, p<.001), total α=.920. Factor 1=Implementation, Factor 2=Reflection, Factor 3=Preparation, Factor 4=Assessment. Full Korean item statements, item mappings, codebook, and simulated response data are available in the Zenodo repository(doi:10.5281/zenodo.21911568).

명성과 내용타당도의 균형을 고려하여 선별되었음을 보여준다.

활용성찰 요인(R1~R5)이 독립된 요인으로 도출된 점은 생성형 AI 교수통합경험이 수업자료 생성이나 평가문항 개발 경험만으로 설명되지 않음을 보여준다. 활용성찰 문항은 수업 준비 방식에 대한 성찰, 교사 역할 변화, AI 결과의 오류와 환각 가능성, 개인정보 · 저작권 · 표절 등 윤리적 쟁점, 동료 교사와의 경험 공유를 포함한다. 이는 생성형 AI 활용에서 교사의 전문적 판단, 윤리적 책임, 평가 공정성이 중요하다는 논의와 연결된다(UNESCO, 2023; Perkins et al., 2023). 따라서 활용성찰 경험은 AI 활용 과정에서 수반되는 위험과 책임을 교수실천 속에서 점검하는 핵심 구성요인으로 해석할 수 있다.

2. 독립 표본 기반 확인적 요인분석

확인적 요인분석은 탐색적 요인분석에 사용된 N=549 표본과 완전히 독립적으로 생성된 별도의 합성 페르소나 표본(N=600)을 사용하여 수행하였다. 이는 EFA에서 선별된 20문항 4요인 구조가 독립 합성 페르소나 기반 모의응답 자료에서도 예비적으로 재현되는지를 점검하기 위한 절차이다. 본 연구에서는 단일요인 모형, 4요인 상관모형, 2차요인 모형을 경쟁모형으로 설정하고, 각 모형의 적합도와 표준화 추정치를 비교하였다.

<Table 4>와 같이, 단일요인 모형은 χ^2(170) =1838.599, CMIN/df=10.815, CFI=.585, TLI=.536, RMSEA=.128, SRMR=.157로 나타나 적합도 기준을 충족하지 못하였다. 이는 생성형 AI 교수통합경험이 하나의 일반요인만으로 설명되기 어렵다는 것을 시사한다. 반면 4요인 상관모형은 χ^2 (164)=191.309, CMIN/df=1.167, CFI=.993, TLI=.992, RMSEA=.017, SRMR=.026으로 양호한 적합도를 보였다. 2차요인 모형도 χ^2 (166)=200.245, CMIN/df=1.206, CFI=.991, TLI=.990, RMSEA=.019, SRMR=.023으로 수용 가능한 적합도를 보였다. 다만 본 연구의 CFA 자료는 실제 교사 응답이 아니라 합성 페르소나 기반 모의응답자료이므로, CFI와 TLI가 높게 나타난 결과는 합성자료의 정제된 응답 패턴이나 생성 과정의 규칙성을 일부 반영했을 가능성이 있다. 따라서 본 결과는 실제 교사 모집단에 대한 최종 검증이 아니라 예비문항 구조의 사전 점검 결과로 제한하여 해석할 필요가 있다.

독립 합성 페르소나 표본의 CFA 결과에서 4요인 상관모형과 2차요인 모형은 모두 양호한 적합도를 보였다. 다만 본 연구에서는 이론적으로 구분한 네 하위요인의 해석 가능성과 모형 간 적합도 차이를 종합적으로 고려하여 4요인 상관모형을 기본 측정모형으로 해석하고, 2차요인 모형은 전체 점수 활용 가능성을 보여주는 보조 근거로 보았다.

4요인 상관모형의 표준화 요인부하량은 AI 수업준비 경험 .624~.774, AI 수업운영 경험 .657~.720, AI 평가 · 피드백 경험 .660~.754, AI 활용성찰 경험 .592~.741로 나타났다. 요인 간 상관은 .332~.523 범위로 나타나 하위요인들이 서로 관련되어 있으나 과도하게 중복되지는 않는 것으로 확인되었다. 수렴타당도와 판별타당도도 보조적으로 검토한 결과, 1차 요인의 합성신뢰도(CR)는 .809~.828로 수용 가능한 수준이었다. 그러나 평균분산추출(AVE)은 .460~.491로 .50 기준에 근

<Table 4> CFA fit indices based on an independent synthetic persona sample (N=600)

Model	χ^2 (df)	CMIN/df	CFI	TLI	RMSEA	SRMR
One-factor	1838.599 (170)	10.815	.585	.536	.128	.157
Correlated four-factor	191.309 (164)	1.167	.993	.992	.017	.026
Second-order	200.245 (166)	1.206	.991	.990	.019	.023

접하였으나 완전히 충족하지는 못하였다. 이는 생성형 AI 교수통합경험 문항이 실제 교사 맥락에서 추가적으로 정교화될 필요가 있음을 시사한다. 한편 Fornell-Larcker 기준을 검토한 결과, 각 요인의 AVE 제곱근은 .678~.701로 요인 간 상관계수(.332~.523)를 상회하였다. 또한 HTMT 값은 .329~.527로 일반적 기준인 .85를 초과하지 않았다. 따라서 합성 페르소나 기반 모의응답자료 내에서는 네 하위요인의 판별 가능성이 확인되었으나, 실제 교사 표본을 통한 수렴·판별타당도 재검증이 필요하다.

2차요인 모형에서의 표준화 부하량은 <Table 5>와 같이, AI 수업준비 경험 .582, AI 수업운영 경험 .686, AI 평가·피드백 경험 .752, AI 활용성찰 경험 .569로 나타났으며, 모든 경로는 통계적으로 유의하였다. 이는 네 하위요인이 상위 개념인 생성형 AI 교수통합경험과 관련될 가능성을 보여준다. 다만 4요인 상관모형의 적합도가 가장 양호하였고, 활용성찰 경험이 윤리적 판단과 전문적 성찰을 포함하는 상대적으로 구분되는 차원이라는 점을 고려하여, 본 연구에서는 전체점수 해석보다 하위요인별 해석을 우선하는 것이 적절하다고 보았다.

<Table 5> Standardized second-order factor loadings based on the independent CFA sample (N=600)

Subfactor	β
AI lesson preparation	.582
AI lesson implementation	.686
AI assessment and feedback	.752
AI use reflection	.569

Note. All paths were statistically significant at $p<.001$.

이상의 결과를 종합하면, 독립 합성 페르소나 표본 기반 CFA에서 단일요인 구조는 지지되지 않았고, 4개 하위요인이 구분되는 다차원 구조가 예비적으로 확인되었다. 2차요인 모형도 수용 가능한 적합도를 보였으므로, 본 척도는 하위요인별 점수 해석을 기본으로 하되 전체 생성형 AI 교수통합경험 점수를 보조적으로 활용할 가능성이 있다. 다만 일부 AVE 지표가 기준에 미달하였고 분석자료가 합성 페르소나 기반 모의응답자료라는 점을 고려할 때, 실제 교사 표본을 활용한 후속 검증 없이는 최종 타당화 근거로 해석하기 어렵다.

Ⅳ. 결 론

본 연구는 생성형 AI 교수통합경험 척도를 예비개발하고, 전문가 내용타당도와 합성 페르소나 기반 모의응답자료를 활용하여 문항 구조를 사전 점검하였다. 연구 결과, 최종 후보 척도는 AI 수업준비 경험, AI 수업운영 경험, AI 평가·피드백 경험, AI 활용성찰 경험의 네 하위요인과 총 20문항으로 구성되었다. 다만 본 연구의 분석자료는 실제 교사 응답이 아니라 합성 페르소나 기반 모의응답자료이므로, 연구 결과는 실제 교사 모집단에 대한 최종 타당화 근거가 아니라 예비문항 구조의 가능성을 점검한 보조적 근거로 해석되어야 한다.

이론적으로 본 연구는 생성형 AI 교수통합경험을 일반적 AI 수용 태도, AI 자기효능감, AI 역량 인식과 구분되는 교수실천 경험으로 개념화하였다. 기존 척도들이 주로 AI를 사용할 수 있다는 신념, AI가 유용하다는 인식, AI를 수용하려는 의도를 측정해 왔다면, 본 척도는 최근 1년간 교사가 생성형 AI를 수업 준비, 수업 운영, 평가·피드백, 활용성찰 과정에 실제로 연결해 본 경험 빈도를 측정하도록 설계되었다. 이는 TPACK의 기술통합 관점을 생성형 AI 시대의 교수활동 경험 차원으로 조작화한 시도라는 점에서 의의가 있다.

특히 활용성찰 경험이 독립된 하위요인으로 도출된 점은 생성형 AI 교수통합경험이 수업자료

생성이나 평가문항 개발의 효율성만으로 설명될 수 없음을 보여준다. 교사는 AI 결과의 오류와 환각 가능성을 검토하고, 개인정보와 저작권, 표절 문제를 고려하며, 생성형 AI 활용 과정에서 교사의 역할 변화를 성찰해야 한다. 따라서 활용 성찰 경험은 생성형 AI 활용의 부가적 요소가 아니라, AI 기반 교수활동에서 전문적 판단과 윤리적 책임을 확인하는 핵심 차원으로 볼 수 있다. 후속연구에서는 실제 현직 교사를 대상으로 환각으로 인한 수업 오류, 부적절한 피드백 생성, 프롬프트 실패, 개인정보 처리 곤란 등 구체적인 오류 및 실패 경험을 질적・양적으로 발굴하여 문항의 스펙트럼을 확장할 필요가 있다.

방법론적으로 본 연구는 전문가 내용타당도 검토, 합성 페르소나 기반 EFA, 독립 합성 페르소나 표본 기반 CFA를 순차적으로 결합하여 실제 인간 표본 수집 이전 단계에서 문항 구조를 예비 점검하는 절차를 제시하였다. 특히 EFA와 CFA 표본을 독립적으로 생성하고, 최종 문항, 문항 매핑표, 모의응답자료, 코드북 및 분석 요약자료를 Zenodo에 공개함으로써 합성자료 활용 연구의 투명성과 재현가능성을 높이고자 하였다. 그러나 이러한 접근은 실제 교사 응답자료를 대체하는 것이 아니라, 예비문항의 구조적 가능성을 탐색하고 문제 문항을 선별하기 위한 보조적 절차로 제한적으로 활용되어야 한다.

본 연구의 한계는 다음과 같다. 첫째, 요인분석 자료가 실제 교사 응답이 아니라 Nemotron-Personas-KOREA 기반 합성 페르소나 모의응답자료라는 점에서, 본 연구 결과를 실제 교사 모집단에 일반화하기 어렵다. 둘째, AVE가 .50 기준에 완전히 도달하지 못한 점은 실제 교사 표본에서 문항 재검토와 수렴타당도 보완이 필요함을 시사한다. 셋째, 본 연구는 내적 구조의 예비 점검에 초점을 두었으므로 AI 교수효능감, AI 리터러시, 기술수용, 실제 AI 활용행동 등 외부 준거와의 관계를 검토하지 못하였다. 넷째, 학교급, 교과, 경력, AI 활용경험 수준에 따른 측정동일성도 분석하지 못하였다. 후속연구에서는 실제 교사 표본을 대상으로 교차타당화, 준거타당도, 수렴・판별타당도, 측정동일성 검증을 수행할 필요가 있다.

직업・전문교육 맥락에서 본 척도는 실습, 안전, 직무상황 기반 수업에서 생성형 AI가 어떻게 활용되는지를 탐색하는 기초 도구로 확장될 수 있다. 특히 수산・해양교육과 같이 산업기술 변화와 현장 실습이 결합되는 분야에서는 생성형 AI가 사례 기반 수업자료 개발, 실습 전 사전학습, 수행평가 피드백, 직무문제 해결 활동 등에 활용될 가능성이 있다. 다만 본 연구는 해당 분야 교사를 대상으로 실증검증을 수행한 것은 아니므로, 분야 특수성을 반영한 실제 교사 표본 연구가 후속적으로 이루어질 필요가 있다.

References

Bae JH(2026). Marginal alignment does not guarantee joint-distribution fidelity: An official-reference audit of Nemotron-Personas-Korea with cross-locale replication. arXiv preprint. http://dx.doi.org/10.48550/arXiv.2606.12433.

Chiu TKF, Ahmad Z and Coban M(2025). Development and validation of teacher artificial intelligence (AI) competence self-efficacy (TAICS) scale. Education and Information Technologies, 30, 6667~6685. http://dx.doi.org/10.1007/s10639-024-13094-z.

Choi HK and Won HH(2026). Generative AI Instructional Integration Experience Scale: Final Items, Mapping Table, Codebook, and Synthetic Persona-Based Simulated Response Data. Zenodo. http://dx.doi.org/10.5281/zenodo.21911568.

Davis LL(1992). Instrument review: Getting the most from a panel of experts. Applied Nursing Research, 5(4), 194~197. http://dx.doi.org/10.1016/S0897-1897(05)80008-4.

Işık S, Çakır R and Korkmaz Ö(2024). Teachers' perception scale towards the use of artificial intelligence tools in education. Participatory Educational Research, 11(6), 80~94.

http://dx.doi.org/10.17275/per.24.95.11.6.

Kasneci E, Sessler K, Kuchemann S, Bannert M, Dementieva D, Fischer F, Gasser U, Groh G, Gunnemann S, Hullermeier E, Krusche S, Kutyniok G, Michaeli T, Nerdel C, Pfeffer J, Poquet O, Sailer M, Schmidt A, Seidel T, Stadler M, Weller J, Kuhn J and Kasneci G(2023). ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences, 103, 102274. http://dx.doi.org/10.1016/j.lindif.2023.102274.

Kim HW, Ryu JH, Lee JH, Ryu HG, Praveen K, Prayaga S, Thadaka K, Jennings W, Sadeghi B, Sharabiani A, Choi YJ and Meyer Y(2026). Nemotron-Personas-Korea: Synthetic Personas Aligned to Real-World Distributions for Korea. Hugging Face Dataset, Retrieved from https://huggingface.co/datasets/nvidia/Nemotron-Personas-Korea on August 13, 2026.

Lawshe CH(1975). A quantitative approach to content validity. Personnel Psychology, 28(4), 563~575. http://dx.doi.org/10.1111/j.1744-6570.1975.tb01393.x.

Mishra P and Koehler MJ(2006). Technological pedagogical content knowledge: A framework for teacher knowledge. Teachers College Record, 108(6), 1017~1054. http://dx.doi.org/10.1111/j.1467-9620.2006.00684.x.

Perkins M, Furze L, Roe J and MacVaugh J(2023). The AI Assessment Scale (AIAS): A framework for ethical integration of generative AI in educational assessment. arXiv preprint. http://dx.doi.org/10.48550/arXiv.2312.07086.

Polit DF, Beck CT and Owen SV(2007). Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. Research in Nursing & Health, 30(4), 459~467. http://dx.doi.org/10.1002/nur.20199.

UNESCO(2023). Guidance for generative AI in education and research. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000386693 on August 13, 2026.

UNESCO(2024). AI competency framework for teachers. UNESCO, Retrieved from https://unesdoc.unesco.org/ark:/48223/pf0000391104 on August 13, 2026.

Wang YY and Chuang YW(2024). Artificial intelligence self-efficacy: Scale development and validation. Education and Information Technologies, 29, 4785~4808. http://dx.doi.org/10.1007/s10639-023-12015-w.

• Received : 14 August, 2026
• Revised : 07 September, 2026
• Accepted : 11 September, 2026