미래사회
Journal of Future Society
2026. Vol. 17, No. 3, 285~299.
https://doi.org/10.22987/jifso.2026.17.3.285

생성형 AI 기반 UI/UX 디자인 교육의 평가 방법 연구:
결과물 중심 평가와 과정 중심 평가의 비교를 중심으로

조 진 경*

연성대학교

본 연구는 생성형 AI 기반 UI/UX 디자인 교육에서 결과물 중심 평가가 학습자의 실제 역량과 AI 활용 과정을 충분히 반영하는지 검토하였다. UI/UX 디자인 실무·교육·연구 경력을 가진 전문가 3인이 학습자 5인의 최종 결과물을 대상으로 1차 블라인드 평가를 실시한 뒤, 프롬프트 수정 이력과 작업 과정 자료를 공개하여 2차 과정 평가를 진행하였다. 그 결과 사용자 적합성과 시각적 위계는 비교적 높게 평가된 반면, 사용성·상호작용·종합 완성도는 낮게 나타났다. 과정 공개 후 프롬프트 구체화와 판단의 타당성은 대체로 적절했으나, 수정의 효과, 전문가 판단과의 관계, AI 의존성에서는 부분 적절 또는 미흡 평가가 다수 확인되었다. 특히 학습자는 AI 결과의 선택·수정을 의사결정의 완결로 인식한 반면, 전문가는 요구사항 충족, 데이터 정합성, 실제 사용성 및 최종 품질까지 평가하였다. 이에 본 연구는 문제 정의부터 최종 품질 보증까지의 8단계 통합 검증 프로세스를 제안하고, 생성형 AI 기반 디자인 교육에서 결과물 평가와 과정 평가를 병행하는 이중 평가체계의 필요성을 제시한다.

주요어: 생성형 AI, UI/UX 디자인 교육, 결과물 평가, 과정 평가, 전문가 평가

* 주저자: 조진경/연성대학교 시각디자인과 조교수/경기도 안양시 만안구 양화로 37길 34
/Tel: 031-441-1374/E-mail: ivysoul@yeonsung.ac.kr

Ⅰ. 서론

1. 연구의 필요성

생성형 AI는 UI/UX 디자인 실습에서 짧은 시간 안에 여러 대안을 제시하여 학습자의 아이디어 발상과 화면 구성을 지원한다. 그러나 산출물의 시각적 완성도가 높아지는 현상과 학습자의 디자인 역량이 향상되는 현상은 동일한 의미가 아니다. 학습자가 생성 결과를 그대로 채택했는지, 문제를 인식하고 프롬프트와 화면을 수정했는지에 따라 동일한 완성도의 결과물이라도 그 안에 담긴 학습자의 판단 수준은 달라질 수 있다. 그럼에도 생성형 AI 활용 디자인 교육의 성과는 여전히 결과물의 완성도나 만족도를 기준으로 평가되는 경우가 많아, 완성된 화면 이면에 있는 프롬프트 수정 과정과 판단 근거는 평가에 반영되기 어렵다.

이 같은 한계는 교육평가 방법론의 문제로 이어진다. 결과물만을 평가 대상으로 삼을 경우, 학습자가 AI 제안을 비판적으로 검토했는지, 요구사항과 데이터의 정합성을 확인했는지, 실제 기기에서 상호작용을 검증했는지는 확인할 수 없다. 반대로 과정 자료만을 근거로 평가할 경우에는 결과물의 실질적 완성도와 사용성을 놓칠 위험이 있다. 특히 생성형 AI가 개입한 산출물에서는 기존의 결과물 중심 평가 기준(사용성, 시각적 위계, 일관성 등)만으로는 학습자가 AI의 제안을 그대로 수용했는지, 비판적으로 재검토했는지, AI가 스스로 제시한 '완료' 신호를 재검증했는지를 구별할 수 없다는 문제가 새롭게 발생한다. 이는 전통적 디자인 산출물 평가에는 존재하지 않던, 생성형 AI 환경에 특화된 판단 지점이다. 이에 본 연구는 결과물과 과정을 이분법적으로 구분하는 데 그치지 않고, 프롬프트 구체화·판단의 타당성·수정의 효과·AI 의존성·전문가 판단과의 관계·최종 의사결정이라는 생성형 AI 활용 과정에 특화된 6개 세부 판단 행동을 구체적인 평가 항목으로, 조작적으로 정의하여 제시한다(Ⅱ장 2. 측정도구, <표 4> 참조). 따라서 생성형 AI를 활용한 디자인 교육에서는 결과물 중심 평가와 과정 중심 평가를 함께 실시하고, 두 평가가 일치하는 지점과 불일치하는 지점을 구체적으로 확인하는 이중 평가체계에 대한 논의가 필요하다.

이 문제는 디자인 교육 내부의 쟁점에 머무르지 않는다. 미래사회에서 생성형 AI는 디자인뿐 아니라 문서 작성, 데이터 분석, 콘텐츠 제작 등 다양한 영역의 산출물 생성에 관여하며, 그 결과를 비판적으로 검증하지 않고 그대로 신뢰하여 채택할 때 발생하는 오류는 이미 여러 실무 영역에서 보고되고 있다. 따라서 AI가 제시한 결과를 어떤 절차로 검증하고 최종 의사결정에 이르는지를 구조화하는 것은 특정 전공 교육을 넘어, 미래사회가 요구하는 AI 리터러시 및 검증 역량 교육의 공통 과제로 볼 수 있다.

이에 본 연구는 생성형 AI 기반 UI/UX 디자인 교육에서 산출된 학습자 결과물에 대해 전문가 3인이 실시한 1차 블라인드 결과물 평가와 2차 과정 평가를 비교하여, 결과물 평가와 과정 평가가 일치·불일치하는 특성을 분석하고, 이를 통합한 최종 검증 및 의사결정 프로세스를 제안하는 것을 목적으로 한다. 본 연구의 목적을 달성하기 위하여 다음과 같은 연구 문제를 설정하였다.

RQ1. 전문가의 블라인드 결과물 평가에서 나타나는 주요 UI/UX 평가 기준은 무엇인가?

RQ2. 학습자의 프롬프트 수정과 작업 과정을 공개한 후 전문가 평가는 어떻게 변화하는가?

RQ3. 결과물 평가와 과정 평가가 일치하거나 불일치하는 사례는 어떠한 특성을 보이는가?

RQ4. 생성형 AI 기반 UI/UX 디자인 교육에서 결과물 평가와 과정 평가를 어떻게 통합할 수 있는가?

2. 이론적 배경

1) 생성형 AI 산출물 평가의 두 관점: 결과물 중심 평가와 과정 중심 평가

전통적인 사용성 평가에서 Nielsen(1994)의 휴리스틱 평가는 소수의 전문가가 정해진 원칙에 따라 인터페이스의 문제를 판정하는 결과물 중심 평가 방법으로 널리 활용되어 왔다. 생성형 AI가 개입한 디자인 산출물에 대해서도 결과물 중심 평가는 여전히 유효하지만, AI가 제안한 대안 중 무엇을 채택하고 무엇을 수정했는지에 대한 학습자의 판단 과정은 설명하지 못한다. Duan 등(2024)은 GPT-4를 활용한 UI 휴리스틱 피드백이 미세한 오류 탐지, 텍스트 개선, UI 의미 해석에 유용할 수 있으나, 반복 과정에서 피드백의 유용성이 감소하고 불완전한 제안도 나타날 수 있음을 보고하였다. 이는 AI 기반 평가 역시 최종 기준이 될 수 없으며, 전문가와 학습자의 판단을 보조하는 자료로 활용되어야 함을 시사한다.

White 등(2023)은 프롬프트 패턴을 역할, 맥락, 제약조건, 출력 형식 및 평가 기준으로 구조화하면 결과의 통제 가능성이 높아진다고 설명하였다. 그러나 프롬프트가 구체적이라고 해서 디자인 품질이 자동으로 확보되는 것은 아니며, 생성 결과가 사용자 요구와 휴리스틱을 충족하는지 판단하기 위해서는 디자인 원칙과 맥락 지식이 필요하다. Amershi 등(2019)은 인간-AI 상호작용에서 사용자가 AI의 행동을 이해하고, 필요할 때 수정·거부·피드백할 수 있도록 하는 18개 설계 지침을 제안하였다. Shneiderman(2022) 역시 인간 통제와 자동화를 상호 배타적으로 보지 않고, 높은 수준의 인간 통제를 유지하면서 자동화를 활용하는 인간 중심 AI 프레임워크를 제시한다. 본 연구는 위의 논의를 바탕으로, 완성된 결과물의 품질을 판정하는 결과물 중심 평가와, 그 결과물에 이르기까지의 판단과 수정 과정을 검토하는 과정 중심 평가를 구분하고, 두 평가를 비교함으로써 생성형 AI 산출물에 대한 인간 검증자의 승인 절차를 구체화하고자 한다. 이러한 결과물-과정 이분법은 교육평가 이론에서 오랫동안 논의되어 온 총괄평가와 형성평가의 구분과 맥을 같이한다. Black과 Wiliam(1998)은 형성평가가 학습이 진행되는 과정에서 수집된 증거를 통해 학습자의 이해 수준과 다음 단계의 조정 방향을 확인하는 데 목적이 있음을 강조하였으며, 이는 완성된 산출물만을 판정하는 총괄평가와 구별되는 지점이다. 수행평가 역시 결과물 자체보다 그 결과에 도달하는 수행 과정에서 나타나는 지식과 기능의 적용 양상을 평가 대상으로 삼는다는 점에서, 본 연구가 제안하는 과정 중심 평가와 이론적으로 맞닿아 있다. 다만 생성형 AI가 개입한 상황에서는 학습자의 수행이 AI와의 상호작용을 매개로 이루어지므로, 기존의 수행평가·형성평가가 전제하는 '학습자 단독 수행'이라는 틀만으로는 프롬프트 조정, AI 제안의 수용·거부 판단과 같은 행동을 온전히 포착하기 어렵다. 이에 본 연구는 총괄평가 성격의 결과물 중심 평가와 형성·수행평가 성격의 과정 중심 평가를 동일한 산출물에 대해 순차적으로 실시하고, 두 평가체계가 일치하는 지점과 불일치하는 지점을 비교함으로써 각 평가체계가 단독으로는 포착하지 못하는 판단 공백을 상호 보완적으로 확인할 수 있다는 점에서 이중 평가체계의 타당성을 확보하고자 하였다.

이처럼 역량을 직접적인 측정 대상이 아니라 해석적 준거틀로 활용하는 접근은 최근 국내 사이버대학 연구에서도 확인된다. 박종선(2026)은 UNESCO 학습자 역량 프레임워크를 학습자 역량을 직접 점수화하기 위한 도구가 아니라, AI 활용이 가능한 학습 환경에서 학습자의 역량이 학습 활동과 성과로 전이되는 구조를 해석하기 위한 이론적 준거틀로 활용하였으며, 구조방정식 모형을 통해 이러한 접근의 분석적 유효성을 확인하였다. 본 연구 역시 결과

물 평가·과정 평가 항목을 UNESCO 프레임워크와 같은 별도의 역량 척도로 직접 측정하기보다, 전문가 판정을 통해 학습자의 판단 행동을 해석하는 질적 준거로 활용한다는 점에서 이와 접근을 같이 한다.

2) 전문가 평가와 질적 연구의 신뢰성

전문가 평가를 결과물의 우열을 가리는 통계적 근거가 아니라 학습자의 의사결정 과정을 검증하는 질적 자료로 다루기 위해서는 신뢰성 확보 절차가 필요하다. Lincoln과 Guba(1985)는 질적 연구의 신뢰성 기준으로 신빙성, 전이 가능성, 의존 가능성, 확증 가능성을 제시하였으며, 본 연구는 이 기준에 따라 원자료와 평가 코드의 연결을 유지하고 상반된 판정 사례도 결과에 포함하였다. 또한 Braun과 Clarke(2006)가 제시한 6단계 주제 분석 절차와 재귀적 분석 원칙을 참고하여 코딩과 범주화를 반복적으로 수행하였다.

Yin(2018/2021)과 Creswell 및 Poth(2018/2021)이 제시한 사례연구 설계 논의는 소수 사례를 빈도가 아니라 판단 근거의 내용과 사례 간 공통 패턴을 중심으로 해석해야 한다는 본 연구의 분석 방향에 근거를 제공하였다.

3) 생성형 AI 검증 역량과 미래사회 리터러시 교육

생성형 AI가 교육 현장 전반으로 확산되면서, 결과물을 산출하는 기술적 조작 능력보다 산출된 결과를 비판적으로 검증하고 책임 있게 수용하는 판단 능력이 미래사회가 요구하는 핵심 리터러시로 논의되고 있다. 임혜진(2023)이 제시한 AI 리터러시 진단도구는 AI 결과에 대한 비판적 검증 역량을 리터러시의 핵심 구성요소로 규정하였으며, 이는 디자인 교육에 국한되지 않고 문서작성, 데이터 분석, 콘텐츠 제작 등 AI가 개입하는 모든 산출 과정에 공통적으로 적용될 수 있는 관점이다.

이와 관련하여 이성태와 이종원(2025)은 사이버대학 학습자의 생성형 AI 활용 양상을 활용 정도에 따라 도구형·의존형·수비형으로 유형화하고, 강의계획서에 AI 허용 범위를 명확히 제시하는 동시에 비판적 사고 및 AI 리터러시 교육을 강화할 필요가 있음을 제안하였다. 이는 본 연구의 2차 과정 평가 항목 중 'AI 의존성' 판정이 적절·부분 적절·미흡 세 등급에 고르게 분산된 결과(Ⅲ장 2절 참조)와 유사한 양상을 보이며, 학습자의 AI 활용 유형에 따라 검증 행동의 질이 달라질 수 있음을 시사한다.

이 관점에서 본 연구가 다루는 결과물 평가와 과정 평가의 비교는 UI/UX 디자인 교육평가 방법론의 문제에 그치지 않는다. 학습자가 AI 제안을 어떤 기준으로 검토하고 어느 단계에서 재확인을 생략했는지를 구조화하는 작업은 특정 전공 영역을 넘어, 미래사회 구성원이 갖추어야 할 AI 활용 및 검증 역량의 교육 모형으로 확장될 수 있다. 이에 본 연구는 디자인 교육 사례를 바탕으로 미래사회에서 요구되는 AI 리터러시 교육을 위한 구체적인 검증 절차와 적용 방향을 제시하였다는 점에서 의의를 갖는다.

<표 1> 생성형 AI 산출물 평가 관련 선행연구

연구자	연구 대상 및 방법	주요 결과
Nielsen (1994)	전문가 소수의 휴리스틱 평가 원칙 제시	결과물 중심 평가의 표준적 절차와 한계 규명
Duan 외 (2024)	GPT-4 기반 UI 목업 휴리스틱 피드백과 인간 전문가 평가 비교	AI 피드백은 반복 수정 후 유용성 감소, 인간 검토 필요
Amershi 외 (2019)	인간-AI 상호작용 가이드라인 제시	AI 제안에 대한 인간의 최종 승인 절차 필요성 강조

연구자	연구 대상 및 방법	주요 결과
임혜진·이문수 (2023)	예비 중등교사 대상 AI 리터러시 진단도구 개발	AI 결과에 대한 비판적 검증 역량을 리터러시 핵심 요소로 규정
김지영 (2025)	디자인 교육 내 생성형 AI 활용 교수-학습 구조 실험연구	AI 도입이 교수·학습 및 평가 구조 변화를 요구함을 확인
장재준·김은영 (2026)	생성형 AI 확산 전후 디자인 의사결정 유형에 대한 문헌고찰·전문가 검증	전문가 검증이 AI 시대 디자인 역량 평가의 핵심 절차임을 제시

<표 1>에서 보듯 선행연구는 공통적으로 생성형 AI 산출물에 대한 평가가 결과의 완성도만으로는 충분하지 않으며, 인간 전문가의 비판적 검증 절차가 함께 설계되어야 함을 지적한다. 그러나 결과물 평가와 과정 평가를 같은 대상에 대해 실시하고 두 평가의 일치·불일치를 비교한 연구는 충분하지 않다. 이에 본 연구는 동일한 학습자 산출물에 대해 결과물 중심 평가와 과정 중심 평가를 순차적으로 실시하여 그 차이를 실증적으로 확인하고자 한다.

Ⅱ. 연구 방법

1. 연구 대상

연구 참여자는 UI/UX 디자인 실무·교육·연구 경력을 보유한 전문가 3인(E1~E3)과, 생성형 AI를 활용하여 UI/UX 과제를 수행한 학습자 5인(S1~S5)으로 구성하였다. 평가 대상은 학습자 5인의 산출물로 구성하였다. 전문가 집단은 결과물을 사용자 경험 관점에서 평가할 수 있는 인원으로 구성하였으며, 학습자의 결과물 평가와 작업 과정 검토라는 두 단계 평가에 모두 참여하였다. 평가 대상이 된 학습자 산출물은 '대학생을 위한 일정관리 모바일 서비스'라는 동일한 과제 조건에서 제작된 홈, 일정 상세, 일정 추가 3개 화면이며, 학습자는 생성형 AI 대화와 Figma를 사용하여 150분 이내에 화면을 완성하였다. 학습자 5인은 모두 UI/UX 디자인 관련 교과목을 수강했거나 기본적인 화면 설계 경험이 있는 참여자였다. 평가 결과의 맥락을 이해하기 위해 학습자별 산출물의 주요 특징을 <표 2>에 제시하였다.

<표 2> 평가 대상 학습자 산출물 개요

학습자	산출물의 주요 특징
S1	정보 위계를 강조한 간결한 UI로 홈 화면의 오늘 일정과 중요 일정을 시각적으로 구분
S2	수업·과제·팀플 등 일정 성격의 차이를 반영해 D-day 카드, To Do, 일정 상세를 구성
S3	화면 크기·헤더·일정 카드·플로팅 버튼까지 구체화한 와이어 프레임형 UI
S4	카드형과 원형 진행도 위젯을 활용해 현재 일정과 다음 일정을 구분한 UI
S5	인디고·퍼플 컬러 체계와 로고·마스코트 등 브랜드 요소를 결합한 UI

2. 측정도구

평가도구는 결과물의 완성도를 판정하는 1차 도구와 결과물에 이르는 과정의 타당성을 판정하는 2차 도구로 구분하여 설계하였다. 1차 블라인드 결과물 평가는 사용자 적합성, 정보구조, 사용성, 상호작용, 시각적 위계, 일관성·접근성, 종합 완성도의 7개 항목을 5점 척도로 측정하고, 각 결과물의 강점·문제점·개선 필요사항을 서술형으로 함께 기록하도록 하였다<표 3>.

<표 3> 1차 블라인드 결과물 평가 항목

평가 항목	내용
사용자 적합성	대학생 일정관리라는 사용자·과제 목적에 부합하는 정도

평가 항목	내용
정보구조	정보 위계 및 화면 구성의 명확성
사용성	기능 이해 가능성 및 조작 편의성
상호작용	화면 간 흐름 및 인터랙션의 논리적 연결성
시각적 위계	색상·타이포그래피·레이아웃을 통한 정보 강조 및 구분
일관성·접근성	디자인 요소의 일관성 및 가독성·접근성
종합 완성도	구현 가능성을 포함한 전반적 완성 수준

2차 과정 평가는 최초·수정 프롬프트, AI 생성 결과, 화면 녹화, 사후 설문 및 최종 Figma 결과물을 함께 제시한 뒤, 프롬프트 구체화, 판단의 타당성, 수정의 효과, AI 의존성, 전문가 판단과의 관계, 최종 의사결정의 6개 항목을 '적절·부분 적절·미흡'의 3단계로 판정하도록 하였다<표 4>.

<표 4> 2차 과정 평가 항목

평가 항목	내용
프롬프트 구체화	문제·조건이 프롬프트에 구체적으로 반영되었는가
판단의 타당성	학습자의 AI 결과 판단 근거가 UI/UX 원칙에 부합하였는가
수정의 효과	프롬프트·화면 수정이 실제 결과 개선으로 이어졌는가
AI 의존성	AI 제안을 비판적으로 재검토했는지, 그대로 수용했는가
전문가 판단과의 관계	학습자 판단이 전문가 판단기준과 일치하는 정도
최종 의사결정	수용·수정·재생성·거부 결정의 적절성

1차 평가의 5점 척도는 학습자 집단의 우열을 통계적으로 일반화하기 위한 추론 통계 자료가 아니라, 전문가 의견의 방향과 사례 간 차이를 확인하기 위한 기술 자료로 활용하였다. 다만 동일한 전문가 3인이 동일한 학습자 결과물을 평가하였으므로, 전문가 평균 점수를 연구결과에 활용할 수 있는지를 확인하기 위해 평가자 간 신뢰도 분석을 별도로 실시하였다.

1차 결과물 평가(5점 척도 평균)와 2차 과정 평가('적절·부분 적절·미흡' 3단계 판정)는 척도 유형이 상이하므로, 두 평가를 직접 비교하기 위해 다음과 같은 조작적 분류 기준을 <표 5>와 같이 사전에 설정하였다. 이 기준은 <표 10>의 유형 분류에 선행 적용되었으며, 연구자 1인의 주관적 해석이 아니라 전문가 판정 빈도에 근거한 규칙으로 운용함으로써 비교의 객관성을 확보하고자 하였다.

<표 5> 결과물 평가–과정 평가 비교기준

유형	판정 기준
일치	1차 평가에서 상대적으로 높게 평가된 항목(예: 사용자 적합성, 시각적 위계)에 대응하는 2차 항목(예: 프롬프트 구체화, 판단의 타당성)에서 전문가 3인 중 2인 이상이 '적절'로 판정한 경우
부분 일치	1차와 2차 판정의 방향이 항목별로 엇갈리거나, 2차 판정이 '적절'과 '부분 적절'에 걸쳐 분산된 경우
불일치	1차 평가에서 상대적으로 높게 평가된 영역이라도, 대응하는 2차 항목(수정의 효과, AI 의존성, 전문가 판단과의 관계)에서 전문가 3인 중 2인 이상이 '부분 적절' 또는 '미흡'으로 판정한 경우

3. 절차

평가는 두 단계로 순차 진행하였다.

1차 평가에서는 참여자 코드로 익명화한 학습자별 최종 UI/UX 화면 3개만 전문가에게 제시하고, 학습자의 이름, 생성형 AI 활용 경험, 프롬프트 수정 기록, 작업 과정 및 사후 설문 내용은 공개하지 않았다. 이는 학습자의 설명이나 AI 활용 방식에 대한 사전 정보가 결과물 평가에 영향을 미치는 것을 최소화하고, 최종 결과물 자체의 품질을 독립적으로

평가하기 위한 절차였다.

2차 평가에서는 1차 평가가 완료된 후, 동일한 전문가에게 각 학습자의 최초·수정 프롬프트, 생성형 AI 결과, 선택·폐기된 대안, 화면 녹화, 사후 설문 및 최종 Figma 결과물을 함께 제공하였다. 전문가는 학습자가 디자인 문제를 어떻게 이해했고, AI 결과를 어떤 기준으로 평가했으며, 수용·부분 수정·재생성·거부·직접 수정 중 어떤 행동을 선택했는지를 시간순으로 검토하였다. 자료수집은 행동 자료(화면 녹화), 생성 자료(프롬프트·AI 결과·Figma 산출물), 성찰 자료(사전·사후 설문), 전문가 평가 자료(1차·2차 평가)를 함께 확보하는 삼각 검증 방식으로 구성하였으며, 신뢰성 확보를 위해 Lincoln과 Guba(1985)가 제시한 신빙성·전이 가능성·의존 가능성·확증 가능성 기준에 따라 원자료와 코드의 연결을 유지하고 상반된 사례도 결과에 포함하였다.

4. 자료분석

수집된 자료는 학습자별로 1차·2차 평가 결과지, 화면 녹화, 프롬프트 기록, 사전·사후 설문을 하나의 사례로 통합한 뒤, Braun과 Clarke(2006)의 주제 분석 절차를 참고하여 평가 대상, 평가 기준, 판정 행동, 판단 근거의 네 범주로 반복적으로 대조하며 코드화하였다. 판단이 모호한 구간은 원자료로 돌아가 전문가의 서술형 의견과 결과물 화면을 함께 재검토하였으며, 이 과정을 통해 결과물 평가와 과정 평가라는 두 판단 체계가 일치하는 지점과 갈라지는 지점을 사례 간 비교로 도출하였다. 코드화는 특정 사례의 점수나 판정을 통계적으로 일반화하기 위한 절차가 아니라, 소수 사례에서 나타나는 판단 근거의 내용과 공통 패턴을 질적으로 해석하기 위한 절차로 활용하였다.

<표 6> 전문가 평가의 평가자 간 신뢰도 분석 결과

평가 단계	신뢰도 분석 및 결과
1차 블라인드 결과물 평가	학습자 5인×전문가 3인 (학습자별 7개 항목 평균) ICC(A,3): Two-way mixed / Absolute agreement / Average measures 결과=.708 → 중등도 수준의 평균평가 일관성
2차 과정 평가	30개 평가단위×전문가 3인(총 90개 판정) Krippendorff's α(ordinal) 결과=.400 → 평가자 간 제한적 일치

전문가 평가의 일관성을 확인하기 위해 평가자 간 신뢰도를 분석하였다. 1차 블라인드 결과물 평가는 동일한 전문가 3인이 학습자 5명의 결과물을 7개 항목으로 평가하였으며, 각 전문가의 학습자별 7개 항목 평균을 종합점수로 산출한 뒤 two-way mixed-effects, absolute agreement, average-measures 방식의 급내상관계수 ICC(A,3)를 적용하였다(Koo & Li, 2016). 2차 과정 평가는 '적절·부분 적절·미흡'의 순서형 판정 자료이므로 학습자 5명×평가항목 6개로 구성된 30개 평가단위를 대상으로 Krippendorff's α(ordinal)를 산출하였다(Krippendorff, 2004; Hayes & Krippendorff, 2007).

<표 6>과 같이 1차 블라인드 결과물 평가의 ICC(A,3)는 .708로 나타나 전문가 3인의 평균 평가가 중등도 수준의 일관성을 보였다. 반면 2차 과정 평가의 Krippendorff's α(ordinal)는 .400으로 나타나 평가자 간 일치가 제한적인 수준이었다. 이에 1차 평가는 전문가 평균 점수를 사례 간 기술적 비교 자료로 활용하였으며, 2차 과정 평가는 단일 합의 점수로 환원하지 않고 항목별 판정 빈도, 전문가의 서술형 의견 및 학습자 자기평가와의 일치·불일치 사례를 함께 분석하였다.

이들 계수는 연구결과의 일반화를 위한 유의성 검정이 아니라, 전문가 평가의 일관성과 불일치 정

도를 확인하기 위한 보조적 신뢰도 지표로 활용하였다.

Ⅲ. 결과

1. 전문가의 1차 블라인드 결과물 평가

전문가 3인에게는 학습자가 누구인지, 어떤 과정을 거쳤는지에 대한 정보를 전혀 제공하지 않고 S1~S5의 최종 UI/UX 결과물만 제시한 뒤 평가를 받았다<표 7>.

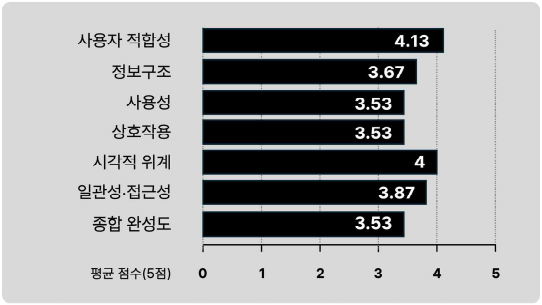
평가자 간 신뢰도를 확인한 결과, 전문가 3인의 학습자별 종합평가에 대한 ICC(A,3)는 .708로 나타났다.

이는 전문가 3인의 평균평가가 중등도 수준의 일관성을 보였음을 의미한다. 다만 평가 대상이 학습자 5명으로 제한되어 있으므로 해당 계수는 전문가 평균 점수 활용의 타당성을 보조적으로 확인하는 지표로 해석하고, 집단 수준의 일반화에는 신중을 기하였다.

<표 7> 학습자별 1차 블라인드 평가 평균

학습자	E1	E2	E3	전체평균
S1	3.71	3.71	4.43	3.95
S2	4.00	3.14	4.00	3.71
S3	4.14	4.29	4.29	4.24
S4	3.43	2.43	3.14	3.00
S5	3.29	3.86	4.43	3.86

다섯 사례의 평균 점수는 S3(4.24점)가 가장 높았고, 이어 S1(3.95점)·S5(3.86점)·S2(3.71점) 순이었으며 S4(3.00점)가 가장 낮았다<표 7>.



[그림 1] 전문가 1차 블라인드 평가 항목별 전체 평균

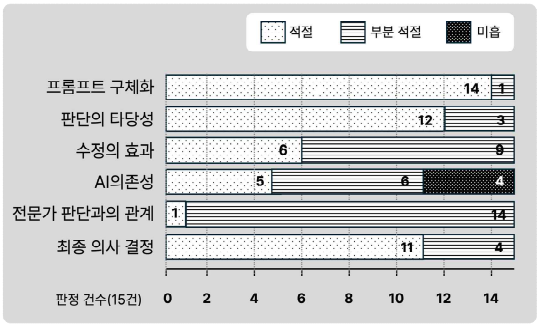
일곱 개 평가항목 가운데는 사용자 적합성(4.13점)과 시각적 위계(4.00점)가 상대적으로 높게, 일관성·접근성(3.87점)과 정보구조(3.67점)가 중간 수준으로 평가되었고, 사용성·상호작용·종합 완성도 세 항목은 나란히 3.53점으로 가장 낮게 평가되었다[그림 1]. 이 점수 분포는 학습자들이 대학생이라는 사용자 맥락과 시각적 콘셉트는 비교적 충실히 구현했으나, 실제 과업 흐름, 기능 누락, 오류 방지, 최종 품질 검증에서는 보완이 필요했음을 보여준다.

사례별 서술형 의견을 정리하면, 다섯 명의 결과물은 저마다 다른 지점에서 강점과 한계를 동시에 나타냈다<표 8>. 시각적 구분과 컬러·아이콘 체계의 통일성은 여러 사례에서 공통적으로 좋은 평가를 받았으나, 화면 이면의 기능 누락이나 실제 데이터와의 불일치처럼 눈에 바로 띄지 않는 요소들이 반복적으로 감점 요인이 되었다는 점이 특징적이다.

<표 8> 학습자별 1차 결과물 평가의 강점과 보완점

학습자	강점	보완점
S1	일정 중요도에 따른 시각적 구분, 아이콘·컬러의 통일감	우선순위 입력, 삭제·복귀 경로 등 일부 기능의 누락
S2	마감 임박 일정을 강한 컬러로 강조한 표현력	D-day·To Do 항목 구분과 버튼 의미의 모호함, 네온 색 사용의 피로도
S3	화면 구조·상호작용·컴포넌트 전반의 일관된 완성도	상단 영역의 표시 오류, AI 산출물에 대한 직접 검증 부족

학습자	강점	보완점
S4	현재·다음 일정을 구분하는 위젯 구성의 참신함	지도 이미지와 실제 장소의 불일치, 학사 일정 카테고리 누락
S5	로고·마스코트를 포함한 브랜드 요소의 통합적 표현	일부 콘텐츠의 자리 표시자 미교체, 카드 표현 반복에 따른 정보 우선순위 약화



[그림 2] 전문가 2차 과정검토 항목별 판정 분포

2. 전문가의 2차 과정 평가

2차 과정 평가의 평가자 간 신뢰도를 확인하기 위해 순서형 판정 자료에 대한 Krippendorff's α(ordinal)를 산출하였다. 학습자 5명×평가항목 6개의 30개 평가단위에 대한 전문가 3인의 판정을 분석한 결과 α=.400으로 나타났다. 이는 과정 평가에서 평가자 간 일치가 제한적인 수준이었음을 보여준다. 따라서 2차 과정 평가 결과는 단일 합의 점수로 환원하지 않고, 항목별 판정 빈도와 전문가의 서술형 의견, 학습자 자기평가와의 일치·불일치 사례를 함께 해석하였다.

2차 과정 평가에서는 생성형 AI 대화와 프롬프트 수정 이력, 학습자의 판단 기록 및 최종 결과물을 함께 제시하였다. 전문가 3인이 학습자 5사례를 6개 항목으로 판정하여 전체 90개의 판정 자료가 생성되었으며, 각 평가항목별로는 15건(전문가 3인×학습자 5사례)의 판정을 집계하였다. 그 결과 프롬프트 구체화 항목은 15건 중 14건이 '적절'로 나타나 가장 높은 비중을 보였다. 판단의 타당성은 12건, 최종 의사결정은 11건이 '적절'로 평가되었다. 반면 수정의 효과는 9건만 '적절'로 판정되었고, 전문가 판단과의 관계는 14건이 '부분 적절'로 나타났다. 이는 학습자가 문제를 언어로 정리하고 대안을 선택하는 단계는 비교적 원활했으나, 수정 결과가 전문가가 요구하는 검증 수준까지 충분히 개선된 사례는 제한적이었음을 보여준다[그림 2].

판정이 가장 크게 갈린 항목은 AI 의존성으로, 적절 5건·부분 적절 6건·미흡 4건으로 세 등급에 고르게 분산되었다. 전문가들은 학습자가 AI 제안에 재질문을 제기하거나 여러 대안을 비교한 점은 긍정적으로 평가했지만, AI가 스스로 제시한 '완료' 보고를 별도의 검증 없이 수용하여 아이콘·텍스트·지도·자리 표시자 등의 오류가 그대로 남은 사례들을 공통적인 취약점으로 지적하였다.

사례별로 보면 타당한 판단과 약한 검증 단계가 학습자마다 서로 다른 지점에서 나타났다<표 9>. S3는 도구별 역할분담과 디자인 원칙 적용이 가장 뚜렷했음에도 AI의 자체 완료 보고를 실제 화면에서 재검증하지 않았고, S1은 정보 위계 재조정 판단이 전문가 의견과 부합했지만 요구사항과 아이콘 의미에 대한 최종 점검은 미흡했다. 이는 학습자가 특정 판단 국면에서는 전문가 수준의 판단을 보이면서도, 이러한 판단이 반드시 별도의 검증 행동으로 이어지는 것은 아님을 보여준다.

<표 9> 학습자별 2차 과정 평가 핵심 결과

학습자	타당한 판단	약한 단계	전문가 관계
S1	정보 위계 재조정 판단	요구사항·아이콘 재검증 부족	부분 일치
S2	사용자 재정의·레이아웃의 비판적 선택	인터랙션의 구현 검증 부족	부분 일치

학습자	타당한 판단	약한 단계	전문가 관계
S3	도구 역할분담과 디자인 시스템 구체화	AI 완료보고 및 실제 기기 레이아웃 재검증 부족	대체로 일치
S4	카드–타임라인 비교와 직접 와이어 프레임 활용	지도·카테고리 데이터 정합성 및 요구사항 검증 부족	부분 일치
S5	Claude.ChatGPT 역할분담과 시각요소 단순화	플레이스홀더·상태 아이콘 최종 QA 부족	부분 일치

3. 결과물 평가와 과정 평가의 일치·불일치

학습자 사후 설문과 전문가 평가를 비교하면, 두 평가가 일치하는 지점과 불일치하는 지점이 뚜렷이 구분된다. 학습자들은 한결같이 AI의 첫 결과를 그대로 사용하지 않고 수정했다고 답했고, 전문가들 역시 프롬프트를 구체화하고 대안을 선택하는 판단 자체에는 무리가 없었다고 평가하였다는 점에서 두 평가는 이 지점에서 일치한다. AI 산출물을 완제품이 아니라 다듬어 가는 재료로 다루었다는 인식이 학습자와 전문가 모두에게서 공통적으로 확인된다. 차이는 그다음 단계에서 발생한다.

한편, 2차 과정 평가의 Krippendorff's α(ordinal)가 .400으로 나타난 점을 고려하면, 전문가 간에도 수정의 효과, AI 의존성 및 최종 검증 수준을 판단하는 기준에는 차이가 존재하였다. 따라서 본 연구에서 제시하는 일치·부분 일치·불일치 유형은 전문가 간 완전한 통계적 합의를 의미하기보다, 전문가의 판정과 서술형 의견 및 학습자의 자기평가를 교차하여 도출한 질적 비교유형으로 해석하는 것이 적절하다.

학습자는 스스로 '필요한 부분만 선별하여 반영했다', '대부분 수용하되 일부는 수정했다'고 응답하며 수정의 실행 여부 자체를 성과로 인식하는 경향을 보였다. 반면 전문가는 수정 이후에도 남아 있는 기능 누락, 텍스트 및 레이아웃의 불일치, 실제와 상이한 지도·카테고리 정보, 미수정된 자리 표시자 등을 근거로 후속 검증이 충분하지 않았다고 평가하였다.<표 10>.

결국 학습자에게 의사결정의 완료 기준은 '수정을 했는가'였던 반면, 전문가에게 그 기준은 '수정한 결과가 요구사항과 데이터, 실제 기기에서의 사용성, 최종 품질 기준을 통과했는가'였다. 이 두 기준 사이의 거리는 과정 인식과 실무 검증 기준 사이의 차이를 보여준다.

<표 10> 학습자 자기평가와 전문가 평가의 관계

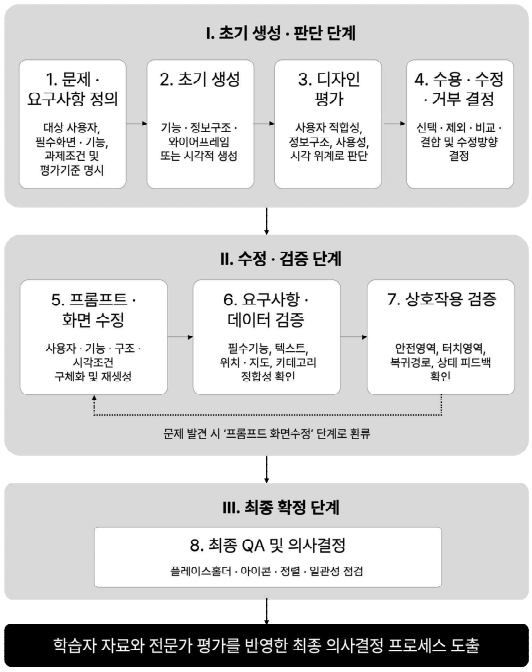
관계 유형	학습자 인식	전문가 검토
일치	AI 결과를 그대로 사용하지 않고 수정·비교했다고 인식	프롬프트 구체화와 판단 근거는 대체로 적절
부분 일치	정보구조·사용자 요구를 고려했다고 인식	일부 화면에는 반영됐으나 기능·상태·시각 규칙이 불완전
불일치	최종안을 충분히 선택·수정했다고 인식	요구사항 대조, 데이터 사실 확인, 실제 기기 검증, 최종 QA가 누락

이는 학습자의 역량 부족이라기보다 '검증'이라는 행위가 무엇을 의미하는지에 대한 암묵적 정의의 차이에서 비롯된 것으로 해석할 수 있다. 학습자에게 검증은 대체로 화면이 의도대로 보이는지를 눈으로 확인하는 행위였던 반면, 전문가에게 검증은 요구사항 대조, 데이터 사실 확인, 실제 기기에서의 조작 가능성까지 포함하는 다단계 절차였다.

따라서 결과물 평가와 과정 평가 간의 불일치를 줄이기 위해서는 학습자 개인의 주의력 향상에 초점을 두기보다, 검증 행위를 요구사항 확인, 데이터 검증, 실제 기기 확인의 세부 절차로 구체화하여 교육 과정에 명시적으로 포함하는 접근이 보다 효과적일 수 있다.

4. 전문가 평가를 반영한 최종 통합 의사결정 프로세스

학습자의 판단·수정 기록만 놓고 보면 문제 이해에서 시작해 프롬프트를 작성하고, AI 결과를 검토해 판단을 내리고, 그 판단을 다시 구체화·재평가 단계로 연결하는 반복 구조로 정리할 수 있다. 그러나 이 흐름은 '수정했다'는 사실까지만 보여줄 뿐, 그 수정이 실제로 결과를 개선했는지, 최종 산출 전에 무엇을 확인해야 하는지는 명시하지 못한다. 전문가 평가에서 확인된 미비점을 반영한 최종 프로세스는 문제·요구사항 정의(①) → 초기 생성(②) → 디자인 평가(③) → 수용, 수정·거부 갈림길(④) → 프롬프트·화면 수정(⑤) → 요구사항·데이터 검증(⑥) → 실제 기기·상호작용 검증(⑦) → 최종 QA 및 의사결정(⑧)이라는 세 단계가 새로 추가된 8단계로 완성된다[그림 3].



[그림 3] 전문가 평가를 반영한 최종 통합 의사결정 프로세스

이 세 단계가 갖는 의미는 각기 다르다. ⑥은 처음 정의한 필수 기능이 실제 화면의 텍스트·위치·카테고리와 어긋나지 않는지를, ⑦은 화면이 실제 기기에서 잘리거나 눌리지 않는지, 터치 영역과 복귀 경로가 정상 작동하는지를, ⑧은 자리 표시자·아이콘 표기·정렬과 컴포넌트가 끝까지 일관적인지를 각각 점검하는 단계이다. 어느 단계에서든 문제가 발견되면 곧바로 ⑤ 프롬프트·화면 수정으로 되돌아가도록 설계되어 있어, 전체 프로세스는 일직선이 아니라 검증이 실패할 때마다 이전 단계로 되돌아가는 순환형 구조를 이룬다.

Ⅳ. 논의

본 연구는 생성형 AI 기반 UI/UX 디자인 교육에서 산출된 학습자 결과물에 대해 전문가 3인이 실시한 1차 블라인드 결과물 평가와 2차 과정 평가를 비교하여, 결과물 평가와 과정 평가의 일치·불일치 특성을 분석하고 이를 통합한 최종 검증 프로세스를 도출하였다.

첫째(RQ1), 전문가의 블라인드 결과물 평가에서는 사용자 적합성과 시각적 위계가 상대적으로 높게 평가된 반면, 사용성·상호작용·종합 완성도는 낮게 평가되어, 학습자들이 사용자 맥락과 시각적 콘셉트는 비교적 충실히 구현했으나 실제 과업 흐름과 최종 품질 검증에서는 보완이 필요했음이 확인되었다.

둘째(RQ2), 학습자의 프롬프트 수정과 작업 과정을 공개한 후 전문가 평가는 프롬프트 구체화와 판단의 타당성에서는 긍정적으로 유지되었으나, 수정의 효과와 전문가 판단과의 관계, AI 의존성 항목에서는 부분 적절 또는 미흡 판정이 늘어, 과정 공개가 평가를 일률적으로 상향시키지 않고 오히려 검증되지 않은 지점을 드러냈다.

셋째(RQ3), 결과물 평가와 과정 평가가 일치하는 지점은 AI 결과를 중간 산출물로 인식하고 반복적

으로 구체화했다는 점이었으나, 불일치는 학습자가 수정의 실행 여부로 의사결정을 완결로 인식한 반면 전문가는 수정 결과가 요구사항·데이터 정합성·실제 기기 사용성·최종 QA 기준을 충족했는지까지 평가한 데서 발생하였다.

넷째(RQ4), 이에 본 연구는 문제·요구사항 정의부터 최종 QA 및 의사결정에 이르는 8단계의 통합 검증 프로세스를 제안하였으며, 그 핵심은 AI 결과의 평가와 최종 의사결정 사이에 요구사항·데이터 검증과 실제 기기·상호작용 검증이라는 세분화된 단계를 배치한 데 있다.

이상의 결과는 다음과 같은 시사점을 갖는다.

실무적으로 생성형 AI를 활용한 산출물 검토에서는 완성도가 높아 보이는 결과물이라도 데이터의 정확성과 일관성, 실제 기기에서의 상호작용을 별도로 검증하는 절차가 필요하다.

교육적으로는 평가 기록 방식을 '수정했다'라는 결과 중심 서술에서 문제 인식, 적용 판단 기준, 선택한 행동, 수정 전후 변화와 재검증 결과를 연결하는 과정 중심 기록으로 전환할 필요가 있으며, 평가 방법 측면에서도 결과물 중심 평가와 과정 중심 평가를 병행해야 한다. 본 연구에서도 S3와 S5는 과정과 시각 체계에서 높은 평가를 받았지만 최종 QA 결함이 확인되었고, S2와 S4는 대화와 기획의 논리성에 비해 결과물 구현이 충분히 연결되지 않아, 결과물만 평가하면 판단 과정이 누락되고 과정만 평가하면 실제 사용성과 구현 결함을 간과할 수 있음을 보여준다.

또한 이 결과는 미래사회 AI 리터러시 교육에도 함의를 갖는다. 특정 전공의 실습 결과물이 아니라 판단 절차 자체를 교육과정에 명시적으로 포함시키고, 학습자가 AI 제안을 수용·수정·거부하는 과정에서 어떤 기준을 사용했는지를 기록하고 공유하도록 하는 것은 디자인 교육을 넘어 다양한 전공에서 활용 가능한 AI 리터러시 교육 모형으로 확장될 수 있다. 특히 요구사항·데이터 검증과 실제 기기·상호작용 검증 단계는 특정 도구나 산출물 형식에 종속되지 않는 일반적 검증 절차로 재구성할 수 있어, 향후 미래사회 교육과정에서 전공을 불문한 공통 AI 리터러시 모듈로 발전시킬 가능성이 있다.

다만 본 연구는 전문가 평가의 일관성을 확인하기 위해 1차 결과물 평가에 ICC(A,3), 2차 과정 평가에 Krippendorff's α(ordinal)를 적용하였으나, 평가자가 3인이고 평가대상 학습자가 5명으로 제한되어 있어 해당 신뢰도 계수를 다른 학습자 집단이나 전문가 집단에 일반화하기 어렵다는 한계를 갖는다. 1차 평균 평가의 ICC(A,3)는 .708로 중등도 수준의 일관성을 보인 반면, 2차 과정 평가의 α는 .400으로 제한적인 일치도를 보였다. 이는 과정 평가에서 '수정의 효과', 'AI 의존성', '전문가 판단과의 관계'와 같은 항목의 행동지표를 보다 구체화할 필요가 있음을 시사한다. 향후 연구에서는 학습자와 평가자 수를 확대하고 평가 기준을 행동지표 수준으로 정교화한 뒤 평가자 간 일치도를 반복 검증할 필요가 있다. 또한 본 연구에서 도출한 8단계 검증 프로세스를 실제 수업에 적용하여 적용 전후의 오류율·수정 횟수·결과물 완성도 변화를 확인하고, 다수 평가자의 비교를 통해 이중 평가체계의 타당성을 검증할 필요가 있다.

본 연구가 제안하는 이중 평가체계와 8단계 검증 프로세스는 결과물 평가에 더해 프롬프트 이력, 화면 녹화, 사후 설문 등 다층적 과정 자료를 함께 검토해야 하므로, 이를 실제 수업 현장에 적용할 경우 교수자 1인이 다수 학습자의 과정 자료를 개별적으로 검토하는 데 따르는 평가 부담(workload)이 현실적인 제약으로 작용할 수 있다. 본 연구에서도 전문가 3인이 학습자 5명의 과정 자료를 검토하는 데에도 상당한 시간이 소요되었으며, 수강 인원이 많은 일반 강의 상황에서 이를 그대로 적용하기는 어렵다. 이러한 양상은 몰입형·상호작용형 학습환경에서 시스템 차원의 AI 지원이 부재할 경우, 내비게

이션 안내·상호작용 중재·인지부하 관리와 같은 역할이 고스란히 교수자에게 전가된다는 최근 연구 결과와도 맥을 같이한다(김은호, 2026). 김은호(2026)는 메타버스 기반 한국어 학습 환경에서 AI 기반 지원 체계의 부재가 교수자를 내비게이션 안내자이자 상호작용 조율자, 기술적 중재자로 만든다는 점을 밝히고, AI를 단순한 보조 도구가 아니라 상호작용을 매개하고 인지적 부담을 분산시키는 '교육 인프라(pedagogicalinfrastructure)'로 재개념화할 필요성을 제안한 바 있다. 이러한 한계를 완화하기 위한 대안으로는 첫째, 전체 학습자를 대상으로 하기보다 결과물 평가에서 등급이 엇갈리거나 경계선에 위치한 사례를 표집하여 과정 평가를 선별적으로 실시하는 방안, 둘째, 프롬프트 로그와 화면 녹화를 자동으로 수집·요약하는 도구를 활용해 교수자가 검토할 자료의 양을 축소하는 방안, 셋째, 본 연구의 6개 과정 평가 항목을 체크리스트 형태의 간이 루브릭으로 재구성하여 평상시 형성평가 단계에서는 약식으로, 학기말 총괄평가 단계에서는 전체 항목으로 적용하는 이원화 방안을 고려할 수 있다. 본 연구가 제안하는 8단계 통합 검증 프로세스 역시, 향후 AI 기반 사전 스크리닝이나 로그 자동요약과 같은 시스템 차원의 지원이 결합될 때 비로소 교수자 개인의 부담에 의존하지 않는 지속 가능한 평가체계로 발전할 수 있을 것이다. 다만 이러한 대안들이 실제 교수자의 평가 부담을 얼마나 경감하면서도 과정 평가의 정보량을 유지할 수 있는지는 본 연구의 범위를 벗어나며, 후속 연구를 통한 실증적 검증이 필요하다.

참고문헌

김은호 (2026). 메타버스 기반 한국어 학습 플랫폼에서의 학습자 참여와 교수자 부담: AI 지원 한국어교육에 대한 시사점. 미래사회, 17(1), 295-307.

김지영 (2025). 디자인 교육에서 생성형 AI 도구의 활용과 교수-학습 구조 변화 연구: 국내 사례 분석 및 실험 연구를 중심으로. 조형미디어학, 28(12) 260-268.

박종선 (2026). AI 활용이 가능한 학습 환경에서 사이버대학 학습자 역량과 학업 성과의 구조적 관계: UNESCO 학습자 역량 프레임워크를 중심으로. 미래사회, 17(12)281-294.

이성태, 이종원 (2025). 사이버대학교 학습자를 위한 생성형 AI 활용 가이드라인 연구. 미래사회, 16(12) 192-206.

임혜진, 이문수 (2023). 예비 중등교사의 인공지능 리터러시 진단도구 개발 및 타당화. 학습자중심교과교육연구, 23(12), 875-892.

장재준, 김은영 (2026). 디자인 역량과 의사결정 유형의 시기별 변화에 관한 연구: 생성형 인공지능 확산 전후의 문헌고찰과 전문가 검증을 중심으로. 조형교육, 98, 345-380.

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E.(2019). Guidelines for human-AI interaction. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Article 3,* 1-13. https://doi.org/10.1145/3290605.3300233

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. Assessment in Education: *Principles, Policy & Practice, 5*(1), 7-74. https://doi.org/10.1080/0969595980050102

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77-101.

Creswell, J. W., & Poth, C. N. (2021). 질적 연구 방법: 다섯 가지 접근(4판) (조흥식, 정선욱, 김진숙, 권지성 역). 학지사. (원저 출판 2018)

Duan, P., Warner, J., Li, Y., & Hartmann, B. (2024). Generating automatic feedback on UI mockups with large language models. *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems.* https://doi.org/10.1145/3613904.3642782

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures, 1*(1), 77–89.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155–163.

Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research, 30*(3), 411–433.

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry. Sage.*

Nielsen, J. (1994). Heuristic evaluation. In J. Nielsen & R. L. Mack (Eds.), *Usability inspection methods (pp. 25–62).* John Wiley & Sons.

Shneiderman, B. (2022). Human-centered AI. *Oxford University Press.*

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with ChatGPT. *Proceedings of the 30th Conference on Pattern Languages of Programs.* https://doi.org/10.5555/3721041.3721046

Yin, R. K. (2021). 사례연구방법(6판) (송민채, 신현섭, 조수현, 서이안, 신경식 역). 한경사. (원저 출판 2018)

투고일자: 2026. 8. 20.
심사일자: 2026. 9. 29.
게재확정일자: 2026. 9. 30.

미래사회
Journal of Future Society
2026. Vol. 17, No. 3, 285~299.
https://doi.org/10.22987/jifso.2026.17.3.285

Study on Evaluation Methods in Generative Artificial Intelligence-Based UI/UX Design Education: Comparison of Outcome-Based and Process-Based Expert Evaluation

JinKyoung Cho

Yeonsung University

This paper presents an investigation of whether evaluating only the finished outcomes of generative artificial intelligence (AI)-based user interface/user experience (UI/UX) design tasks adequately captures learners' actual competence and their process of working with AI. Three UI/UX experts first conducted a blind outcome evaluation of the final deliverables produced by five learners and then conducted a process evaluation after the learners' prompt revision histories and work materials were disclosed. User fit and visual hierarchy received relatively high ratings, whereas usability, interaction, and overall completeness received relatively low ratings. After disclosure, most judgments regarding prompt elaboration and rational validity were appropriate, but revision effectiveness, alignment with expert judgment, and AI dependency frequently showed partially appropriate or insufficient ratings, revealing a gap between the two evaluation types. Learners regarded revising AI outputs as completing their decision-making, whereas experts additionally assessed whether the results satisfied the requirements, maintained data consistency, ensured on-device usability, and met final quality assurance criteria. Accordingly, this paper presents an eight-stage integrated verification process and discusses the need for a dual evaluation system that combines outcome-based and process-based evaluations in generative AI-based design education

Key words: Generative AI, UI/UX Design Education, Outcome-Based Evaluation, Process-Based Evaluation, Expert Evaluation