



AIESAD

Ried

**Revista Iberoamericana de
Educación a Distancia**

La Revista Iberoamericana de la Educación Digital

VOL. 29 N° 1 ENERO, 2026
ISSN: 1138-2783

RIED

Esta publicación de periodicidad semestral está dirigida a los estudiosos e investigadores del ámbito educativo, docentes universitarios y público interesado en su objeto de estudio. La RIED centra su atención en la difusión de ensayos, trabajos de carácter científico y experiencias innovadoras dentro del ámbito de la educación a distancia en cualesquiera de sus formulaciones y de las tecnologías aplicadas a la educación.

La RIED se gestiona íntegramente a través del Open Journal System (OJS), tanto para la edición como para la relación con los autores y revisores, así como para la difusión electrónica en abierto.

La RIED, además de su formato impreso, se publica en formato electrónico en dos sedes: OJS en UNED de España: <http://revistas.uned.es/index.php/ried>

INTERCAMBIOS y SUSCRIPCIONES:

RIED. Revista Iberoamericana de Educación a Distancia.

UTPL – SAN CAYETANO ALTO, s/n

Loja (Ecuador)

ried@utpl.edu.ec

Consejo Directivo de AIESAD (Asesor en RIED)

- **Presidente:** Jaime Leal Afanador, Rector Magfco. Universidad Nacional Abierta y a Distancia UNAD, Colombia
- **Directora Ejecutiva:** Constanza Abadía García, Vicerrectora. Universidad Nacional Abierta y a Distancia – UNAD. Colombia.
- **Vicepresidente Primero:** Ricardo Mairal Usón, Rector Magfco. Universidad Nacional de Educación a Distancia UNED. España
- **Vicepresidente Segundo:** Rodrigo Arias Camacho, Rector Magfco. Universidad Estatal a Distancia UNED. Costa Rica
- **Vicepresidente Tercero:** Santiago Acosta Aide, Rector Magfco. Universidad Técnica Particular de Loja. Ecuador
- **Vicepresidente Cuarto:** David Mejía Rodríguez, Rector Magnífico. Universidad Internacional de La Rioja – UNIR. México.
- **Vocales:**
 - Alfredo Alonso, Rector Magfco. Universidad Nacional de Quilmes UNQ. Argentina
 - Carla Padrel de Oliveira, Rectora Magfca. Universidad Aberta UAB. Portugal
 - Alejandra Hernández Acosta, Rectora Magnífica. Universidad Abierta para Adultos – UAPA. República Dominicana
 - Fray Álvaro José Arango Restrepo, O.P., Rector Magnífico. Universidad Santo Tomás – USTA. Colombia
- **Secretaría permanente y Tesorería:** Laura Alba Juez, Vicerrectora de Internacionalización. Universidad Nacional de Educación a Distancia – UNED. España.

Director/Editor (Director/Editor-in-Chief)

- Dr. Lorenzo García Aretio, Universidad Nacional de Educación a Distancia (UNED), España

Directora/Editora Adjunta (Deputy Director/Editor)

- Dra. Cristina Díaz de la Cruz, UTPL (Ecuador)

Consejo Editorial y Científico (Editorial and Scientific Board)

- Jordi Adell Segura. Univ. Jaime I (España).
- José Ignacio Aguaded Gómez. Univ. de Huelva (España). Editor “Comunicar”
- Luisa Aires. Univ. Aberta (Portugal).
- Terry Anderson. Athabasca Univ. (Canadá). Editor Emeritus IRRODL.
- Manuel Area Moreira. Univ. de La Laguna (España).
- Elena Barberá Gregori. UOC (España).
- Elena Bárcena Madera. UNED (España)
- Antonio Bartolomé Pina. Univ. de Barcelona (España).
- Julio Cabero Almenara. Univ. de Sevilla (España). Editor Pixelbit.
- Isabel Cantón Mayo. Univ. de León (España).
- Linda Castañeda. Univ. de Murcia (España).
- Manuel Castro Gil. UNED (España).
- M. Elena Chan Núñez. Univ. de Guadalajara (México).
- Cristóbal Cobo. Univ. of Oxford (R. Unido). The World Bank.
- César Coll Salvador. Univ. de Barcelona (España).
- Grainne Conole. e4innovation (R.Unido).
- Laura Czerniewicz. Univ. of Cape Town (Sudáfrica). CILT.
- Carlos Delgado Kloos. Univ. Carlos III (España). Cátedra UNESCO Educación Digital.
- Frida Díaz Barriga. UNAM (México).
- M. Esther del Moral Pérez. Univ. de Oviedo (España).
- Pierre Dillenbourg. EPFL (Suiza). Coordinator CHILI Lab.
- Josep M. Duarte. UOC (España). Editor ETHE Journal. EDEN President.
- Rubén Edel Navarro. Univ. Veracruzana (México).
- María Jesús Gallego-Arrufat. Univ. de Granada (España).
- Francisco J. García Peñalvo. Univ. de Salamanca (España). Editor EKS.

- Ana García-Valcárcel Muñoz-Repiso. Univ. de Salamanca (España).
- Mercè Gisbert Cervera. Univ. Rovira i Virgili (España).
- Carina Soledad González González. Univ. de La Laguna (España).
- Mercedes González Sanmamed. Univ. de A Coruña (España).
- Jaime Leal Afanador. UNAD (Colombia). Presidente AIESAD.
- M. del Carmen Llorente-Cejudo. Univ. de Sevilla (España).
- Ricardo Mairal Usón. Rector UNED (España). EADTU Ex-President.
- Carlos Marcelo García. Univ. de Sevilla (España).
- João Mattar. Pontificia Univ. Católica de São Paulo (Brasil).
- Rory McGreal. Athabasca Univ. Editor de IRRODL (Canadá). UNESCO Chair in OER.
- Daniel Mill. Univ. Federal de São Carlos (Brasil). Grupo Horizonte.
- António Moreira Teixeira. Univ. Aberta (Portugal). EDEN Ex-President.
- Lina Morgado. Univ. Aberta (Portugal).
- Jaime Muñoz Arteaga. Univ. Autónoma Aguascalientes (México).
- Sara Osuna-Acedo. UNED (España).
- Adolfinia Pérez Garcías. Univ. Islas Baleares (España).
- Mar Pérez-Sanagustín. Univ. Paul Sabatier (Francia).
- Teresa Pessoa. Univ. de Coimbra (Portugal).
- M. Paz Prendes Espinosa. Univ. de Murcia (España).
- Claudio Rama. IESAL/UNESCO (Venezuela).
- M. Soledad Ramírez Montoya. TEC Monterrey (México). Cátedra UNESCO-ICDE.
- Rosabel Roig. Univ. de Alicante (España). Editora NAER.
- Jesús Salinas Ibáñez. Univ. Islas Baleares (España). Editor EDUTEC.
- Ángeles Sánchez-Elvira. UNED (España) Cátedra UNESCO EaD.
- Juana Sancho Gil. Univ. de Barcelona (España).
- Albert Sangrá. UOC (España). Cátedra UNESCO E-learning.
- Alan W. Tait. The OU (R.Unido). EDEN Ex-President. Ex-Editor Open Learning.
- Hernán Thomas. Univ. Nacional de Quilmes (Argentina).
- Javier Tourón Figueroa. UNIR (España).
- Martín James Weller. The OU (R. Unido). Director OER Hub. Editor JIME.
- Miguel Zapata Ros. Univ. de Murcia (España). Editor RED.

Delegación de AIESAD en RIED (AIESAD's Delegate in RIED)

- Santiago Acosta Aide, Univ. Técnica Particular de Loja (UTPL), Ecuador
- Cristina Díaz de la Cruz. UTPL, Ecuador

Secretaría Técnica (Technical Secretariat)

- José Luis García Boyé, AIESAD, España

Consejo de Redacción (Editing Board)

- Martha Vanessa Agila-Palacios , UTPL (Ecuador).
- Laura T. Alonso Díaz, Univ. Extremadura (España).
- Julia Ángel Osorio, UNAD (Colombia).
- Analía Chiecher, Univ. Río Cuarto (Argentina).
- Cristina Díaz de la Cruz, UTPL (Ecuador)
- Paz Díez Arcón, UPM (España)
- Anna Engel, Univ. de Barcelona (España).
- Francesc Esteve Mon, Univ. Jaime I (España).
- Alicia García Holgado, Univ. Salamanca (España)
- Montse Guitert Catasús, UOC (España).
- M. José Hernández Serrano, Univ. Salamanca (España)
- Agustín Lagunes Domínguez, Univ. Veracruzana (México).
- Andrea López Ingelmo, Univ. Salamanca (España)
- Verónica Marín Díaz, Univ. Córdoba (España).
- Victoria I. Marín, Univ. Lleida (España)
- Elena M. Martín Monje, UNED (España).
- Rafael Morales Gamboa, Univ. Guadalajara (México).
- João Paz, Univ. Aberta (Portugal).
- Neuza Pedro, Univ. Lisboa (Portugal)
- Iliana Ramírez Asanza, UTPL (Ecuador).
- Enrique Ruiz Velasco, UNAM, (México).
- José M. Sáez López, UNED (España).
- Cassio Santos, Univ. Lisboa (Portugal).
- Carolina Schmitt Nunes, Univ. Sta. Catarina (Brasil).
- Paola Sarango-Lapo, UTPL (Ecuador)
- Juan José Sosa Alonso, Univ. La Laguna (España).
- Romero Tori, Univ. Sao Paulo (Brasil).
- Gemma Tur Ferrer, Univ. Islas Baleares (España).
- José Carlos Vázquez Parra, TEC Monterrey (México).

.Redes Sociales (Social Networks)

- José Luis García Boyé
- Carla Netto
- Carolina Schmitt

Apoyo Técnico (Technical Assistance)

- Alexis Moreno-Pulido, Responsable de Biblioteca (UNED)

Soporte OJS y Publicación digital

- Servicio Publicación y Difusión Digital - BIBLIOTECA, UNED

LA REVISTA IBEROAMERICANA DE EDUCACIÓN A DISTANCIA (RIED) SE ENCUENTRA INDIZADA ACTUALMENTE EN LAS SIGUIENTES BASES DE DATOS Y CATÁLOGOS:

BASES DE DATOS Y PLATAFORMAS DE EVALUACIÓN

- BASE. Bielefeld Academic Search Engine
- CAPES
- CARHUS Plus+
- CCHS-CSIC
- CEDAL (Instituto Latinoamericano de Comunicación Educativa (ILCE) de México)
- CIRC (Clasificación Integrada de Revistas Científicas)
- CiteFactor – Academic Scientific Journals
- CREDI- OEI (Centro de Recursos de la OEI)
- Crossref (Metadata Search)
- Dialnet (Alertas de Literatura Científica Hispana)
- DICE (Difusión y Calidad Editorial de Revistas)
- EI Compendex
- EBSCO. Fuente Académica Premier
- ERA. Educational Research Abstracts
- ERIH-Plus. European Reference Index for the Humanities and Social Sciences.
- EZB-Electronic Journals Library Genamics JournalSeek
- HEDBIB (International Bibliographic Database on Higher Education)
- IN-RECS (Índice de Impacto de Revistas Españolas de Ciencias Sociales)
- IRESIE (Índice de Revistas de Educación Superior e Investigación Educativa)
- ISOC (CSIC/CINDOC)
- JournalTOCs
- MIAR (Matriz para Evaluación de Revistas)
- ProQuest-CSA
- Psycodoc
- REDIB. Red Iberoamericana de Innovación y Conocimiento Científico
- REDALYC. Red de Revistas Científicas de América Latina y el Caribe, España y Portugal
- REDINED. Red de Información Educativa
- RESH - Revistas Españolas de Ciencias Sociales (CSIC/CINDOC)
- ResearchBib. Academic Resource Index
- Scopus
- Web of Science (SSCI)
- WEBQUALIS

DIRECTORIOS Y BUSCADORES

- DOAJ
- Dulcinea
- Google Scholar
- LATINDEX (Publicaciones Científicas Seriadadas de América, España y Portugal)
- Recolecta
- Sherpa/Romeo
- Scirus
- Ulrich's Periodicals (CSA)

PORTALES Y REPOSITARIOS ESPECIALIZADOS

- Actualidad Iberoamericana
- Asociación Internacional de Estudios en comunicación social
- CLARISE - Comunidad Latinoamericana Abierta Regional de Investigación Social y Educativa
- Educ.ar
- Enlaces educativos en español de la Universitat de València
- e-sPacio-UNED. Repositorio institucional de la UNED
- Institut Français de l'éducation
- Plataforma de revistas 360º
- Red Iberoamericana de Revistas de Comunicación y Cultura
- REDIAL & CEISAL
- Universia. Biblioteca de recursos

CATÁLOGOS DESTACADOS DE BIBLIOTECA

- 360grados
- British Library
- Buz
- Catàleg Col·lectiu de les Universitats de Catalunya
- Catálogo Colectivo de Publicaciones
- Periódicas Español CCPP
- Catálogo de la Biblioteca de Educación (Ministerio de Educación, Cultura y Deporte)
- Catálogo del CSIC (CIRBIC)
- CENDOC
- CIDE
- CISNE
- COMPLUDOC
- COPAC (Reino Unido)
- ICDL
- INRP
- IOE (Institute of Education. University of London)
- Library of Congress (LC)
- KINGS
- MIGUEL DE CERVANTES
- Observatorio de revistas científicas de Ciencias Sociales
- REBIUN
- SUDOC (Francia)
- UBUCAT
- UIB
- WORDLCAT (OCLC)
- ZDB (Alemania)

La Asociación Iberoamericana de Educación Superior a Distancia (AIESAD) es una entidad sin ánimo de lucro, constituida por universidades o instituciones de educación superior que imparten sus ofertas educativas en esta modalidad de enseñanza y promueve el estudio e investigación del modelo de enseñanza superior a distancia. *RIED. Revista Iberoamericana de Educación a Distancia* es el instrumento de la AIESAD para la difusión internacional de los avances en la investigación e innovación dentro del ámbito de la enseñanza y aprendizaje abiertos y a distancia.



RIED. Revista Iberoamericana de Educación a Distancia es una publicación científica que se edita semestralmente los meses de enero y julio. Promueve el intercambio institucional con otras revistas de carácter científico. La *RIED* no se hará responsable de las ideas y opiniones expresadas en los trabajos publicados. La responsabilidad plena será de los autores de los mismos.



"Los textos publicados en esta revista están sujetos a una licencia "Reconocimiento-No comercial 3.0" de Creative Commons. Puede copiarlos, distribuirlos, comunicarlos públicamente, siempre que reconozca los créditos de la obra (autor, nombre de la revista, instituciones editoras) de la manera especificada en la revista."



AIESAD

Ried

Revista Iberoamericana de
Educación a Distancia

VOL. 29 N° 1

Enero, 2026

Index

Research and Case Studies

Generative artificial intelligence and educational autonomy: historical metaphors and ethical principles for pedagogical transformation

Marc Alier-Forment, M. José Casañ-Guerrero, J. Antonio Pereira-Varela, Fco José García-Peñalvo, Faraón Llorens-Largo

Artificial intelligence and collective intelligence in digital higher education: a quasi-experimental study using the Kampal platform

Alicia Martínez-De la muela, Juan Pablo Ruiz-Fuentes, José L. González-Geraldo

Impact of generative AI on university students' digital competences: experimental evidence based on the DigComp framework

Concepción González García, Nina Pallarés

AI-ED-SAT: design and validation of a questionnaire for self-assessment of teaching skills in educational AI

Andresa Sartor-Harada, Juliana Azevedo Gomes

Self-perception and usefulness of generative artificial intelligence among pre-service teachers

Ana María Pinto-Llorente, Vanessa Izquierdo-Álvarez, Marta M. Dolcet-Negre

Teachers and science communication on social media: development and initial validation of assessment instruments

Dieter Reynaldo Fuentes-Cancell, Odiel Estrada-Molina, Mónica Gutiérrez-Ortega

Immersive technologies in higher education: faculty profiles and barriers to integration

Julio Cabero-Almenara, Antonio Palacios-Rodríguez, Julio Barroso-Osuna, Carmen Siles-Rojas

Development and validation of the AI-enhanced Self-Regulated Learning (AI-SRL) scale

Mehmet Ali Ayaz, Fatih Karataş, Erkan Yüce

The influence of learning motivation and procrastination on ChatGPT dependence

Héctor Galindo-Domínguez, Martín Sainz-de-la-Maza, Lucía Campo, Daniel Losada-Iglesias

Feedback on learning with generative artificial intelligence in university students

María Verónica Leiva-Guerrero, Ignacio Araya Zamorano, Rafael Escobar Collins, Francisca Silva Castro

PathRAG application in adaptive learning with generative AI for inclusive and sustainable education

Rubén Juárez Cádiz

Smart leveling: an AI-driven adaptive learning strategy in higher education

Patricia Aldape-Valdes, Elvira G. Rincon-Flores, Leticia Castano, Sadie Guerrero

A language model-based recommender assessment system for short-answer questions in the intellectual property domain

David Bañeres Besora, Ana-Elena Guerrero Roldán, M. Elena Rodríguez González

Measuring writing skills in Spanish as a foreign language with generative artificial intelligence

M. Victoria Cantero Romero, María-Teresa Martín-Valdivia, Ana María Ortiz-Colón, Salud M. Jiménez-Zafra

Participatory methodology for creating microlearning digital resources with generative AI

Bell Manrique-Losada, Sandra Isabel Arango-Vásquez, Mariutsi Alexandra Osorio-Sanabria

Active learning in higher engineering education: a decalogue for the design of post-lecture online quizzes

Dubravka Novkovic, Ricardo Torres Kompen, Josep Petchamé

Generative artificial intelligence and educational autonomy: historical metaphors and ethical principles for pedagogical transformation

Inteligencia artificial generativa y autonomía educativa: metáforas históricas y principios éticos para la transformación pedagógica



- Marc Alier-Forment - *Universitat Politècnica de Catalunya, UPC (Spain)*
 María José Casañ-Guerrero - *Universitat Politècnica de Catalunya, UPC (Spain)*
 Juan Antonio Pereira-Varela - *Universidad del País Vasco, UPV/EHU (Spain)*
 Francisco José García-Peñalvo - *Universidad de Salamanca, USAL (Spain)*
 Faraón Llorens-Largo - *Universidad de Alicante, UA (Spain)*

ABSTRACT

This article examines the integration of generative artificial intelligence in education from a critical, historical, and ethical perspective. It highlights growing concerns about the opacity of current artificial intelligence tools, particularly in learning systems. The study adopts a metaphor-based approach to explore how technological narratives influence the adoption of educational innovations. It reviews historical metaphors used to describe educational technologies, from Multivac and Matrix to the free software Bazaar and the App Store, and proposes new conceptual frameworks that may better reflect the current context in which artificial intelligence is entering the educational sphere. Based on this metaphorical analysis, the article outlines seven fundamental ethical principles for the safe adoption of generative artificial intelligence in education, focusing on privacy, pedagogical alignment, human oversight, and technological transparency. These principles are illustrated through a practical application: the LAMB (Learning Assistant Manager and Builder) environment, an open-source software framework that enables the ethical and contextualized design of artificial intelligence-based learning assistants. The article presents real-world cases of LAMB implementation in higher education, including a controlled experience with students that demonstrates significant improvements in student autonomy and pedagogical coherence. Finally, it emphasizes how LAMB embodies the proposed ethical principles and responds to the identified critical metaphors, offering a model for technology integration centered on teacher autonomy, alignment with institutional values and practices, and meaningful student learning that prioritizes pedagogical control over technological determinism.

Keywords: educational technology; generative artificial intelligence; adaptive learning; learning assistant; ethics; instructional design.

RESUMEN

Este artículo analiza la integración de la inteligencia artificial generativa en educación desde una perspectiva crítica, histórica y ética. Se identifica una creciente preocupación por la opacidad de las herramientas de inteligencia artificial actuales, especialmente en sistemas de aprendizaje. El trabajo utiliza un enfoque basado en metáforas para entender cómo las narrativas tecnológicas influyen en la adopción de innovaciones educativas. Se revisan metáforas históricas en las tecnologías aplicadas a la educación, desde Multivac y Matrix hasta el Bazar del software libre y la App Store, y se proponen nuevas imágenes conceptuales que podrían aplicarse al contexto actual en el que la inteligencia artificial irrumpe en la educación. A partir de este análisis metafórico, se plantean siete principios éticos para una adopción segura de la inteligencia artificial generativa en educación, centrados en la privacidad, la alineación pedagógica, la supervisión humana y la transparencia tecnológica. Estos principios se ejemplifican con el entorno LAMB (Learning Assistant Manager and Builder), un marco de código abierto que permite diseñar asistentes de aprendizaje basados en inteligencia artificial de forma ética y contextualizada. Se presentan casos reales de aplicación de LAMB en educación superior, incluyendo una experiencia controlada con estudiantes que muestran mejoras significativas en autonomía y coherencia pedagógica. Finalmente, se destaca cómo LAMB encarna los principios éticos propuestos y responde a las metáforas críticas identificadas, ofreciendo un modelo de integración tecnológica centrado en la autonomía de los docentes, la alineación con los principios y prácticas de la institución educativa y el aprendizaje significativo de los estudiantes.

Palabras clave: tecnología educativa; inteligencia artificial generativa; aprendizaje adaptativo; asistente de aprendizaje; ética; diseño instruccional.

How to cite: Alier-Forment, M., Casañ-Guerrero, M. J., Pereira-Varela, J. A., García-Peñalvo, F. J., & Llorens-Largo, F. (2026). Generative artificial intelligence and educational autonomy: historical metaphors and ethical principles for pedagogical transformation [Inteligencia artificial generativa y autonomía educativa: metáforas históricas y principios éticos para la transformación pedagógica]. *RIED-Revista Iberoamericana de Educación a Distancia*, 29(1), 9-28.
<https://doi.org/10.5944/ried.45536>

INTRODUCTION

In September 2024, conversations were held with ten experts in educational innovation and technology regarding the "Manifesto for Safe AI in Education" (Alier, García-Peñalvo, Casañ et al., 2024). These discussions centered around a proposed evaluation framework for educational technologies based on generative artificial intelligence (GenAI). This framework resulted in a questionnaire for evaluating GenAI-based educational initiatives. The associated questionnaire focused on key aspects: data privacy, access control within institutions, and above all, the ability of tools to align with pedagogical objectives. Surprisingly, more than half of the experts considered that no currently available GenAI-based solution would meet these criteria or that it would be possible to meet them.

From these exchanges emerged a relevant observation: many technologists with extensive experience in the educational field do not view GenAI, and particularly adaptive learning systems based on AI that promise to personalize the student experience, as clear and malleable tools, but rather as powerful yet opaque artifacts, difficult to understand or control. This phenomenon led us to propose the metaphor of the *Palantír* (Tolkien, 1954-1955).

In J.R.R. Tolkien's narrative universe, the *Palantír* is a magical sphere that allows seeing through time and space and communicating with users of other *Palantírs*. However, using the *Palantír* also entails significant risks, as illustrated by the corruption of the wizard Saruman, which Sauron exercises through this object. This metaphor, applied to GenAI in education, captures two very present sentiments: fascination with its apparent power of personalization and prediction, and fear or skepticism about who controls it and for what purposes.

We will call this the *Palantír* metaphor: GenAI understood as a centralized, powerful, inevitable tool, but ultimately beyond the control of its real users (teachers, students, and institutions). This image helps explain an increasingly common attitude in the educational field: accepting the implementation of adaptive AI systems as they come, without deeply understanding how they are trained, what biases they may contain, or how decisions that affect individual learning are made.

For the purposes of this work, we use the term "GenAI-based learning assistants" to refer specifically to conversational tools that provide personalized support to students through natural language processing, distinguishing them from traditional adaptive systems that automatically adjust learning paths based on performance analysis.

The idea of technology viewed as "magic" has been widely discussed in the past. For example, Arthur C. Clarke explained in 1962 that sufficiently advanced technology could become something similar to magic (Clarke, 1962).

However, a distinction is necessary. Although the *Palantír* metaphor is useful for describing certain current narratives, it also has limitations. Artificial intelligence is not magic, and describing it as such may reinforce a deterministic view of technology that inhibits critical capacity and understanding of it.

Metaphors, when explicitly clarified, are valuable analytical tools. They reflect and influence how technologies are imagined, designed, implemented, and above all, how they are disseminated. As Weller and Hofstadter point out, metaphors and analogies are not just stylistic resources, but frameworks of thought that guide our decisions (Weller, 2022; Hofstadter & Sander, 2018).

For example, the metaphor of "digital natives" shaped educational policies for years, often with erroneous results. Similarly, the idea of an "Uber of education" inspired business models that ignored the complexity of learning (Adell-Segura et al., 2018).

Currently, there is growing pressure to integrate GenAI into educational systems. Faced with this situation, this work presents in section two a critical journey through the history of educational technology, with the objective of analyzing how different innovations have been introduced, what challenges they have posed, and what conceptual frameworks, often in the form of metaphors, have guided their understanding and use. Based on this historical and conceptual analysis, section three identifies fundamental ethical principles that the authors consider necessary for responsible incorporation of GenAI in educational contexts. Subsequently, section four proposes a practical application that translates these principles into concrete guidelines for the design and use of adaptive learning systems based on artificial intelligence. Finally, the article's conclusions are presented.

ANALYTICAL FRAMEWORK: METAPHORICAL EVOLUTION OF EDUCATIONAL TECHNOLOGIES

Throughout history, the introduction of new technologies in the educational field has been guided not only by technical advances, but also by narratives, metaphors, and imaginaries that have given meaning to their adoption and modeled their applications. This section explores how certain metaphors have shaped the evolution of educational technology in recent decades and what lessons can be extracted from this history to ultimately explain the current approach to GenAI (Llorens-Largo, 2019). As Weller (2022) reminds us, "Technology often comes wrapped in its own metaphors" (p. 59), and those initial metaphors can have profound consequences on how that technology develops and is applied. Along these lines, Mason (2018) empirically demonstrates how underlying metaphors structure educational technology discourse, suggesting that while it is necessary to use metaphors to understand new digital phenomena, some of the specific metaphors commonly employed may impede more reflective approaches to conceptualizing and implementing new technologies in the educational field.

This section has been divided into four key moments in the evolution of educational technologies, analyzing the dominant metaphors that accompanied each stage and their implications for learning forms, especially in relation to personalization and adaptation.

From Multivac to the Educational Mainframe: The Metaphor of Education as a Centralized System

Isaac Asimov, in his stories about Multivac, imagined a supercomputer capable of making crucial decisions for all humanity. This vision reflects a powerful idea: technology as a centralized entity that collects data, processes massive information, and directs society's actions (Asimov, 1955-1976).

This centralized and authoritarian vision of computers was also reflected in the first proposals for Computer-Assisted Education (CAE) or Computer-Assisted Instruction. B. F. Skinner, a key figure in behaviorism, proposed in his work *The Technology of Teaching* (Skinner, 1968) that computers could optimize learning

through systematic application of behaviorist principles. These systems did not conceive learning as an open or exploratory process, but as a series of trajectories programmed and controlled from a single center.

Applied to GenAI-based learning systems, Multivac represents a vision where the computer system dictates the pace and content presented to the student, relegating adaptability to a rigid supervision model. The optimization of predefined results was prioritized, not genuine personalization based on interests, contexts, or individual motivations.

This metaphor remains relevant today when certain systems tend to prioritize instructional efficiency over student autonomy, reproducing a logic of centralized control rather than self-directed learning.

From Matrix to Educational Pills: The Metaphor of Learning as Instantaneous Knowledge Transfer

In the film *The Matrix* (Wachowski & Wachowski, 1999), characters acquire complex skills, such as piloting helicopters, through immediate download of programs into their brains. This metaphor of "instantaneous knowledge transfer" became very influential during the CD-ROM era and the first e-learning platforms during the second half of the 1990s.

In education, it translated into the development of closed and standardized learning units, the so-called "educational pills," which aimed to offer rapid, modular, and autonomous training. Under this logic, learning became "consuming" discrete information packages, not actively constructing knowledge.

The relationship with contemporary generative AI tools is clear: many AI-based approaches continue to assume that learning is a matter of administering the appropriate dose of content, adjusted to the student's speed or level, but without necessarily considering the critical, reflective, or social dimension of learning.

The Matrix metaphor feeds the promise of GenAI as a magic solution, but also hides the risks of simplifying learning to a merely receptive, passive, and decontextualized process.

From the Bazaar to the Virtual Campus: The Metaphor of Collaborative Knowledge Construction

The next revolution in educational technology came with the web, and the web cannot be understood without free software culture. This movement was born in MIT laboratories, when Richard Stallman, in response to restrictions imposed by software user licenses (EULA), which were previously shared freely among programmers, published the *GNU Manifesto* (Stallman, 1985) and founded the *Free Software Foundation (FSF)* in 1985. The FSF provided a fundamental legal framework: the GPL license, which allowed developers to share their software while guaranteeing four essential freedoms to users: to use, study, share, and improve the software. This philosophy, summarized in the slogan "free as in freedom, not as in free beer," facilitated collaboration and source code exchange on an unprecedented scale (Ceruzzi, 2003).

In contrast to previous centralized visions, the free software movement can be synthesized in the Bazaar metaphor (Raymond, 2001), which introduced a new

paradigm in learning systems: open systems, where multiple actors collaborate, innovate, and adapt solutions in a decentralized and democratic manner.

In his influential essay *The Cathedral and the Bazaar* (Raymond, 2001), Eric S. Raymond observes that free software communities, where knowledge and code are shared and members gain influence based on their contributions rather than their status or position, offer a solution to the "endemic software crisis" described by Pressman in *Software Engineering: A Practitioner's Approach* (Pressman, 1991). This practical application of free software principles generates a split in the movement with the emergence of the *Open Source Initiative (OSI)*. The OSI proposed the term "open source software" as an alternative to "free software," introducing more flexible licenses (such as Apache or MIT) that were more attractive to large companies.

This philosophy inspired the creation of environments like Moodle and Sakai, learning management platforms that, being open source, allowed teachers, students, institutions, and developers to adapt tools to their specific needs. It was not just about consuming content, but about building dynamic, shared, and configurable learning environments.

In this context, the Bazaar metaphor points to the possibility of designing systems that are not black boxes, but open spaces where adaptations are understood, negotiated, and collectively reconfigured. This perspective reinforces the idea that adaptability should not only be algorithmic, but also pedagogical and social.

From the App Store to the Educational Cloud: The Metaphor of Invisible Dependency

With the emergence of smartphones and app stores (Apple's App Store and Google Play) the metaphor that "there's an app for everything" became established. The promise was to have immediate, personalized, and accessible solutions, but in exchange for almost total dependence on proprietary and opaque infrastructures, platforms, that sell security and convenience at the price of control. As Ben Thompson reminds us when explaining the true engine of these dynamics: "the reality is that platforms are not a chicken-and-egg problem: the first thing that matters is users, who attract developers despite obstacles" (Thompson, 2025). In other words, rather than each individual app, the platform itself is the center of power.

In education, there was a proliferation of educational apps whose information and management no longer occurs in systems controlled by institutions, but in "the cloud" (a new metaphor), offering customized learning experiences, but in exchange for a loss of control over data, processes, and personalization criteria. Algorithms decide what content to recommend, what routes to follow, what evaluations to apply, almost always without transparency or real participation from educators or institutions. Note how "the cloud" and the "algorithm" become subjects of analysis and criticism, when they are actually instruments operated by the companies that design and control them.

This metaphor invites us to ask: who adapts for whom? What commercial interests, what cultural or technical biases model the paths that students travel?

The platform metaphor, the App Store, warns that personalized learning can become simulated personalization, predefined by corporate interests, if mechanisms of technological and pedagogical sovereignty are not established.

Alternative Metaphors for AI-Powered Adaptive Learning

While the Palantír metaphor is useful for representing the fascination and fear surrounding the use of artificial intelligence in educational contexts, we need frameworks that allow us to think in more critical and pedagogically useful ways about integrating these technologies. For this purpose, the authors propose two alternative metaphors that can help imagine learning support systems with GenAI that are more open, controlled by educational communities, and focused on meaningful learning: the Lego building blocks set and the Prometheus dilemma.

The Lego building blocks metaphor can be used to explain that, instead of conceiving generative AI as a closed and magical artifact, the Lego metaphor poses a more accessible, modular, and comprehensible vision. Under this approach, artificial intelligence is not an "oracle" that offers automatic and unquestionable learning routes, but a set of reusable pieces that can be combined in multiple ways to design personalized, contextual learning experiences with pedagogical foundations.

Applied to GenAI-based learning tools, this implies that:

- GenAI-based tools can be configured by teachers to respond to the real needs of their students.
- The "algorithm" does not replace the educator, but offers materials with which to build alternative routes.
- Adaptations occur not only based on performance or speed, but also on interests, languages, motivations, and diverse trajectories.

This approach connects with a constructivist and participatory vision of learning, where technology does not replace pedagogical foundations, but amplifies them. In this sense, it can be linked to the *construction* metaphor, proposed by Weller (2022), which understands educational technology as a process in which users (teachers and students) actively construct their tools and learning environments, promoting autonomy and personalization. Additionally, it allows imagining futures in which systems are transparent, auditable, and created by educational communities, thus avoiding dependence on opaque platforms or automated decisions that cannot be explained.

A second relevant metaphor is the Prometheus dilemma. In Greek mythology, Prometheus steals fire from the gods and gives it to humans, an act that represents both an advance for civilization and potential punishment for humanity. "In the stalk of a reed I hid the spark, the fountain of stolen fire, which for mortals is the teacher of all arts; and for this crime I now pay the penalty, nailed with chains under the ether." (Jonas, 1984).

GenAI-powered tools can be understood as systems with the capacity to offer "personalized" recommendations, adapt content in real time, and predict educational trajectories. All of this represents a form of "fire," powerful and transformative, but also a risk if used without considering ethics and transparency.

This metaphor warns about:

- The possibility of excessively delegating educational judgment to automated algorithms, displacing the teacher's role.
- The risk of generating limited or closed trajectories for students, guided by historical data patterns, instead of opening new paths.

- The implications of reducing learning to consumption or efficiency patterns, ignoring more complex dimensions such as creativity, conflict, or exploration.

Thus, while the Lego metaphor points to the possibility of emancipatory use, the Prometheus metaphor reminds us that all powerful technology carries ethical and political tensions, especially when deployed on a large scale in educational systems.

ETHICAL PRINCIPLES FOR SAFE INTRODUCTION OF GENAI IN EDUCATION

The growing pressure to integrate GenAI into educational systems poses not only opportunities, but also ethical, legal, and pedagogical challenges of great magnitude. At the regulatory level, in the European Union, GenAI systems must guarantee safe adoption that respects the European Union's privacy regulations (European Parliament, & Council of the European Union, 2016). Additionally, its recent emergence has driven the development of specific legislation on artificial intelligence in various regions, such as the European Union's AI Act (European Parliament, 2024) or emerging regulations in China (Cyberspace Administration of China et al., 2023).

To guarantee safe adoption that respects privacy regulations and is aligned with institutional values, strategies, and practices, the authors propose in this section a series of principles to guide the evaluation and deployment of GenAI applications in education. These principles seek to ensure that GenAI-based technologies: 1) Align with institutions' educational objectives. 2) Maintain adequate levels of security, accuracy, and ethical integrity. 3) Favor quality and equitable learning, minimizing risks associated with misuse, privacy, or misinformation.

Following these principles, educational institutions will be able to take advantage of the opportunities offered by GenAI while mitigating the risks associated with its implementation in adaptive learning contexts.

Principles of Safe GenAI in Education

Beyond legal compliance, ethical integration of artificial intelligence requires practical and actionable principles that allow evaluating both the pedagogical adequacy and technological security of adopted solutions.

Based on an analysis of legal frameworks and the technical nature of GenAI, Alier, García-Peñalvo y Camba (2024) have proposed a series of principles that were subsequently expanded to seven specific guidelines for educational environments (García-Peñalvo et al., 2024). These principles allow for technical, pedagogical, and ethical evaluation of GenAI integration strategies in education. These principles are presented below.

- **(SAIE1) Confidentiality guarantee:** GenAI systems must strictly protect student data, ensuring the privacy of their identities, records, and educational interactions.
- **(SAIE2) Alignment with educational strategies:** GenAI tools must support institutional objectives and adhere to technological governance policies, facilitating learning and knowledge creation without fostering bad practices such as plagiarism or evasion of academic controls.

- **(SAIE3) Adjustment to didactic practices:** GenAI applications must operate under defined pedagogical parameters. They should not automatically solve problems or tasks, but guide the student in reasoning processes aligned with instructional design.
- **(SAIE4) Accuracy and error minimization:** AI systems must prioritize the truthfulness and relevance of their responses, especially to minimize risks of erroneous information derived from biases or model hallucinations.
- **(SAIE5) Comprehensible interface and appropriate behavior:** GenAI must present its functionalities and limitations clearly to teachers and students, avoiding false expectations or erroneous interpretations of its reliability level.
- **(SAIE6) Human supervision and responsibility:** Decisions based on GenAI must be supervised by people. Faculty must maintain control over educational processes, ensuring the possibility of appeal and review of automated decisions.
- **(SAIE7) Ethical training and transparency:** GenAI models must be trained ethically, guaranteeing transparency about their data sources and methodologies, and minimizing biases to ensure fair and explainable results.

The application of these principles has important practical consequences for educational institutions:

- **Confidentiality (SAIE1):** Institutions must guarantee control over AI platforms used, avoiding mandatory registration in external tools like ChatGPT or Gemini, which could violate student privacy.
- **Strategic alignment (SAIE2):** General-purpose tools, such as Large Language Models (LLM), present institutional integration difficulties due to their complexity of use (Willison, 2023) and risks associated with academic integrity (González-Geraldo & Ortega-López, 2024). Additionally, adding complexity to the learning process is not good pedagogical practice, as it increases students' cognitive load (Chen et al., 2023).
- **Didactic adaptation (SAIE3):** Applications must integrate coherently with educational methodologies. For example, in environments such as medical training (Hwang et al., 2024), GenAI systems must act as assistants in clinical reasoning, not as automatic substitutes for students' analytical processes.
- **Accuracy and error minimization (SAIE4):** In educational environments, it is fundamental to design adaptive AI systems capable of citing sources and providing information validation mechanisms (Tonmoy et al., 2024).
- **Interface and transparency (SAIE5):** Interface design must favor understanding of GenAI's scope and limitations, avoiding the presentation of speculative responses as absolute truths.
- **Supervision and responsibility (SAIE6-7):** Human supervision not only ensures educational quality, but protects equity in learning and guarantees that adaptive systems do not reproduce biases or systematic exclusions.

These principles have led to the development of the Safe AI Manifesto (Alier, García-Peñalvo, Casañ et al., 2024), a living document that gathers the necessary commitments for ethical and responsible integration of GenAI in education, available for consultation and subscription at <https://manifesto.safeaieducation.org>. Additionally, the above principles allow for technical, pedagogical, and ethical

evaluation of GenAI integration strategies in education, as developed in (García-Peñalvo et al., 2024).

Relationship between Safe GenAI Principles in Education and Educational Metaphors

Each of the proposed safe GenAI principles responds, in some way, to the risks or aspirations represented by the historical metaphors analyzed in the previous section:

- **(SAIE1) Confidentiality guarantee** combats the risk of the App Store and the educational cloud: The App Store metaphor shows how dependence on opaque platforms puts privacy at risk. Confidentiality requires reversing that logic, ensuring that student data is protected and under institutional control.
- **(SAIE2) Alignment with educational strategies** responds to the danger of Matrix. Faced with the vision of learning as passive download of predefined information, aligning GenAI with educational strategies allows building systems that foster active learning processes, instead of automating content transmission.
- **(SAIE3) Adjustment to didactic practices** connects with the Bazaar metaphor. As in the free software Bazaar, the principle seeks that GenAI tools be configurable and adaptable by teachers to specific pedagogical practices, not closed products that impose standard methodologies.
- **(SAIE4) Accuracy and error minimization** combats the illusions of the Prometheus vision. The Prometheus dilemma reminds us that the "fire" of GenAI can be dangerous if not managed properly. Minimizing errors and hallucinations helps control that risk inherent to systems as powerful as they are fragile.

LAMB: A PRACTICAL PROPOSAL FOR INTEGRATING SAFE GENAI IN EDUCATION

After analyzing the evolution of metaphors that have accompanied educational technologies and defining the principles of safe GenAI (SAIE), it becomes necessary to show concrete examples that include these principles in their design and application. This section presents LAMB (Learning Assistant Manager and Builder), a software framework designed specifically to create GenAI-based learning assistants in a safe, controlled manner that is pedagogically aligned with the educational institution's principles. First, the concept of GenAI-based learning assistants is explained. Next, LAMB is defined and how this framework allows creating GenAI-based learning assistants. After that, LAMB use cases are presented as well as how this framework is aligned with the principles presented in section 3 and finally its relationship with educational metaphors.

GenAI-Based Learning Assistants

A GenAI-based learning assistant is a conversational tool that interacts with students to offer personalized educational support. Unlike a generic chatbot, a learning assistant is designed to fulfill specific functions within an educational context (Kochmar et al., 2020; Wollny et al., 2021). The emergence of tools like ChatGPT has

marked a turning point in this field, transforming the possibilities for personalization and educational interaction (Bettayeb et al., 2024; Chan & Lee, 2023):

- Respond to questions about course content.
- Guide problem-solving by applying defined pedagogical methods.
- Suggest relevant support resources.
- Promote reasoning and self-explanation in students.
- Accompany formative assessment processes by providing adjusted feedback.

These assistants do not seek to replace faculty, but to amplify their capacity for accompaniment and feedback, especially in environments with large numbers of students or in hybrid and distance modalities.

LAMB

LAMB is a software environment that allows educational institutions and teachers to create their own learning assistants, adapted to each subject, course, or specific need. This framework also allows integrating these assistants simply into learning management platforms (LMS) like Moodle, through the IMS LTI (Learning Tools Interoperability) standard, so that students access learning assistants through the institution's LMS (IMS Global Learning Consortium [IMS GLC], 2014). Additionally, LAMB allows controlling and configuring the knowledge base that feeds the assistant's responses, ensuring pedagogical alignment with the subject matter. Additionally, this framework allows managing the security, privacy, and supervision of created assistants.

Alier, Pereira et al. (2024) present the technical description, as well as LAMB's detailed architecture. In summary and from a technical perspective, LAMB combines:

- Large Language Models (LLMs) that generate responses in natural language.
- Retrieval-Augmented Generation (RAG) techniques, which limit the assistant's responses to verified information from sources controlled by the institution.
- A system of interaction templates and pedagogical configurations that the teacher defines previously.

In this way, LAMB allows leveraging the power of generative AI, but under conditions of control, transparency, and human responsibility.

Examples of Assistants Created with LAMB

To illustrate LAMB's real application in educational contexts, some of the experiences carried out are presented.

First, it is important to highlight the experience developed in the Economics and Economic Environment subject, a mandatory second-year subject in the Computer Engineering Degree at the Barcelona Faculty of Informatics.

Before addressing pilot experiences with students, it is important to contextualize LAMB's development process through the assistant called "Macroeconomics Study Coach," which functioned as a prototype for the design and refinement of the LAMB framework. This assistant was not subjected to formal

evaluation with students, but served as a technical and pedagogical testbed during LAMB system development iterations.

This assistant was created from 30 course lecture videos. Its main functionality consisted of answering student questions based on video transcriptions and complementary PDF documents, providing specific citations and links to videos with precise moments where the queries made were addressed.

Once framework stability was consolidated through this case, controlled experiences with students were designed and implemented, with the PESTLE assistant being the first of these empirical validations.

The PESTLE case learning assistant was designed as a resource for the Business Administration and Management Fundamentals subject, taught in the Computer Engineering degree at EPSEVG (Vilanova i la Geltrú Higher Polytechnic Engineering School). The created assistant helped students work on an evaluation case of a technology-driven business project. The assistant had been equipped with a specific knowledge base for the case. Students, organized in teams, had to analyze the case using the PESTLE methodology (Political, Economic, Social, Technological, Legal, and Environmental), which implied they had to pose questions from various dimensions before writing their final report. They posed these questions to the learning assistant created with LAMB. In previous courses, students simulated being experts with help from search engines, specific documents, and even ChatGPT in autumn 2023 (Casañ et al., 2024).

The experience with the assistant to help students in the PESTLE case was developed during the 2023-2024 academic year. The activity involved a total of 47 second-year degree students organized in groups of 6-8 members.

The experience was structured following a case study methodology focused on analyzing Tesla's Optimus humanoid robot, using the PESTLE framework (Political, Economic, Social, Technological, Legal, Environmental). The activity was developed during two 2-hour class sessions, with a one-week interval between them to allow students to continue analysis autonomously.

In the first session, the case was presented and the PESTLE methodology explained, collaborative group work to analyze the six PESTLE dimensions was conducted, as well as using the assistant to obtain expert information. During the week, students had time to continue with the analysis using the assistant from home. In the second session, SWOT analysis integrated with PESTLE was introduced. Next, students were asked to categorize the elements or aspects they had found in each PESTLE dimension as "highly relevant," "relevant," or "not very relevant." Finally, students prepared a final report with conclusions.

The data collection methodology to evaluate the assistant's effectiveness was administered through a non-mandatory 5-question questionnaire using a 5-point Likert scale (1=strongly disagree, 5=strongly agree). The questionnaire obtained 27 responses from a total of 47 students (57.4% response rate). The questionnaire questions were as follows:

1. The assistant has helped me find relevant information for the case more quickly than if I had had to do it myself using the internet
2. The additional questions suggested by the system are good prompts to search for more information about the case
3. The answers to the questions suggested by the system provide useful information

4. Being able to consult data sources has been useful
5. The assistant's interface is simple and easy to use

The results show a mostly positive reception of the assistant by students:

- **Question 1** (Efficiency in information search): 83.3% of students scored 4 or 5 (17 favorable responses out of 27 total)
- **Question 2** (Quality of suggested questions): 77.7% scored 4 or 5
- **Question 3** (Usefulness of responses): 85.2% scored 4 or 5
- **Question 4** (Access to sources): 81.5% scored 4 or 5
- **Question 5** (Interface usability): 92.6% scored 4 or 5

The results indicate that no student scored 1 or 2 on questions 1, 2, and 4, and only a small percentage did so on questions 3 and 5, evidencing a significant absence of negative evaluations.

Qualitative and quantitative results confirm several pedagogical benefits:

In general, as positive feedback in the experience, student responses highlighted benefits such as consistency and alignment. The assistants' ability to stick to the provided knowledge base was highlighted, without generating irrelevant information or hallucinations. Regarding workload reduction, the usefulness of assistants to answer repetitive queries about course procedures was valued. In the specific case of the assistant to help students resolve an evaluation of a business project, an improvement in the quality of student reports was observed thanks to access to expert knowledge.

Among the challenges and limitations, it is worth highlighting excessive dependence on these types of tools, as there was a warning about the risk of students using the assistant as a shortcut, without deepening the course materials provided by faculty. Future improvements to LAMB were also proposed for more fluid integration with the LMS, specifically to allow sending queries to the teacher when the assistant has no response. Finally, the need to support minority languages such as Catalan or Basque was pointed out.

In previous courses, students simulated being experts using search engines, specific documents, and, in the autumn quarter of 2023, ChatGPT in an uncontrolled manner. The implementation of the LAMB assistant provided a more controlled and pedagogically aligned environment, with information verified and contextualized specifically for the case study.

Relationship between LAMB and Safe GenAI Principles in Education

LAMB's design directly reflects compliance with the seven safe AI principles outlined in section three:

- **SAIE1 (Confidentiality):** LAMB guarantees student data privacy by operating on institutional servers or through controlled APIs. Sensitive information is not exposed to unauthorized external platforms.
- **SAIE2 (Alignment with educational strategies):** Assistants created with LAMB only use sources approved by the institution, ensuring that content and interactions are aligned with institutional educational objectives and values.

- **SAIE3 (Adjustment to didactic practices):** LAMB allows teachers to define interaction templates and specific knowledge bases for each course or activity, integrating coherently with instructional designs, instead of imposing generic interactions.
- **SAIE4 (Accuracy and error minimization):** Through Retrieval-Augmented Generation (RAG) techniques, assistants base their responses on reliable reference documents, reducing the risk of hallucinations or errors typical of generalist LLMs.
- **SAIE5 (Comprehensible interface and appropriate behavior):** LAMB provides assistants with predictable and delimited behaviors, clarifying their function and limitations, avoiding the illusion of infallibility that often accompanies many AI systems.
- **SAIE6 (Human supervision and responsibility):** Faculty design and supervise the contents, strategies, and behaviors of the assistant, preserving their central role as pedagogical mediator. This approach is consistent with practical experiences where LLMs act as assistants that automate repetitive tasks (code review, project monitoring), but maintaining the teacher in the role of guide and supervisor of the educational process (Pereira et al., 2025).
- **SAIE7 (Ethical training and transparency):** While underlying models are from third parties, LAMB allows institutions to build their own interaction bases, which in the future can be used to adjust or train models ethically adapted to their contexts.

LAMB and Educational Metaphors

The practical implementation of LAMB and learning assistants in real contexts has revealed difficulties that significantly, validate the warnings contained in the metaphorical analysis developed in the first part of this article. This convergence between critical theory and empirical experience reinforces the importance of maintaining active vigilance over the narratives that accompany technological integration in education.

One of the most evident concerns observed in LAMB experiences is the risk of excessive dependence by students, who may use the assistant as a shortcut that prevents them from deepening the course materials provided by faculty. This practical difficulty materializes precisely the dangers anticipated in the Palantír metaphor.

This observation does not invalidate the usefulness of assistants, but reinforces the need for clear pedagogical frameworks that guide their use. As one of the consulted experts noted: "it must be clearly defined that its 'knowledge' does not go beyond the corresponding content and does not address content beyond the teacher's needs." The Palantír metaphor reminds us that the power of technology must be accompanied by structures of control and ethical reflection.

The integration difficulties with learning management systems (LMS) observed in practical experiences reflect the tension between the Bazaar and App Store metaphors analyzed earlier. Students have requested improvements for more fluid integration with the Moodle platform (the LMS used in the educational institution), especially the possibility of making queries to the teacher when the assistant has no adequate response.

This demand illustrates the inherent limitation of closed systems (App Store metaphor). When technological tools operate in isolation, without capacity for

interconnection and collaboration, frustration is generated and pedagogical potential is limited. In contrast, the Bazaar metaphor suggests the need for open, modular, and collaborative systems where different tools and actors (students, teachers, systems) can interact fluidly.

LAMB, by design, attempts to respond to this tension through its modular architecture and integration with open standards like LTI. However, the practical difficulties observed underscore that technical openness is not sufficient: pedagogical openness is also required that allows connection between artificial intelligence and the human intelligence of faculty.

The identified need for support of minority languages such as Catalan and Basque constitutes a concrete manifestation of the limitations of the Matrix metaphor discussed in the conceptual framework. The Matrix vision suggests that knowledge can be "downloaded" instantaneously and universally, ignoring specific contextual, cultural, and linguistic realities.

Practical experience with LAMB demonstrates that this apparent universality is illusory. Language models, however advanced they may be, reflect the biases and limitations of their training data, which tend to favor majority languages.

From a conceptual perspective, LAMB rejects the logic of the App Store and the Palantír, where users depend on closed and opaque systems. Instead, it aligns with the Lego metaphor: assistants are modular, customizable, and transparent tools that are adaptable.

In summary, LAMB offers a realistic example of how institutions can appropriate educational GenAI as a didactic resource and not as a threat to their autonomy or pedagogical agency.

CONCLUSIONS

This article has presented a critical perspective on the incorporation of GenAI in education, combining historical, ethical, and conceptual analysis. Through the use of metaphors, it has shown how visions of educational technology have oscillated between centralized control and participatory autonomy. Metaphors such as the Palantír, Matrix, or App Store help understand current risks: opacity, technological dependence, and pedagogical misalignment. In response, alternative frameworks have been proposed, such as Lego blocks or the Prometheus dilemma, which invite critical, transparent integration controlled by the educational community.

Based on this foundation, seven principles for safe GenAI have been formulated, centered on confidentiality, institutional alignment, human supervision, and transparency. These principles are not theoretical, but have been applied in practice through LAMB, a software environment that allows institutions to design and deploy configurable and auditable learning assistants. The analyzed cases show clear benefits in pedagogical coherence, student autonomy, and reduction of repetitive tasks, but also point to challenges such as dependency or integration with educational institutions' LMS.

Collectively, this work advocates for the need to recover pedagogical control over technology, to ensure that artificial intelligence in education serves meaningful learning and not external or opaque interests.

REFERENCES

- Adell-Segura, J., Castañeda, L., & Esteve-Mon, F. M. (2018). ¿Hacia la ubersidad? Conflictos y contradicciones de la universidad digital. *RIED-Revista Iberoamericana de Educación a Distancia*, 21(2), 51-68. <https://doi.org/10.5944/ried.21.2.20669>
- Alier, M., García-Peñalvo, F. J., & Camba, J. D. (2024). Generative artificial intelligence in education: From deceptive to disruptive. *International Journal of Interactive Multimedia and Artificial Intelligence*, 8(5), 5-14. <https://doi.org/10.9781/ijimai.2024.02.011>
- Alier, M., García-Peñalvo, F. J., Casañ, M. J., Pereira, J. A., & Llorens-Largo, F. (2024). *Safe AI in Education Manifesto (v 0.4.0)*. <https://manifesto.safeaieducation.org>
- Alier, M., Pereira, J., García-Peñalvo, F. J., Casañ, M. J., & Cabré, J. (2024). LAMB: An open-source software framework to create artificial-intelligence assistants deployed and integrated into learning-management systems. *Computer Standards & Interfaces*, 92, 103940. <https://doi.org/10.1016/j.csi.2024.103940>
- Asimov, I. (1955-1976). *Historias de Multivac* [Serie de relatos]. Diversas publicaciones.
- Bettayeb, A. M., Abu Talib, M., Sobhe Altayasinah, A. Z., & Dakalbab, F. (2024). Exploring the impact of ChatGPT: Conversational AI in education. *Frontiers in Education*, 9, 1379796. <https://doi.org/10.3389/feduc.2024.1379796>
- Casañ, M. J., Alier, M., Pereira, J., & García-Peñalvo, F. J. (2024). Asistentes de aprendizaje basados en inteligencia artificial: Principios de seguridad y experiencias de implementación en educación superior. In M. Navarro-Granados, J. J. Sánchez Amate, P. Berbel Oller, & C. Rodríguez Jiménez (Eds.), *Investigación y conocimientos en la educación actual* (pp. 13-35). Dykinson.
- Ceruzzi, P. E. (2003). *A history of modern computing* (2nd ed.). MIT Press.
- Chan, C. K. Y., & Lee, K. K. W. (2023). Students' voices on generative AI: Perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education*, 20(1), 43. <https://doi.org/10.1186/s41239-023-00411-8>
- Chen, O., Paas, F., & Sweller, J. (2023). A cognitive-load theory approach to defining and measuring task complexity through element interactivity. *Educational Psychology Review*, 35(2), 63. <https://doi.org/10.1007/s10648-023-09782-w>
- Cyberspace Administration of China, National Development and Reform Commission, Ministry of Education, Ministry of Science and Technology, Ministry of Industry and Information Technology, Ministry of Public Security, & National Radio and Television Administration. (2023). *Interim measures for the management of generative artificial intelligence services*. Promulgadas el 13 de julio de 2023, en vigor desde el 15 de agosto de 2023. https://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm
- Clarke, A. C. (1962). *Profiles of the future: An inquiry into the limits of the possible*. Harper & Row.
- European Parliament. (2024, 13 de junio). *Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (published in OJ on 12.7.2024)*. <https://data.europa.eu/eli/reg/2024/1689/oj>
- European Parliament, & Council of the European Union. (2016, 27 de abril). *Regulation (EU) 2016/679 of the European Parliament and of the Council on the protection of natural persons with regard to the processing of personal data*

- and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (OJ L 119, 4.5.2016, pp. 1-88). <https://data.europa.eu/eli/reg/2016/679/oj>
- García-Peñalvo, F. J., Alier, M., Pereira, J., & Casañ, M. J. (2024). Safe, transparent, and ethical artificial intelligence: Keys to quality sustainable education (SDG 4). *International Journal of Educational Research and Innovation*, 22, 1-21. <https://doi.org/10.46661/ijeri.11036>
- González-Geraldo, J. L., & Ortega-López, L. (2024). Can AI fool us? University students' ability to detect ChatGPT. *Education in the Knowledge Society*, 25, e31760. <https://doi.org/10.14201/eks.31760>
- Hofstadter, D. R., & Sander, E. (2018). *La analogía: El motor del pensamiento* (R. Musa, Trad.). Tusquets.
- Hwang, G.-J., Tang, K.-Y., & Tu, Y.-F. (2024). How artificial intelligence supports nursing education: Profiling the roles, applications and trends of AI in nursing education research (1993–2020). *Interactive Learning Environments*, 32(1), 373-392. <https://doi.org/10.1080/10494820.2022.2086579>
- IMS Global Learning Consortium. (2014). *IMS Learning Tools Interoperability (LTI) implementation guide v2.0*.
- Jonas, H. (1984). *The imperative of responsibility: In search of an ethics for the technological age*. University of Chicago Press. <https://doi.org/10.7208/chicago/9780226850337.001.0001>
- Kochmar, E., Vu, D. D., Belfer, R., Gupta, V., Serban, I. V., & Pineau, J. (2020). Automated personalized feedback improves learning gains in an intelligent tutoring system. In I. I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, & E. Millán (Eds.), *Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6-10, 2020, Proceedings, Part II* (Lecture Notes in Computer Science, Vol. 12164, pp. 140-146). Springer. https://doi.org/10.1007/978-3-030-52240-7_26
- Llorens-Largo, F. (2019, 13 de febrero). Las tecnologías en la educación: Características deseables, efectos perversos. *Universidad, sí*. <https://www.universidadsi.es/las-tecnologias-en-la-educacion-caracteristicas-deseables-efectos-perversos/>
- Mason, J. (2018). A critical metaphor analysis of educational technology research in the social studies. *Contemporary Issues in Technology and Teacher Education*, 18(3). <https://citejournal.org/volume-18/issue-3-18/social-studies/a-critical-metaphor-analysis-of-educational-technology-research-in-the-social-studies/>
- Pereira, J., López-Gil, J. M., & Alier, M. (2025). The AI-powered classroom: LLMs as teacher assistants for enhanced software-engineering learning experiences. In R. Molina Carmona, C. J. Villagrà Arnedo, P. Compañ Rosique, F. García-Peñalvo, & A. García-Holgado (Eds.), *Proceedings of TEEM 2024: The Twelfth International Conference on Technological Ecosystems for Enhancing Multiculturality* (pp. 115-124). Springer. (Lecture Notes in Educational Technology). https://doi.org/10.1007/978-981-96-5658-5_12
- Pressman, R. S. (1991). *Software engineering: A practitioner's approach* (3rd ed.). McGraw-Hill.
- Raymond, E. S. (2001). *The cathedral & the bazaar*. O'Reilly.
- Skinner, B. F. (1968). *The technology of teaching*. Appleton-Century-Crofts.
- Stallman, R. M. (1985). *The GNU manifesto*. Free Software Foundation. <https://www.gnu.org/gnu/manifesto.html>
- Thompson, B. (2025, 5 de mayo). *Platform power is underrated*. Stratechery. <https://stratechery.com/2025/platform-power-is-underrated/>
- Tolkien, J. R. R. (1954-1955). *The lord of the rings* (Vols. 1-3). George Allen & Unwin.
- Tonmoy, S. M. T. I., Mehedi Zaman, S. M., Jain, V., Rani, A., Rawte, V., Chadha, A., & Das, A. (2024). A comprehensive survey of hallucination-mitigation techniques in

large language models. *arXiv preprint arXiv:2401.01313*.

<https://doi.org/10.48550/arXiv.2401.01313>

Wachowski, L., & Wachowski, L. (1999). *The Matrix* [Película]. Warner Bros.

Weller, M. (2022). *Metaphors of ed tech*. Athabasca University Press.

<https://doi.org/10.15215/aupress/9781771993500.01>

Willison, S. (2023). *Prompt injection attacks: It's only going to get worse*.

<https://simonwillison.net/2023/Apr/25/prompt-injection>

Wollny, S., Schneider, J., Di Mitri, D., Weidlich, J., Rittberger, M., & Drachsler, H. (2021). Are we there yet? A systematic literature review on chatbots in education. *Frontiers in Artificial Intelligence*, 4, 654924.

<https://doi.org/10.3389/frai.2021.654924>

Date of reception: 1 June 2025

Date of acceptance: 27 August 2025

Date of approval for layout: 26 September 2025

Date of publication in OnlineFirst: 16 October 2025

Date of publication: 1 January 2026

Artificial intelligence and collective intelligence in digital higher education: a quasi-experimental study using the Kampal platform

Inteligencia artificial e inteligencia colectiva en la educación superior digital: estudio cuasi-experimental con la plataforma Kampal



 Alicia Martínez-De la muela - *Universidad Internacional de La Rioja, UNIR (Spain)*

 Juan Pablo Ruiz-Fuentes - *Universidad Internacional de La Rioja, UNIR (Spain)*

 José L. González-Geraldo - *Universidad de Castilla-La Mancha, UCLM (Spain)*

ABSTRACT

The emergence of Artificial Intelligence (AI) in higher education poses the challenge of integrating personalized learning with collaborative learning. Collective intelligence is presented in this study as a tool that allows both approaches to be articulated, transforming individual contributions into shared knowledge through AI-mediated technologies. Within this framework, the perception of the Kampal platform as a learning environment that simultaneously promotes student autonomy and collective knowledge construction is analyzed. A pretest-posttest quasi-experimental design was used with an initial sample of 399 digital distance higher education students and a final matched sample of 25 cases. Three dimensions were evaluated through a structured questionnaire: familiarity with the tool, perception of its effectiveness, and expectations of use or future willingness to use similar technologies. The results show a significant increase in familiarity with Kampal and willingness to use collective intelligence tools in the future, although no significant changes were found in expectations or perceptions of immediate effectiveness. Factor analysis revealed two main dimensions: familiarity and effectiveness/expectations. The findings suggest that these tools promote a favorable attitude toward their future use, although without significant changes in perceived effectiveness. This highlights the need for intentional teacher mediation. When properly supervised, AI empowers hybrid environments that integrate personalization and collaboration, where collective intelligence operates as a nexus, facilitating complex dynamics such as swarm intelligence in higher education.

Keywords: artificial intelligence; AI-mediated learning; higher education; collective intelligence; distance education; Kampal.

RESUMEN

La irrupción de la Inteligencia Artificial (IA) en educación superior plantea el desafío de integrar el aprendizaje personalizado con aprendizaje colaborativo. La inteligencia colectiva se sitúa en este estudio como una herramienta que permite articular ambos enfoques, al transformar aportaciones individuales en conocimiento compartido mediante tecnologías mediadas por IA. En este marco, se analiza la percepción del uso de la plataforma Kampal como entorno de aprendizaje que favorece simultáneamente la autonomía del estudiante y la construcción colectiva del conocimiento. Se empleó un diseño cuasi-experimental pretest-posttest con una muestra inicial de 399 estudiantes de educación superior a distancia digital y una muestra final emparejada de 25 casos coincidentes. A través de un cuestionario estructurado, se evaluaron tres dimensiones: familiaridad con la herramienta, percepción de eficacia de esta y expectativas de uso o disposición futura a utilizar tecnologías similares. Los resultados muestran un incremento significativo en la familiaridad con Kampal y en la disposición a emplear herramientas de inteligencia colectiva en el futuro, aunque no se hallaron cambios significativos en las expectativas ni en la percepción de eficacia inmediata. El análisis factorial reveló dos dimensiones principales: familiaridad y eficacia/expectativas. Los hallazgos sugieren que estas herramientas fomentan una actitud favorable hacia su uso futuro, aunque sin cambios significativos en la percepción de eficacia. Esto resalta la necesidad de una mediación docente intencionada. Supervisada adecuadamente, la IA potencia entornos híbridos que integran personalización y colaboración, donde la inteligencia colectiva opera como nexo, facilitando dinámicas complejas como la inteligencia de enjambre en educación superior.

Palabras clave: inteligencia artificial; aprendizaje mediado por IA; educación superior; inteligencia colectiva; educación a distancia; Kampal.

How to cite: Martínez-De la muela, A., Ruiz-Fuentes, J. P., & González-Geraldo, J. L. (2026). Artificial intelligence and collective intelligence in digital higher education: a quasi-experimental study using the Kampal platform [Inteligencia artificial e inteligencia colectiva en la educación superior digital: estudio cuasi-experimental con la plataforma Kampal]. *RIED-Revista Iberoamericana de Educación a Distancia*, 29(1), 29-51. <https://doi.org/10.5944/ried.45560>

INTRODUCTION

Reference to Artificial Intelligence (AI), especially within the educational field, has become risky but also significantly synonymous with ChatGPT or, failing that, with other Large Language Models (LLM) such as Copilot, Gemini, or Claude, to mention some of the main LLMs, all of them proprietary, although there are some open source alternatives. However, as becomes evident, as soon as we observe the evolution of the use and implementation not only of AI, but also of Information and Communication Technologies (ICT) in education and especially in distance education (Bozkurt, 2023), we realize that—without having to resign ourselves to the merely generative text (as also occurred before with other technological advances, for example, MOOC/NOOC)—AI requires rethinking current educational models both urgently and profoundly as pointed out by García-Aretio (UNED, 2025).

Strictly speaking, the concept of AI has been with us for almost seventy years, having evolved irregularly—known as the winters of AI—to the present day (Russell & Norvig, 2020). Without detailing the different subfields that comprise it (*machine learning*, for example), as well as the different AIs that exist depending on their specificity (weak-strong AI), we can comprehend that under the concept of AI, we refer to the use of algorithms that allow performing and streamlining intelligent work normally associated with human beings. In this sense, it should also be noted that, especially in education, it is not so much a matter of seeing what can be done *by* us through AI substitution but rather what AI can do *for* us—provision, and even *with* us—collaboration.

It is also appropriate to consider the roadmap proposed by OpenAI towards artificial general intelligence, which establishes five progressive levels (Table 1) that guide the current and future development of these technologies (Duenas & Ruiz, 2024).

Table 1
OpenAI's Five-Level Framework for AI Development

Level	Name	Key features
1	Chatbots	Conversational Language Skills
2	Reasoning [skills]	Problem-solving at a human level
3	Agents	Autonomous action
4	Innovative [actions]	Supporting invention and creativity
5	Organizations	Complex organizational management

Source: Duenas & Ruiz (2024, p. 3)

This line addresses the potential of AI in digital distance education, going beyond LLM or Level 1, while, as we will see, reflecting on some of the challenges we currently face, whether we like it or not. In this sense, the European Union Regulation on Artificial Intelligence (European Union, 2024), commonly referred to as the "AI Act," must be a key piece that establishes ethical and responsible guidelines and a roadmap for these new platforms to effectively catalyze educational improvement, rather than simply delegate responsibilities. Obviously, such responsibilities must always remain not only in human hands but also in expert hands, especially in sensitive areas such as justice, health, and, in the case at hand, education—particularly higher education.

For these reasons, platforms such as the one we are dealing with, focused on the use of AI for a collective and pedagogical purpose, would allow us to act as an extension of group memory, optimize collective attention, strengthen collective reasoning, carry out tutoring on a personalized scale and, consequently and in parallel, save time on routine tasks (Riedl & De Cremer, 2025). Of course, as these and other authors point out, these benefits are not without risks: biases, dependency, digital gaps, resistance to change, etc. Among them, as we will see in more detail in the final part of the article, is the potential reduction of intellectual diversity (Riedl & Bogert, 2024). In contrast, collaborative work in virtual environments enhances the development of intrinsic motivation and strengthens self-esteem by reducing feelings of isolation and allowing each participant to perceive their contribution as essential to the collective achievement (Marrón Luna, 2021).

This article is framed within this critical need, with a particular focus on a key issue in the current pedagogical debate: how to integrate personalized learning with collaborative learning in AI-mediated environments, including a balance between AI-mediated interaction and the human factor in learning contexts and the collective construction of knowledge (Msambwa et al., 2025). While current applications tend to prioritize individualized learning through conversational systems, there is still little development of tools that promote collective processes of knowledge construction. Despite the exponential advancement of everything related to AI, this scarcity extends to the lack of evidence in the literature regarding platforms that go beyond language generation models, particularly concerning the concept of AI-assisted collective intelligence.

Collective intelligence is therefore presented in this study as a tool that supports both personalized learning, through the review of contributions and the monitoring of individual behavior, and collective learning, revealing a potential to share knowledge and reach joint results through a digital platform. Kampal Collective Learning (KCL) is a collective intelligence platform developed by Kampal Data Solutions, a spin-off company from the University of Zaragoza. Its main objective is to facilitate collaborative problem-solving by combining individual and collective contributions in real time. The platform operates through structured phases where participants can propose, review, and modify ideas, promoting dynamic and collaborative learning.

Before continuing, it is essential not only to point out that Kampal is not a *chatbot*, but instead, the use of AI is far removed from the control and knowledge of the users themselves and, in this case, also its beneficiaries. In other words, neither teachers nor students are aware of what and how the algorithms that underlie Kampal work, and which make up the work groups and networks on which this Collective Intelligence is built. We are, therefore, dealing with a platform that considers AI from the prism of Big Data (BD) and Machine Learning (ML), and which, as we will also discuss in the final part of these lines, among other issues, faces the elusive traceability of AI models typified as "black box" models. Hence, to complement and differentiate our work from the avalanche of studies on the relationship between AI (LLM) and education, and to explore the extent to which other platforms are accepted, their impact, and potential use in our educational programs, we selected the Kampal tool (<https://ic.kampal.com/?locale=es>). The ultimate objective of Kampal is to enable and enhance the use of Collective Intelligence (CI, also known as *Thinkhub*) for collaborative problem-solving through the use of AI.

Although the literature on the use of Kampal is not very extensive, we are dealing with a medium that is certainly adequate for mass education and whose benefits go beyond the strictly academic (Gonzalo et al., 2023) or merely informational, thus addressing moral dilemmas and not only knowledge (Bautista-Alcaine et al., 2025). The genesis of some of this knowledge stems from the deficient implementation of AI, such as the prevalence of fake news (Cebollero-Salinas et al., 2024). In any case, as we observe in the methodology of these same investigations, it is suitable for quality education, and therefore higher education, which may well be distance education.

Collective intelligence in the context of digital distance higher education

As García Aretio (2020) points out, distance education has undergone a semantic and functional transformation, evolving into more complex models (face-to-face, distance, or mixed/combined), which can be integrated into the concept of Digital Distance Education (DDE) in the field of higher education. In this context, far from being limited to an individualistic learning model, DDE increasingly directs its efforts towards strengthening personalized learning within the educational process. After a systematic review of the literature, Bayly-Castaneda et al. (2024) note that AI is primarily focused on the field of higher education, where the use of adaptive learning technologies for developing personalized learning paths is predominant. This paradigm aims to integrate the personalization of learning with the construction of a community of knowledge, in which all its members share responsibility for the generation and consolidation of knowledge. This view is reinforced by the idea of using visualization tools to facilitate the interpretation of the contributions in collective intelligence environments (Ullmann et al., 2019).

We highlight the contribution of Fidalgo-Blanco et al. (2017), who developed a theoretical-practical model of collective intelligence in the presential university classroom, based on the management of knowledge generated by students through a private social network. Although their experience takes place in a face-to-face environment, the underlying principles and cooperative approach based on the collective's individual intelligence are extended to the use of platforms that automate and scale these interactions through AI. According to the analysis by Zawacki-Richter et al. (2019), it is noted that the use of this technology in most higher education applications developed to date has relegated teachers to a peripheral role. According to these authors, AI-based solutions focus on automatic content design, outcome prediction, or performance monitoring, but they neglect the pedagogical role of the educator as a mediator of learning, designer of meaningful contexts, and promoter of critical and collective thinking.

Digital education should integrate this approach with the collective construction of knowledge. In this sense, tools like Kampal represent a hybrid proposal of notable pedagogical value because they allow combining the individual traceability of the learning process with active participation in experiences of collective intelligence. The algorithms used allow student contributions to be organized according to their relative popularity, and they also make the processes of editing, copying, and transforming ideas visible, thus creating an environment where the individual and the collective feed each other. In pedagogical terms, collective intelligence prevents personalized learning from being limited to the adaptation of content or rhythms. Instead, it allows personalized learning to expand towards the personalization of cognitive contributions, which can then be put at the service of the group. In turn, collaborative

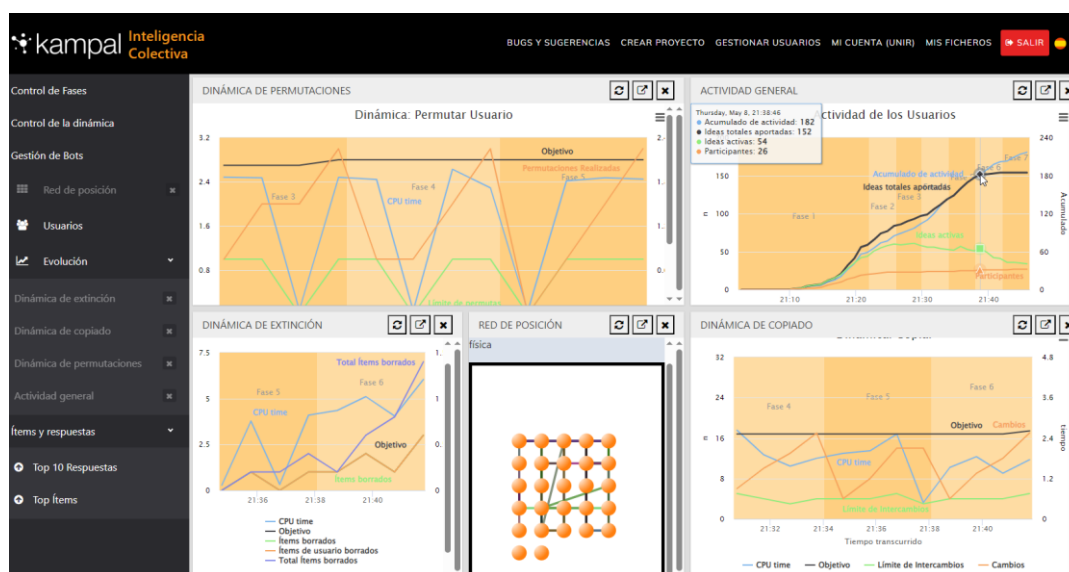
learning is enriched with a broader and more diverse base, where each student participates not only as a receiver but as a producer and evaluator of shared knowledge. This dynamic is strengthened by the traceability offered by the tool, allowing the teacher to monitor participation, detect patterns of interaction, and make more informed training decisions.

In this way, collective intelligence is configured as a meeting point between the singularity of individual thought and the transformative power of group dialogue, which goes through a series of phases that allow editing, eliminating, or copying ideas until a collective response is achieved. Its implementation, mediated by AI, not only optimizes teaching-learning processes but also promotes a more inclusive, reflective pedagogical culture committed to the construction of situated and shared knowledge. This technological solution, specifically designed to promote collective intelligence processes, enables the students' individual contributions—which reflect their own learning styles, previous knowledge, and levels of development—to be shared in a digital environment that fosters the analysis, synthesis, and collaborative evaluation of these ideas. Using analysis algorithms based on AI (mainly BD and ML) and data visualization, the platform generates indicators such as the three or ten most popular responses, facilitating the identification of emerging consensus and the most significant contributions according to the majority's preference.

Kampal is, therefore, an online tool that, through a process ordered by AI, helps the teacher to monitor individual behavior in the development of ideas in the students' collective learning groups. According to the Kampal Data Solutions company, "the platform generates an internal dynamic using Artificial Intelligence, Complex Networks, etc., to energize the interaction between the participants and guarantee the survival of the most popular responses among all those emitted" (p. 5). Figure 1 presents an example of the different dynamics during the development of the activity, where collective intelligence is empirically shown through the convergence of individual responses towards a shared synthesis.

Figure 1

Visualization panel of dynamic and structural metrics of the Kampal platform in real time during a collective intelligence activity



Source: Prepared by the authors.

One of this tool's underlying benefits is that "all individuals start from equity, and each idea should be valued for its content and not for who proposes it" (Bautista-Alcaine et al., 2025, p. 2). For their part, Orejudo et al. (2022) conducted an investigation involving 900 students, demonstrating the potential of collective intelligence to enhance group collaboration and how a high-quality response is guided through the phases.

Taken together, this evidence indicates that collective intelligence, mediated by technological tools such as Kampal, represents a qualitative evolution in the teaching-learning models of digital distance higher education. Not only does it respond to the increasing demands for personalization in virtual environments, but it also enables the incorporation of deep collaboration dynamics in a structured and measurable way, conceiving of the student as an active co-creator of knowledge. Collective intelligence emerges as a legitimate approach to rebalancing the social dimension of online learning, valuing both individual contributions and the emerging synergies of groups. In this sense, its implementation constitutes not only a methodological innovation but also a response to the pedagogical, technological, and ethical challenges posed by digital higher education in the twenty-first century.

This study analyzes the impact of Kampal use on the perception of collaborative learning in digital higher education, assessing variations in familiarity, perceived efficacy, and expectations about AI-mediated technologies.

METHODOLOGY

In accordance with the nature of the problem, a quantitative quasi-experimental design (Hernández Sampieri et al., 2010) was considered adequate, as it offers greater possibilities of generalization to real everyday contexts than experimental designs. At the same time, it is the most viable and ethical alternative, providing relevant information on the impact of a treatment or the evolution of a change process. For this purpose, a pre-post-test measurement was performed in a single group, using the Kampal collective intelligence tool as the intervention. This methodological structure enables the observation of effects in a non-randomized sample, allowing for comparisons of results obtained before and after treatment.

The students of the "Technologies for educational innovation" subject in the Official Master's Degree in Educational Innovation at the International University of La Rioja (UNIR) participated in solving a collaborative activity within an academic forum. The collective intelligence tool Kampal was used to analyze and synthesize the students' responses to a practical case concerning the ethical use of AI in the classroom. The first intervention consisted of responding to three key questions, which were addressed synchronously and collectively through Kampal. The collaborative forum activity took place over three weeks, allowing for scaled and reflective participation. Kampal was used for one hour, during which the students underwent seven phases generated by automatic dynamics to promote collaboration. They could analyze their "neighbor's" information, modify their own responses, and/or copy the responses of one of their classmates. Kampal's objective is to generate a collective response based on the students' contributions in the first intervention, which served as a basis for subsequent reflections and improvements. Specifically, the third intervention had a direct relationship with the contents that emerged from the collective intelligence in the virtual face-to-face session. Given the formative nature of the activity, integrated into the curricular development of the subject and linked to a formative assessment, it

was not possible to establish a control group. This constitutes a methodological limitation to be considered.

In this study, three essential conceptual dimensions were addressed: familiarity, efficacy, and expectations of use. Familiarity refers to intuitive recognition and perceived ease of use, which are determining factors in the initial acceptance of AI platforms in collaborative university environments (Chan & Hu, 2023). Effectiveness refers to the quality and pedagogical relevance of collaborative contributions facilitated by AI (Wang et al., 2025). Expectations of use refer to the future willingness to employ AI technologies in academic activities, aligning with technology adoption models that link positive experiences with a greater intent to use (Khlaif et al., 2024).

Based on the quasi-experimental pretest-posttest approach with a single group, three research hypotheses were formulated that guided the analysis of the effects of the use of Kampal. We expected:

- A significant increase in students' familiarity with collective intelligence technologies and with the Kampal platform itself compared to their level before the intervention.
- A significant improvement in the perception of the effectiveness of collaborative work and the ability to relate and improve individual contributions after using Kampal, compared to previous perceptions.
- An increase in students' future willingness to employ collective intelligence-based tools in academic settings after participation in Kampal-mediated collaborative activities.

These hypotheses were tested through non-parametric statistical analyses, which are appropriate for the type of data and the structure of the sample, to assess the perceived impact of the tool in a digital distance higher education environment.

Context

The research was conducted at the International University of La Rioja (UNIR), which provides 100% online higher education on an official basis. Law 3/2008, of October 13, legally recognized the International University of La Rioja. The Parliament of the autonomous community thereby approved its legal personality in the form of a public limited company, characterized by distance learning, based on ICT, without a physical campus, with synchronous and asynchronous classes (UNIR, 2024). It offers university education and provides public service in higher education through research, teaching, and scholarly study.

Population

The sample for this study ($n = 399$) comprises students of the Master's Degree in Educational Innovation at UNIR enrolled in the subject "Technology for Educational Innovation". The invitation guaranteed the confidentiality of the data, leaving a total of $n = 152$ subjects at pre-test and $n = 77$ subjects at post-test. After filtering the data, the final sample was $n = 147$ at pre-test and $n = 62$ at post-test. When analyzing the university students' perception of the use of the Kampal collective intelligence application between pre-test and post-test, a sufficiently matched sample was identified concerning the proposed objectives ($n = 25$). The loss of continuity in the

anonymous identifiers assigned to each student, together with the voluntary nature and participation in the live session of the tool, may have influenced this final sample.

In all measurements, most of the participants identified as women, and this trend was maintained both at pre-test (65.31%) and post-test (66.13%), and in the matched cases (60%). Male representation was lower at all times, and non-binary or undeclared categories were very scarce. Regarding the distribution by country (Table 2), most of the sample came from Colombia, both at pre-test (76.87%) and post-test (70.97%), and in the matched cases (64%). Although other countries, such as Spain, Ecuador, and Argentina, had a lower presence, a slight relative increase in their representation was observed at post-test and in the matched cases. The participation of other countries was marginal, and some, like Chile or the United States, only appeared in one of the phases or were absent in the matched data.

Table 2

Distribution of the sample according to country

Country	Pre-test		Post-test		Matched Pre-Post	
	n	%	n	%	n	%
Colombia	113	76.87	44	70.97	16	64.00
Spain	9	6.12	7	11.29	4	16.00
Ecuador	5	3.40	5	8.06	2	8.00
Peru	5	3.40	3	4.84	0	0.00
Chile	5	3.40	0	0.00	0	0.00
Argentina	3	2.04	2	3.23	2	8.00
Rest of Latin America	2	1.36	1	1.61	1	4.00
Other	4	2.72	0	0.00	0	0.00
United States	1	0.68	0	0.00	0	0.00
TOTAL	147	100	62	100	25	100

Source: Prepared by the authors.

Regarding teaching experience (Table 3), the profile with more than 10 years of experience predominated, both at pre-test (42.18%) and post-test (37.10%), and in the matched cases (40%). However, a higher relative representation was observed of teachers with less than 3 years of experience at post-test (27.42%) and in the matched cases (24%).

Table 3

Distribution of the sample according to level of teaching experience

Category	Pre-test		Post-test		Matched Pre-Post	
	n	%	n	%	n	%
More than 10 years	62	42.18	23	37.10 %	10	40.00 %
6 to 10 years	34	23.13	10	16.13 %	6	24.00 %
Less than 3 years	29	19.73	17	27.42 %	6	24.00 %
3 to 5 years	16	10.88	9	14.52 %	2	8.00 %
No experience	6	4.08	3	4.84 %	1	4.00 %
TOTAL	147	100.00	62	100.00 %	25	100.00 %

Source: Prepared by the authors.

Instrument

The assessment instrument "Questionnaire to analyze university students' perception of the use of the Kampal Collective Intelligence application" (CPIC) was administered. The questionnaire had been previously validated by a committee of three experts with diverse academic backgrounds, all of whom were specialists in pedagogy, before the intervention began. The data were migrated to the *Microsoft Forms* application to facilitate their distribution and collection. The results of the Aiken V-index were higher than .7 in all the items except for three, and in no case less than .6, thus obtaining adequate content validity (Penfield & Giacobbi, 2004).

The questionnaire's first item is an informed consent form, in compliance with the regulations for the approval of research by the Research Ethics Committee (Ref. 037/2025) of the International University of La Rioja (UNIR), following all the protocols in terms of confidentiality and anonymity (UNIR, 2025). The instrument consists of four blocks:

- Block I: Biographical information. Comprising Items 2 to 5, it includes questions to obtain personal identity information: sex (male, female, none of the above, I prefer not to answer), country (Spain, Colombia, Ecuador, Peru, Mexico, Argentina, Chile, United States, the rest of Latin America, others), and level of teaching experience in years (no experience, less than 3 years, 3 to 5 years, 6 to 10 years, more than 10 years).
- Block II: Familiarity. It includes Items 6 to 8, which aim to assess the students' familiarity with collective intelligence applications, specifically Kampal.
- Block III: Efficiency. It integrates Items 9 to 12, aimed at assessing the perception of the effectiveness of collective intelligence among students.
- Block IV: Use expectations. It presents Items 12 to 14, which obtain usability expectations in terms of learning, collaboration, and availability of collective intelligence applications.

Table 4 shows the questions in each block by item:

Table 4
CPIC blocks, items, and questions

Items and questions		
Block I: Biographical Information	1.	Personal ID.
	2.	Sex.
	3.	Country.
	4.	Indicate level of experience in years.
Block II: Familiarity	5.	Do you think that the level of use of the application of Artificial Intelligence has increased after the experience?
	6.	Do you think that the level of use of collective intelligence applications has increased after the experience?
	7.	Do you think the level of use with the Kampal app has increased after the experience?

Items and questions	
Block III: Efficacy	8. To what extent do you think the use of a Collective Intelligence application could improve the quality of collaborative work?
	9. Has the use of the Collective Intelligence application helped you to change, improve, or discard your initial answers to the questions posed in the activity?
	10. Do you think that the responses arising from Collective Intelligence have helped you relate contributions to the forum?
Block IV: Use expectations	11. I think this Collective Intelligence application will help me in my learning process.
	12. I think this Collective Intelligence application will help me in terms of collaboration.
	13. Would you be willing to use collective intelligence applications in future academic activities?

Source: Prepared by the authors.

After the respondents completed the questionnaire, the data were imported into the statistical package for the social sciences, IBM SPSS Statistics 25, to obtain the appropriate psychometric properties and to perform the relevant statistical analyses.

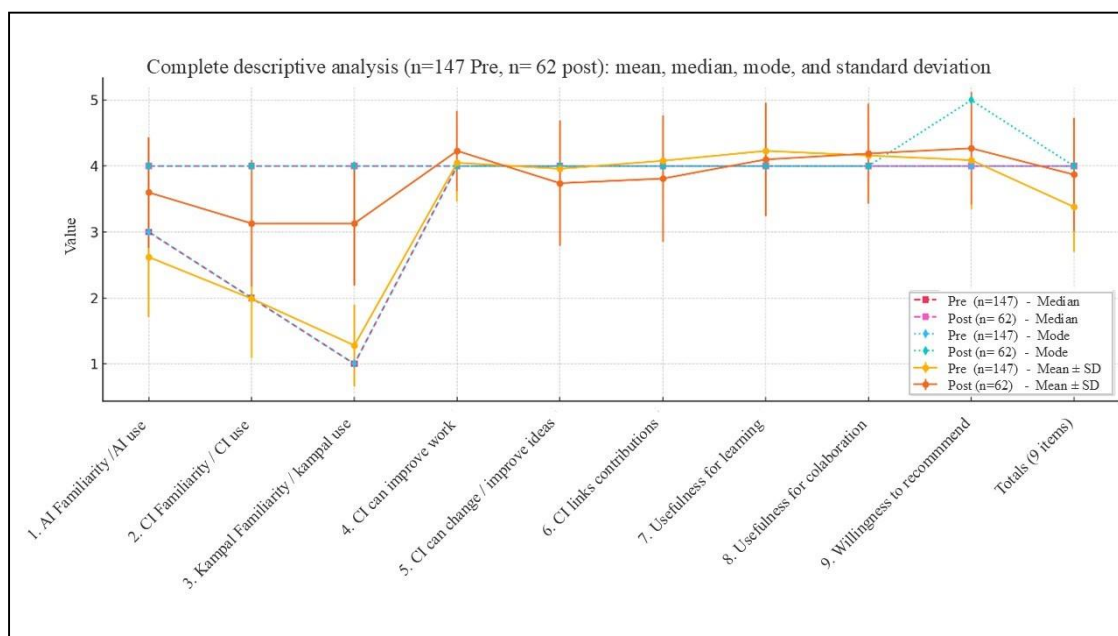
ANALYSIS AND RESULTS

Within the framework of incorporating collaborative technologies into the educational field, a descriptive statistical analysis was conducted, along with an examination of the psychometric properties and perceived usefulness of a collective intelligence tool. Using a quantitative approach, internal consistency tests, exploratory factor analysis, and non-parametric techniques were applied to assess the instrument's adequacy in capturing significant perceptions. The results obtained provide relevant evidence regarding the validity, reliability, and perceived usefulness of a collective intelligence tool applied in the educational setting.

Descriptive statistical analysis

From pre-test to post-test, there was clear progress in the adoption of the technologies (Figure 2): mean familiarity with artificial intelligence registered an increase of 2.62 to 3.74, for collective intelligence of 1.99 to 3.60, and for Kampal of 1.28 to 3.13. The belief that collective intelligence improves collaborative work also increased from 4.05 to 4.23, while expectations about its impact on ideas and the relation of contributions decreased slightly. The ratings of usefulness for learning and collaboration remained around 4.10 and 4.20, and the willingness to recommend the tool increased from 4.09 to 4.27, changing the mode from 4 to 5. Overall, the global mean of the nine items increased from 3.38 to 3.87, with standard deviations always less than 1.

Figure 2
Results of descriptive statistical analysis



Source: Prepared by the authors.

Overall, there was a global improvement in familiarity with Artificial Intelligence, Collective Intelligence, and the Kampal tool. Ratings of its usefulness and willingness to recommend it also increased slightly, while perceptions of its contribution to learning and collaboration remained stable. The responses showed consistency, with low variability in scores.

Inferential statistical analysis

Following Hair et al. (2019), the instrument demonstrated high internal consistency ($\alpha = .856$), which supports its reliability and suitability in capturing participants' perceptions of the analyzed tool.

For the exploratory factor analysis, orthogonal Varimax rotation was used with Items 5 to 13. Sampling adequacy was moderate ($KMO = .764$), but acceptable for the procedure (Field, 2009), supported by Bartlett's significant sphericity test ($p < .000$). The analysis identified two factors, which explained 59.82% of the construct variability. The first factor explained 39.72% and the second explained 20.10%. In contrast, for the subsequent interpretation of the factors, we employed the initial rotated component matrix (Table 5). This matrix determined different factorial loadings for the selection of the items included in each of the two factors and provided the magnitude of the correlation between the item and the factors, ordered by size and eliminating low coefficients with an absolute value below .400, based on the suggestion of Stevens (2002).

Table 5
Rotated component matrix

	Component
Familiarity AI	.738
Familiarity Collective Intelligence	.862
Familiarity Kampal Intelligence	.663
To what extent do you think that the use of a Collective Intelligence application can improve the quality of collaborative work?	.828
To what extent do you think Collective Intelligence can help change, improve, or discard your initial idea of certain content?	.754
To what extent do you think that the responses that emerge from a collective intelligence will help you relate contributions and improve them?	.762
I think this Collective Intelligence application will help me in my learning process.	.795
I think this Collective Intelligence application will help me in terms of collaboration.	.815
Would you be willing to use collective intelligence applications in future academic activities?	.621

Source: Prepared by the authors.

The name of the factors found was determined by the elements that constituted it and the content to which each of them referred, finally called: Factor 1, Familiarity, and Factor 2, Efficacy/Expectations. Although the factorial results were optimal, it is worth noting that there may be some conceptual overlap between the two factors, as an increase in familiarity could influence perceptions of efficacy and use expectations.

The evaluation of the normality of the data, by means of the Kolmogórov-Smirnov and Shapiro-Wilk indices, showed significant deviations from the normal distribution in all variables ($p < .05$). This justified the use of non-parametric tests, which are more appropriate and robust for asymmetric distributions or when there are outliers.

Comparisons by sex were then made using the Mann-Whitney *U*-test (Table 6). The results yielded no statistically significant differences between men and women in any of the dimensions evaluated.

Table 6
Test results of the Mann-Whitney U-test by sex

	pre_ familiarity_ Factor1	post_ familiarity_ Factor1	pre_ effectiveness/ expectations Factor2	post_ effectiveness/ expectations Factor2
Mann-Whitney <i>U</i> -test	53.000	67.000	51.000	52.500
Wilcoxon <i>W</i>	173.000	187.000	171.000	172.500
Z	-1.253	-.459	-1.354	-1.262
Sig.(2-tailed)	.210	.646	.176	.207
Exact. Sig. (2-tailed)	.238	.683	.196	.216

Source: Prepared by the authors.

Subsequently, the differences according to teaching experience were explored using the Kruskal-Wallis test. The results (Table 7) yielded no statistically significant differences, indicating a positive transversal assessment of the tool, regardless of the level of professional experience.

Table 7

Test results of the Kruskal-Wallis test according to teaching experience

	pre_ familiarity_ Factor1	post_ familiarity_ Factor1	pre_ effectiveness/ expectations Factor2	post_ effectiveness/ expectations Factor2
Kruskal-Wallis <i>H</i>	1.281	3.655	1.799	.524
df	4	4	4	4
Sig.(2-tailed)	.865	.455	.773	.971

Source: Prepared by the authors.

Finally, a pre-post analysis was carried out using the Wilcoxon signed-rank test. The results revealed statistically significant improvements in familiarity with the tool ($z = -4.32$, $p < .001$, $r = .68$) and future willingness to use collective intelligence-based technologies ($z = -4.26$, $p < .001$, $r = .67$), both with large effect sizes. In contrast, no significant changes were observed in expectations about outcomes ($z = -0.29$, $p = .770$, $r = .05$) or the perception of self-efficacy ($z = -0.75$, $p = .453$, $r = .12$).

DISCUSSION AND CONCLUSIONS

Currently, AI in education has been predominantly limited to applications based on LLMs, such as ChatGPT, which operate according to the levels delimited by OpenAI, at a conversational level 1. This narrow focus has raised concerns about its impact on assessment and learning processes in higher education (Consuegra-Fernández et al., 2024). In this study, we have addressed a second higher level of processing thanks to the collective intelligence approach proposed by Kampal, which transcends simple conversational interaction. In contrast, the work is performed at Level 3 (agents) because IA makes structural decisions on behalf of the user (formation of groups, selection of ideas, etc.), although without full autonomy. However, the most significant convergence is located at Levels 4 (innovative) and 5 (organizations), because the Kampal platform enables emerging collaborative solutions, complex and self-organized dynamics, and structures that learn and evolve.

Although the study started with a defined population (399 students), the progressive reduction to a matching sample of 25 subjects in the pretest-posttest measurement introduces some limitations concerning external validity. However, as pointed out by Creswell and Creswell (2017), this type of exploratory study with non-probabilistic samples provides relevant information in real contexts, especially when seeking to evaluate the acceptance and perception of new educational technologies. The final sample, which had a predominance of women, high geographical representation from Colombia and high teaching experience offer a relevant perspective within the framework of digital higher education, although its results are not directly generalizable to other populations. Likewise, the pre-test-posttest design without a control group, typical of quasi-experimental approaches, limits the possibility of establishing definitive causal inferences, as it cannot completely isolate

the effects of external variables. However, as Shadish et al. (2002) point out, this type of design is still appropriate in the initial stages of applied research, where the interest lies more in exploring patterns of change than in testing causal hypotheses with maximum controllability. From an analytical point of view, the choice of non-parametric tests—Kolmogórov-Smirnov and Shapiro-Wilk for normality—the Mann-Whitney *U*-test, and Kruskal-Wallis and Wilcoxon tests were justified by the small sample size, the asymmetry of the distributions, and the ordinal nature of the Likert-type data (Sullivan & Artino, 2013).

The results obtained in this study show a positive impact on students' familiarity with the Kampal tool, as well as a high future willingness to use collective intelligence technologies in educational settings, confirming the first two hypotheses, which suggest that contact with well-designed digital environments can transform students' attitudes towards more collaborative forms of learning (Hogan et al., 2023). This positive disposition is especially relevant in digital distance higher education environments, where opportunities for peer-to-peer interaction may be limited without an intentional design of collective dynamics. However, concerning the third hypothesis, there were no significant changes in the expectations of results or the perception of immediate effectiveness, suggesting the need for a review of its pedagogical potential and the assessment of the responses that emerged collectively as the most popular, but not necessarily the most suitable, complex, or desired. This lack of changes may be due to various circumstances, among which we could highlight the short time of exposure to the tool or the identification of the latent potential benefits yet to be discovered. From a pedagogical viewpoint, these reflections encourage us to establish greater monitoring and accompaniment not only during its use, but also at the end of the process, incorporating strategies such as longitudinal evaluations, controlled designs, objective performance indicators, and the replication of the study in different institutions and training contexts.

These findings suggest that the use of the Kampal tool can facilitate processes of approach, exploration, and openness towards its implementation, although it does not necessarily immediately transform beliefs about its academic benefits, reveal its ability to promote active participation, the collective construction of knowledge, and a positive disposition towards the use of collaborative technologies mediated by AI in similar educational settings. Likewise, the tool's use presents some limitations because it requires the students' synchronous participation, which can restrict its integration into more flexible dynamics typical of distance education. However, collectively generated outputs could serve as a basis for asynchronous reflection in other interaction spaces, such as collaborative forums. This suggests a dual reading: on the one hand, direct intervention allows each student to perceive their contribution as a constituent part of the collective response; on the other hand, it raises the need to question whether the solutions emerging from the collective intelligence process effectively represent the most pertinent responses or simply the most popular ones.

In this sense, from a critical point of view, it is necessary to contrast these results with those that suggest that the use of these and other AI tools entails a reduction in intellectual diversity (Riedl & Bogert, 2024), causing a negative effect of excessive homogenization by attempting to speed up decision-making processes where the most frequent response may not be what we are really seeking from the pedagogical perspective. Kosmyna et al. (2025) coined the term "cognitive debt" when observing a significant decrease in neural connectivity between individuals who used LLM and

those who performed tasks without AI assistance, thus corroborating the danger of incorrect automated use.

In the case of Kampal, the responses that emerge as "collective" are those that have achieved greater acceptance or popularity within the group, but this does not guarantee that they are the most complex, pertinent, or desirable from a formative perspective. Concerning Mulgan's (2018) statement that, despite the facilitating role of artificial intelligence, "human collectives and human intelligence must continue to be at the center" (p. 235), the study takes into account that people activate the pedagogical role of these tools. This underscores the need for supervision—not only human but also expert—when working with any platform based on AI algorithms. It is recommended that teachers act as critical mediators, guiding reflection and participation in the dynamics of collective intelligence. Additionally, in this sense, although Kampal indeed has sufficient tools that allow the traceability of the proposed interactions, we should not lose sight of the fact that AI models, especially those based on neural networks, tend to be included in the so-called "black box" models. Therefore, it is essential not only to supervise the result achieved but also—and perhaps above all, as Kampal is a pedagogical tool—the path followed to reach it. One of the central debates that emerges from this study is that of the opacity of AI systems, commonly referred to as the "black box" problem.

As a future line of research, we propose the constitution of assessment committees to analyze the algorithmically selected responses, contrasting them with explicit pedagogical, ethical, and epistemological criteria. An expert validation of the emerging products of collective intelligence would allow determining whether what is most accepted coincides with what is most aligned with the defined training purposes, thus opening a necessary debate on what is meant by the "best response" in a collaborative learning context. This approach would contribute to balancing technical efficiency with educational demands, granting greater legitimacy to the automated collective process.

Placing the focus on the expert human implies training the competencies of the facilitators of collective intelligence, as proposed by Broome and Hogan (2020), through a three-level progressive trajectory: foundational, performative, and master competencies. This approach requires not only technical skills, but also pedagogical competencies that allow collaborative processes to be guided using tools like Kampal. The facilitator should be able to guide the group in generating new perspectives, managing interventions mediated by technologies such as "bots" with AI, and assessing group dynamics by controlling the dynamics or network positions. In particular, "bots" with AI act as "mediating agents," which would enhance the link between personalized and collective learning. However, we should be aware that this type of element would also increase "black box" risks.

Thus, under the planned precautions and based on the results achieved along with the anticipation of the improvement of the instruments employed—as well as of those in the cited literature—, we observe how using Kampal within digital higher education can facilitate the convergence between the individual and the collective, offering an environment where each student contributes content from their interests and cognitive styles, while incorporating them into a collective dynamic aimed at providing shared knowledge. One line of work is to explore future developments where AI not only analyzes responses, but also learns to facilitate dynamics similar to Artificial Swarm Intelligence (ASI), "because they allow a very large group of individuals to meet for a very short time and make effective and relatively accurate

decisions” (Baltzersen, 2022, p. 104). Within the educational context, this implies the responsible and efficient use of AI, with the capacity to orchestrate, in real-time, collective decision-making processes among large groups of students, promoting the rapid convergence of ideas, consensus, and collaborative resolution of complex problems. Its responsible application in educational contexts, such as digital distance higher education, could enable the generation of spaces for dynamic deliberation, resolution of complex problems, and shared decision-making that meet both inclusion and quality training criteria. Structured learning experiences should always be designed with intentional teacher mediation, using authentic tasks and traceability mechanisms that ensure the transparency and pedagogical value of the collective process.

Declaration of conflicts of interest and transparency

The authors state that there are no economic or institutional links with the company that owns Kampal.

REFERENCES

- Baltzersen, R. K. (2022). *Cultural-historical perspectives on collective intelligence: Patterns in problem solving and innovation*. Cambridge University Press.
<https://doi.org/10.1017/9781108981361>
- Bayly-Castaneda, K., Ramirez-Montoya, M.-S., & Morita-Alexander, A. (2024). Crafting personalized learning paths with AI for lifelong learning: A systematic literature review. *Frontiers in Education*, 9, 1-12.
<https://doi.org/10.3389/feduc.2024.1424386>
- Bautista-Alcaine, P., Íñiguez-Berrozpe, T., & Fernández-Albás, J. (2025). Collective intelligence online experiments: A learning method and its pedagogical applications through online collaborative work platforms. In V. Goglio & C. Biancalana (Eds.), *Mapping the evolution of platform society* (pp. 13-26). Emerald Publishing Limited.
<https://doi.org/10.1108/978-1-83608-028-220251002>
- Bozkurt, A. (2023). Generative artificial intelligence (AI) powered conversational educational agents: The inevitable paradigm shift. *Asian Journal of Distance Education*, 18(1), 198-204.
<https://doi.org/10.5281/zenodo.7716416>
- Broome, B., & Hogan, M. (2020). Core competencies of collective intelligence facilitators. *Systems Research and Behavioral Science*, 38, 940-944.
<https://doi.org/10.1002/sres.2734>
- Cebollero-Salinas, A., Elboj-Saso, C., Íñiguez-Berrozpe, T., & Bautista-Alcaine, P. (2024). Confronting fake news through collective intelligence in adolescents. *International Journal of Human-Computer Interaction*. Advance online publication.
<https://doi.org/10.1080/10447318.2024.2405567>
- Chan, C. K. Y., & Hu, W. (2023). Students' voices on generative AI: Perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education*, 20, 43.
<https://doi.org/10.1186/s41239-023-00411-8>
- Consuegra-Fernández, M., Sanz-Aznar, J., Burguera-Serra, J. G., & Caballero-Molina, J. J. (2024). ChatGPT: El dilema sobre la autoría de las actividades evaluables en educación universitaria. *Revista de Investigación Educativa*, 42(2).
<https://doi.org/10.6018/rie.565391>
- Creswell, J. W., & Creswell, J. D. (2017). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). SAGE Publications.

- Duenas, T., & Ruiz, D. (2024). *The path to superintelligence: A critical analysis of OpenAI's five levels of AI progression* [Preprint]. ResearchGate. <https://doi.org/10.13140/RG.2.2.33794.70085>
- European Union. (2024). Reglamento del Parlamento Europeo y del Consejo de 13 de marzo de 2024 por el que se establecen normas armonizadas en materia de inteligencia artificial y se modifican determinados actos legislativos de la Unión. *Diario Oficial de la Unión Europea*, L 1689, 12 de julio de 2024. <https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX:32024R1689>
- Fidalgo-Blanco, Á., Sein-Echaluce, M. L., & García-Peñalvo, F. J. (2017, 4–6 October). Inteligencia colectiva en el aula. Un paradigma cooperativo [Conference presentation]. *IV Congreso Internacional sobre Aprendizaje, Innovación y Competitividad*, Zaragoza, España. <https://doi.org/10.26754/CINAIC.2017.000001.125>
- Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). SAGE Publications.
- García Aretio, L. (2020). Bosque semántico: ¿educación/enseñanza/aprendizaje a distancia, virtual, en línea, digital, eLearning...? *RIED-Revista Iberoamericana de Educación a Distancia*, 23(1), 9-28. <https://doi.org/10.5944/ried.23.1.25495>
- Gonzalo, A., Sanz-García, F., Pelacho, M., Tarancón, A., Rivero, A., Varela, O., & Moreno, A. (2023). Collective intelligence to find solutions to the challenges posed by the Sustainable Development Goals. *Citizen Science: Theory and Practice*, 8(1), 47. <https://doi.org/10.5334/cstp.587>
- Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate data analysis* (8th ed.). Cengage.
- Hernández Sampieri, R., Fernández Collado, C., & Baptista Lucio, M. P. (2010). *Metodología de la investigación* (5.ª ed.). McGraw-Hill.
- Hogan, M. J., Barton, A., Twiner, A., James, C., Ahmed, F., Casebourne, I., Hamilton, P., Shi, S., Zhao, Y., Harney, O., & Wegerif, R. (2023). Education for collective intelligence. *Irish Educational Studies*, 44(1), 137-166. <https://doi.org/10.1080/03323315.2023.2250309>
- Khlaif, Z. N., Ayyoub, A., Hamamra, B., Bensalem, E., Mitwally, M. A. A., Ayyoub, A., Hattab, M. K., & Shadid, F. (2024). University teachers' views on the adoption and integration of generative AI tools for student assessment in higher education. *Education Sciences*, 14(10), 1090. <https://doi.org/10.3390/educsci14101090>
- Kosmyna, N., Hauptmann, E., Yuan, Y. T., Situ, J., Liao, X.-H., Beresnitzky, A. V., Braunstein, I., & Maes, P. (2025, 10 June). *Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing task* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2506.08872>
- Ley Orgánica 3/2008, de 13 de octubre, de reconocimiento de la Universidad Internacional de La Rioja. *Boletín Oficial del Estado*, núm. 261, 42805-42806. <https://www.boe.es/eli/es-ri/l/2008/10/13/3>
- Marrón Luna, Y. A. (2021). Trabajo colaborativo: una opción en la educación en línea. *Revista SUAyED Psicología*, (2), 26–31. https://www.psicologia.unam.mx/documentos/pdf/publicaciones/revista_suayed_psicologia_2021_2.pdf
- Msambwa, M. M., Wen, Z., & Daniel, K. (2025). The impact of AI on the personal and collaborative learning environments in higher education. *European Journal of Education*, 60(1). <https://doi.org/10.1111/ejed.12909>
- Mulgan, G. (2018). *Big mind: How collective intelligence can change our world*. Princeton University Press. <https://doi.org/10.1515/9781400888511>
- Orejudo, S., Cano-Escoriaza, J., Cebollero-Salinas, A. B., Bautista, P., Clemente-Gallardo, J., Rivero, A., Rivero, P., & Tarancón, A. (2022). Evolutionary emergence of collective intelligence in large groups of students. *Frontiers in Psychology*, 13, 1-16. <https://doi.org/10.3389/fpsyg.2022.848048>

- Penfield, R. D., & Giacobbi, P. R. (2004). Applying a score confidence interval to Aiken's item content-relevance index. *Measurement in Physical Education and Exercise Science*, 8(4), 213-225. https://doi.org/10.1207/s15327841mpee0804_3
- Riedl, C., & Bogert, E. (2024). Effects of AI feedback on learning, the skill gap, and intellectual diversity [Preprint]. *arXiv*. <https://doi.org/10.48550/arXiv.2409.18660>
- Riedl, C., & De Cremer, D. (2025). AI for collective intelligence. *Collective Intelligence*, 4(2). <https://doi.org/10.1177/26339137251328909>
- Russell, S. J., & Norvig, P. (2020). *Artificial intelligence: A modern approach* (4th ed.). Pearson.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.
- Stevens, J. P. (2002). *Applied multivariate statistics for the social sciences*. Routledge. <https://doi.org/10.4324/9781410604491>
- Sullivan, G. M., & Artino, A. R. (2013). Analyzing and interpreting data from Likert-type scales. *Journal of Graduate Medical Education*, 5(4), 541-542. <https://doi.org/10.4300/JGME-5-4-18>
- Ullmann, T. D., De Liddo, A., & Bachler, M. (2019). Un panel de visualización para la inteligencia colectiva controvertida: analíticas de aprendizaje para la mejora de creación de ideas en grupos de discusión. *RIED-Revista Iberoamericana de Educación a Distancia*, 22(1), 41-80. <https://doi.org/10.5944/ried.22.1.22294>
- UNED. (2025, 20 February). Lorenzo García Aretio: un gran legado en la educación a distancia. <https://www.uned.es/universidad/inicio/noticias/mostrarnoticia.html?noticia=lorenzo-garcia-aretio-un-gran-legado-en-la-educacion-a-distancia>
- Universidad Internacional de La Rioja. (2024). *Manual de calidad*. UNIR. https://static.unir.net/calidad/Manual_calidad.pdf
- Universidad Internacional de La Rioja. (2025). *Centro de Estudios sobre Innovación (CEI)*. UNIR. <https://www.unir.net/universidad-online/vicerrectorados/vicerrectorado-investigacion/cei/>
- Wang, H., Wang, C., Zhan, C., Liu, F., Bao, C., & Xianlong, X. (2025). Impact of AI-agent-supported collaborative learning on the learning outcomes of university programming courses. *Education and Information Technologies*, 30(12), 17717-17749. <https://doi.org/10.1007/s10639-025-13487-8>
- Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education: Where are the educators? *International Journal of Educational Technology in Higher Education*, 16(1), 1-27. <https://doi.org/10.1186/s41239-019-0171-0>

Date of reception: 1 June 2025

Date of acceptance: 24 August 2025

Date of approval for layout: 24 September 2025

Date of publication in OnlineFirst: 15 October 2025

Date of publication: 1 January 2026

Impact of generative AI on university students' digital competences: experimental evidence based on the DigComp framework

Impacto de la IA generativa en competencias digitales universitarias: evidencia experimental basada en el marco DigComp



 Concepción González García - *Universidad Católica de Murcia, UCAM (Spain)*
 Nina Pallarés - *Universidad Católica de Murcia, UCAM (Spain)*

ABSTRACT

This study analyzes the impact of the formative use of generative artificial intelligence (AI) on the development of digital competencies in university students. The intervention was implemented through a randomized controlled trial research design. The experimental group received training aimed at strategically using generative AI models to complete academic tasks, while the control group carried out the same activities without specific AI guidance. The impact was assessed using a difference-in-differences model with fixed effects, based on pre- and post-intervention questionnaires. Competences were analyzed according to the European DigComp 2.2 framework, covering four main competence areas: information and data literacy, communication and collaboration, safety, and problem solving. The results show statistically significant improvements in information and data literacy and in problem solving, both in their functional and metacognitive dimensions. Differential effects were also identified depending on the initial level of digital competence, with more pronounced gains among students with lower prior proficiency, who showed significant progress across all evaluated competencies. These findings suggest a compensatory effect of the didactic use of AI, capable of reducing gaps and promoting more equitable and inclusive learning processes. The study supports the guided integration of emerging technologies in higher education to strengthen digital competencies.

Keywords: artificial intelligence; digital competences; higher education; autonomous learning; DigComp.

RESUMEN

Este estudio analiza el impacto del uso formativo de inteligencia artificial (IA) generativa en el desarrollo de competencias digitales en estudiantes universitarios. La intervención se implementó mediante un ensayo controlado aleatorizado. El grupo experimental recibió formación orientada a utilizar estratégicamente modelos de IA generativa para la realización de tareas, mientras que el grupo de control completó las mismas actividades sin orientación específica sobre IA. El impacto se evaluó mediante un modelo de diferencias en diferencias con efectos fijos, basado en cuestionarios pre y postintervención. Las competencias se analizaron según el marco europeo DigComp 2.2, contemplándose cuatro áreas: alfabetización en información y datos, comunicación y colaboración, seguridad y resolución de problemas. Los resultados muestran mejoras estadísticamente significativas en alfabetización en información y datos y en resolución de problemas, tanto en su dimensión funcional como metacognitiva. Asimismo, se identificaron efectos diferenciales según el nivel inicial de competencia digital, siendo más pronunciados entre estudiantes con menor dominio previo, quienes presentan avances significativos en todas las competencias evaluadas. Estos hallazgos sugieren un efecto compensatorio del uso didáctico de la IA, capaz de reducir brechas y promover aprendizajes más equitativos. El estudio respalda la integración guiada de tecnologías emergentes en la educación superior para fortalecer las competencias digitales.

Palabras clave: inteligencia artificial; competencias digitales; educación superior; aprendizaje autónomo; DigComp.

INTRODUCTION

The emergence of generative artificial intelligence (AI) tools has brought about a significant transformation in teaching and learning processes in higher education (Hwang et al., 2020; Lee et al., 2024). These tools have reshaped the ways in which students access, analyze, and evaluate information, potentially influencing the improvement of their digital competences. In this context, and given the growing adoption of these technologies, it is essential to analyze their effects, particularly on digital competences, a key element for meeting the demands of a constantly evolving labor market (Van Laar et al., 2020).

These competences encompass the set of knowledge and skills required to effectively use digital technologies (Ferrari & Punie, 2013; Laupichler et al., 2022). Currently, they include not only information search and management, but also the ability to critically understand and evaluate digital sources (Polizzi, 2020). Previous studies have highlighted their importance in various domains such as education, employment, and social relations (Bastian et al., 2023; Buchholz et al., 2020; Morgan et al., 2022). Moreover, the European Commission defines them as one of the eight key competences for lifelong learning, emphasizing their role in the safe, critical, and responsible use of digital technologies for learning, work, and active participation in society (European Commission, 2019). Therefore, promoting their development has become a priority in educational and policy agendas.

Recent research has shown that the incorporation of technologies for educational purposes fosters the development of digital competences, as well as self-regulation and a positive perception of learning among students (Blau et al., 2020; Lin et al., 2025). Furthermore, when these technologies are integrated intentionally and with a clear orientation, they not only strengthen such competences but can also contribute to autonomous learning (Ting, 2015). Along the same lines, Prior et al. (2016) show that a higher level of digital competences, together with positive attitudes toward technology, is associated with higher levels of self-efficacy, which positively influences autonomous learning and participation.

The use of generative AI tools in higher education has been widely investigated, with studies highlighting both their benefits and associated risks (Dimitriadou & Lanitis, 2023; Zhai et al., 2021). García Peñalvo et al. (2024) emphasize that, although generative AI tools offer important benefits, such as personalized learning, the automation of repetitive tasks, the generation of educational content, and support for the development of critical thinking, they also present significant limitations. Likewise, challenges have been identified related to equity of access, the protection of personal data, and the need to train both educators and students in the ethical and critical use of these technologies. Popenici and Kerr (2017) point out that the incorporation of artificial intelligence-based technologies is profoundly transforming teaching and learning processes in universities, which poses new challenges for institutions and requires a reconsideration of traditional approaches.

Despite the growing interest, few studies directly analyze how the formative use of generative AI contributes to the development of digital competences. A systematic review conducted by Zhao et al. (2021), which analyzed 33 articles published between 2015 and 2021, reveals that only about 15% of the published works examine factors that may influence their acquisition. In contrast, recent research has focused on the inverse relationship: how a higher level of digital competence favors the adoption of these technologies. Specifically, Moravec et al. (2024) show that a higher level of digital

competences is associated with more frequent use of ChatGPT for testing, entertainment, and the acquisition of new knowledge. Furthermore, recent studies have found that greater digital literacy is associated with more favorable attitudes toward the use of AI-powered technologies in educational contexts (Saklaki & Gardikiotis, 2024), reinforcing the idea of a dynamic interaction between competences and technological disposition. Although some of the recent works provide valuable insights into the skills required in AI-mediated environments, there remains a shortage of research that systematically and structurally evaluates the acquisition of specific digital competences among students in higher education, using standardized frameworks such as DigComp.

This study makes three main contributions. First, it offers, to the best of our knowledge, the first empirical analysis of the impact of the didactic integration of generative AI on the development of specific digital competences within the DigComp framework, based on a quasi-experimental design. Second, it quantifies the magnitude of this impact, identifying significant progress in fundamental aspects such as digital self-regulation and interaction with technological tools. Third, by analyzing heterogeneous effects, it reveals a potential leveling effect: the greatest progress is observed among students with lower initial levels, suggesting that, when applied with didactic intentionality, generative AI can contribute to greater equity in the development of digital competences.

The results show that an intentional didactic implementation of generative AI can not only enhance learning but also reduce pre-existing inequalities. Its strategic use contributes to building more equitable and inclusive environments, strengthening key competences to operate critically and autonomously in increasingly complex digital contexts.

The remainder of the article is structured into five sections. First, the relevant literature on digital competences and the use of generative AI in educational contexts is reviewed. Next, the methodology employed is described, comprising the experimental design and the estimation strategy, based on a difference-in-differences model with fixed effects. Then, the data used are described and a descriptive analysis is provided. In the following section, the main results are presented and discussed, considering both average effects and heterogeneous effects. Finally, the article concludes with a reflection on the implications derived from the findings.

LITERATURE REVIEW

The development of digital competences has become a necessity to face the challenges and benefit from the opportunities offered by an increasingly digitalized environment. In the educational field, it is essential for students to acquire these competences, since the ability to effectively use technologies and access information constitutes a key factor for their academic success, as well as for their professional and personal performance (OECD, 2023).

This need becomes particularly relevant in higher education, where students are commonly referred to as “digital natives” (Prensky, 2001), under the premise that, having grown up in digital environments, they possess innate skills in the use of technological tools. However, belonging to this generation does not guarantee competent, critical, or learning-oriented use, as various studies have pointed out (Gallardo-Echenique et al., 2015; Li & Ranieri, 2010; Ng, 2012; Selwyn, 2009). This highlights the importance of educating students in digital competences that enable

them to actively adapt to a constantly evolving technological environment and to develop an open attitude towards innovation (Barak, 2018). Therefore, educators must actively promote the strengthening of these competences in order to foster student learning and take advantage of the educational potential offered by digital technologies.

In this context, the concept of digital competence has been widely addressed in the specialized literature, giving rise to a diversity of approaches and definitions. Among the most widely accepted definitions of digital competences today is the one that conceives them as an interrelated set of essential skills for successfully operating in digital environments (List, 2019). This notion is present in the European Digital Competence Framework for Citizens (DigComp), initially developed by Ferrari and Punie (2013). This framework defines digital competences as a set of knowledge, skills, and attitudes that enable citizens to effectively use digital technologies for work, learning, and social participation. The model identifies five key areas: information and data literacy, communication and collaboration, digital content creation, safety, and problem-solving, with the most recent version being DigComp 2.2 (Vuorikari et al., 2022).

Although numerous studies have adopted the DigComp framework as a reference for evaluating digital competences in higher education, the majority have concentrated on assessing students' levels of proficiency. In contrast, relatively few studies analyze the impact of didactic interventions on the effective development of these competences. Among the most relevant contributions in this line is the study by Gutiérrez Porlán and Serrano Sánchez (2016), which demonstrates how a didactic intervention mediated by digital technologies can significantly enhance competences across the five DigComp areas. These findings reinforce the importance of integrating intentional and sustained strategies to promote critical and functional digital literacy.

Despite the growing interest and the rapid adoption of generative AI tools, the empirical literature examining their impact on the improvement of university students' digital competences remains limited. Kasneci et al. (2023) highlight the importance of complementing the use of tools such as ChatGPT with didactic strategies that enhance the critical verification of information, encourage the use of reliable complementary sources, and integrate activities aimed at advancing students' cognitive development. Moreover, the design of effective prompts, the contextualized interpretation of AI responses, and reflection on its use are key elements for achieving meaningful interaction with these technologies (Lee & Palmer, 2025). Promoting these practices in the classroom can contribute to strengthening digital competences. Benvenuti et al. (2023) argue that artificial intelligence represents a valuable resource for educators in promoting skills such as creativity, critical thinking, and problem-solving. For their part, Arseven and Bal (2025) emphasize that this potential lies in its capacity to facilitate metacognitive processes that are fundamental for the development of these competences; however, both point out that educational experiences integrating AI for these purposes remain scarce in the literature. Consequently, it is necessary to analyze the extent to which a didactic and structured integration of these technologies contributes to the strengthening of such competences among students.

A first contribution from this perspective is the work of Dalgıç et al. (2024), which demonstrates that the integration of ChatGPT into education has a positive impact both on learning outcomes and on the development of students' digital competences. The study points out that digital literacy is an essential mediating variable and that, in order to fully benefit from the advantages of artificial intelligence,

students must possess and continuously strengthen their digital competences. It further concludes that the use of ChatGPT directly promotes skills such as information search and analysis, problem-solving, and learning in virtual environments, thus acting not only as a tool that requires digital competences but also as a resource that actively contributes to their development. Therefore, the authors emphasize the importance of educational institutions prioritizing training in digital competences in order to maximize the potential of advanced technologies such as ChatGPT. A second relevant contribution is the study by Naamati-Schneider and Alt (2024), who claim to be the first to empirically evaluate the hypothesis that the use of ChatGPT may render certain digital competences obsolete by taking over tasks previously carried out by students, particularly those related to information access and analysis. Nevertheless, their findings also reveal that the incorporation of this technology into the educational process can enhance the competence of critical evaluation, as it requires students to verify the validity and reliability of AI-generated content.

The analysis of previous studies shows a growing need to promote digital competences in higher education, while also revealing a still limited empirical basis regarding the impact of generative artificial intelligence tools on their development, particularly due to the lack of a unified reference framework that facilitates comparison across studies. Although these technologies show considerable potential to strengthen such competences, their effectiveness largely depends on how they are integrated into the teaching and learning process. In this context, the present research aims to quantify the effect of a structured didactic intervention based on generative AI, with the objective of providing evidence of its contribution to the development of students' digital competences within the DigComp framework, as well as analyzing its possible heterogeneous effects.

METHODOLOGY

To evaluate the impact of the formative use of generative artificial intelligence (AI) on learning, we implemented a randomized controlled trial (RCT) with students enrolled in the Microeconomics course. The students belonged to the degree programs in Business Administration and in Marketing and Commercial Management at the Catholic University of Murcia. Both programs offered the course in two modalities: face-to-face and online. In addition, in the Business Administration program, the face-to-face modality was available in both Spanish and English. Thus, a total of five groups (classrooms) were formed: three face-to-face and two online.

The assignment of the treatment was carried out at the group (classroom) level, so that each of the five groups was entirely allocated either to the treatment or to the control group. Randomization was stratified by modality (face-to-face or online) to ensure an equitable distribution of the treatment across modalities.¹

To maintain a consistent learning environment across all modalities, instructors standardized the teaching materials and methods, and established an evaluation protocol to ensure that all groups received identical course content (materials, resources, and activities), delivered uniformly and evaluated with the same criteria, regardless of the program. All course material was available through CANVAS, the university's institutional platform.

The intervention consisted of two components: (1) training focused on the strategic application of generative AI models, offered only to the treatment group, and (2) a series of online self-assessments, administered to both groups, composed of

García, C. G., & Pallarés, N. (2026). Impact of generative AI on university students' digital competences: experimental evidence based on the DigComp framework [Impacto de la IA generativa en competencias digitales universitarias: evidencia experimental basada en el marco DigComp]. *RIED-Revista Iberoamericana de Educación a Distancia*, 29(1), 53-77.

<https://doi.org/10.5944/ried.45533>

multiple-choice questions covering both the theoretical and practical contents of each course unit. The self-assessments included immediate feedback, which was differentiated according to the assigned group.

The training focused on techniques for designing effective prompts and on the use of large language models (LLMs) powered by AI, with the overall objective of improving students' learning and digital competences. To this end, the following specific objectives were established, aimed at promoting effective interaction with LLMs:

- Contextualize the question – Encourage students to describe the entire problem before formulating specific questions.
- Define key terms and variables – Teach students to make explicit the economic principles, formulas, or assumptions relevant to their queries.
- Break down complex problems – Advise structuring queries into step-by-step segments to obtain more organized responses.
- Use iterative questioning – Emphasize the importance of reformulating, refining, and expanding AI-generated responses to deepen understanding.
- Verify their own solutions – Encourage students to attempt solving the problems themselves before resorting to AI.

These objectives were addressed in a training workshop delivered exclusively to the treatment group during the first day of class. The session focused mainly on the use of ChatGPT, although other tools, such as Copilot and Gemini, were also briefly introduced. The educators provided an introductory explanation, accompanied by practical examples, after which the students interacted with the tools themselves and were able to raise their questions.

In the following weeks, the students were required to complete the online self-assessments at the end of each unit, through the Canvas platform of the Microeconomics course. Each test had a time limit of one hour, and upon completion, students received a numerical grade. A maximum of five attempts per assessment was allowed, with a minimum interval of one day between them. In each attempt, the questions were randomly generated from the item bank. Feedback in the self-assessments was intentionally designed to differ between the treatment groups. In the treatment group, the feedback provided step-by-step guidance, with personalized hints and prompts aligned with the specific objectives taught for effective interaction with LLMs. This formative feedback was designed to guide students toward the correct answer without explicitly providing it, fostering deeper understanding and enhancing their ability to solve problems autonomously. Instead of indicating only whether the answer was correct or incorrect, the feedback was adaptive: students could interact with different components of the feedback and customize their learning experience according to their needs. This allowed them to explore alternative explanations and refine their understanding based on their performance and learning gaps. In contrast, the control group students received only basic feedback indicating whether their answers were correct or incorrect, without explanations, hints, or additional instructional resources. The materials and data to replicate the intervention are available in an open repository.²

To evaluate the impact of the intervention, two questionnaires were administered: one on the first day of class, before the AI training workshop (pre), and another after the last self-assessment (post-intervention). Each questionnaire included

García, C. G., & Pallarés, N. (2026). Impact of generative AI on university students' digital competences: experimental evidence based on the DigComp framework [Impacto de la IA generativa en competencias digitales universitarias: evidencia experimental basada en el marco DigComp]. *RIED-Revista Iberoamericana de Educación a Distancia*, 29(1), 53-77.
<https://doi.org/10.5944/ried.45533>

two parts: (i) demographic, family, and economic literacy data ³, and (ii) an assessment of students' self-reported digital competences, based on the framework developed by Vuorikari et al. (2022). The first version of this framework, DigComp 1.0, was developed by Ferrari and Punie (2013), with the aim of identifying and defining the digital competences relevant for all citizens living and working in Europe. Currently, the DigComp model has been consolidated as one of the most widely used reference frameworks for the design of instruments measuring digital competences in higher education contexts (Mattar et al., 2022).

Table 1 presents a detailed description of the items answered by students in relation to the digital competences assessed. These were aligned with the European DigComp 2.2 model and included four key areas: information and data literacy, communication and collaboration, safety, and problem-solving. In the latter, two dimensions were evaluated: on the one hand, interaction with technological tools or functional dimension (Comp3), and on the other, digital self-regulation, understood as the ability to reflect metacognitively on one's own learning in digital environments (Comp4).

Table 1
Assessed digital competences

DigComp			Item
Area 1	Information and Data Literacy	Comp1	Identify my information needs and find data and content through a simple search in digital environments, such as Google.
Area 2	Communication and Collaboration	Comp2	Achieve effective communication using digital tools, such as email, discussion forums, etc.
Area 5	Problem Solving	Comp3	Provide simple instructions to a computer system to help me solve a basic problem or task, such as databases, spreadsheet software, presentation software, AI tools, etc.
	Problem Solving	Comp4	Recognize what I need to improve or update in my digital competences.
Area 4	Safety	Comp5	Evaluate the benefits and risks before allowing third parties to process personal data when I use digital environments or AI tools.

The items, written in self-report format with closed-ended options, began with "I am able to...", to assess the perceived skill level. They were formulated in a neutral manner, so that the response indicated the degree of proficiency without implying difficulty (Clifford et al., 2020). Their elaboration was based on the definitions and examples of the DigComp 2.2 framework, adapted to the objectives of the study. In addition, only five items were included to measure digital competences, in order to avoid making the questionnaire excessively long, thus reducing the risk of dropout and favoring the quality of responses, a strategy recommended in this field (Vuorikari et al., 2025).

Students rated their level of agreement with each statement using a 5-point Likert scale, where 1 corresponded to the lowest level (basic) and 5 to the highest (advanced). This format, based on self-reported items, is one of the most widely used methods to assess digital competence (Laanpere, 2019).

The reliability of the instrument was estimated using Cronbach's alpha and McDonald's omega coefficients, which yielded the same values in both the pretest (0.75) and the posttest (0.80). These coefficients indicate acceptable internal consistency at the initial measurement and good internal consistency after the intervention (Hair et al., 2010).

Estimation Method

The estimation is based on a randomized experiment combined with a difference-in-differences (DiD) model with fixed effects. First, randomization allows participants to be assigned to the treatment group and the control group exogenously, which ensures comparability and minimizes selection bias. Based on this assignment, a difference-in-differences strategy was implemented to compare the evolution of competence levels between both groups before and after the intervention.

Given that the treatment was assigned at the group (classroom) level, standard errors were clustered at that unit. To correct for potential biases arising from the small number of clusters (five), statistical significance was evaluated using the wild cluster bootstrap-t method, as detailed in the robustness section.

The econometric specification is as follows:

$$Y_{it} = \delta(Treat_i \times Post_t) + \beta Post_t + \gamma_i + \epsilon_{it}$$

Where:

- Y_{it} : dummy variable that takes the value 1 if the student reports the maximum level (value 5) in the competence analyzed.
- $Treat_i \times Post_t$: interaction between $Treat_i$ and $Post_t$.
- $Treat_i$: dummy variable that takes the value 1 if the student is in the treatment group.
- $Post_t$: dummy variable that takes the value 1 for observations after the intervention (post-intervention questionnaire).
- γ_i : individual fixed effects

The dichotomization of the dependent variable facilitates a more direct interpretation of the treatment effect by focusing on the probability of reaching the value 5, understood as a clear and demanding threshold of full competence. This definition mitigates social desirability bias by requiring the maximum response, filtering out possible moderate overestimations. Although it entails a partial loss of variance, the robustness of the results is contrasted with ordinal models (ordinal logit), presented in the corresponding section.

Table 2
Baseline characteristics

Variable	(1) All	(2) Treated	(3) Control	(4) P-value (2)-(3)
A. Student Characteristics				
Female	0.355 (0.037)	0.356 (0.047)	0.354 (0.060)	0.968
Age	20.911 (0.328)	20.808 (0.452)	21.077 (0.453)	0.823
(Age)^2	455.325 (20.568)	454.039 (29.661)	457.385 (24.936)	0.962
Working	0.266 (0.034)	0.269 (0.044)	0.262 (0.055)	0.903
Exchange student	0.053 (0.017)	0.058 (0.023)	0.046 (0.026)	0.691
Repeating student	0.077 (0.021)	0.096 (0.029)	0.046 (0.026)	0.360
Online student	0.089 (0.022)	0.067 (0.025)	0.123 (0.041)	0.760
Western nationality	0.716 (0.035)	0.625 (0.048)	0.862 (0.043)	0.512
Pre-course economic literacy	3.840 (0.14)	3.875 (0.176)	3.785 (0.232)	0.800
Pre-course digital competence ⁴	0.000 (0.123)	-0.297 (0.155)	0.475 (0.189)	0.141
B. Household Characteristics				
Income <20000€	0.095 (0.023)	0.096 (0.029)	0.092 (0.036)	0.853
Income 20000-40000€	0.243 (0.033)	0.269 (0.044)	0.2 (0.055)	0.124
Income 40000-60000€	0.266 (0.034)	0.269 (0.044)	0.262 (0.055)	0.846
Income >60000€	0.32 (0.036)	0.288 (0.045)	0.369 (0.06)	0.548
Mother - Primary education	0.065 (0.019)	0.087 (0.028)	0.031 (0.022)	0.301
Mother - Secondary education	0.420 (0.038)	0.337 (0.047)	0.554 (0.062)	0.118
Mother - Tertiary education	0.515 (0.039)	0.577 (0.049)	0.415 (0.062)	0.157
Father - Primary education	0.065 (0.019)	0.077 (0.026)	0.046 (0.026)	0.467

Variable	(1) All	(2) Treated	(3) Control	(4) P-value (2)-(3)
Father - Secondary education	0.432 (0.038)	0.394 (0.048)	0.492 (0.062)	0.534
Father - Tertiary education	0.503 (0.039)	0.529 (0.049)	0.462 (0.062)	0.652
Observations	169	104	65	

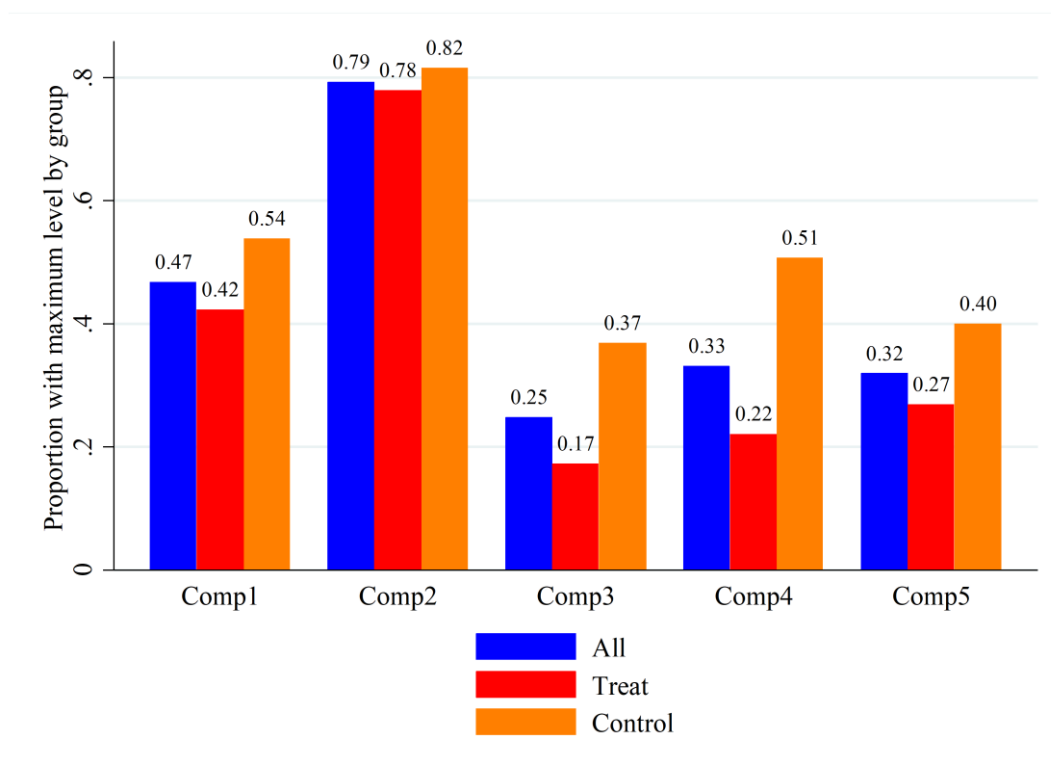
Note: Robust standard errors clustered at the classroom level in parentheses. The p-value in column (4) corresponds to the difference in means and was calculated adjusting for the small number of clusters using the wild cluster bootstrap-t method.

DATA AND DESCRIPTIVE STATISTICS

As described in the previous section, the data used come from two surveys conducted before and after the intervention.⁵ Table 2 presents the baseline characteristics of the students randomly assigned to the treatment and control groups. The total sample consists of 169 students, of whom 104 were randomly assigned to the treatment group and 65 to the control group according to the group (classroom) to which they belonged. No statistically significant differences were detected between the groups in any of the demographic, academic, or socioeconomic variables considered, which suggests that the randomization process was appropriate and that there is an adequate balance in the initial conditions.

Figure 1

Proportion of students with the maximum level (value 5) in digital competences before the intervention



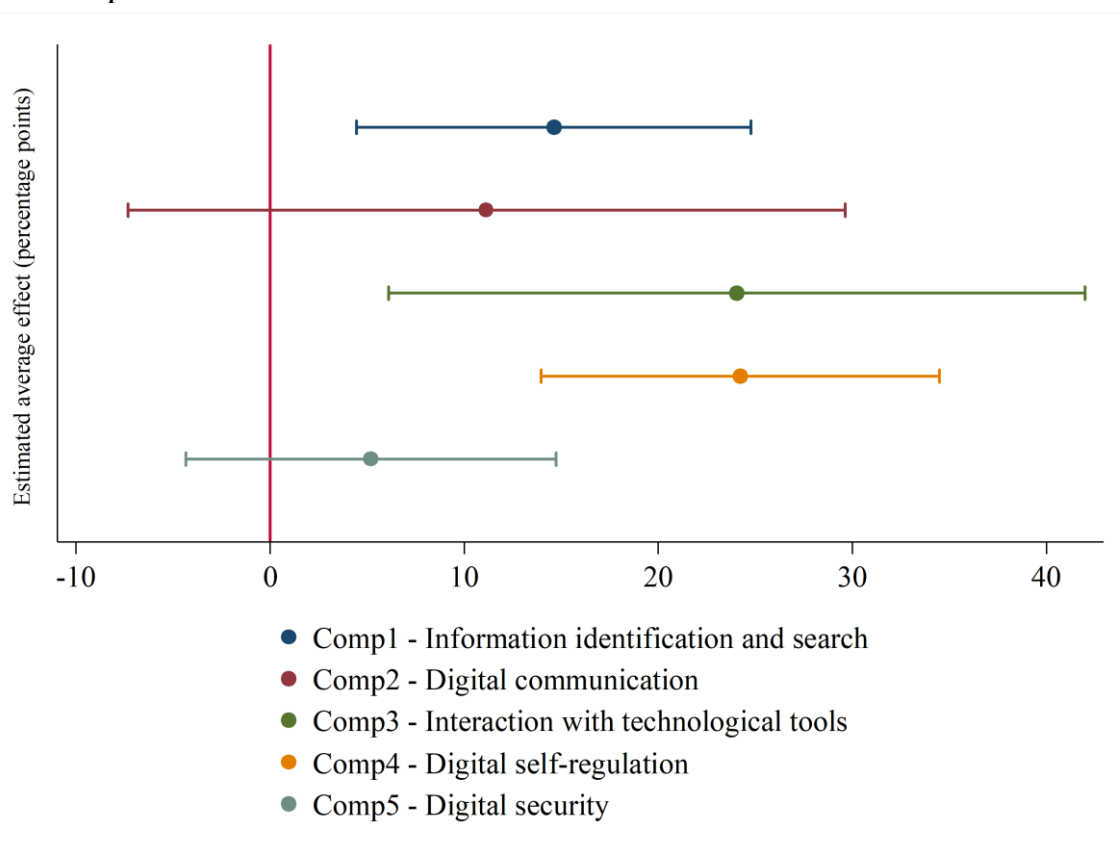
Additionally, the dimension-based analysis presented in Figure 1 reveals that the initial competence levels were heterogeneous across areas. The proportions of students who reported having reached the maximum level (value 5) ranged from 0.25 (Comp3) to 0.79 (Comp2) in the total sample, and from 0.17 to 0.78, respectively, within the treatment group. This indicates a considerable margin for improvement, especially in functional dimensions such as interaction with technological tools (Comp3), as well as in metacognitive areas related to digital self-regulation skills (Comp4) or informed decision-making in digital security (Comp5). As previous studies have pointed out, low levels of general digital competence are often associated with lower confidence and practice in digital security (Tomczyk, 2020), which could help explain the relatively lower values in these dimensions.

RESULTS AND DISCUSSION

The estimated models employ a dichotomous dependent variable that indicates whether the student reached the highest level of digital competence, which allows the coefficients (multiplied by 100) to be interpreted as changes in the probability, in percentage points (p.p.), of reaching that level.

Figure 2

Average treatment effects on the probability of reaching the highest level of digital competence



Note: Estimated coefficients of Treat*Post, in percentage points ($\times 100$), with 95% confidence intervals. These correspond to Panel A of Table 3.

Figure 2 visually illustrates the estimated average treatment effects on the probability of reaching the highest level of digital competence. It is clearly observed that the largest gains are concentrated in the dimensions of digital self-regulation (Comp4) and interaction with technological tools (Comp3), followed by competence in information identification and search (Comp1). In contrast, competences associated with digital communication (Comp2) and digital security (Comp5) show positive estimates, although with smaller magnitude and greater statistical uncertainty. This graphical representation provides an integrated visualization of the magnitude and statistical precision of the estimated effects for each competence.

These results correspond to the coefficients presented in Table 3, where a positive and significant impact of the treatment is confirmed in three of the five dimensions analyzed: Comp1, Comp3, and Comp4. The strongest impact is observed in Comp3 and Comp4, with increases of 24 p.p., followed by Comp1 with an increase of 14.6 p.p. In contrast, competences related to digital communication (Comp2) and digital security (Comp5) show positive but not significant effects at the average level, although differentiated patterns do emerge when the initial competence level is considered.

To explore the existence of distributive effects, models with interactions were estimated based on an indicator variable that identifies students whose initial digital competence was above the median (Q2). The results reveal heterogeneity in the impacts⁶, with the most pronounced benefits concentrated in the subgroup with lower initial levels (i.e., below the median), as shown in Panel B of Table 3. Among students with lower initial levels of digital competence, significant improvements were observed in all competences, with increases ranging from 18 to 38 p.p. In contrast, among those with higher initial digital competence, statistically significant effects were positive in the case of digital self-regulation (Comp4), with an effect of 14 p.p., but negative in the case of digital security (Comp5). This negative result could reflect an adjustment in the self-assessment of students with higher initial levels, who, after the intervention, may have developed a more critical view of their own competences. Although the study does not provide direct empirical evidence in this regard, this pattern is consistent with the theoretical implications of the Dunning–Kruger effect, according to which increased knowledge may be accompanied by greater awareness of one's own limitations (Dunning, 2011; Kruger & Dunning, 1999).

All of this points to a compensatory effect of the intervention in the context studied, consistent with research showing that technologies, when implemented with didactic orientation, can reduce gaps and foster more equitable learning. Recent research on generative AI highlights its potential to strengthen digital competences. Naamati-Schneider and Alt (2024) show that, in structured environments such as problem-based learning, ChatGPT promotes skills such as formulating effective prompts and critically evaluating results, emphasizing that these skills require deliberate instructional implementation. Otherwise, fundamental digital competences may be neglected. Along the same lines, Güner and Er (2025) observe that, without guidance, students use ChatGPT superficially, whereas training in prompting and strategic use stimulates more reflective and autonomous interactions. Finally, experimental studies and meta-analyses (Wang et al., 2025; Wang & Fan, 2025) confirm that personalized feedback from ChatGPT improves motivation, reflection, and performance in complex tasks, supporting the development of advanced cognitive and digital skills.

García, C. G., & Pallarés, N. (2026). Impact of generative AI on university students' digital competences: experimental evidence based on the DigComp framework [Impacto de la IA generativa en competencias digitales universitarias: evidencia experimental basada en el marco DigComp]. *RIED-Revista Iberoamericana de Educación a Distancia*, 29(1), 53-77.

<https://doi.org/10.5944/ried.45533>

The findings are especially relevant for the improvement of metacognitive competences such as digital self-regulation (Comp4). Lin et al. (2025) point out that the guided use of tools such as ChatGPT facilitates autonomous planning, the organization of ideas, and the formulation of questions by mitigating limitations in working memory and information structuring. However, they caution that such support may generate dependence on automated feedback, limiting critical autonomy. Hence the need to accompany AI with didactic strategies that promote active appropriation.

Although studies such as Hillmayr et al. (2020) focus on lower educational levels, their findings on the positive effects of technologies on self-regulation and problem-solving could be extrapolated to higher education, particularly among students with lower levels of competence.

The pattern observed in our results, with greater benefits among students with lower initial levels, suggests that these technologies, when implemented with didactic intentionality, can play a compensatory role. This leveling effect is particularly evident in key dimensions such as digital self-regulation (Comp4) and interaction with technological tools (Comp3), which are essential in higher education contexts that demand increasing autonomy and environments oriented toward the acquisition of digital competences linked to the critical use of AI. This is consistent with the literature arguing that generative AI, as personalized support, can perform functions equivalent to intelligent tutoring systems without replacing the educator (Hwang et al., 2020).

In the current educational landscape, the need to strengthen competences that enable students to adapt to dynamic and uncertain environments stands out, particularly those related to problem-solving, considered key to lifelong learning (Csapó & Funke, 2017; OECD, 2013; Rahman, 2019).

Table 3

Results of the linear difference-in-differences models with fixed effects (Estimated coefficients on the probability of reaching the highest level of digital competence)

	(1) Comp1	(2) Comp2	(3) Comp3	(4) Comp4	(5) Comp5
A. Average					
Treat*Post	0.146** (0.037)	0.112 (0.067)	0.240** (0.065)	0.242*** (0.037)	0.052 (0.034)
R ²	0.017	0.012	0.055	0.038	0.004
B. Pre-course digital competences above the median Q2					
Treat*Post	0.374*** (0.041)	0.209** (0.05)	0.294*** (0.045)	0.318*** (0.057)	0.178*** (0.02)
Treat*Post*Q2	-0.539*** (0.105)	-0.230* (0.108)	-0.126 (0.082)	-0.179 (0.095)	-0.298*** (0.009)
Treat*Post+Treat*Post*Q2	-0.165 (0.092)	-0.021 (0.133)	0.168 (0.097)	0.139* (0.065)	-0.12** (0.027)
R ²	0.151	0.039	0.061	0.05	0.043
N	338	338	338	338	338

Note: Robust standard errors clustered at the classroom level in parentheses. * $p < 0.10$ ** $p < 0.05$ *** $p < 0.01$

Robustness of the results

To assess the robustness of the results of the linear DiD models (Table 3), ordinal logit models were estimated under the same strategy. These models, which account for the ordinal nature of the dependent variable, allow the calculation of marginal effects on the probability of reaching the highest level of digital competence (value 5). The results (Table 4) confirm the validity of the estimates by showing consistent effects.

On average (Panel A), the treatment shows positive and significant effects across all dimensions (13–28 p.p.), especially in Comp4 and Comp3. The pattern is accentuated among students with lower initial levels (Panel B), who exhibit greater improvements (21–36 p.p.), whereas higher-level students also benefit, albeit to a lesser extent (17–27 p.p.). The negative and significant interactions reinforce this difference.

The differences between models stem from the way the dependent variable is specified: the linear model operates on a dichotomous version, whereas the ordinal model considers the entire scale and captures non-linear effects. Both coincide with the key patterns, with positive effects especially among students with lower initial levels. The linear model is preferred for its interpretative clarity and control of unobserved heterogeneity through fixed effects.⁷ The coherence across approaches reinforces the robustness of the conclusions.

Table 4

Results of the ordinal logit difference-in-differences models (Marginal effects on the probability of reaching the highest level of digital competence)

	(1) Comp1	(2) Comp2	(3) Comp3	(4) Comp4	(5) Comp5
A. Average					
Treat*Post	0.144*** (0.000)	0.126*** (0.002)	0.175*** (0.008)	0.281*** (0.003)	0.160*** (0.008)
R ²	0.02	0.02	0.021	0.028	0.02
B. Pre-course digital competences above the median Q2					
Treat*Post	0.211*** (0.001)	0.209*** (0.000)	0.227*** (0.004)	0.364*** (0.003)	0.302*** (0.003)
Treat*Post*Q2	-0.247*** (0.001)	-0.253*** (0.002)	-0.151*** (0.002)	-0.233*** (0.095)	-0.330*** (0.005)
Treat*Post+Treat*Post*Q2	0.188*** (0.011)	0.170*** (0.012)	0.185*** (0.009)	0.273*** (0.008)	0.183*** (0.017)
R ²	0.111	0.081	0.098	0.042	0.102
N	338	338	338	338	338

Note: Robust standard errors clustered at the classroom level in parentheses. * $p < 0.10$ ** $p < 0.05$ *** $p < 0.01$

In addition to the consistency observed in the alternative models (ordinal logit), the robustness of the results has been verified from the perspective of statistical inference. As detailed in the methodological section, since treatment assignment was carried out at the group (classroom) level, with a total of five clusters, the statistical

significance of the coefficients was evaluated using the wild cluster bootstrap-t method proposed by Cameron et al. (2008), implemented with Webb (2014) weights and 1,000 replications.⁸ The main results remain stable under this procedure and the overall conclusions are not affected, which supports the validity of the estimates.

CONCLUSIONS AND RECOMMENDATIONS

This study provides empirical evidence of the positive impact of a didactic intervention, focused on the use of generative artificial intelligence and specifically aimed at developing digital competences among higher education students. The analysis reveals significant average effects in three of the five dimensions evaluated within the DigComp framework: information identification and search (Comp1), interaction with technological tools (Comp3), and digital self-regulation (Comp4), with increases ranging from 14.6 to 24 p.p.

These advances are relevant not only because of their magnitude but also due to the type of competences involved. In particular, the increase in “information and data literacy” (Comp1) stands out in a context where the abundance of content generated by automated systems requires critical skills to filter, verify, and cross-check information. This finding reinforces the central role of this competence in the digital citizenship and autonomous learning frameworks proposed by Vuorikari et al. (2025).

Regarding Comp3 and Comp4, the results show significant average improvements in both the functional use of technology and the metacognitive dimension of digital learning. These competences are particularly critical in higher education contexts that require greater student autonomy. Likewise, the analysis of heterogeneous effects reveals a consistent pattern: the greatest benefits are concentrated among those starting from lower initial levels, suggesting that the intervention had a leveling effect. In these cases, significant improvements were observed across all competences, reinforcing the compensatory potential of such strategies.

The findings suggest the need to revisit the widespread idea that university students, due to their familiarity with technology, possess well-developed digital skills. As previous research has emphasized, being a ‘digital native’ does not guarantee competent or critical use of technological tools. On the contrary, the results highlight the need for explicit, well-designed, and guided didactic interventions that systematically address the acquisition of these skills (Gallardo-Echenique et al., 2015; Ng, 2012).

Taken together, this study contributes to the emerging field of generative AI in education by providing quantitative evidence of how its didactic integration can strengthen both functional and metacognitive competences. These results support the intentional use of emerging technologies not only as instructional resources but also as tools to reduce inequalities and foster more equitable and inclusive learning environments across similar contexts, aligned with the principles of lifelong learning and the European standards for digital competences.

Finally, the results underscore the importance of educators actively participating in the didactic integration of these technologies. This requires educators with strong digital competences and an open attitude toward change, capable of adapting their practices to evolving educational contexts (Inamorato dos Santos et al., 2023). Moreover, given that the impact varies according to students’ initial levels, their implementation must be guided by instructional criteria that ensure effective and

García, C. G., & Pallarés, N. (2026). Impact of generative AI on university students’ digital competences: experimental evidence based on the DigComp framework [Impacto de la IA generativa en competencias digitales universitarias: evidencia experimental basada en el marco DigComp]. *RIED-Revista Iberoamericana de Educación a Distancia*, 29(1), 53-77.

<https://doi.org/10.5944/ried.45533>

ethical use aimed at closing gaps. Therefore, harnessing the possibilities offered by generative AI in education not only entails introducing new tools, but also rethinking the conditions that enable these technologies to contribute to a more equitable education, better adapted to contemporary challenges.

Nevertheless, these results must be interpreted in light of certain limitations. First, the sample is composed exclusively of students from a single university, which limits the generalizability of the findings to other educational contexts. Future research could broaden the sample by including different institutions and courses, which would allow for the evaluation of the transferability and scalability of the intervention. Second, although assignment to the treatment was random, it was carried out at the classroom level for didactic reasons and to avoid spillovers, which reduced the number of clusters and may entail some risk of ex-ante self-selection. To address this, the analyses incorporated resampling-based inference, as well as prior balance tests (Table 2), which show a balanced distribution of observable characteristics across conditions. Individual fixed effects were also included in the difference-in-differences models to control for potential biases associated with time-invariant unobserved factors. This approach strengthens causal interpretation, although it does not replace the evidence that could be provided by additional studies. Third, the dependent variables come from self-reported questionnaires, which may introduce perception or social desirability biases. To mitigate these potential biases, we defined the variable of interest strictly (only those who reported the maximum level of competence), to avoid moderate overestimations. Future studies could complement this with more objective measures, such as performance rubrics or the analysis of digital artifacts.

NOTES

1. Randomization led to the following allocation of groups (classrooms): in Business Administration, the face-to-face group in English and the online group were assigned to the treatment; in Marketing and Commercial Management, the face-to-face group was also assigned to the treatment. The remaining groups acted as the control group: the face-to-face group in Business Administration and the online group in Marketing and Commercial Management.
2. https://figshare.com/projects/Impacto_de_la_IA_Generativa_en_Competencias_Digitales_Universitarias_Evidencia_Experimental_Basada_en_el_Marco_DigComp/262561
3. Economic literacy was assessed in the questionnaire through 10 multiple-choice questions, adapted from the Test of Economic Literacy (TEL-4) and the microeconomics section of the Test of Understanding in College Economics (TUCE-4) (Walstad et al., 2007, 2013).
4. Each student's digital competence prior to the course is estimated through a principal component analysis based on the five assessed competences. The resulting index ranges from -5.662 to 2.341.
5. Additionally, administrative enrollment information was used (gender, age, exchange student, repeating student, and online student).
6. Moreover, beyond the initial level of digital competence, the heterogeneity of the treatment effect was analyzed according to gender, age, and prior knowledge in economics. Stronger effects were observed among men and among students with lower levels of economic literacy at the beginning of the course. For women, digital self-regulation (Comp4) showed significant improvements, a dimension that displays positive effects across all subgroups. Age, by contrast, does not show a clear differential effect.
7. Although, in formal terms, it is possible to include individual dummies in nonlinear models such as the ordinal logit, this approach does not reproduce the estimation properties of linear fixed effects and may generate biases or identification problems, especially when the number of observations per unit is limited (Greene, 2004).
8. This approach, suitable for fewer than ten clusters, improves the accuracy of p-values and remains robust even with five clusters.

REFERENCES

- Arseven, T., & Bal, M. (2025). Critical literacy in artificial intelligence-assisted writing instruction: A systematic review. *Thinking Skills and Creativity*, 57, 101850. <https://doi.org/10.1016/j.tsc.2025.101850>
- Barak, M. (2018). Are digital natives open to change? Examining flexible thinking and resistance to change. *Computers & Education*, 121, 115-123. <https://doi.org/10.1016/j.compedu.2018.01.016>
- Benvenuti, M., Cangelosi, A., Weinberger, A., Mazzoni, E., Benassi, M., Barbaresi, M., & Orsoni, M. (2023). Artificial intelligence and human behavioral development: A perspective on new skills and competences acquisition for the educational context. *Computers in Human Behavior*, 148, 107903. <https://doi.org/10.1016/j.chb.2023.107903>
- Bastian, A., Liza, L. O., & Efastri, S. M. (2023). Revolutionizing education: How digital literacy is transforming inclusive classrooms in post-COVID-19. *Journal of Public Health*, 45(3), e609-e610. <https://doi.org/10.1093/pubmed/fdado58>
- Blau, I., Shamir-Inbal, T., & Avdiel, O. (2020). How does the pedagogical design of a technology-enhanced collaborative academic course promote digital literacies, self-regulation, and perceived learning of students? *The Internet and Higher Education*, 45, 100722. <https://doi.org/10.1016/j.iheduc.2019.100722>
- Buchholz, B. A., DeHart, J., & Moorman, G. (2020). Digital citizenship during a global pandemic: Moving beyond digital literacy. *Journal of Adolescent & Adult Literacy*, 64(1), 11-17. <https://doi.org/10.1002/jaal.1076>
- Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics*, 90(3), 414-427. <https://doi.org/10.1162/rest.90.3.414>
- Dalgıç, A., Yaşar, E., & Demir, M. (2024). ChatGPT and learning outcomes in tourism education: The role of digital literacy and individualized learning. *Journal of Hospitality, Leisure, Sport & Tourism Education*, 34, 100481. <https://doi.org/10.1016/j.jhlste.2024.100481>
- Dimitriadou, E., & Lanitis, A. (2023). A critical evaluation, challenges, and future perspectives of using artificial intelligence and emerging technologies in smart classrooms. *Smart Learning Environments*, 10(1), Article 12. <https://doi.org/10.1186/s40561-023-00231-3>
- Dunning, D. (2011). The Dunning–Kruger effect: On being ignorant of one's own ignorance. In J. M. Olson & M. P. Zanna (Eds.), *Advances in Experimental Social Psychology* (Vol. 44, pp. 247-296). <https://doi.org/10.1016/B978-0-12-385522-0.00005-6>
- Clifford, I., Kluzer, S., Troia, S., Jakobsone, M., Zandbergs, U., Vuorikari, R., Punie, Y., Castaño-Muñoz, J., Mediavilla, C., O'Keeffe, W., & Cabrera, M. (2020). *DigCompSAT: A self-reflection tool for the European Digital Competence Framework for citizens* (JRC Technical Report). Publications Office of the European Union. <https://doi.org/10.2760/77437>
- Csapó, B., & Funke, J. (Eds.). (2017). *The nature of problem solving: Using research to inspire 21st century learning*. OECD Publishing. <https://doi.org/10.1787/9789264273955-en>
- European Commission, Directorate-General for Education, Youth, Sport and Culture (2019). *Key competences for lifelong learning*. Publications Office of the European Union. <https://doi.org/10.2766/569540>
- Ferrari, A., & Punie, Y. (2013). *DIGCOMP: A framework for developing and understanding digital competence in Europe*. Publications Office of the European Union. <https://doi.org/10.2788/52966>

- García Peñalvo, F. J., Llorens-Largo, F., & Vidal, J. (2024). La nueva realidad de la educación ante los avances de la inteligencia artificial generativa. *RIED-Revista Iberoamericana de Educación a Distancia*, 27(1), 9-39. <https://doi.org/10.5944/ried.27.1.37716>
- Gallardo-Echenique, E. E., de Oliveira, J. M., Marqués-Molias, L., Esteve-Mon, F., Wang, Y., & Baker, R. (2015). Digital competence in the knowledge society. *MERLOT Journal of Online Learning and Teaching*, 11(1). https://jolt.merlot.org/vol11no1/Gallardo-Echenique_0315.pdf
- Greene, W. (2004). The behaviour of the maximum likelihood estimator of limited dependent variable models in the presence of fixed effects. *The Econometrics Journal*, 7(1), 98-119. <https://doi.org/10.1111/j.1368-423X.2004.00123.x>
- Güner, H., & Er, E. (2025). AI in the classroom: Exploring students' interaction with ChatGPT in programming learning. *Education and Information Technologies*, 30, 12681-12707. <https://doi.org/10.1007/s10639-025-13337-7>
- Gutiérrez Porlán, I., & Serrano Sánchez, J. L. (2016). Evaluation and development of digital competence in future primary school teachers at the University of Murcia. *Journal of New Approaches in Educational Research*, 5(1), 51-56. <https://doi.org/10.7821/naer.2016.1.152>
- Hair, J. F., Jr., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis* (7th ed.). Prentice Hall.
- Hillmayr, D., Ziernwald, L., Reinhold, F., Hofer, S. I., & Reiss, K. M. (2020). The potential of digital tools to enhance mathematics and science learning in secondary schools: A context-specific meta-analysis. *Computers & Education*, 153, 103897. <https://doi.org/10.1016/j.compedu.2020.103897>
- Hwang, G. J., Xie, H., Wah, B. W., & Gašević, D. (2020). Vision, challenges, roles and research issues of artificial intelligence in education. *Computers and Education: Artificial Intelligence*, 1, 100001. <https://doi.org/10.1016/j.caeai.2020.100001>
- Inamorato dos Santos, A., Chinkes, E., Carvalho, M. A. G., Solórzano, C. M. V., & Marroni, L. S. (2023). The digital competence of academics in higher education: Is the glass half empty or half full? *International Journal of Educational Technology in Higher Education*, 20, Article 9. <https://doi.org/10.1186/s41239-022-00376-0>
- Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., ... Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 102274. <https://doi.org/10.1016/j.lindif.2023.102274>
- Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, 77(6), 1121-1134. <https://doi.org/10.1037/0022-3514.77.6.1121>
- Laanpere, M. (2019). *Recommendations on assessment tools for monitoring digital literacy within UNESCO's Digital Literacy Global Framework* (Information Paper No. 56; UIS/2019/LO/IP/56). UNESCO Institute for Statistics. <https://unesdoc.unesco.org/ark:/48223/pf0000366740>
- Laupichler, M. C., Aster, A., Schirch, J., & Raupach, T. (2022). Artificial intelligence literacy in higher and adult education: A scoping literature review. *Computers and Education: Artificial Intelligence*, 3, 100101. <https://doi.org/10.1016/j.caeai.2022.100101>
- Lee, D., & Palmer, E. (2025). Prompt engineering in higher education: A systematic review to help inform curricula. *International Journal of*

- Educational Technology in Higher Education*, 22, Article 7. <https://doi.org/10.1186/s41239-025-00503-7>
- Lee, D., Arnold, M., Srivastava, A., Plastow, K., Strelan, P., Ploeckl, F., Lekkas, D., & Palmer, E. (2024). The impact of generative AI on higher education learning and teaching: A study of educators' perspectives. *Computers and Education: Artificial Intelligence*, 6, 100221. <https://doi.org/10.1016/j.caeai.2024.100221>
- Li, Y., & Ranieri, M. (2010). Are 'digital natives' really digitally competent?—A study on Chinese teenagers. *British Journal of Educational Technology*, 41(6), 1029-1042. <https://doi.org/10.1111/j.1467-8535.2009.01053.x>
- Lin, C. J., Lee, H. Y., Wang, W. S., Huang, Y. M., & Wu, T. T. (2025). Enhancing reflective thinking in STEM education through experiential learning: The role of generative AI as a learning aid. *Education and Information Technologies*, 30(5), 6315-6337. <https://doi.org/10.1007/s10639-024-13072-5>
- List, A. (2019). Defining digital literacy development: An examination of pre-service teachers' beliefs. *Computers & Education*, 138, 146-158. <https://doi.org/10.1016/j.compedu.2019.03.009>
- Mattar, J., Ramos, D. K., & Lucas, M. R. (2022). DigComp-based digital competence assessment tools: Literature review and instrument analysis. *Education and Information Technologies*, 27(8), 10843-10867. <https://doi.org/10.1007/s10639-022-11034-3>
- Moravec, V., Hynek, N., Gavurová, B., & Rigelský, M. (2024). Who uses it and for what purpose? The role of digital literacy in ChatGPT adoption and utilisation. *Journal of Innovation & Knowledge*, 9(4), 100602. <https://doi.org/10.1016/j.jik.2024.100602>
- Morgan, A., Sibson, R., & Jackson, D. (2022). Digital demand and digital deficit: Conceptualising digital literacy and gauging proficiency among higher education students. *Journal of Higher Education Policy and Management*, 44(3), 258-275. <https://doi.org/10.1080/1360080X.2022.2030275>
- Naamati-Schneider, L., & Alt, D. (2024). Beyond digital literacy: The era of AI-powered assistants and evolving user skills. *Education and Information Technologies*, 29(16), 21263-21293. <https://doi.org/10.1007/s10639-024-12694-z>
- Ng, W. (2012). Can we teach digital natives digital literacy? *Computers & Education*, 59(3), 1065-1078. <https://doi.org/10.1016/j.compedu.2012.04.016>
- OECD. (2013). *PISA 2012 assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy*. OECD Publishing. <https://doi.org/10.1787/9789264190511-en>
- OECD. (2023). *OECD Digital Education Outlook 2023: Towards an effective digital education ecosystem*. OECD Publishing. <https://doi.org/10.1787/c74f03de-en>
- Polizzi, G. (2020). Digital literacy and the national curriculum for England: Learning from how the experts engage with and evaluate online content. *Computers & Education*, 152, 103859. <https://doi.org/10.1016/j.compedu.2020.103859>
- Popenici, S. A. D., & Kerr, S. (2017). Exploring the impact of artificial intelligence on teaching and learning in higher education. *Research and Practice in Technology Enhanced Learning*, 12(1), 22. <https://doi.org/10.1186/s41039-017-0062-8>
- Prensky, M. (2001). Digital natives, digital immigrants. *On the Horizon*, 9(5), 1-6. <https://doi.org/10.1108/10748120110424816>
- Prior, D. D., Mazanov, J., Meacheam, D., Heaslip, G., & Hanson, J. (2016). Attitude, digital literacy and self-efficacy: Flow-on effects for online learning behavior. *The Internet and Higher Education*, 29, 91-97.

- <https://doi.org/10.1016/j.iheduc.2016.01.001>
- Rahman, M. M. (2019). 21st century skill “problem solving”: Defining the concept. *Asian Journal of Interdisciplinary Research*, 2(1), 64-74. <https://doi.org/10.34256/ajir1917>
- Saklaki, A., & Gardikiotis, A. (2024). Exploring Greek students’ attitudes toward artificial intelligence: Relationships with AI ethics, media, and digital literacy. *Societies*, 14(12), 248. <https://doi.org/10.3390/soc14120248>
- Selwyn, N. (2009). The digital native—myth and reality. *Aslib Proceedings: New Information Perspectives*, 61(4), 364-379. <https://doi.org/10.1108/00012530910973776>
- Ting, Y.-L. (2015). Tapping into students’ digital literacy and designing negotiated learning to promote learner autonomy. *The Internet and Higher Education*, 26, 25-32. <https://doi.org/10.1016/j.iheduc.2015.04.004>
- Tomczyk, Ł. (2020). Skills in the area of digital safety as a key component of digital literacy among teachers. *Education and Information Technologies*, 25(1), 471-486. <https://doi.org/10.1007/s10639-019-09980-6>
- Van Laar, E., van Deursen, A. J. A. M., van Dijk, J. A. G. M., & de Haan, J. (2020). Determinants of 21st-century skills and 21st-century digital skills for workers: A systematic literature review. *Sage Open*, 10(1), 1-14. <https://doi.org/10.1177/2158244019900176>
- Vuorikari, R., Kluzer, S., & Punie, Y. (2022). *DigComp 2.2: The digital competence framework for citizens – with new examples of knowledge, skills and attitudes* [EUR 31006 EN; JRC128415]. Publications Office of the European Union. <https://doi.org/10.2760/115376>
- Vuorikari, R., Pokropek, A., & Castaño Muñoz, J. (2025). Enhancing digital skills assessment: Introducing compact tools for measuring digital competence. *Technology, Knowledge and Learning*. Advance online publication. <https://doi.org/10.1007/s10758-025-09825-x>
- Walstad, W. B., Rebeck, K., & Butters, R. B. (2013). *Test of Economic Literacy: Examiner’s manual* (4th ed.). Council for Economic Education. <https://www.econedlink.org/wp-content/uploads/2018/09/TEL-Manual.pdf>
- Walstad, W. B., Watts, M., & Rebeck, K. (2007). *Test of Understanding of College Economics: Examiner’s manual* (4th ed.). Council for Economic Education. <https://www.econedlink.org/wp-content/uploads/2018/09/TUCE-4th.pdf>
- Wang, J., & Fan, W. (2025). The effect of ChatGPT on students’ learning performance, learning perception, and higher-order thinking: Insights from a meta-analysis. *Humanities and Social Sciences Communications*, 12, 621. <https://doi.org/10.1057/s41599-025-04787-y>
- Wang, W.-S., Lin, C.-J., Lee, H.-Y., Huang, Y.-M., & Wu, T.-T. (2025). Integrating feedback mechanisms and ChatGPT for VR-based experiential learning: Impacts on reflective thinking and AIoT physical hands-on tasks. *Interactive Learning Environments*, 33(2), 1770-1787. <https://doi.org/10.1080/10494820.2024.2375644>
- Webb, M. D. (2014). *Reworking wild bootstrap based inference for clustered errors* (QED Working Paper No. 1315). Queen’s University, Department of Economics. https://qed.econ.queensu.ca/working_papers/papers/qed_wp_1315.pdf
- Zhai, X., Chu, X., Chai, C. S., Jong, M. S. Y., Istenič, A., Spector, J. M., Liu, J., Yuan, J., & Li, Y. (2021). A review of artificial intelligence (AI) in education from 2010 to 2020. *Complexity*, 2021, 8812542. <https://doi.org/10.1155/2021/8812542>
- Zhao, Y., Pinto Llorente, A. M., & Sánchez Gómez, M. C. (2021). Digital competence in higher education research: A systematic literature review. *Computers & Education*, 168, 104212. <https://doi.org/10.1016/j.compedu.2021.104212>

Date of reception: 1 June 2025

Date of acceptance: 9 September 2025

Date of approval for layout: 6 October 2025

Date of publication in OnlineFirst: 24 October 2025

Date of publication: 1 January 2026

AI-ED-SAT: design and validation of a questionnaire for self-assessment of teaching skills in educational AI

AI-ED-SAT: diseño y validación de un cuestionario para autoevaluar competencias docentes en IA educativa



 Andresa Sartor-Harada - *Universidad Internacional de La Rioja, UNIR (Spain)*

 Juliana Azevedo Gomes - *Universidad Internacional de La Rioja, UNIR (Spain)*

ABSTRACT

The accelerated integration of Artificial Intelligence (AI) in education poses new challenges for teacher training and assessment. This study presents the design, validation, and psychometric analysis of the AI-ED-SAT questionnaire, a self-assessment tool designed for teachers to diagnose their level of preparedness in the pedagogical, ethical, and curricular use of AI. The AI-ED-SAT questionnaire was validated within the Design-Based Research approach. It comprises four dimensions aligned with UNESCO's current AI competency frameworks (2024a, 2024b): conceptual understanding of AI, pedagogical use of intelligent tools, ethical and critical reflection, and curricular integration. Its construction was based on an exhaustive theoretical review and an iterative validation process using the Delphi method with 12 experts in educational technology, AI, and teacher training. Subsequently, in a pilot test involving a sample of 128 teachers from various educational levels, the reliability (Cronbach's $\alpha = 0.93$), content validity, and construct validity were analyzed. The results of the exploratory factor analysis (EFA) confirmed the grouping of items into four theoretical factors, while the confirmatory factor analysis (CFA) showed excellent fit indices (CFI = 0.96; RMSEA = 0.045). The AI-ED-SAT is a robust and up-to-date tool that is useful in both research and teacher training programs. Its self-reflective approach helps strengthen teachers' critical literacy and professional agency in the face of the challenges posed by AI in education.

Keywords: artificial intelligence; education; teacher competencies; questionnaire; instrument validation.

RESUMEN

La integración acelerada de la Inteligencia Artificial (IA) en el ámbito educativo plantea nuevos retos para la formación y evaluación de competencias docentes. Este estudio presenta el diseño, validación y análisis psicométrico del cuestionario AI-ED-SAT, una herramienta de autoevaluación dirigida a docentes para diagnosticar su grado de preparación en el uso pedagógico, ético y curricular de la IA. Desarrollado bajo un enfoque de investigación basado en diseño (Design-Based Research), el instrumento se estructuró en torno a cuatro dimensiones alineadas con los actuales marcos de competencias en IA de la UNESCO (2024a, 2024b): comprensión conceptual de la IA, uso pedagógico de herramientas inteligentes, reflexión ética y crítica e integración curricular. Su construcción se apoyó en una revisión teórica exhaustiva y un proceso iterativo de validación mediante el método Delphi con 12 expertos en tecnología educativa, IA y formación docente. Posteriormente, en una prueba piloto a una muestra de 128 docentes de distintos niveles educativos, se analizó su fiabilidad (α de Cronbach = 0.93), validez de contenido y validez de constructo. Los resultados del análisis factorial exploratorio (AFE) confirmaron la agrupación de los ítems en cuatro factores teóricos, mientras que el análisis factorial confirmatorio (AFC) mostró índices de ajuste excelentes (CFI = 0.96; RMSEA = 0.045). El AI-ED-SAT se configura como una herramienta robusta y actualizada, útil tanto en la investigación como en programas de formación docente. Su enfoque autorreflexivo contribuye a fortalecer la alfabetización crítica y la agencia profesional del profesorado frente a los desafíos de la IA en la educación.

Palabras clave: inteligencia artificial; educación; competencias docentes; cuestionario; validación de instrumento.

INTRODUCTION

The emergence of Artificial Intelligence (AI) in educational ecosystems is transforming not only teaching and learning methodologies, but also the ethical, pedagogical, and organizational frameworks that govern teaching practice. While technological advances have enhanced the development of adaptive virtual environments, intelligent tutoring systems, and automated assessment processes, their implementation in school contexts requires specific, situated, and reflective teacher preparation (Holmes et al., 2019; UNESCO, 2021; UNESCO, 2024b).

Beyond access to digital devices or platforms, education professionals must develop critical AI competencies that enable them to understand how AI works, assess its social and ethical implications, and design pedagogical practices consistent with the principles of equity, inclusion, and educational value. In this context, AI teaching competencies should be understood as a multidimensional set of knowledge, skills, and attitudes that go beyond traditional digital literacy (Celik, 2023; Chen et al., 2020).

However, the literature reveals a shortage of specific, validated instruments that enable the assessment of these skills in relation to AI in a rigorous, contextualized, and formative manner. Most existing models focus on generic indicators of digital competence, which limits their ability to address the particular challenges posed by the integration of AI in education (Zawacki-Richter et al., 2019). This gap hinders both empirical research and the design of teacher training programs that cater to new technological demands.

In response to this need, this study presents the design, validation, and psychometric analysis of the AI-ED-SAT (Artificial Intelligence in Education – Self-Assessment Tool), a self-assessment questionnaire aimed at teachers at different educational levels, with the purpose of diagnosing their level of competence in the use of AI in the school environment. The instrument is structured around four key dimensions—conceptual understanding of AI, pedagogical use of tools, ethics and considerations in the classroom, and curriculum integration—and has been developed using a mixed methodological approach, combining theoretical review, expert judgment validation (Delphi method), and exploratory and confirmatory factor analyses. Its objective is twofold: to provide a rigorous tool for educational research and, at the same time, to serve as a training resource that promotes professional reflection and continuous improvement among teachers in the face of the challenges posed by AI in education.

Although the questionnaire was developed entirely in Spanish, it was decided to retain an English acronym to facilitate its visibility and adoption in future international research. This decision responds to the growing presence of English as the *lingua franca* in the field of educational AI, which may favor its inclusion in academic databases, global collaboration networks, and translation and cross-cultural adaptation processes.

Education and emerging technologies: the new teaching paradigm

The advancement of emerging technologies, particularly AI, is profoundly reshaping the contemporary educational landscape. Recent research highlights the urgent need to redefine teaching skills in light of this transformation (Celik, 2023; Holmes et al., 2019; UNESCO, 2023). Far from being limited to technical mastery, these competencies include a critical understanding of the impact of technology on

teaching and learning processes, the ethical use of data, ensuring inclusion, and designing personalized and meaningful experiences.

The digital transformation has brought about significant structural changes in education, affecting not only the tools available and pedagogical dynamics but also regulatory frameworks and the professional roles of teachers. In this context, technologies such as augmented reality, big data, blockchain, and AI itself not only open up new teaching possibilities but also impose the need for a profound reconfiguration of teachers' professional competencies (Bo, 2024; UNESCO, 2024b).

This new paradigm requires teachers to perform more complex functions: designing learning experiences, acting as critical mediators between technology and the curriculum, and managing digital knowledge ethically. Consequently, technological integration in the classroom cannot be reduced to instrumental or technical use but requires a contextualized pedagogical approach capable of assessing its social, ethical, and educational implications and adapting them to the needs and trajectories of students (Cabero-Almenara et al., 2020).

Various international initiatives, such as the DigCompEdu framework (Ghomi & Redecker, 2019) and UNESCO's guidelines for teacher competence in ICT, as well as those currently specific to the use of AI, have established benchmarks for the development of digital competences in teachers. However, these frameworks tend to take a generalist perspective, focusing on cross-cutting skills and established technologies, without addressing in sufficient depth the specific challenges posed by emerging technologies, particularly AI. This limitation underscores the need to develop models that integrate specific dimensions of critical appropriation of these technologies, aiming to promote new digital literacies that position teachers as active agents of transformation within contemporary educational ecosystems.

Although DigCompEdu (Ghomi & Redecker, 2019) is a widely recognized framework for assessing teachers' digital competence, its generalist orientation limits its ability to address the specificities of Artificial Intelligence in depth. This instrument describes cross-cutting areas of competence applicable to various digital technologies, without differentiating between established and emerging tools, or explicitly integrating ethical and curricular dimensions specific to AI. In contrast, AI-ED-SAT focuses exclusively on the educational use of Artificial Intelligence, incorporating indicators linked to the conceptual understanding of algorithms, contextualized pedagogical application, ethical reflection, and strategic curricular integration. Thus, the AI-ED-SAT does not seek to replace frameworks such as DigCompEdu, but rather to complement them, offering a level of detail adapted to the unique challenges and opportunities presented by AI in educational settings.

Artificial Intelligence in education: potential, risks, and controversies

The inclusion of Artificial Intelligence into education represents one of the most significant transformations in recent years, with direct implications for teaching practices, pedagogical models, and teaching-learning processes. Its presence manifests itself in multiple forms: intelligent tutoring systems, adaptive platforms, learning analytics, chatbots, recommendation engines, and automated feedback tools (Holmes et al., 2019; Zawacki-Richter et al., 2019).

This technological expansion offers significant opportunities to personalize learning, optimize assessment processes, and alleviate teachers' administrative tasks. However, it also poses substantial challenges that cannot be ignored. These include

algorithmic biases, the opacity of automated systems, privacy violations, and the risk of delegating fundamental pedagogical decisions to systems that operate with logic alien to human judgment (Chesterman, 2021; Ferrante, 2021; Huang, 2023; Selwyn, 2019).

In this sense, AI cannot be conceived solely as a functional tool, but rather as a complex socio-technical phenomenon that involves political, pedagogical, and ethical decisions (Flores-Vivar & García-Peñalvo, 2023). Its implementation in education requires teachers not only to acquire operational skills, but also to develop a critical literacy that enables them to understand how algorithms work, identify their potential biases, and assess their social and educational implications (García Peñalvo et al., 2024; Lane et al., 2023; Luckin, 2017).

A recent and relevant ethical reference is the Safe AI Education Manifesto (Alier Forment et al., 2024), which sets out clear recommendations for the responsible use of AI in educational contexts. Its fundamental principles include transparency in systems, accountability, protection of student privacy, promotion of equity, and fostering critical literacy in the face of algorithmic technologies. These principles reinforce the ethical dimension of the AI-ED-SAT, particularly in items related to bias identification, automated decision-making, and educational impacts, thereby consolidating the instrument as a tool that integrates updated ethical principles relevant to contemporary teaching practice.

Despite this outlook, teacher training in AI remains in its incipient stages. Existing proposals tend to focus on instrumental aspects, without incorporating a comprehensive view that considers both its pedagogical and ethical dimensions (Delgado et al., 2024; Holmes & Tuomi, 2022). Given this gap, it is essential to design valid and reliable instruments that allow for the evaluation not only of teachers' familiarity with AI, but also of their ability to integrate it in a critical, contextualized, and pedagogically meaningful way. To this end, it is necessary to articulate rigorous methodological processes that combine expert judgment with robust statistical analysis, as proposed by Sampieri (2018) and Muñoz and Fonseca-Pedrero (2019).

Teaching skills in AI: beyond digital literacy

The emergence of Artificial Intelligence (AI) in educational settings has consolidated its position as a disruptive technology, with applications ranging from personalized learning to the automation of assessment and administrative tasks (Holmes et al., 2019; Owan et al., 2023). Its implementation poses significant challenges that require teachers to have a deep understanding of its technical foundations and to critically reflect on the associated risks, such as algorithmic biases, opacity in decision-making, technological dependence, and inequalities in access (Celik, 2023; Sharma et al., 2024).

In this context, developing specific AI skills is a strategic priority for education systems. Unlike generic digital skills, which focus on managing digital resources or virtual environments, teaching skills in AI requires a multidimensional understanding that integrates technical, ethical, pedagogical, and curricular aspects. Various conceptual documents have begun to develop this new competency profile, including UNESCO's AI competency frameworks for teachers and students (2024a, 2024b), UNESCO's AI and the Future of Teaching document (2021), the update of the DigCompEdu framework to the AI context (Caena & Redecker, 2019), and the studies by Chen et al. (2020), Chiu et al. (2024), and Ng et al. (2023).

Khreisat et al. (2024) propose a structure based on four interrelated areas:

1. Understanding the principles and functioning of AI.
2. Ability to use AI tools for educational purposes.
3. Critical knowledge of its ethical and social implications.
4. Curriculum integration from an interdisciplinary approach.

This proposal reflects a broader conception of the role of teachers, which transcends the instrumental use of technology. Rather than merely operating tools, teachers must position themselves as reflective and critical agents, capable of evaluating the pedagogical meaning, social consequences, and curricular relevance of AI in their educational practices. As Holmes et al. (2019), Zawacki-Richter et al. (2019), and Ayuso del Puerto and Gutiérrez Esteban (2022) warn, many existing frameworks remain descriptive or eminently technical, which limits their ability to guide truly transformative professional development.

From this perspective, it is necessary to create assessment tools that respond to this complexity. Such tools must extend beyond measuring declarative knowledge to encompass dimensions such as pedagogical competence, critical thinking, and curricular integration. The AI-ED-SAT instrument, developed within the framework of this study, follows this logic, as it is structured around four key dimensions: conceptual understanding, pedagogical application, ethical dimension, and curricular integration. This approach aims to capture the complexity of teacher professional development in the era of AI, acknowledging the need for critical literacy that enables educators to navigate the challenges posed by emerging technologies proactively.

Assessment of teaching skills: development and validation of instruments

In a context of educational transformation driven by emerging technologies, it is crucial to have valid and reliable instruments for diagnosing the level of teaching competence in relation to artificial intelligence. In this sense, self-assessment is a formative and reflective strategy that enables the identification of strengths and areas for improvement from a situated perspective (Panadero et al., 2019; Rachbauer et al., 2025).

Although there are numerous instruments for measuring teacher digital competence (Cabero-Almenara et al., 2020; Ghomi & Redecker, 2019), most focus on generic technical skills, without specifically addressing the pedagogical, ethical, and curricular complexities posed by AI. This limitation has created an urgent need for assessment models that capture not only teachers' technological familiarity but also their ability to integrate AI in a critical, contextualized, and pedagogically meaningful way (Celik, 2023; Delgado et al., 2024).

The construction of instruments with these characteristics requires a rigorous methodological approach that combines theoretical, empirical, and practical criteria. In recent years, a mixed methodology has been established that includes systematic literature review for item development, validation through expert judgment (such as the Delphi method), and psychometric analysis (content and construct validity, internal reliability) (Diefes-Dux et al., 2010; Muñoz & Fonseca-Pedrero, 2019; Sampieri, 2018). In this process, participatory and interactive approaches, such as Design-Based Research (DBR), have proven particularly useful in ensuring the semantic and pedagogical relevance of the instrument.

Self-assessment has also established itself as an effective tool for encouraging professional reflection, providing data for the design of personalized training plans, and promoting teacher agency (Borge et al., 2005). Unlike external assessment models, it places teachers at the center of the diagnosis and improvement process.

The AI-ED-SAT instrument aligns with this framework, as it is designed to assess teaching competencies related to AI in educational contexts comprehensively. Its development is based on a critical review of the specialized literature and recent reference frameworks that address the pedagogical, ethical, and curricular implications of AI (Celik, 2023; Holmes et al., 2019; Zawacki-Richter et al., 2019).

Based on this analysis, four key dimensions were identified that make up a competency profile geared toward a critical and formative appropriation of AI:

1. **Conceptual understanding of AI**
This includes basic knowledge of algorithms, machine learning, and adaptive systems, with a focus on their educational applications. This dimension seeks to position teachers as informed actors in the face of dominant technological discourses (Holmes et al., 2019).
2. **Pedagogical use of AI tools**
This assesses the ability to apply AI tools for clear educational purposes, taking into account the level of the students and the academic context. It encompasses activity design, personalized learning, and automated assessment, extending beyond mere technical use (Chen et al., 2020).
3. **Ethics and critical considerations regarding AI**
This encompasses the identification of risks such as algorithmic biases, privacy issues, and decision automation. This dimension promotes critical literacy, positioning the teacher as an ethical mediator (Celik, 2023).
4. **Curricular integration of AI**
Measures the ability to incorporate AI into teaching and curriculum planning through interdisciplinary projects, active methodologies, and adaptation to new technological scenarios (Rachbauer et al., 2025).

It should be noted that, although some competencies may appear to overlap—particularly between the pedagogical use of AI tools and their curricular integration—the two are distinguished by their level of application: while the second dimension refers to educational use in the classroom (microplanning), the fourth focuses on the strategic incorporation of AI into general curriculum planning (macroplanning), including interdisciplinary approaches and institutional alignment. This differentiation avoids ambiguities and captures different levels of teacher appropriation.

These dimensions are articulated in a complementary manner, enabling a holistic assessment of the teaching role within the digital ecosystem. Unlike fragmented or overly technical models, the AI-ED-SAT proposes an integrated framework where pedagogical action, conceptual knowledge, curricular vision, and ethical commitment are intertwined in the same assessment device.

Figure 1
Diagram of the instrument

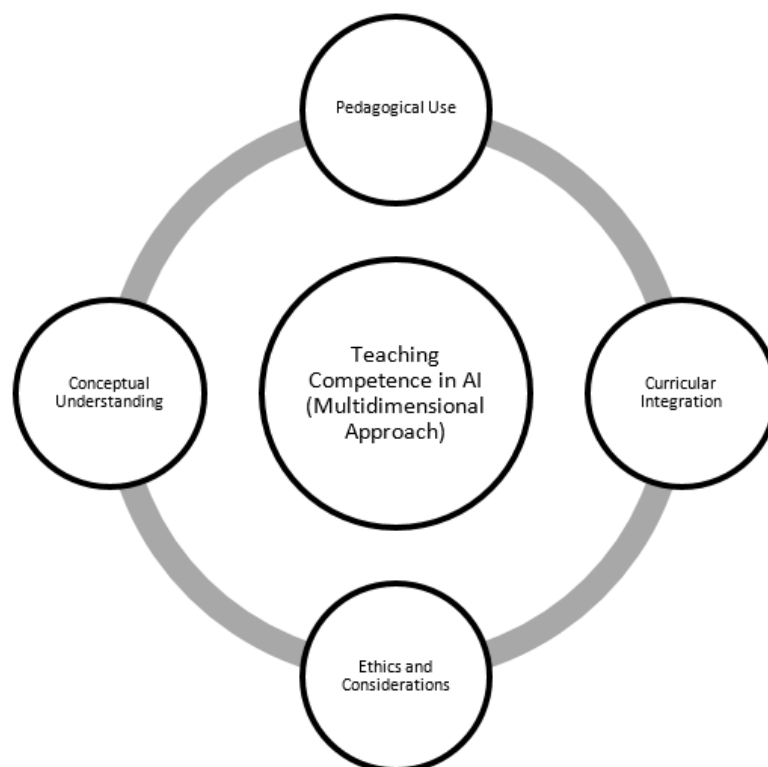


Figure 1 presents a schematic representation of the AI-ED-SAT model, highlighting the interrelationship between its dimensions and its theoretical foundation. The graphic order of the dimensions —Pedagogical Use, Conceptual Understanding, Curriculum Integration, and Ethics and Considerations— does not imply a hierarchy, but rather reflects a circular, non-linear arrangement that emphasizes their interdependence. The dimensions of the model do not follow a predetermined order of acquisition but rather form a multidimensional framework that can be developed in parallel, complementary, or combined ways, depending on the teacher's profile, previous experience, and educational context.

In short, the incorporation of AI in the educational field requires a comprehensive understanding on the part of teachers, combining conceptual mastery with critical, ethical, and pedagogical appropriation. Despite advances in the theoretical definition of these competencies, a lack of specific, valid, and reliable instruments remains, hindering the rigorous and contextualized diagnosis of these competencies.

The AI-ED-SAT is conceived as a response to this need. Its methodological design, based on a combination of qualitative and quantitative approaches, including Design-Based Research, expert judgment, and factor analysis, seeks to ensure the theoretical and psychometric soundness of the instrument, as well as its educational usefulness in improving teaching practice.

METHODOLOGY

Methodological design of the instrument

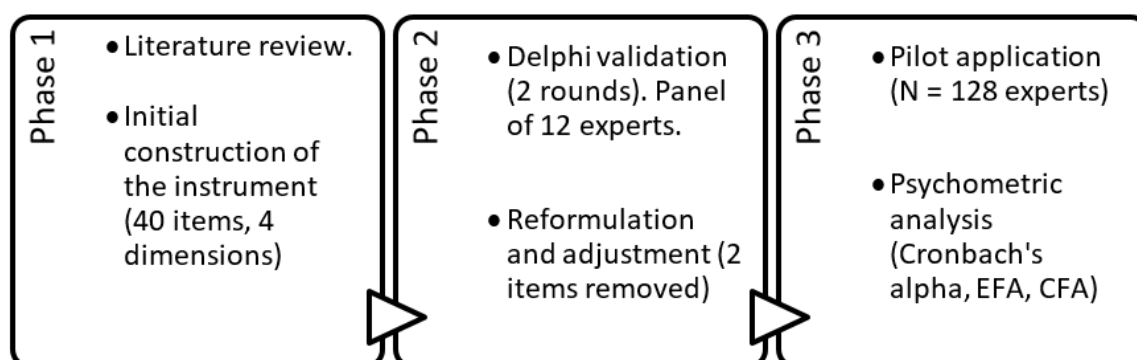
This study is part of a quantitative, descriptive, and correlational approach aimed at designing, validating, and conducting psychometric analysis of a teacher self-assessment instrument on the use of Artificial Intelligence (AI) in education. Methodologically, the principles of Design-Based Research (DBR) (Brown, 1992) were integrated with current guidelines for the validation of educational instruments (Muñiz & Fonseca-Pedrero, 2019; Sampieri, 2018).

The process was structured in three main phases:

1. Construction of the questionnaire, based on a systematic review of recent literature.
2. Content validation through expert judgment, using the Delphi method.
3. Pilot application and psychometric analysis of the instrument.

Figure 2

Stages of the development and validation process



A pilot test was carried out as an intervention, in which the instrument was applied to individuals with characteristics similar to those of the target sample. The purpose of this test was to adjust those items that might require modifications in terms of operability, wording, or other aspects, as well as to optimize the appearance of the instrument (Hernández-Sampieri et al., 2006).

The literature review on teaching competencies in the use of Artificial Intelligence (AI) in education led to the configuration of the first structure of the self-assessment questionnaire around two fundamental dimensions that cover the spectrum of analysis of teaching practice: (1) the pedagogical integration of AI in the classroom and (2) teaching competency in AI and its impact on teaching.

The first draft was constructed using a process based on the design-based research (DBR) methodology, as described by Brown (1992). The process was structured in three phases: in the first, a review of the literature on AI in education and teaching competencies in digital environments was carried out; in the second, the dimensions were defined and the items that would make up the questionnaire were developed; and in the third, the instrument was validated using the Delphi method

with a panel of experts and learning (Chen et al., 2020; Holmes et al., 2019; Luckin, 2017; UNESCO, 2024b).

The instrument, called AI-ED-SAT (Artificial Intelligence in Education – Self-Assessment Tool), was designed to assess the level of teacher competence in four key dimensions:

1. General knowledge of AI and its educational application
2. Pedagogical use of AI tools
3. Ethics and considerations regarding AI in the classroom
4. Curricular integration of AI

Initially, 40 items were developed, written as statements and distributed evenly across the four dimensions (10 items per dimension). Responses were structured on a 5-point Likert scale:

1 = Strongly disagree | 2 = Disagree | 3 = Neither agree nor disagree | 4 = Agree | 5 = Strongly agree

The decision to include a balanced number of 10 items per dimension was based on the criteria of structural symmetry and comparability between dimensions, seeking to ensure a balanced psychometric analysis. Although the nature of the dimensions could have allowed for an extension or reduction of items, an initial equitable distribution was chosen as a starting point for validation, in line with common practices in similar instruments (Muñiz & Fonseca-Pedrero, 2019). Subsequently, the validity analysis enabled the instrument to be refined, resulting in a final version comprising 38 items.

This scale allows for the capture of differentiated levels of self-assessment without overloading decision-making and is suitable for factor analysis. The validation of the instrument sought to ensure that the indicators adequately reflected the integration of AI in education and the teaching skills necessary for its effective implementation.

The final version of the questionnaire, comprising 38 items distributed across four dimensions, is available in full in Appendix 1. The inclusion of the complete instrument provides a detailed overview of the indicators used to assess teaching competence in AI, facilitating its possible replication in subsequent studies or training programs.

Instrument content validation: the Delphi method

The second phase of the study aimed to validate the content of the preliminary draft of the questionnaire that had been previously developed. To this end, the Delphi method was used, widely recognized in the field of Education Sciences for its ability to clarify complex issues through a structured process of consultation among experts organized in a panel (Reguant-Álvarez & Torrado-Fonseca, 2016). This iterative technique fosters informed consensus through successive rounds of evaluation, enabling a progressive convergence of informed opinions (Linstone & Turoff, 2002; Reguant-Álvarez & Torrado-Fonseca, 2016).

The instruments used for data collection consisted of questionnaires designed to assess specific dimensions and items, including sections for qualitative observations.

The collection process was organized with a flexible yet limited timeframe to facilitate the effective participation of a panel of experts, who were distributed geographically and had varying availability.

At this stage, the panel of experts was formed, prioritizing the representativeness and professional diversity of the participants over the absolute number of participants. To this end, selection criteria were defined to ensure a balance between up-to-date knowledge on the integration of Artificial Intelligence (AI) in educational contexts and the teaching skills required for its effective application.

Although the panel included a higher proportion of university educators, this choice was made in response to the need for experts with a systematic view of teacher development and experience in continuing education, which is crucial for assessing cross-cutting skills, such as those related to AI. However, the presence of primary and secondary school professionals with direct experience in applying AI technologies in the classroom was ensured, as well as accredited specialists in artificial intelligence with scientific output and participation in educational innovation projects.

Based on the preliminary structure of the instrument, the central criterion was to include professionals from three key profiles: (1) teachers from different educational levels with experience in the pedagogical use of emerging technologies; (2) specialists in AI applied to education; and (3) teacher trainers in the field of educational technology. This strategy enabled the integration of a multidisciplinary perspective, consistent with the study's objectives.

Although there is no normative consensus on the ideal size of Delphi panels, various authors suggest practical guidelines. Generally, it is considered that a minimum of ten participants ensures the stability of the consensus, as samples of fewer than seven may compromise representativeness (Reguant-Álvarez & Torrado-Fonseca, 2016). In line with this premise, Skulmoski et al. (2007) argue that, for relatively homogeneous panels, a sample of between ten and fifteen experts can be methodologically sound.

Twelve experts who met the established criteria were invited to participate in the study. The invitation included a detailed description of their professional profile, information about their teaching and research roles, and a self-assessment of their suitability for the study's purpose.

All the experts accepted the invitation, forming a panel with a high level of specialization. Seventy-five percent of the group held doctoral degrees, comprising seven university trainers in teacher training programs, three basic education teachers (from primary and secondary education) with direct experience in using AI in the classroom, and two specialists with combined experience in academic research and teacher training in educational AI. Both experts have published indexed works on AI in education, served as advisors on technology integration projects in educational institutions, and collaborate in international networks focused on academic innovation and digital ethics.

In addition, a balanced gender distribution was sought in the composition of the panel, which included seven women (58%) and five men (42%), in line with the representativeness and diversity criteria established in the methodological design.

Common inclusion criteria included:

- A minimum of five years' experience in educational technology, teacher training, or applied research on AI.
- Participation in innovation initiatives or digital skills training projects.

- Representation of different levels of the education system (basic and higher education).

Table 1*Composition of the Delphi expert panel*

Category	Number
Total number of experts	12
With doctoral degrees	9
University trainers	7
Primary and secondary school teachers	3
Specialists in AI and teacher training	2

The researchers in charge supervised data collection during the validation phase, which was conducted through an iterative process of consulting experts, utilizing email as the primary channel of communication. The procedure was structured in two successive rounds, which was considered adequate to facilitate the convergence of opinions and reach a methodologically sound consensus (Linstone & Turoff, 2002). From the outset, participants were informed about this structure and the degree of involvement required, with the aim of ensuring transparency and commitment on the part of the panel (Reguant-Álvarez & Torrado-Fonseca, 2016).

The primary purpose of the first round was to evaluate the quantitative and qualitative aspects of the items in the instrument, with a focus on two fundamental dimensions: conceptual relevance and linguistic clarity. At the quantitative level, the experts were asked to rate each item on a five-point Likert scale, where one represented "not relevant" or "unclear," and five indicated "very relevant" or "very clear." At the same time, space was provided for open comments, which allowed for the collection of qualitative suggestions aimed at improving the wording, accuracy, or appropriateness of the items. This mixed approach, based on previous work (Gallant & Luthy, 2020; Grand-Guillaume-Perrenoud et al., 2023; Paulin et al., 2024; Ramírez, 2019; Ramírez-Montoya & Lugo-Ocando, 2020), enriches the instrument from an interpretive perspective.

Based on the analysis of this first phase, six items were reformulated based on qualitative observations, and two were eliminated because they were considered redundant or conceptually weak. These adjustments led to the design of the form for the second round, which focused on re-evaluating the modified items. Again, a five-point Likert scale was used to assess their relevance and clarity after the modifications. The objective of this second iteration was to confirm the validity of the changes introduced and consolidate a final version that reflected the panel's consensus. The results obtained enabled the process to be closed with a definitive structure comprising 38 items, distributed across four theoretical dimensions.

The results of both rounds were analysed using a combination of quantitative techniques:

- Descriptive statistics: calculation of mean, standard deviation, and percentiles for each item.
- Kendall's coefficient of concordance (W): used to measure the degree of agreement among experts.
- Acceptance criterion based on the 80th percentile: established as the minimum threshold for considering items accepted, according to Mauri et al. (2007).

This analytical approach enabled the accurate identification of items that achieved high levels of consensus, as well as those that required adjustments or elimination.

Pilot study and psychometric analyses

Once the expert judgment validation process was complete, the revised version of the instrument underwent a pilot study with a sample of 128 teachers from various educational contexts in Spain and Latin America. The selection was intentional, based on criteria of accessibility and institutional diversity.

Sample profile:

- Primary education: 34%
- Secondary education: 31%
- Higher education: 35%
- Average teaching experience: 9.6 years
- Gender distribution: 69% women, 31% men

The application was carried out online. The sample consisted of teachers from seven Spanish-speaking countries: Spain ($n = 49$), Mexico ($n = 27$), Argentina ($n = 18$), Colombia ($n = 12$), Chile ($n = 9$), Peru ($n = 7$), and Uruguay ($n = 6$). This distribution was based on criteria of institutional accessibility and geographical diversity, ensuring a balanced representation of different educational contexts. In terms of academic level, the proportion of participants remained similar across all countries, with representation from all three levels: primary, secondary, and tertiary education.

All participants signed a digital informed consent form before accessing the questionnaire. The data collected were processed using IBM SPSS v28 and AMOS v24 software and underwent three types of statistical analysis:

a) Internal consistency

The internal reliability of the questionnaire was estimated using Cronbach's alpha coefficient, calculated individually for each dimension and for the instrument as a whole. This analysis made it possible to evaluate the internal consistency of the items grouped around the different theoretical constructs:

- General knowledge about AI: $\alpha = 0.88$
- Pedagogical use of AI tools: $\alpha = 0.90$
- Ethics and considerations: $\alpha = 0.87$
- Curricular integration: $\alpha = 0.91$
- Total (AI-ED-SAT): $\alpha = 0.93$

These values demonstrate high internal consistency across all dimensions, exceeding the threshold of 0.70 recommended by specialized literature.

In addition to Cronbach's alpha coefficient, McDonald's omega coefficients were calculated, which exceeded the threshold of 0.85 in all cases (total $\omega = 0.94$). This metric, considered a more robust estimate of reliability in scales with multidimensional

structures (McDonald, 1999), reinforces the evidence of the instrument's high internal consistency. Although no temporal reliability analyses (test-retest) or factor invariance tests by subgroups (educational level, gender, or country) were applied in this phase, due to the cross-sectional nature of the design and the sample size, the relevance of these techniques for future validation research with longitudinal designs or larger samples is recognized.

Although the number of participants can be considered adequate in relation to the total number of items in the instrument, various methodological studies (Costello & Osborne, 2005; Lloret-Segura et al., 2014; MacCallum et al., 1999) agree that sample adequacy does not depend solely on absolute size, but on other factors such as item communality, the number of expected factors, and the robustness of statistical indicators. Along the same lines, Hogarty et al. (2005) point out that the quality of factorial solutions is more influenced by the average communality and overdetermination of factors than by the sample size itself. In this case, the high internal consistency of the instrument ($\alpha = 0.93$), the KMO sample adequacy index = 0.91, the significance of Bartlett's sphericity test, and factor loadings greater than 0.60 in all items allow us to consider the sample of 128 teachers as methodologically sound for the factor analysis performed.

b) Exploratory factor analysis (EFA)

In order to identify the underlying structure of the questionnaire, an EFA was performed using the maximum likelihood method and Varimax rotation. The adequacy of the sample was confirmed by:

- KMO index = 0.91
- Bartlett's sphericity test: $\chi^2 (df = 703) = 3982.23, p < .001$

Both indicators showed excellent conditions for the application of factor analysis. Factor loadings were greater than 0.60 for all retained items, and the total variance explained reached 67.8%, distributed evenly among the four proposed dimensions.

In accordance with methodological recommendations and to provide greater transparency in the results, Annex 2 includes the complete matrix of factor loadings, communalities, and specific variances corresponding to the 38 items retained in the exploratory factor analysis. This information enables a detailed examination of the factor structure of the instrument and supports the robustness of the model obtained.

c) Confirmatory factor analysis (CFA)

To verify the empirical adequacy of the theoretical model derived from the EFA, a CFA was performed using a two-factor model, evaluating the quality of the fit using the following indicators:

- $\chi^2/df = 2.34$
- CFI = 0.96
- TLI = 0.95
- RMSEA = 0.045

All the values obtained indicate an excellent fit of the factorial model, confirming the structural validity of the AI-ED-SAT instrument as a tool for teacher self-assessment on the use of Artificial Intelligence in educational settings.

Although it is methodologically recommended to apply EFA and CFA on independent samples, several studies (Brown, 2015; Worthington & Whittaker, 2006; Lloret-Segura et al., 2014) recognize that in applied research contexts with limited samples, it is acceptable to use both procedures on the same sample, provided that the statistical results are robust and the limitation is made explicit. In this study, this strategy was chosen for exploratory and initial validation purposes, recognizing the need for future replication in broader contexts.

RESULTS

Validation by expert judgment (Delphi method) and concordance analysis

During the first round of the Delphi method, the 40 items that made up the questionnaire were evaluated by the panel of experts in relation to two fundamental aspects: conceptual relevance and linguistic clarity. Although most items received high ratings, six were reformulated and two were eliminated because they did not meet the minimum threshold of the 80th percentile in the relevance dimension.

In the second round, the experts reevaluated the adjusted items. The results were analyzed using Kendall's coefficient of agreement (W), whose values showed a statistically significant level of agreement in both dimensions evaluated. This finding supports the existence of consensus among experts, despite the diversity of their profiles and approaches.

Table 2

Expert agreement – Kendall's coefficient

Aspect evaluated	Kendall's W	Chi-square	df	Sig. (p)
Relevance	0.27	84.44	26	< .001
Clarity	0.24	74.53	26	< .001

To ensure content validity, the approach of Mauri et al. (2007) was applied, adapting the cutoff point to the 80th percentile instead of the arithmetic mean. Additionally, a conditional relationship between relevance and clarity was established: the clarity of an item was only evaluated if it had previously been considered relevant. In this way, conceptually weak items were avoided from a formal perspective. The concordance coefficients obtained reflect an adequate consensus, which justifies the incorporation of the reformulated items and the exclusion of those that did not meet the established criteria.

Although Kendall's coefficients are moderate ($W = 0.27$ for relevance and $W = 0.24$ for clarity), these values are consistent with the diverse nature of the expert panel and the complementary qualitative approach of the Delphi method. Previous studies have highlighted that intermediate levels of agreement may be methodologically acceptable in heterogeneous panels, particularly when combined with qualitative analyses that enrich interpretation and facilitate informed adjustments to the items (Grand-Guillaume-Perrenoud et al., 2023; Reguant-Álvarez & Torrado-Fonseca, 2016). Therefore, the magnitude of W reflects a reasonable balance between diversity

of perspectives and sufficient consensus to support the content validity of the instrument.

Internal consistency

Internal reliability analysis was performed on the pilot sample ($N = 128$ teachers) using Cronbach's alpha coefficient, both for the individual dimensions and for the instrument as a whole. All coefficients greatly exceeded the reference value of 0.85, indicating high internal consistency between the items in each dimension and in the questionnaire as a whole:

Table 3
Internal consistency (Cronbach's alpha)

Dimension	Cronbach's Alpha
General knowledge about AI	0.88
Pedagogical use of AI	0.9
Ethics and considerations	0.87
Curricular integration of AI	0.91
Total AI-ED-SAT	0.93

In addition to the internal consistency analysis, descriptive statistics were calculated for the scores obtained in the pilot test. The means per dimension ranged from 3.2 to 4.1 (on a 1-5 Likert scale), indicating a generally positive perception among teachers. The dimension with the highest score was "Pedagogical use of AI" ($M = 4.1$, $SD = 0.6$), followed by "General knowledge about AI" ($M = 3.8$, $SD = 0.7$). The dimensions "Curricular integration" and "Ethics and considerations" had means of 3.5 ($SD = 0.8$) and 3.2 ($SD = 0.9$), respectively, suggesting areas with greater scope for professional development among teachers.

This result supports the internal consistency of the instrument and confirms that the items are grouped logically and consistently within their respective constructs.

Exploratory factor analysis (EFA)

To explore the underlying structure of the questionnaire, exploratory factor analysis (EFA) was performed using the maximum likelihood method with Varimax rotation. The sample adequacy was excellent ($KMO = 0.91$) and Bartlett's sphericity test was significant (χ^2 ($df = 703$) = 3982.23, $p < .001$), which justified the application of the factorial model.

Considering the geographical diversity of the sample, special attention was paid to possible lexical and usage differences between Spain and Latin America. During the design of the items, idiomatic expressions and regional technical terms were avoided, prioritizing clear, neutral, and accessible language. Both the Delphi phase and the pilot test included spaces for linguistic observations, which were analyzed qualitatively and, where necessary, led to the reformulation of potentially ambiguous terms. This care contributed to improving the semantic validity of the instrument and its applicability in diverse Spanish-speaking educational contexts.

The EFA revealed four main factors, in line with the proposed theoretical structure, which together explained 67.8% of the total variance. Factor loadings

remained above 0.60 in all dimensions, confirming an adequate level of item saturation in their respective factors.

During this process, the 40 original items were used. For the presentation in Table 4, a reduced set of 10 representative items (2-3 per dimension) was selected to illustrate the most prominent factor loadings concisely and facilitate readability. The complete matrix, which includes the 38 final items after wording adjustments and statistical refinement, along with their specific communalities and variances, is presented in Appendix 2.

Table 4

Factor loadings of selected items (EFA)

Item	Factor loading (EFA)
Item 1	0.72
Item 2	0.81
Item 3	0.75
Item 4	0.77
Item 5	0.8
Item 6	0.69
Item 7	0.73
Item 8	0.78
Item 9	0.74
Item 10	0.82

These results show that the selected items maintain a clear and consistent structure in each dimension and are representative of the factorial pattern obtained for the instrument as a whole.

Confirmatory factor analysis (CFA)

To empirically confirm the structural validity of the questionnaire, a confirmatory factor analysis (CFA) was conducted on the theoretical model derived from the EFA. The model evaluated was a two-factor model, which allowed for the verification of both the general common variance and the specific variance attributable to each dimension.

The fit indices obtained were:

- $\chi^2/df = 2.34$
- CFI = 0.96
- TLI = 0.95
- RMSEA = 0.045

These values fall within the ranges established as indicators of excellent fit (Hu & Bentler, 1999), confirming that the structure of the AI-ED-SAT instrument adequately reflects the proposed theoretical model. Specifically, the CFI (Comparative Fit Index) and TLI (Tucker-Lewis Index) exceed the threshold of 0.95, while the RMSEA remains below 0.06, indicating the model's validity from a multidimensional perspective.

This analysis provides solid evidence that the questionnaire has a robust factorial structure, capable of accurately capturing the four theoretical components

proposed: knowledge about AI, pedagogical use, ethics and considerations, and curricular integration.

In line with the Delphi method-based methodological process, and with the aim of ensuring the clarity, relevance, and theoretical consistency of the items, the instrument was refined after qualitative and quantitative analysis of the validation rounds. In this process, two items were eliminated because they did not meet the established relevance threshold, and six more were reformulated due to issues of clarity, ambiguity, or suitability for the teaching profile. Table 5 presents a summary of the actions taken on the items during this expert review phase.

Table 5
Items eliminated and reformulated after the Delphi process

Original item	Action taken	Justification	New version (if applicable)
Item 5 – “I understand the mathematical and logical foundations of AI.”	Removed	Low level of relevance for the general teaching profile; excessively technical content; did not exceed the 80th percentile.	—
Item 12 – “I have designed learning activities that incorporate AI.”	Removed	Uncommon level of implementation in the average teaching profile; low relevance as perceived by experts.	—
Item 4 – “I am familiar with terms such as machine learning and neural networks.”	Reformulated	Technical language lacking sufficient clarity; more accessible wording suggested.	I am familiar with general terms related to AI, such as “machine learning” or “neural networks.”
Item 7 – “I can explain to other teachers how basic AI works.”	Reformulated	Ambiguity in the term “basic AI”; need for greater precision.	I can explain the fundamental concepts of AI to other teachers in a simple way.
Item 16 – “I use chatbots or virtual assistants to answer students' questions.”	Reformulated	Technical ambiguity and diversity of interpretation; imprecise wording.	I use tools such as chatbots or virtual assistants to help answer students' questions.
Item 27 – “I identify the social impacts of AI in education.”	Reformulated	Item too broad; specifying the focus is recommended.	I identify how AI can affect access, equity, and social interaction in educational contexts.
Item 33 – “I am familiar with curriculum frameworks that incorporate AI in education.”	Reformulated	Ambiguity about the availability of frameworks; context not easily generalizable.	I am familiar with examples of curriculum proposals or policies that promote the inclusion of AI in education.
Item 36 – “I use AI-based simulations to reinforce concepts in the classroom.”	Reformulated	Unclear terminology; more practical and concrete wording is suggested.	I use AI-supported simulations or interactive tools to reinforce student learning.

DISCUSSION

The results obtained in the validation of the AI-ED-SAT instrument confirm the conceptual and psychometric soundness of its design, allowing it to be considered a robust tool for assessing teacher competence in the use of Artificial Intelligence (AI) in educational settings. In line with the reference frameworks proposed by international organizations such as UNESCO (2021) and the OECD (2022), the questionnaire presents a comprehensive, up-to-date, and contextualized approach to the training requirements of teachers in response to the rise of AI in education.

The high internal consistency obtained in all dimensions (Cronbach's α ranging from 0.87 to 0.93) supports the internal coherence of the items around well-defined constructs. Likewise, both exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) provided clear empirical evidence of the structural validity of the proposed model, with fit indices (CFI = 0.96; RMSEA = 0.045) that far exceed the minimum standards required in the psychometric literature (Hu & Bentler, 1999). These findings support the conclusion that the AI-ED-SAT assesses differentiated and complementary dimensions of teaching knowledge and practice in relation to AI.

Beyond psychometric indicators, the results underscore the relevance of the AI-ED-SAT as a diagnostic and formative tool in contexts where teacher literacy in AI is becoming increasingly strategic. As Celik (2023) warns, the development of teaching competencies in AI must go beyond the technical management of tools, also encompassing the capacity for critical analysis, meaningful pedagogical design, and curricular adaptation to new technological scenarios. The instrument addresses this need by integrating four key dimensions: conceptual understanding, pedagogical use, an ethical approach, and curricular integration. This segmentation is not only consistent with previous studies (Chen et al., 2020; Zawacki-Richter et al., 2019) but also allows for differentiated and specific feedback, which is highly valuable for guiding teacher professional development processes.

Potential of AI-ED-SAT for university teacher training in digital and hybrid environments

The growing incorporation of digital and hybrid modalities in higher education requires university teacher training to integrate specific skills for teaching mediated by emerging technologies, including Artificial Intelligence. AI-ED-SAT, with its four-dimensional structure and self-reflective nature, offers an operational framework for diagnosing and developing these skills in diverse academic contexts. In digital environments, its results can guide the selection of AI-based resources, strategies, and pedagogical approaches, promoting the adaptation of content and methodologies to virtuality. In hybrid modalities, the instrument facilitates the identification of practices that coherently integrate the potential of AI in both physical classrooms and virtual environments, promoting pedagogical coherence and digital inclusion. Its application in university professional development programs also enables the design of personalized, evidence-based training itineraries, enhancing adaptive learning and the continuous updating of teachers in response to technological advances.

One of the distinctive strengths of the tool lies in its self-assessment format, which encourages personal reflection and positions teachers as active protagonists in their continuing education. Unlike assessment models focused on generic digital competencies, the AI-ED-SAT focuses on a specific technology—AI—and its

contextualized pedagogical use, which represents an innovative and necessary contribution in the face of the rapid integration of AI systems into educational platforms, virtual learning environments, and automated assessment devices (Baltazar, 2023; Owan et al., 2023).

From a methodological perspective, the combination of the Design-Based Research (DBR) approach with the Delphi technique provided rigor, flexibility, and contextual sensitivity in the development of the instrument. This approach enabled iterative and consensual development, articulating the theoretical foundation through practical validation by a panel of experts. Previous research has supported the effectiveness of this methodological combination in developing complex educational instruments (Diefes-Dux et al., 2010; Rachbauer et al., 2025), especially when a balance between technical soundness and pedagogical appropriateness is required.

These findings are consistent with previous studies that have identified greater teacher proficiency in the technical use of digital tools compared to more complex ethical or curricular dimensions (Celik, 2023; Chen et al., 2020). Compared to similar instruments focused on general digital competencies, the AI-ED-SAT presents slightly higher means on items related to the specific pedagogical application of AI. This difference can be attributed to the instrument's contextualized approach, which is designed specifically to capture teachers' perceptions of the use of artificial intelligence in real educational contexts.

Despite the promising results, it is essential to acknowledge certain methodological limitations that affect the interpretation of the findings. First, the pilot sample was selected for convenience, which introduces potential bias and limits the generalization of the results to other educational contexts. Second, the instrument was validated exclusively in Spanish-speaking countries, which restricts its cross-cultural applicability. These limitations should be taken into account when evaluating the validity and utility of the questionnaire in other educational settings.

In this sense, the results should be understood as a preliminary validation of the instrument. It is recommended that the study be replicated in larger and more diverse samples, particularly in higher education contexts and virtual learning environments, to strengthen the generalizability of the findings and refine the instrument's sensitivity to different educational realities.

CONCLUSIONS

The development and validation of the AI-ED-SAT questionnaire represent a significant contribution to the field of educational innovation and teacher training, offering a rigorous, relevant, and up-to-date tool for self-assessing professional competencies related to the use of Artificial Intelligence (AI) in school contexts. Its design is based on a robust theoretical approach, a participatory construction informed by expert consensus, and a comprehensive psychometric validation process, ensuring its usefulness for both research and training improvement purposes.

Among the main contributions of the instrument are:

- Its structure in four key dimensions—conceptual understanding, pedagogical application, ethical approach, and curricular integration—comprehensively reflects the current challenges facing teaching practice in the face of advances in AI.

- Its formative and self-reflective nature allows teachers to identify strengths, areas for improvement, and professional development needs autonomously and in context.
- Its versatility for use in various training contexts, both face-to-face and virtual, makes it an adaptable tool for initial and continuing training programs, with the potential to personalize training itineraries.

In addition to its immediate contributions, the AI-ED-SAT opens up various lines of future research and development, among which the following are recommended:

- Expand its empirical validation with larger, more representative, and culturally diverse samples.
- Perform factorial invariance analyses to verify the stability of its structure at different educational levels, geographic regions, or groups of teachers with heterogeneous profiles.
- Develop linguistically and culturally adapted versions for non-Spanish-speaking contexts, considering both technical translation and semantic adaptation.
- Integrate the questionnaire into digital teacher training platforms, where the results can be linked to adaptive learning paths, promoting automated formative feedback.

In a global scenario where AI is redefining not only teaching and learning methods but also the ethical, social, and political frameworks that underpin educational decision-making, it is becoming increasingly urgent to provide teachers with tools that facilitate their critical literacy, professional autonomy, and transformative agency. In this sense, the AI-ED-SAT aspires to be more than just a measurement tool; it seeks to become a catalyst for reflection, empowerment, and ethical commitment among teachers in the face of the technological challenges of the 21st century.

REFERENCES

- Alíer Forment, M., García Peñalvo, F. J., Casañ Guerrero, M. J., Pereira, J. A., & Llorens-Largo, F. (2024, 8 October). *Safe AI in Education Manifesto* (Version 0.4.0). Safe AI in Education Manifesto. <https://manifesto.safeaieducation.org>
- Ayuso del Puerto, D., & Gutiérrez Esteban, P. (2022). La inteligencia artificial como recurso educativo durante la formación inicial del profesorado. *RIED-Revista Iberoamericana de Educación a Distancia*, 25(2), 347-362. <https://doi.org/10.5944/ried.25.2.32332>
- Baltazar, C. (2023). Herramientas de IA aplicables a la educación. *Technology Rain Journal*, 2(2), e15-e15. <https://doi.org/10.55204/trj.v2i2.e15>
- Bo, N. S. W. (2024). OECD digital education outlook 2023: Towards an effective education ecosystem. *Hungarian Educational Research Journal*, 15(2), 284-289. <https://doi.org/10.1556/063.2024.00340>
- Borge, R., García, J., Oliver, R., & Salomón, L. (2005). *Competencias y diseño de la evaluación continua y final en el Espacio Europeo de Educación Superior*. Dirección General de Universidades, MEC.
- Brown, A. L. (1992). Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings.

- Journal of the Learning Sciences*, 2(2), 141-178.
https://doi.org/10.1207/s15327809jls02_02_2
- Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). The Guilford Press.
- Cabero-Almenara, J., Barroso-Osuna, J., Palacios-Rodríguez, A., & Llorente-Cejudo, C. (2020). Marcos de competencias digitales para docentes universitarios: Su evaluación a través del coeficiente competencia experta. *Revista Electrónica Interuniversitaria de Formación del Profesorado*, 23(3), 17-34.
<https://doi.org/10.6018/reifop.414501>
- Caena, F., & Redecker, C. (2019). Aligning teacher competence frameworks to 21st century challenges: The case for the European Digital Competence Framework for Educators (DigCompEdu). *European Journal of Education*, 54(3), 356-369.
<https://doi.org/10.1111/ejed.12345>
- Celik, I. (2023). Towards Intelligent-TPACK: An empirical study on teachers' professional knowledge to ethically integrate artificial intelligence (AI)-based tools into education. *Computers in Human Behavior*, 138, 107468.
<https://doi.org/10.1016/j.chb.2022.107468>
- Chen, L., Chen, P., & Lin, Z. (2020). Artificial intelligence in education: A review. *IEEE Access*, 8, 75264-75278.
<https://doi.org/10.1109/ACCESS.2020.2988510>
- Chesterman, S. (2021). Through a glass, darkly: Artificial intelligence and the problem of opacity. *The American Journal of Comparative Law*, 69(2), 271-294.
<https://doi.org/10.1093/ajcl/avabo12>
- Chiu, T. K., Ahmad, Z., Ismailov, M., & Sanusi, I. T. (2024). What are artificial intelligence literacy and competency? A comprehensive framework to support them. *Computers and Education Open*, 6, 100171.
<https://doi.org/10.1016/j.caeo.2024.100171>
- Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research, and Evaluation*, 10(1), 1-9.
<https://doi.org/10.7275/jvj1-4868>
- Delgado, N., Carrasco, L. C., de la Maza, M. S., & Etxabe-Urbiet, J. M. (2024). Aplicación de la inteligencia artificial (IA) en educación: Los beneficios y limitaciones de la IA percibidos por el profesorado de educación primaria, educación secundaria y educación superior. *Revista Electrónica Interuniversitaria de Formación del Profesorado*, 27(1), 207-224.
<https://doi.org/10.6018/reifop.577211>
- Diefes-Dux, H. A., Zawojewski, J. S., & Hjalmarson, M. A. (2010). Using educational research in the design of evaluation tools for open-ended problems. *International Journal of Engineering Education*, 26(4), 807-819.
- Ferrante, E. (2021). Inteligencia artificial y sesgos algorítmicos: ¿Por qué deberían importarnos? *Nueva Sociedad*, (294), 27-36.
<https://nuso.org/articulo/inteligencia-artificial-y-sesgos-algoritmicos/>
- Flores Vivar, J. M., & García Peñalvo, F. J. (2023). Reflexiones sobre la ética, potencialidades y retos de la inteligencia artificial en el marco de la Educación de Calidad (ODS4). *Comunicar: Revista Científica de Comunicación y Educación*, (74), 37-47.
<https://doi.org/10.3916/C74-2023-03>
- Gallant, D. J., & Luthy, N. (2020). Mixed methods research in designing an instrument for consumer-oriented evaluation. *Journal of MultiDisciplinary Evaluation*, 16(34), 21-43.
<https://doi.org/10.56645/jmde.v16i34.583>
- García Peñalvo, F. J., Llorens-Largo, F., & Vidal, J. (2024). La nueva realidad de la educación ante los avances de la inteligencia artificial generativa. *RIED-Revista Iberoamericana de Educación a Distancia*, 27(1), 9-39.
<https://doi.org/10.5944/ried.27.1.37716>
- Ghomi, M., & Redecker, C. (2019, May). Digital competence of educators (DigCompEdu): Development and evaluation of a self-assessment

- instrument for teachers' digital competence. In *Proceedings of the 11th International Conference on Computer Supported Education (CSEDU 2019)* (Vol. 1, pp. 541-548). SCITEPRESS. <https://doi.org/10.5220/0007679005410548>
- Grand-Guillaume-Perrenoud, J. A., Geese, F., Uhlmann, K., Blasimann, A., Wagner, F. L., Neubauer, F. B., Huwendiek, S., Hahn, S., & Schmitt, K.-U. (2023). Mixed methods instrument validation: Evaluation procedures for practitioners developed from the validation of the Swiss Instrument for Evaluating Interprofessional Collaboration. *BMC Health Services Research*, 23, 83. <https://doi.org/10.1186/s12913-023-09040-3>
- Hernández-Sampieri, R., Fernández-Collado, C., & Baptista-Lucio, P. (2006). *Metodología de la investigación* (4th ed.). McGraw-Hill Interamericana.
- Hogarty, K. Y., Hines, C. V., Kromrey, J. D., Ferron, J. M., & Mumford, K. R. (2005). The quality of factor solutions in exploratory factor analysis: The influence of sample size, communalities, and overdetermination. *Educational and Psychological Measurement*, 65(2), 202-226. <https://doi.org/10.1177/0013164404267287>
- Holmes, W., Bialik, M., & Fadel, C. (2019). *Artificial intelligence in education: Promises and implications for teaching and learning*. Center for Curriculum Redesign.
- Holmes, W., & Tuomi, I. (2022). State of the art and practice in AI in education. *European Journal of Education*, 57(4), 542-570. <https://doi.org/10.1111/ejed.12533>
- Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1-55. <https://doi.org/10.1080/10705519909540118>
- Huang, L. (2023). Ethics of artificial intelligence in education: Student privacy and data protection. *Science Insights Education Frontiers*, 16(2), 2577-2587. <https://doi.org/10.15354/sief.23.re202>
- Khreisat, M. N., Khilani, D., Rusho, M. A., Karkkulainen, E. A., Tabuena, A. C., & Uberas, A. D. (2024). Ethical implications of AI integration in educational decision making: Systematic review. *Educational Administration: Theory and Practice*, 30(5), 8521-8527. <https://doi.org/10.53555/kuey.v30i5.4406>
- Lane, M., Williams, M., & Broecke, S. (2023). The impact of AI on the workplace: Main findings from the OECD AI surveys of employers and workers (OECD Social, Employment and Migration Working Papers, No. 288). OECD Publishing. <https://doi.org/10.1787/ea0a0fe1-en>
- Linstone, H. A., & Turoff, M. (Eds.). (2002). *The Delphi method: Techniques and applications*. Addison-Wesley.
- Lloret-Segura, S., Ferreres-Traver, A., Hernández-Baeza, A., & Tomás-Marco, I. (2014). El análisis factorial exploratorio de los ítems: Una guía práctica, revisada y actualizada. *Anales de Psicología*, 30(3), 1151-1169. <https://doi.org/10.6018/analesps.30.3.199361>
- Luckin, R. (2017). Towards artificial intelligence-based assessment systems. *Nature Human Behaviour*, 1(3), 0028. <https://doi.org/10.1038/s41562-016-0028>
- MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, 4(1), 84-99. <https://doi.org/10.1037/1082-989X.4.1.84>
- Mauri, T., Coll, C., & Onrubia, J. (2007). La evaluación de la calidad de los procesos de innovación docente universitaria: Una perspectiva constructivista. *Revista de Docencia Universitaria*, 5(1). <https://doi.org/10.4995/redu.2007.6290>
- McDonald, R. P. (1999). *Test theory: A unified treatment*. Lawrence Erlbaum Associates. <https://doi.org/10.4324/9781410601087>
- Muñiz, J., & Fonseca-Pedrero, E. (2019). Diez pasos para la construcción de un test. *Psicothema*, 31(1), 7-16.

- <https://doi.org/10.7334/psicothema2018.291>
- Ng, D. T. K., Leung, J. K. L., Su, J., Ng, R. C. W., & Chu, S. K. W. (2023). Teachers' AI digital competencies and twenty-first century skills in the post-pandemic world. *Educational Technology Research and Development*, 71(1), 137-161. <https://doi.org/10.1007/s11423-023-10203-6>
- OECD. (2022). *Trends shaping education 2022*. OECD Publishing. <https://doi.org/10.1787/6ae8771a-en>
- Owan, V. J., Abang, K. B., Idika, D. O., Etta, E. O., & Bassey, B. A. (2023). Exploring the potential of artificial intelligence tools in educational measurement and assessment. *Eurasia Journal of Mathematics, Science and Technology Education*, 19(8), em2307. <https://doi.org/10.29333/ejmste/13428>
- Panadero, E., Broadbent, J., Boud, D., & Lodge, J. M. (2019). Using formative assessment to influence self- and co-regulated learning: The role of evaluative judgement. *European Journal of Psychology of Education*, 34, 535-557. <https://doi.org/10.1007/s10212-018-0407-8>
- Paulin, A. M., Barriga-Arceo, F. D., Mendiola, M. S., & González, A. M. (2024). Evidencias de validez de un instrumento para evaluar la competencia digital docente en educación médica. *Investigación en Educación Médica*, 13(51), 82-92. <https://doi.org/10.22201/fm.20075057e.2024.51.23584>
- Rachbauer, T., Graup, J., & Rutter, E. (2025). Digital literacy and artificial intelligence literacy in teacher training. *Forum for Education Studies*, 3(1), 1842. <https://doi.org/10.59400/fesi1842>
- Ramírez, J. L. M. (2019). El proceso de elaboración y validación de un instrumento de medición documental. *Acción y Reflexión Educativa*, 44, 50-63. https://revistas.up.ac.pa/index.php/accion_reflexion_educativa/article/view/673
- Ramírez-Montoya, M. S., & Lugo-Ocando, J. (2020). Revisión sistemática de métodos mixtos en el marco de la innovación educativa. *Comunicar: Revista Científica de Comunicación y Educación*, 28(65), 9-20. <https://doi.org/10.3916/C65-2020-01>
- Reguant-Álvarez, M., & Torrado-Fonseca, M. (2016). El método Delphi. *REIRE. Revista d'Innovació i Recerca en Educació*, 9(1), 87-102. <https://doi.org/10.1344/reire2016.9.1916>
- Sampieri, R. H. (2018). *Metodología de la investigación: Las rutas cuantitativa, cualitativa y mixta*. McGraw-Hill México.
- Selwyn, N. (2019). *Should robots replace teachers? AI and the future of education*. John Wiley & Sons.
- Sharma, D. M., Ramana, K. V., Jothilakshmi, R., Verma, R., Maheswari, B. U., & Boopathi, S. (2024). Integrating generative AI into K-12 curriculums and pedagogies in India: Opportunities and challenges. In M. S. P. Subathra & G. V. S. N. R. V. Prasad (Eds.), *Facilitating global collaboration and knowledge sharing in higher education with generative AI* (pp. 133-161). IGI Global. <https://doi.org/10.4018/979-8-3693-0487-7.ch006>
- Skulmoski, G. J., Hartman, F. T., & Krahn, J. (2007). The Delphi method for graduate research. *Journal of Information Technology Education: Research*, 6(1), 1-21. <https://doi.org/10.28945/199>
- UNESCO. (2021). *Artificial intelligence and the futures of learning: Towards an ethical and inclusive approach*. UNESCO. <https://www.unesco.org/en/digital-education/ai-future-learning>
- UNESCO. (2023). *Generative AI and the future of education*. UNESCO. <https://doi.org/10.54675/HOYG8740>
- UNESCO. (2024a). *AI competency framework for students*. UNESCO. <https://doi.org/10.54675/JKJB9835>
- UNESCO. (2024b). *AI competency framework for teachers*. UNESCO. <https://doi.org/10.54675/ZJTE2084>
- Worthington, R. L., & Whittaker, T. A. (2006). Scale development research: A content analysis and recommendations for best practices. *The Counseling Psychologist*, 34(6), 806-838.

<https://doi.org/10.1177/0011000006288127>

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher

education. *International Journal of Educational Technology in Higher Education*, 16(1), 1-27. <https://doi.org/10.1186/s41239-019-0171-0>

APPENDIX 1

AI-ED-SAT Tool: First Version

Dimension 1: General knowledge of AI and its educational application

1. I am familiar with the basic concepts of AI and its applications in different sectors.
2. I understand how AI can influence education and learning.
3. I can identify AI tools that are applicable in the field of education.
4. I am familiar with terms such as machine learning and neural networks.
5. I understand the mathematical and logical foundations of AI.
6. I can distinguish between different types of AI and their applications in education.
7. I can explain to other teachers how basic AI works.
8. I know examples of AI used in learning platforms.
9. I can identify the advantages and limitations of AI in the educational context.
10. I have researched current trends in AI applied to education.

Dimension 2: Pedagogical use of AI tools

11. I use AI tools to improve my teaching practice.
12. I have designed learning activities that incorporate AI.
13. I am familiar with AI platforms that can support personalized teaching.
14. I evaluate the impact of AI tools on my students' learning.
15. I apply AI to the generation of teaching materials.
16. I use chatbots or virtual assistants to answer students' questions.
17. I implement AI for automatic task assessment.
18. I explore AI tools that promote adaptive learning.
19. I develop strategies to integrate AI into collaborative activities.
20. I participate in training on new AI tools in education.

Dimension 3: Ethics and considerations regarding AI in the classroom

21. I am aware of potential biases in AI systems.
22. I promote the ethical and responsible use of AI in the classroom.
23. I teach my students about privacy and security in the use of AI.
24. I reflect on the ethical challenges of automation in education.
25. I analyze how AI can influence decision-making.
26. I discuss with students the issues of transparency in AI algorithms.
27. I identify the social impacts of AI in education.
28. I propose practices to ensure the equitable use of AI.
29. I research AI regulations and standards in education.
30. I facilitate spaces for reflection on the impact of AI on society.

Dimension 4: Curricular integration of AI

31. I have integrated AI-related activities into my lesson plans.
32. I design teaching strategies that include AI as a learning tool.
33. I am familiar with curriculum frameworks that incorporate AI into education.

34. I participate in training on AI applied to teaching.
35. I develop interdisciplinary projects that include AI.
36. I use AI-based simulations to reinforce concepts in the classroom.
37. I coordinate activities with other teachers to integrate AI into the curriculum.
38. I research active methodologies that use AI in the classroom.
39. I design complete teaching units focused on AI.
40. I promote AI literacy among my students.

APPENDIX 2

Complete matrix of factor loadings, communalities (h^2), and specific variances for the AI-ED-SAT (EFA).

Item	Factor loading	Communal (h^2)	Specific variance ($1 - h^2$)
Item 1	0.72	0.52	0.48
Item 2	0.81	0.66	0.34
Item 3	0.75	0.56	0.44
Item 4	0.77	0.59	0.41
Item 5	0.80	0.64	0.36
Item 6	0.69	0.48	0.52
Item 7	0.73	0.53	0.47
Item 8	0.78	0.61	0.39
Item 9	0.74	0.55	0.45
Item 10	0.82	0.67	0.33
Item 11	0.76	0.58	0.42
Item 12	0.71	0.50	0.50
Item 13	0.79	0.62	0.38
Item 14	0.70	0.49	0.51
Item 15	0.73	0.53	0.47
Item 16	0.68	0.46	0.54
Item 17	0.75	0.56	0.44
Item 18	0.77	0.59	0.41
Item 19	0.72	0.52	0.48
Item 20	0.69	0.48	0.52
Item 21	0.70	0.49	0.51
Item 22	0.76	0.58	0.42
Item 23	0.74	0.55	0.45
Item 24	0.79	0.62	0.38
Item 25	0.71	0.50	0.50
Item 26	0.73	0.53	0.47
Item 27	0.69	0.48	0.52
Item 28	0.77	0.59	0.41
Item 29	0.75	0.56	0.44
Item 30	0.80	0.64	0.36
Item 31	0.78	0.61	0.39
Item 32	0.73	0.53	0.47
Item 33	0.76	0.58	0.42
Item 34	0.74	0.55	0.45
Item 35	0.72	0.52	0.48
Item 36	0.70	0.49	0.51
Item 37	0.69	0.48	0.52
Item 38	0.81	0.66	0.34

Note: The values follow the reported psychometric pattern, with high loadings and solid explained variance (67.8%).

Date of reception: 1 June 2025

Date of acceptance: 21 August 2025

Date of approval for layout: 23 September 2025

Date of publication in OnlineFirst: 15 October 2025

Date of publication: 1 January 2026

Self-perception and usefulness of generative artificial intelligence among pre-service teachers

Autopercepción y utilidad de la inteligencia artificial generativa en docentes en formación



 Ana María Pinto-Llorente - *Universidad de Salamanca, USAL (Spain)*

 Vanessa Izquierdo-Álvarez - *Universidad de Salamanca, USAL (Spain)*

 Marta M. Dolcet-Negre - *Universidad de Salamanca, USAL (Spain)*

ABSTRACT

The advent of Generative Artificial Intelligence (GenAI) in education presents opportunities, but it also raises ethical and pedagogical challenges. In this context, it is imperative to comprehend how pre-service teachers perceive this technology. The present study analysed the self-perception of 174 pre-service teachers regarding the application of GenAI in education. Seven dimensions (Familiarity, Relevance, Practical Skills, Barriers, Confidence, Ethical-Social Impact, and Expectations) were measured in relation to GenAI. In addition, the usefulness of ChatGPT as a tool for designing Learning Situations (LSs) was assessed after a training experience with this system. Descriptive statistics and Spearman correlations were calculated, and a network of correlations between the seven dimensions was visualised. Differences between degrees were also explored. The findings indicated medium-to-high levels of self-perception, suggesting a very positive evaluation of ChatGPT's usefulness and a high level of satisfaction with its use. Confidence emerged as a central node in the correlation network, exhibiting close associations with Relevance, Barriers, Ethical-social impact, and Expectations. This underscores its pivotal role in the adoption of these technologies. Similarly, most participants adopted a critical stance towards GenAI, checking the responses generated by ChatGPT rather than passively accepting them. In conclusion, while there is a favourable attitude towards integrating GenAI into education, future teachers demand specific training to use it pedagogically and express concern about the ethical implications of such integration.

Keywords: generative artificial intelligence; ChatGPT; teacher training; self-perception; usefulness; education.

RESUMEN

La irrupción de la Inteligencia Artificial Generativa (IA-Gen) en el ámbito educativo ofrece oportunidades, pero también plantea desafíos éticos y pedagógicos. En este contexto, resulta fundamental comprender la percepción de los docentes en formación hacia esta tecnología. Este estudio analizó la autopercepción de 174 docentes en formación sobre la IA-Gen aplicada a la educación. Se midieron siete dimensiones (Familiaridad, Relevancia, Habilidades prácticas, Barreras, Confianza, Impacto ético-social y Expectativas) en referencia a la IA-Gen y se valoró la utilidad de ChatGPT como herramienta para diseñar Situaciones de Aprendizaje (SdAs) tras una experiencia formativa con este sistema. Se calcularon estadísticos descriptivos, correlaciones de Spearman, se visualizó una red de correlaciones entre las siete dimensiones y se exploraron diferencias entre las titulaciones. Los resultados revelan niveles medios-altos de autopercepción con valoración muy positiva de la utilidad de ChatGPT y un alto nivel de satisfacción con su uso. La Confianza emergió como un nodo central en la red de correlaciones, vinculándose estrechamente con la Relevancia, Barreras, Impacto ético-social y Expectativas, lo que resalta su papel clave en la adopción de estas tecnologías. Asimismo, la mayoría de los participantes adoptó una actitud crítica ante la IA-Gen, contrastando las respuestas generadas por ChatGPT en lugar de aceptarlas pasivamente. En conclusión, aunque se observa una disposición favorable hacia la integración de la IA-Gen en educación, los futuros docentes demandan formación específica para su uso pedagógico y expresan preocupación por las implicaciones éticas de dicha integración.

Palabras clave: inteligencia artificial generativa; ChatGPT; formación docente; autopercepción; utilidad; educación.

INTRODUCTION

Since the launch of ChatGPT, Generative Artificial Intelligence (GenAI) has undergone rapid development, leading to a notable increase in scientific output across various disciplines, including health (Currie, 2025; Tai-Han et al., 2024; Denny et al., 2024; Gozalo-Brizuela & Garrido-Merchán, 2024; Storey, 2025), computer science (Gozalo-Brizuela & Garrido-Merchán, 2024; Storey, 2025), or education (Carranza Alcántar, 2024; Haroud & Saqri, 2025; Sánchez-Prieto et al., 2025), demonstrating its cross-cutting impact in multiple fields of knowledge.

In education, GenAI has emerged as a disruptive technology capable of redefining teaching and learning processes. Its gradual incorporation into various contexts has fostered applications such as personalised learning experiences allowing the adaptation of content, activities, and explanations to the level and cognitive development of students, promoting more active, autonomous, and contextualised learning (Carranza Alcántar et al., 2024). Furthermore, it is employed in the design of teaching materials, lesson planning, and task diversification, optimizing time, augmenting the resources available (Bayly-Castaneda et al., 2024) and contributing to educational innovation and Universal Design for Learning (Alba Pastor, 2022). Likewise, its use in academic writing is also noteworthy (Moorhouse & Kohnke, 2024) as well as the provision of automated feedback on student performance, promoting more autonomous and reflective learning (Haroud & Saqri, 2025). Finally, it is noteworthy its applicability in developing digital and metacognitive competences, which are essential to prepare students for a work environment in which GenAI already has a significant presence (Islam & Greenwood, 2024).

The emergence of GenAI in education has opened up important debates on academic integrity and the possible delegation of cognitive functions. In response, several studies propose ethical and pedagogical frameworks that promote a formative and responsible use (Haroud & Saqri, 2025). Considering this challenge, the role of teachers and their AI competences are crucial to maximize its educational potential and mitigate its risks (Alasadi & Baiz, 2023; Nyaaba & Zhai, 2024; Zhang & Villanueva, 2023). In this regard, the AI competency framework for teachers (AI CFT) (Miao & Cukurova, 2024) is particularly relevant, as it aims to safeguard teachers' rights and foster their continuing professional development in the era of AI. This model organises AI competences into five key dimensions: human-centred mindset, Ethics of AI, AI foundations and applications, AI pedagogy, and AI for professional development. Its goal is to ensure that teachers are prepared to use AI safely, effectively, and ethically, while minimising potential risks to students and society.

The current challenge lies in confronting a technology capable of mimicking human behaviour even to the point of questioning human agency (Mouta et al., 2025). Through the analysis of large amounts of data, AI can replicate, or even replace, human decision-making, including teachers' professional autonomy. This reality demands solid preparation and continuous support for teachers to ensure proper educational use. Consequently, it is essential to incorporate AI training into pre-service teacher education curricula (Ishmuradova et al., 2025), providing them with theoretical foundations and practical competences applicable in the classroom, and preparing them to face current and future educational challenges (Laupichler et al., 2022). Educational policies, therefore, play a key role in this transformation (Al-Abdullatif & Alsubaie, 2024).

Recent research highlights the growing interest in understanding how future teachers perceive and integrate GenAI into their training. Studies concur in emphasising its potential as a support tool for the design of teaching proposals and materials (Ishmuradova et al., 2025; Lozano & Blanco, 2023; Markos et al., 2024), and underline benefits associated with the development of professional competences, conceptual understanding, pedagogical adaptation, and critical awareness of its educational use (Moorhouse et al., 2024). However, common concerns also emerge regarding the reliability of generated information, potential technological dependence, the impact on basic cognitive skills, and ethical risks related to plagiarism or academic integrity (Nyaaba & Zhai, 2024), underscoring the need for critical, ethical, and pedagogically grounded training. Despite these advances, previous studies still show relevant limitations, as descriptive or exploratory approaches predominately analyse individual perceptions, with little attention to the relational structure among the different dimensions involved in the pedagogical integration of GenAI. In this regard, the present study seeks to overcome this fragmented logic by adopting a structural modelling approach through network analysis, enabling the modelling of interrelationships among key variables such as familiarity, relevance, barriers, confidence, and ethical impact, thereby providing a more comprehensive and structured understanding of the phenomenon.

Research emphasises the need to conduct studies in different educational institutions and cultural contexts that explore future teachers' perceptions of their knowledge of GenAI in order to gain a better understanding of the role of this technology (Markos et al., 2024). The present study focuses on analysing pre-service teachers' self-perceptions in relation to their Familiarity with GenAI, its educational Relevance, the associated Practical Skills, the perceived Barriers, their level of Confidence, the Ethical and Social Impact, and their Usage Expectations. Self-perception is understood as the assessment and understanding that one has of oneself in a specific domain (Aravena Castillo, 2013). Here it refers to the evaluation that pre-service teachers make of their familiarity, competences, beliefs, and expectations regarding the use of GenAI in education. This approach allows representing the interrelationships among self-perception dimensions, offering an integrated and novel perspective that identifies which factors occupy central positions and how they are clustered, adding value to the state of the art. Within this framework, the study presents a pilot experience based on the application of ChatGPT in the design of Learning Situations (LSs) to assess its usefulness.

The article is organised into five sections. The first outlines the theoretical foundation of the study. The Method section then describes the approach, participants, instrument, and the procedures for data collection and analysis. The results are subsequently presented, followed by a discussion of the findings. Finally, the article concludes with recommendations for teachers and institutions, theoretical and practical implications, and possible future lines for research and knowledge transfer.

METHOD

This study adopts a quantitative approach, appropriate for the collection and analysis of measurable data (Hernández-Sampieri & Mendoza-Torres, 2023). It is framed within an evaluative research design, which is relevant when the aim is to assess processes or interventions in educational contexts (Arias, 2012), in this case,

pre-service teachers' self-perception of GenAI and their satisfaction with the experience of using ChatGPT in the design of LSs.

The following objectives (O) are set out:

- O1. To analyse pre-service teachers' self-perception of GenAI in relation to Familiarity, Relevance, Practical Skills, Barriers, Confidence, Ethical and Social Impact, and Expectations.
- O2. To determine how pre-service teachers rate the usefulness of ChatGPT as a support tool for LS design.

The first responds to the need, identified in previous studies, to gain a comprehensive understanding of how future teachers conceive GenAI across various dimensions, beyond isolated cases (Markos et al., 2024). The second arises from the growing incorporation of GenAI tools in education and the lack of evidence regarding how pre-service teachers perceive their pedagogical usefulness (Nyaaba & Zhai, 2024).

Based on these objectives, the following research questions (RQ) were formulated:

- RQ1. What is the self-perception of pre-service teachers regarding GenAI in relation to their Familiarity with it, its perceived Relevance, their Practical Skills, the Barriers they face, their Confidence, its Ethical and Social Impact, and their Expectations?
- RQ2. How do pre-service teachers rate the usefulness of ChatGPT as a support tool for LS design?

The experience developed aimed to explore the pedagogical potential of ChatGPT in designing LSs for Early Childhood and Primary Education stages. Participants mainly used ChatGPT-3.5, with limited access to ChatGPT-4 depending on service availability. The experience was organised into three two-hour practical sessions held over three weeks. The first had an introductory character, aimed at familiarising students with the educational use of GenAI and the formulation of effective prompts. Prompts were provided in logical sequences that facilitated the progressive construction of proposals, as well as review and analysis prompts to examine, adjust, and improve the generated productions, fostering pedagogical coherence, evidence-based decision-making, and continuous improvement of the design. Teaching support was provided throughout to address technical issues and promote critical thinking and informed decision-making. Quality criteria were established to evaluate the generated responses, focusing on discursive coherence, accuracy of information, and curricular alignment. Through an active learning approach, students explored the educational use of GenAI and strengthened their digital competences in an innovative environment.

Participants

A non-probability convenience sampling method was employed for the selection of the sample, justified by the direct access to participants and the relevance of the content. Students who did not complete the experience were excluded. This type of sampling may involve certain biases, particularly those related to motivation or technological familiarity. The sample size ($n = 185$) was determined by the number of

enrolled students. A total of 174 students from the University of Salamanca, enrolled in education-related degree programmes, participated in the study. Specifically, 37 were enrolled in the bachelor's degree in Early Childhood Education (21.26%), 58 in the bachelor's degree in Primary Education (33.33%), 19 were enrolled in a double degree programme in Primary Education and Early Childhood Education (10.91%), 27 were enrolled in a master's degree in Innovation in Specific Teaching Methods for Early Childhood and Primary Education (15.51%), and 33 were enrolled in a master's degree in ICT in Education: Analysis and Design of Processes, Resources and Training Practices (18.96%).

Student participation took place across different courses related to technology and educational innovation. Bachelor's degree students participated in the course ICT in Education; students in the master's degree in Innovation in the course Advances in Digital Technologies for Educational Innovation; and students of the master's degree in ICT in Education in the course Research Lines in Educational Technology.

In terms of participant characteristics, the group comprised 38 men (21.84%) and 136 women (78.16%), aged between 17 and 47, with an average age of 21.02 years and a standard deviation of 4.13 years.

Instrument

The instrument utilised to collect data was a questionnaire that had been validated in the research conducted by Espinoza-San Juan et al. (2024). It was developed using Google Forms and distributed to participants electronically via the virtual campus (Table 1). It was divided into three sections. The initial section collected sociodemographic data, including educational qualifications, age and sex. The second scale comprised 22 items and seven dimensions associated with self-perception of GenAI in education: Familiarity, Relevance, Practical Skills, Barriers, Confidence, Ethical and Social Impact, and Expectations. Each dimension comprised between two and four Likert-type items, with a response scale ranging from 1 (strongly disagree) to 5 (strongly agree). The third section amassed assessments of the practicality of ChatGPT in the design of LSs through five items: two Likert-type questions on Satisfaction and Perceived usefulness, and three multiple-choice items on the Quality of the responses generated, the Attitude adopted by the participant, and Training.

Table 1
Questionnaire items

Second section		
Dimension	Item	Text
Level of familiarity with GenAI	FO1	I am familiar with the basic concepts of GenAI (machine learning, neural networks, algorithms, etc.).
	FO2	I consider myself to have a broad and detailed understanding of how GenAI can be applied in education.
	FO3	During my time at university, I had the opportunity to learn more about using GenAI in education.
	FO4	I have acquired knowledge of GenAI through online courses, workshops, and self-learning methods.
Perception of the relevance of GenAI in education	RO1	Incorporating GenAI is important for updating teaching and learning methods.
	RO2	GenAI-based tools have the potential to improve the quality of education.
	RO3	Integrating GenAI into education is a current trend.

Practical skills with GenAI tools	S01	I am confident in my practical skills to use specific GenAI tools in education (educational chatbots, content recommendation systems, etc.).
	S02	I have had successful practical experience using GenAI tools in education (educational chatbots, content recommendation systems, etc.).
	S03	I believe I can teach others to use GenAI tools in education (educational chatbots, content recommendation systems, etc.).
	S04	I have shared my practical experience of using GenAI tools in education with others (educational chatbots, content recommendation systems, etc.).
Perceived barriers to the integration of GenAI	B01	The lack of specific and specialised training is an obstacle to integrating GenAI into education.
	B02	Concerns about the privacy and security of personal information act as a barrier to GenAI adoption.
	B03	Some educators' resistance to change is a challenge to the adoption of GenAI.
Confidence in integrating GenAI into teaching	Co1	I am confident in my ability to integrate GenAI into my future classes.
	Co2	I believe that GenAI can be a valuable teaching tool.
Ethical and social impact of AI in education.	Io1	It is important to consider the ethical implications of GenAI in education, such as bias, privacy, and transparency.
	Io2	The development of GenAI tools should prioritise inclusivity for all users.
	Io3	The adoption of GenAI in education requires the constant evaluation of its ethical and social impact.
Future expectations for GenAI in the field of education	Eo1	I expect GenAI to play a central and decisive role in education in the future.
	Eo2	GenAI can be used to tailor learning to the individual needs of each student.
	Eo3	I intend to pursue further training in utilising GenAI tools and techniques, specifically for educational purposes.
Third section		
Dimension	Item	Text
Satisfaction	S01	Indicate your level of satisfaction with the experience held.
Perceived usefulness	U01	Do you consider Chat GPT to be a good tool to help teachers design LSs?
Quality of responses generated	Q01	Do you consider the responses obtained with ChatGPT to be mostly:
		• Incorrect • Partially incorrect • Incoherent • Coherent • Incomplete • Complete • Partially correct • Correct
Attitude	A01	Regarding the answers you obtained from ChatGPT: • You fully trusted their accuracy and used them. • You verified the accuracy of the answer by comparing it with other sources and modified it partially. • You verified the accuracy of the answer by comparing it with other sources and modified it completely.
Training	T01	If we wanted ChatGPT to be used by teachers, you believe that: • Students and teachers should receive training on the specific tool and on GenAI in general. • We could receive a brief introduction to ChatGPT in particular, its context and certain warnings about its use. • The tool is easy to use and easy to adapt to. No training is necessary.
Source: Prepared by the authors.		

Data collection and analysis

Data collection took place between October 2024 and January 2025 with voluntary and anonymous participation and prior informed consent.

Descriptive and relational analyses were carried out to statistically characterise the dimensions evaluated in relation to self-perception and the usefulness of GenAI in education.

Each dimension of the questionnaire was treated as a composite variable, summarised by its mean (μ) and standard deviation (σ), and its normality was assessed using the Shapiro–Wilk test to guide the choice between parametric and non-parametric techniques. Although means and standard deviations are reported, as is common practice in educational research, the relationship between the age of the participants and the different dimensions was analysed using Spearman's correlation coefficient (ρ), given its robustness, as it does not require normality or linearity assumptions, and its suitability for the type of scale used.

Differences between degree programme groups in the questionnaire dimensions were explored using the Kruskal–Wallis test for independent samples, as the assumptions of normality (Shapiro–Wilk) and homogeneity of variances (Levene) were not met. The results are presented as the median [interquartile range], together with the corresponding p-value for the comparison between groups.

Exploratory and structural representation analyses were performed using networks to visualise the relationships between the questionnaire dimensions. Based on a correlation analysis between composite dimensions, a hierarchical dendrogram was generated using $1 - \rho$ (Spearman) distances, which allowed identifying conceptual clusters between the dimensions and detecting the existence of interrelated thematic axes in participants' self-perceptions, such as knowledge, assessment, and barriers. This structure facilitated the visual organization of a correlation network, providing an intuitive and relational representation of the links between the evaluated constructs. All analyses were performed using the free software R (version 4.4.3) (R Core Team, 2024). To promote transparency and reproducibility, the code used for the analyses is available in the OSF repository: <https://osf.io/c83dh> (Pinto-Llorente et al., 2025). As the dataset contains sensitive student information, it will be made available upon request from the authors.

RESULTS

The results are presented below, organised according to the analyses conducted and the research questions.

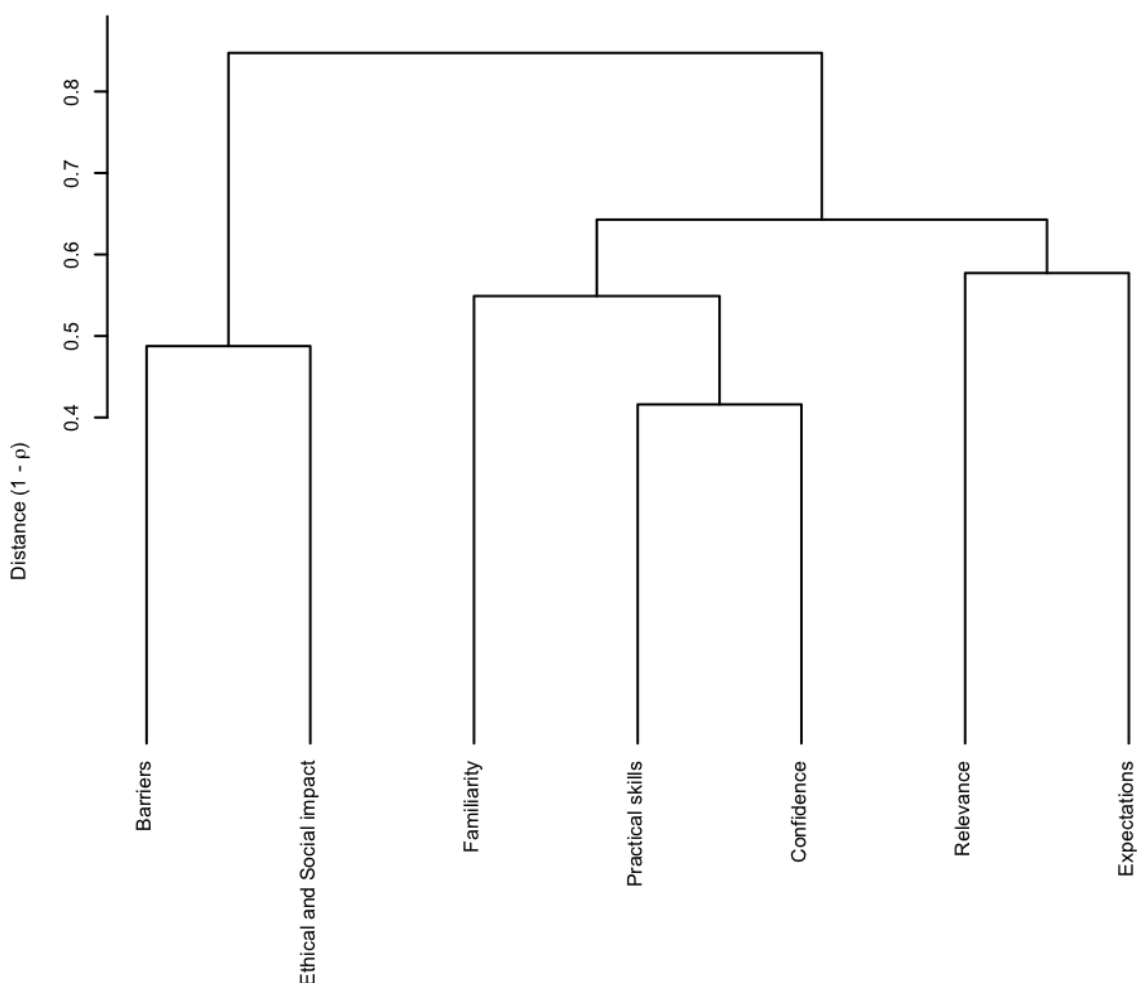
RQ1. What is the self-perception of pre-service teachers regarding GenAI in relation to their Familiarity with it, its perceived Relevance, their Practical Skills, the Barriers they face, their Confidence, its Ethical and Social Impact, and their Expectations?

The descriptive results show medium-high levels of self-perception regarding GenAI. A moderate level is observed in the dimensions: Confidence ($\mu = 3.91$; $\sigma = 0.68$), Expectations ($\mu = 3.72$; $\sigma = 0.68$), Familiarity ($\mu = 3.56$; $\sigma = 0.73$), and Practical Skills ($\mu = 3.56$; $\sigma = 0.76$) and high levels in Ethical and Social Impact ($\mu = 4.34$; $\sigma = 0.65$), Barriers ($\mu = 4.18$; $\sigma = 0.60$), and Relevance ($\mu = 4.08$; $\sigma = 0.57$).

The associations between the age of participants and the questionnaire dimensions were analysed. The results indicated positive and statistically significant Spearman correlations between age and the dimensions of Relevance ($\rho = 0.30$; $p < 0.01$), Expectations ($\rho = 0.20$; $p < 0.01$), and Ethical and Social Impact ($\rho = 0.40$; $p < 0.001$). Differences in the dimensions related to the self-perception of GenAI were also analysed according to the degree programme that the participants were enrolled in. Significant differences ($p < 0.05$) were found concerning the Expectations dimension. The master's degree in ICT in Education obtained the highest median (4 [3.67; 4.33]), while the other degrees presented slightly lower values, reflecting moderate variability between groups. Conversely, in the Familiarity dimension, a difference close to the threshold of statistical significance was detected ($p = 0.051$). Familiarity was lower for the master's degree in Innovation in Specific Teaching Methods (median 3.25 [3; 3.75]) and higher for the bachelor's degree in Early Childhood Education and the master's degree in ICT in Education (both with a median of 3.75). Regarding the Barriers dimension, the differences did not reach statistical significance ($p > 0.05$), but the master's degree in Innovation in Specific Teaching Methods showed a greater perception of barriers (4.67 [4.33; 4.67]) than the bachelor's degrees and the master's degree in ICT in Education (median 4.0). No significant differences were found concerning the Confidence dimension ($p > 0.05$). However, the double degree in Education had the lowest median score (3.50 [3.50; 4]), while the other degrees were all around 4.0. No statistically significant differences were observed between degrees with regard to Practical Skills ($p > 0.05$). The medians remained between 3.25 and 3.75, with no clear pattern emerging. Finally, statistically significant differences ($p < 0.001$) were observed between degrees in the Ethical and Social Impact dimension. The highest median was obtained by the master's degree in Innovation in Specific Teaching Methods (5 [4.67; 5]), followed by the master's degree in ICT in Education (4.67 [4; 5]), in contrast to the lower values obtained by the undergraduate degrees.

Effect sizes were calculated for each inferential comparison by degree programme. The variable Age ($\eta^2_H = 0.678$) showed the largest effect size, with a marked difference between programmes. Among the questionnaire dimensions, Ethical and Social Impact ($\eta^2_H = 0.142$) displayed a moderate effect size. In contrast, dimensions such as Expectations ($\eta^2_H = 0.036$), Relevance ($\eta^2_H = 0.036$), and Familiarity ($\eta^2_H = 0.032$) presented small effect sizes, although consistent with p values close to or below 0.05. Barriers ($\eta^2_H = 0.022$) and Confidence ($\eta^2_H = 0.021$) showed even smaller effects, while no relevant effects were observed for Practical Skills ($\eta^2_H = -0.003$). Regarding the categorical variable Sex, the effect size was moderate (Cramer's $V = 0.372$), indicating an unequal distribution across programmes.

The dendrogram generated from Spearman's correlations (Figure 1) showed a clear hierarchical organisation between the evaluated dimensions. A first cluster was identified, called Knowledge, composed of Familiarity, Practical Skills and Confidence. A second cluster, Assessment, grouped the dimensions Relevance and Expectations. Finally, Ethical and Social Impact, and Barriers formed the Obstacles cluster.

Figure 1*Hierarchical grouping of self-perception dimensions on GenAI*

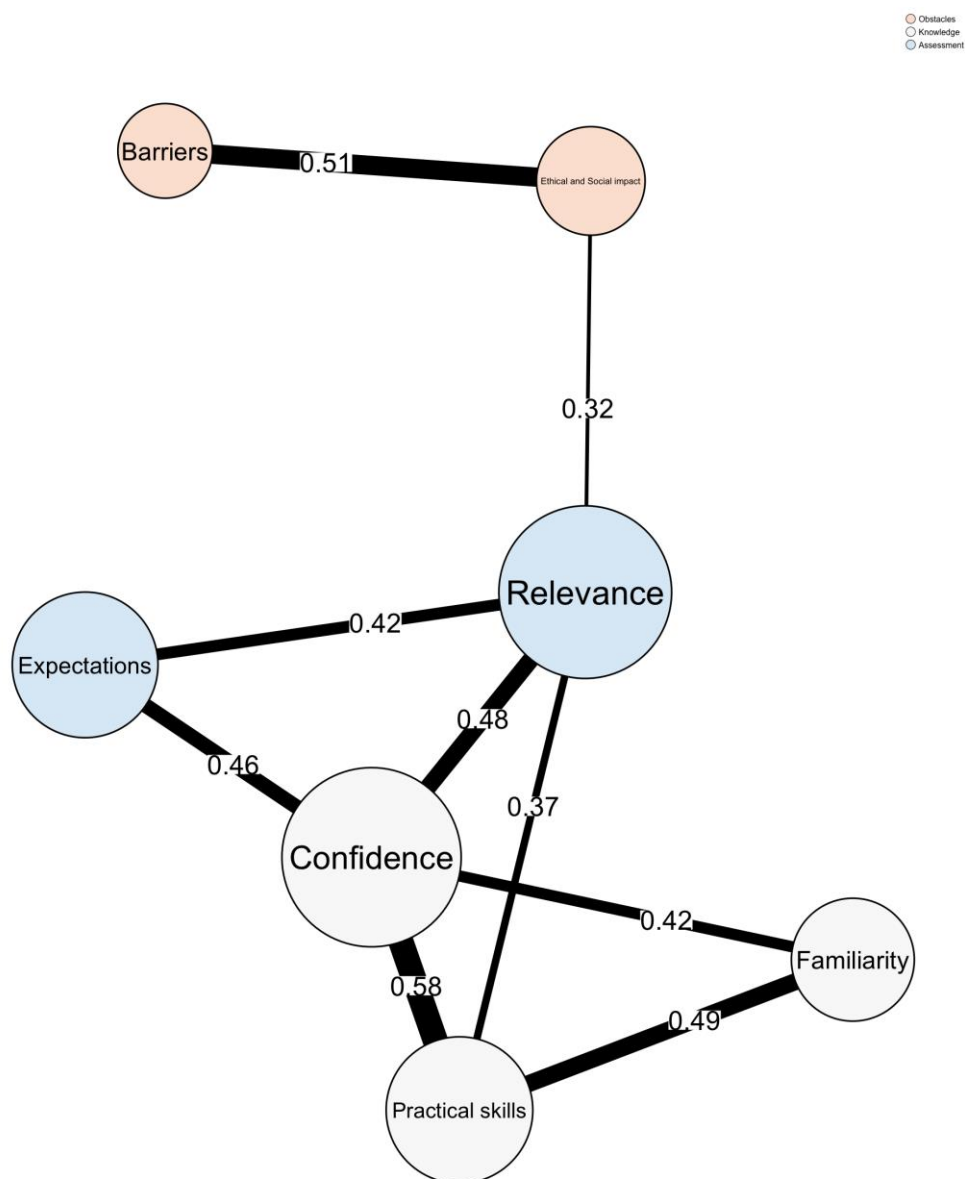
The dendrogram was used to represent a network of correlations between the dimensions (Figure 2), which was constructed from a Spearman's correlation matrix. The correlation network represents each dimension as a node. The nodes are arranged in a way that accurately reflects the clusters that emerged from the hierarchical analysis (Knowledge in white, Assessment in blue and Obstacles in orange), thereby reinforcing the empirical validity of the conceptual organisation adopted. The thickness of each line represents the intensity of the correlation.

Centrality metrics of the dimensions in the correlation network were calculated, considering a threshold of $|\rho| \geq 0.3$ to define the presence of edges. The measures included degree (number of significant connections), betweenness, and closeness.

The dimensions Relevance and Confidence stood out for having the highest degree (4 connections each), as well as the highest betweenness values (8.5 and 3.0, respectively), indicating that they act as central nodes in the network. Practical Skills also showed moderate connectivity (degree = 3). In contrast, dimensions such as Barriers, Familiarity, and Expectations displayed lower degree and low betweenness, suggesting a more peripheral role.

Figure 2

Correlation network of self-perception dimensions on GenAI in education



RQ2. How do pre-service teachers rate the usefulness of ChatGPT as a support tool for LS design?

ChatGPT was generally assessed very positively as a support tool for LS design. Satisfaction reached a value of 4.45 ($\sigma = 0.68$), and Perceived usefulness reached 4.24 ($\sigma = 0.84$). The quality of the generated responses was assessed using a dichotomous item ('Yes'/'No'). Only the results from participants who responded affirmatively are reported. The majority rated the responses as coherent ($n = 82$; 66.70%), while a smaller percentage considered them incorrect ($n = 20$; 16.30%), complete ($n = 19$; 15.40%), partially correct ($n = 17$; 13.80%), partially incorrect ($n = 14$; 11.40%), incomplete ($n = 14$; 11.40%), or incoherent ($n = 7$; 5.70%). None were rated as completely correct.

Regarding their Attitude towards the generated responses, the majority ($n = 138$, 79.30%) indicated that they checked and partially modified the responses. A smaller proportion ($n = 27$, 15.50%) indicated that they used the responses directly without making any changes. Only a very small group ($n = 9$, 5.20%) completely rejected the response provided by GenAI.

Consistent with this critical and reflective approach, a significant proportion of the sample ($n = 117$, 67.20%) stated that training in the use of ChatGPT and other GenAI tools is necessary for teachers and students. A total of 27.60% ($n = 48$) indicated that such training should be aimed exclusively at teachers, while only 5.20% ($n = 9$) considered it unnecessary.

Correlations were explored to identify factors associated with a more positive assessment. The results revealed positive and statistically significant correlations between perceived Usefulness of ChatGPT and the following dimensions:

- Confidence ($\rho = 0.36$; $p < 0.001$)
- Relevance ($\rho = 0.23$; $p = 0.0019$)
- Practical Skills ($\rho = 0.20$; $p = 0.0088$)

These correlations suggest that individuals who have greater Confidence, possess better Practical Skills, and consider GenAI to be more relevant in education tend to rate the Usefulness of ChatGPT for LS design more highly.

McDonald's internal reliability coefficient ω was analysed for the seven dimensions of the second section of the questionnaire. All showed adequate levels of internal consistency: Familiarity ($\omega = 0.80$), Relevance ($\omega = 0.72$), Practical Skills ($\omega = 0.85$), Barriers ($\omega = 0.70$), Ethical and Social Impact ($\omega = 0.81$), and Expectations ($\omega = 0.74$). For the Confidence dimension, composed of only two items, it was not possible to calculate the coefficient. Similarly, the four dimensions of the third section, each consisting of a single item, did not allow its estimation. Regarding the dimension Quality of the generated responses, $\omega = 0.65$ was obtained. The literature considers values between 0.70 and 0.90 acceptable, including values above 0.65 (Ventura-León & Caycho-Rodríguez, 2017).

DISCUSSION

This research focused on understanding pre-service teachers' self-perceptions of GenAI and the usefulness of ChatGPT in designing LSs. Regarding the first research question, the dimensions with the highest levels were Ethical and Social Impact, Barriers, and Relevance, which suggests a critical awareness of GenAI. With respect to Ethical and Social Impact, numerous studies report similar challenges relating to bias, privacy and transparency (Ishmuradova et al., 2025; Lozano & Blanco, 2023; Markos et al., 2024), as well as the lack of access to GenAI, which could exacerbate educational disparities arising from unequal access (Kasneji et al., 2023) and even limit access to knowledge (UNESCO, 2023). Furthermore, Jo (2024) found that privacy concerns are negatively associated with its use. Previous studies also emphasise the need to integrate GenAI into educational contexts as a tool to support teachers and for providing teachers with specific training in its use (Ishmuradova et al., 2025; Lozano & Blanco, 2023; Markos et al., 2024). With regard to the dimension Relevance, previous studies highlight the significant benefits that GenAI can offer for enhancing the teaching and learning process (Ishmuradova et al., 2025; Markos et al., 2024).

The dimensions Confidence, Expectations, Familiarity, and Practical Skills showed moderate levels. These results are consistent with those of Kelly et al. (2023), who found that confidence and familiarity with GenAI are often limited, increasing with experience. Other studies suggest that demonstrating the benefits of using AI chatbots in education can favour their adoption and effective use (Jo, 2024). Regarding Expectations, the results align with other research highlighting the capacity of GenAI to adapt learning to students' needs, thereby reinforcing its role in current and future education (Ruiz-Rojas et al., 2023; Sánchez, 2024; Sánchez-Prieto et al., 2025). In terms of Practical Skills, the results are consistent with those of Ng et al. (2023), who emphasise that future teachers require further experience to become familiar with, use, and teach with GenAI tools. It should be noted that, given the correlational nature of the study, these relationships do not imply causality.

The research revealed a hierarchical organisation of the dimensions studied, grouped into three clusters: Knowledge, Evaluation, and Barriers regarding GenAI. This organisation can be useful for guiding future initiatives aimed at pre-service teachers, as it provides a clear conceptual framework of their perceptions and needs.

The network results show that the dimensions Relevance and Confidence occupy central positions in the questionnaire structure, acting as key connecting nodes that facilitate articulation between blocks such as Familiarity, Practical Skills, and Expectations. In interpretative terms, this suggests that those who feel confident in integrating GenAI into their teaching practice (high Confidence) and who perceive it as a valuable tool (high Relevance) also tend to show greater prior knowledge, higher practical skills, and greater expectations for future use. Some reports indicate that confidence in GenAI increases the intention to use it (Ivanov et al., 2024). Their articulating role, confirmed by their nodal position in this research, had already been noted as a bridge between favourable attitudes and critical concerns. Overall, the results show a balance between positive attitudes and critical awareness. High scores in Expectations and Relevance coexist with equally high values in Ethical–social impact and practical Barriers. This pattern reflects that respondents recognise the potential of GenAI to improve teaching (Kasneci et al., 2023), without ignoring risks and limitations. The perception of obstacles is moderated by high expectations, indicating an attitude that strikes a balance between well-founded optimism and critical caution.

The inclusion of effect sizes enabled a more precise interpretation of the practical relevance of differences between degree programmes. Although some comparisons did not reach statistical significance, small effect sizes in dimensions such as Expectations, Relevance, and Familiarity suggest subtle but potentially meaningful differences in educational contexts. Clear patterns were observed depending on the type of degree. Undergraduate students showed lower levels of Expectations and Familiarity with GenAI, as well as a slight tendency to perceive more Barriers than master's students. This suggests that postgraduate programmes promote greater confidence and expectations, possibly due to a more advanced curriculum or broader prior experience.

The Ethical and Social Impact dimension stands out in particular, with master's students clearly outperforming undergraduates. Overall, these findings reinforce the need to consider the actual magnitude of differences in order to guide educational proposals tailored to students' profiles. In undergraduate programmes, it is advisable to foster interest and familiarity, whereas in postgraduate programmes, the focus should be on strengthening a critical and reflective view of GenAI.

Previous studies indicate that familiarity and predisposition towards innovative technologies tend to increase with higher levels of education (Saihi et al., 2025). Therefore, training interventions should be tailored to the profile: in undergraduate programmes, reinforcing motivation and providing early exposure to GenAI; and in postgraduate programmes developing critical and collaborative competences (Saihi et al., 2025; Wang & Wang, 2022). However, this differentiation should be interpreted with caution, as the sample was self-selected and no qualitative triangulation was applied, limiting generalisation, particularly across educational levels. These findings are consistent with previous research that has highlighted the importance of critical digital literacy (Wang et al., 2024) and the need for training support to facilitate the responsible adoption of emerging technologies in education (Bearman & Ajjawi, 2023).

These reflections are particularly relevant in online or hybrid education contexts, where the use of tools such as ChatGPT can support the design and development of LSs, favouring personalisation and pedagogical adaptation in non-face-to-face modalities.

The implications for teacher training are clear: programmes must integrate technical, critical, and ethical components in a balanced way. Technical competences should be developed through practical instruction with tools such as ChatGPT, a basic understanding of its algorithms, and specific teaching activities (Wang & Wang, 2022). Furthermore, critical thinking and digital literacy should be fostered, for example, through training to evaluate the quality of content generated by GenAI and by promoting strategies to verify information and analyse biases (Kasneci et al., 2023). It is also important to encourage ethical and professional reflection through debates and by raising awareness of ethical challenges (privacy, equity, technological dependence), in line with the call for an ethical pedagogical approach to GenAI (Kasneci et al., 2023).

In response to the second research question, high scores were obtained in both Satisfaction and perceived Usefulness with ChatGPT. These results suggest that this technology could play a significant role in lesson planning and content production. A notable finding was the critical attitude adopted by students who mostly chose to verify and adapt the information, thereby demonstrating metacognitive competence in the use of GenAI. This is particularly relevant given that GenAI can generate erroneous or biased responses, confirming that the effective use of ChatGPT also depends on pedagogical and ethical judgement. Furthermore, most participants considered specific training to be necessary, supporting the idea that the integration of GenAI tools requires training support, consistent with previous studies that emphasise the need for ethical, pedagogical, and operational frameworks for the appropriate use of GenAI (Ivanov et al., 2024; Kasneci et al., 2023).

In summary, the emergence of GenAI has generated considerable interest in understanding the needs of pre-service teachers for its effective incorporation into the classroom (Ogunleye et al., 2024). The evidence reveals a positive disposition and awareness of the educational benefits (Buyakova et al., 2024). Whitbread et al. (2025) emphasise the importance of establishing policies to ensure teacher training and to address the associated challenges. This includes AI literacy training programmes that enable teachers to integrate these technologies with confidence and effectiveness (Al-Abdullatif, 2024; Lozano & Blanco, 2023; Omar et al., 2024), promoting an informed and responsible use (Nyaaba & Zhai, 2024; Markos et al., 2024; Okunade, 2024).

CONCLUSIONS

This quantitative study, adopting an evaluative design, aimed to analyse the self-perception of pre-service teachers regarding GenAI and to assess their experience of using ChatGPT in the design of LSs. To this end, a questionnaire was administered after a training activity, which enabled the collection of empirical data on key dimensions, as well as on the perceived usefulness of the tool. The results show a favourable attitude towards its integration, with high levels of Familiarity, Practical Skills, positive Expectations, and appreciation of Ethical and Social Impact. The Confidence dimension emerges as a key structural axis, acting as a link between Knowledge, Attitude, and willingness to use.

Analysis of the correlation network reveals that teacher Confidence reflects not only self-efficacy but also a predisposition towards critically adopting emerging technologies. At the same time, a duality is observed between enthusiasm for the potential of GenAI and awareness of its ethical and practical risks. This highlights the need for training programmes that combine practical training with opportunities for critical reflection. The findings also suggest that the perception of barriers is not primarily associated with technical deficits, but with ethical and pedagogical sensitivities. The peripheral position of this dimension in the network underscores the importance of addressing it through contextualised approaches rather than exclusively through technological training.

Based on the scientific literature and exploratory findings obtained through network analysis, we propose a conceptual model that articulates the expected relationships among the different dimensions of self-perception of GenAI in pre-service teachers.

In this model, Familiarity and Practical Skills are grouped into a cluster termed Prior knowledge, which acts as an antecedent and facilitates the development of Confidence in the use of GenAI tools. This Confidence is conceived as a central mediating dimension, influencing both the perception of Relevance and a lower perception of Barriers to its implementation. In turn, both influence the construction of future Expectations regarding the educational use of this technology.

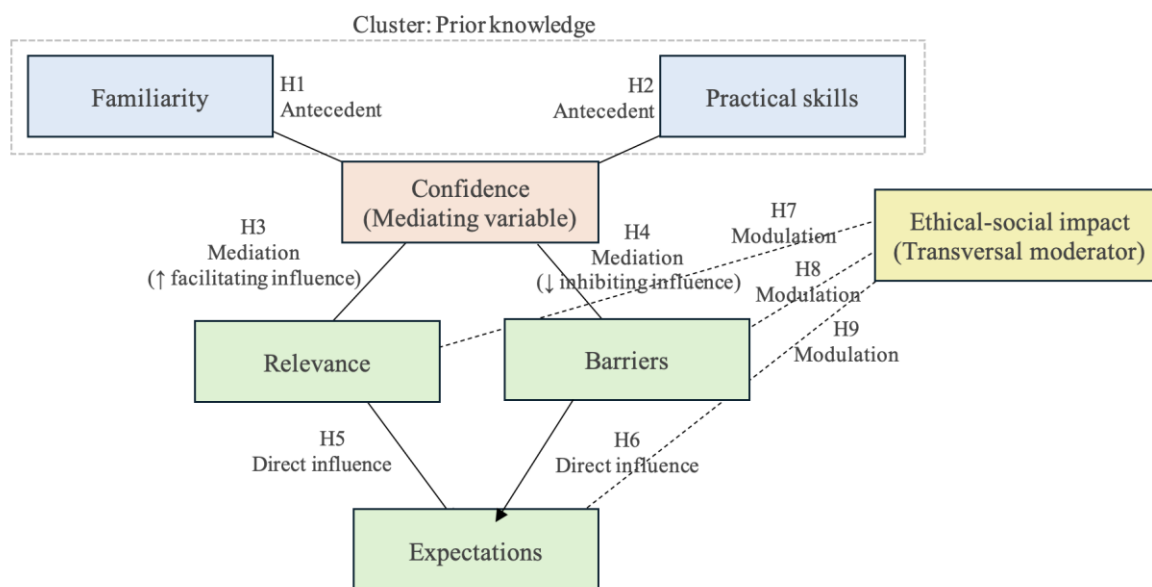
Ethical and Social Impact is positioned as a transversal dimension that critically modulates attitudes towards the adoption of GenAI, particularly among those who perceive greater relevance and confidence, but who are also more aware of ethical and social challenges. This critical disposition is associated with a reflective attitude observable in the use of tools such as ChatGPT.

The model anticipates that Confidence operates as an articulating node between prior knowledge and future willingness towards GenAI, whereas Ethical impact acts as a critical regulator of these attitudes.

This model aligns with the Technology Acceptance Model (TAM), centred on perceived usefulness and attitudes towards technology, and with the AI-CFT, which integrates ethical and pedagogical competences for the responsible use of AI. In summary, the model articulates the hypothetical relationships represented in Figure 3.

Figure 3

Conceptual model: Hypothetical relationships among self-perception dimensions of GenAI



Based on these results, five action lines are proposed to effectively integrate GenAI into teacher training: (1) ongoing training that combines technical use and ethical implications; (2) hands-on learning through applied activities; (3) the development of critical thinking; (4) initial support to strengthen teacher confidence; and (5) ongoing evaluation of ethical and social impact. Likewise, institutions are urged to integrate these principles into their policies and curricula, promoting hybrid and collaborative environments that facilitate equitable use of GenAI and strengthen communities of practice for teacher professional development.

GenAI offers a strategic opportunity to transform education, enriching teaching, fostering innovation, and democratising access to knowledge. Its implementation requires profound pedagogical reflection, clear policies, and ongoing training, focusing not only on technological capabilities but also on its pedagogical integration adapted to each context, while promoting a critical digital culture. From this perspective, the findings deepen the understanding of pre-service teachers' attitudes towards a critical adoption of GenAI and establish a foundation for future teachers to integrate this technology into their classrooms, preparing compulsory-level students for its ethical and creative use.

This study has some limitations. Its cross-sectional design prevents causal inferences, and it relies solely on self-reported data without objective measures of actual use, which may affect accuracy. In addition, it does not include comparisons with other technologies or longitudinal follow-up, limiting the analysis of the specific value of ChatGPT and the evolution of perceptions over time. The sample is not representative of all educational levels, and there is a lack of qualitative triangulation that could provide greater depth.

Therefore, it is recommended that future research adopts longitudinal and inter-institutional designs, integrate mixed methods with qualitative triangulation (interviews, focus groups) and objective measures of use, and systematically compare

GenAI with other educational technologies to identify its specific advantages and limitations.

Acknowledgements

This research was carried out within the framework of the Teaching Innovation Project: Personalised and adaptive learning through generative artificial intelligence systems in Education Sciences (ID2024/260), supported by the SA217P23 project funded by the Regional Government of Castile and León.

REFERENCES

- Al-Abdullatif, A. M. (2024). Modeling teachers' acceptance of generative artificial intelligence use in higher education: The role of AI literacy, intelligent TPACK, and perceived trust. *Education Sciences*, 14(11), 1209. <https://doi.org/10.3390/educsci14111209>
- Al-Abdullatif, A. M., & Alsubaie, M. A. (2024). ChatGPT in learning: Assessing students' use intentions through the lens of perceived value and the influence of AI literacy. *Behavioral Sciences*, 14(9), 845. <https://doi.org/10.3390/bs14090845>
- Alasadi, E. A., & Baiz, C. R. (2023). Generative AI in education and research: Opportunities, concerns, and solutions. *Journal of Chemical Education*, 100(8), 2965-2971. <https://doi.org/10.1021/acs.jchemed.3c00323>
- Alba Pastor, C. (2022). *Enseñar pensando en todos los estudiantes: El modelo de diseño universal para el aprendizaje (DUA)*. Ediciones SM.
- Aravena Castillo, F. (2013). Developing the collaborative model in the initial formation: The autoperception of the professional performance of the beginner teacher in action. *Estudios Pedagógicos*, 39(1), 27-44. <https://doi.org/10.4067/S0718-07052013000100002>
- Arias, F. G. (2012). *El proyecto de investigación*. Episteme.
- Bayly-Castaneda, K., Ramírez-Montoya, M.-S., & Morita-Alexander, A. (2024). Crafting personalized learning paths with AI for lifelong learning: A systematic literature review. *Frontiers in Education*, 9, Article 1424386. <https://doi.org/10.3389/feduc.2024.1424386>
- Bearman, M., & Ajjawi, R. (2023). Learning to work with the black box: Pedagogy for a world with artificial intelligence. *British Journal of Educational Technology*, 54(5), 1160-1173. <https://doi.org/10.1111/bjet.13337>
- Buyakova, K. I., Dmitriev, Ya. A., Ivanova, A. S., Feshchenko, A. V., & Yakovleva, K. I. (2024). Students' and teachers' attitudes towards the use of tools with generative artificial intelligence at the university. *The Education and Science Journal*, 26(7), 160-193. <https://doi.org/10.17853/1994-5639-2024-7-160-193>
- Carranza Alcántar, M. del R., Macías González, G. G., Gómez Rodríguez, H., Jiménez Padilla, A. A., & Jacobo Montes, F. M. (2024). Percepciones docentes sobre la integración de aplicaciones de IA generativa en el proceso de enseñanza universitario. *REDU. Revista de Docencia Universitaria*, 22(2), 21-40. <https://doi.org/10.4995/redu.2024.22027>
- Currie, G. M. (2025). Generative artificial intelligence in nuclear medicine education. *Journal of Nuclear Medicine Technology*, 53(1), 72-79. <https://doi.org/10.2967/jnmt.124.268323>
- Denny, P., Prather, J., Becker, B. A., Finnie-Ansley, J., Hellas, A., Leinonen, J., Luxton-Reilly, A., Reeves, B. N., Santos, E. A., & Sarsa, S. (2024). Computing education in the era of generative AI. *Communications of the ACM*, 67(2), 56-67. <https://doi.org/10.1145/3624720>

- Espinoza-San Juan, J., Raby, M. D., & Sagredo-Lillo, E. (2024). Validación de un cuestionario sobre las percepciones y usos de la IA-Gen entre estudiantes de pedagogía. *RISTI: Revista Ibérica de Sistemas e Tecnologias de Informação*, (70), 574-585.
- Gozalo-Brizuela, R., & Garrido-Merchán, E. (2024). A survey of generative AI applications. *Journal of Computer Science*, 20(8), 801-818. <https://doi.org/10.3844/jcssp.2024.801.818>
- Haroud, S., & Saqri, N. (2025). Generative AI in higher education: Teachers' and students' perspectives on support, replacement, and digital literacy. *Education Sciences*, 15(4), 396. <https://doi.org/10.3390/educsci15040396>
- Hernández-Sampieri, R., & Mendoza-Torres, C. P. (2023). *Metodología de la investigación*. McGraw-Hill.
- Ishmuradova, I. I., Zhdanov, S. P., Kondrashev, S. V., Erokhova, N. S., Grishnova, E. E., & Volosova, N. Y. (2025). Pre-service science teachers' perception on using generative artificial intelligence in science education. *Contemporary Educational Technology*, 17(3), ep579. <https://doi.org/10.30935/cedtech/16207>
- Islam, G., & Greenwood, M. (2024). Generative artificial intelligence as hypercommons: Ethics of authorship and ownership. *Journal of Business Ethics*, 192(3), 659-663. <https://doi.org/10.1007/s10551-024-05741-9>
- Ivanov, S., Soliman, M., Tuomi, A., Alkathiri, N. A., & Al-Alawi, A. N. (2024). Drivers of generative AI adoption in higher education through the lens of the theory of planned behaviour. *Technology in Society*, 77, 102521. <https://doi.org/10.1016/j.techsoc.2024.102521>
- Jo, H. (2024). From concerns to benefits: A comprehensive study of ChatGPT usage in education. *International Journal of Educational Technology in Higher Education*, 21, 35. <https://doi.org/10.1186/s41239-024-00471-4>
- Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., ... Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 102274. <https://doi.org/10.1016/j.lindif.2023.102274>
- Kelly, A., Sullivan, M., & Strampel, K. (2023). Generative artificial intelligence: University student awareness, experience, and confidence in use across disciplines. *Journal of University Teaching & Learning Practice*, 20(6). <https://doi.org/10.53761/1.20.6.12>
- Laupichler, M. C., Aster, A., Schirch, J., & Raupach, T. (2022). Artificial intelligence literacy in higher and adult education: A scoping literature review. *Computers and Education: Artificial Intelligence*, 3, 100101. <https://doi.org/10.1016/j.caeai.2022.100101>
- Lozano, A., & Blanco Fontao, C. (2023). Is the education system prepared for the irruption of artificial intelligence? A study on the perceptions of students of primary education degree from a dual perspective: Current pupils and future teachers. *Education Sciences*, 13(7), 733. <https://doi.org/10.3390/educsci13070733>
- Markos, A., Prentzas, J., & Sidiropoulou, M. (2024). Pre-service teachers' assessment of ChatGPT's utility in higher education: SWOT and content analysis. *Electronics*, 13(10), 1985. <https://doi.org/10.3390/electronics13101985>
- Miao, F., & Cukurova, M. (2024). *AI competency framework for teachers*. UNESCO. <https://doi.org/10.54675/ZJTE2084>
- Moorhouse, B. L., & Kohnke, L. (2024). The effects of generative AI on initial language teacher education: The perceptions of teacher educators. *System*, 122, 103290. <https://doi.org/10.1016/j.system.2024.103290>

- Moorhouse, B. L., Wan, Y., Wu, C., Kohnke, L., Ho, T. Y., & Kwong, T. (2024). Developing language teachers' professional generative AI competence: An intervention study in an initial language teacher education course. *System*, 125, 103399. <https://doi.org/10.1016/j.system.2024.103399>
- Mouta, A., Torrecilla-Sánchez, E. M., & Pinto-Llorente, A. M. (2025). Comprehensive professional learning for teacher agency in addressing ethical challenges of AIED: Insights from educational design research. *Education and Information Technologies*, 30, 3343-3387. <https://doi.org/10.1007/s10639-024-12946-y>
- Ng, D. T. K., Leung, J. K. L., Su, J., Ng, R. C. W., & Chu, S. K. W. (2023). Teachers' AI digital competencies and twenty-first century skills in the post-pandemic world. *Educational Technology Research and Development*, 71(1), 137-161. <https://doi.org/10.1007/s11423-023-10203-6>
- Nyaaba, M., & Zhai, X. (2024). Generative AI professional development needs for teacher educators. *Journal of AI*, 8(1), 1-13. <https://doi.org/10.61969/jai.1385915>
- Ogunleye, B., Zakariyyah, K. I., Ajao, O., Olayinka, O., & Sharma, H. (2024). A systematic review of generative AI for teaching and learning practice. *Education Sciences*, 14(6), 636. <https://doi.org/10.3390/educsci14060636>
- Okunade, A. I. (2024). The role of artificial intelligence in teaching of science education in secondary schools in Nigeria. *European Journal of Computer Science and Information Technology*, 12(1), 57-67. <https://doi.org/10.37745/ejcsit2013/vol12n15767>
- Omar, A., Shaqour, A. Z., & Khlaif, Z. N. (2024). Attitudes of faculty members in Palestinian universities toward employing artificial intelligence applications in higher education: Opportunities and challenges. *Frontiers in Education*, 9, 1414606. <https://doi.org/10.3389/feduc.2024.1414606>
- Pinto-Llorente, A. M., Izquierdo-Álvarez, V., & Dolcet-Negre, M. M. (2025). *Autopercepción y utilidad de la inteligencia artificial generativa en docentes en formación* [Open Science Framework]. <https://osf.io/c83dh>
- R Core Team. (2024). *R: A language and environment for statistical computing* (Version 4.4.3) [Computer software]. R Foundation for Statistical Computing. <https://www.r-project.org/>
- Ruiz-Rojas, L. I., Acosta-Vargas, P., De-Moreta-Llovet, J., & Gonzalez-Rodriguez, M. (2023). Empowering education with generative artificial intelligence tools: Approach with an instructional design matrix. *Sustainability*, 15(15), Article 11524. <https://doi.org/10.3390/su151511524>
- Saihi, A., Ben-Daya, M., & Hariga, M. (2025). The moderating role of technology proficiency and academic discipline in AI-chatbot adoption within higher education: Insights from a PLS-SEM analysis. *Education and Information Technologies*, 30, 5843-5881. <https://doi.org/10.1007/s10639-024-13023-0>
- Sánchez Vera, M. del M. (2024). La inteligencia artificial como recurso docente: Usos y posibilidades para el profesorado. *Educar*, 60(1), 33-47. <https://doi.org/10.5565/rev/educar.1810>
- Sánchez-Prieto, J. C., Izquierdo-Álvarez, V., del Moral-Marcos, M. T., & Martínez-Abad, F. (2025). Inteligencia artificial generativa para autoaprendizaje en educación superior: Diseño y validación de una máquina de ejemplos. *RIED-Revista Iberoamericana de Educación a Distancia*, 28(1), 59-81. <https://doi.org/10.5944/ried.28.1.41548>
- Storey, V. C., Yue, W. T., Zhao, J. L., & Lukyanenko, R. (2025). Generative artificial intelligence: Evolving technology, growing societal impact, and opportunities for information systems research. *Information Systems Frontiers*. <https://doi.org/10.1007/s10796-025-10581-7>
- Tai-Han, L., Hsing-Yi, C., Ming-Jr, J., Chih-Kai, C., Cherng-Lih, P., Guo-Shiou, L., Jyh-Cherng, Y., Ming-Shen, D., Cheng-Ping, Y., & Hung-Sheng, S. (2024). An

- advanced machine learning model for a web-based artificial intelligence-based clinical decision support system application: Model development and validation study. *Journal of Medical Internet Research*, 26, e56022. <https://doi.org/10.2196/56022>
- UNESCO. (2023). *Education 2030 agenda*. <https://www.unesco.org/en/digital-education/artificial-intelligence>
- Ventura-León, J. L., & Caycho-Rodríguez, T. (2017). El coeficiente omega: Un método alternativo para la estimación de la confiabilidad. *Revista Latinoamericana de Ciencias Sociales, Niñez y Juventud*, 15(1), 625-627.
- Wang, K., Ruan, Q., Zhang, X., Fu, C., & Duan, B. (2024). Pre-service teachers' GenAI anxiety, technology self-efficacy, and TPACK: Their structural relations with behavioral intention to design GenAI-assisted teaching. *Behavioral Sciences*, 14(5), 373. <https://doi.org/10.3390/bs14050373>
- Wang, Y. Y., & Wang, Y. S. (2022). Development and validation of an artificial intelligence anxiety scale: An initial application in predicting motivated learning behavior. *Interactive Learning Environments*, 30(4), 619-634. <https://doi.org/10.1080/10494820.2019.1674887>
- Whitbread, M., Hayes, C., Prabhakar, S., & Upsher, R. (2025). Exploring university staff's perceptions of using generative artificial intelligence at university. *Education Sciences*, 15(3), 367. <https://doi.org/10.3390/educsci15030367>
- Zhang, C., & Villanueva, L. E. (2023). Generative artificial intelligence preparedness and technological competence. *International Journal of Education and Humanities*, 11(2), 164-170. <https://doi.org/10.54097/ijeh.v11i2.13753>

Date of reception: 1 June 2025

Date of acceptance: 18 August 2025

Date of approval for layout: 22 September 2025

Date of publication in OnlineFirst: 14 October 2025

Date of publication: 1 January 2026

Teachers and science communication on social media: development and initial validation of assessment instruments

Docentes y divulgación científica en redes sociales: desarrollo y validación inicial de instrumentos de evaluación



 Dieter Reynaldo Fuentes-Cancell - *Universidad de Valladolid, UVa (Spain)*

 Odiel Estrada-Molina - *Universidad de Valladolid, UVa (Spain)*

 Mónica Gutiérrez-Ortega - *Universidad de Valladolid, UVa (Spain)*

ABSTRACT

In the current context of digital education and open science, university professors are not only knowledge creators through their scientific and professional output but are also expected to disseminate it to both academic and non-specialist audiences. While related to digital competence, media literacy provides a critical-communicative perspective essential for effective science dissemination, particularly via digital social networks (DSNs). This study designed and validated two constructs to assess teachers' media competence (TMC): one focused on general DSN use and the other specifically on LinkedIn. Both models integrate the Common Framework for Digital Teaching Competence 2.2 and the media literacy model by Ferrés and Piscitelli (2012). The validation process combined theoretical review, expert judgment ($n = 30$), and exploratory factor analysis (EFA). Data suitability was confirmed through KMO (> 0.80) and Bartlett's test of sphericity, with both instruments showing high reliability ($\alpha > 0.85$; $\omega > 0.87$). The EFA identified six theoretical dimensions, explaining 78.2% of the variance for DSN and 78.8% for LinkedIn. These findings provide an initial approximation of the TMC structure and its potential for diagnostic use in teacher training, highlighting LinkedIn's strategic role as a professional environment for academic dissemination. The incorporation of communicative and media strategies in teacher education is underscored. Future research should include confirmatory factor analysis with larger samples to consolidate these initial results.

Keywords: media competence; university teachers; digital social networks; assessment; content validity; reliability.

RESUMEN

En el contexto actual de educación digital y ciencia abierta, el profesorado universitario no solo genera conocimiento mediante su producción científica y profesional, sino que además debe saber divulgarlo a públicos académicos y no especializados. Aunque vinculada a la competencia digital, la competencia mediática aporta un enfoque crítico-comunicativo necesario para este proceso, especialmente en redes sociales digitales (RSD). Este estudio diseña y valida dos constructos para evaluar la competencia mediática del profesorado (CMP): uno centrado en el uso general de RSD y otro específico para LinkedIn. Ambos modelos integran el Marco Común de Competencia Digital Docente 2.2 y el modelo de alfabetización mediática de Ferrés y Piscitelli (2012). La validación se realizó mediante revisión teórica, juicio de expertos ($n = 30$) y análisis factorial exploratorio (AFE). La idoneidad de los datos se confirmó mediante índice KMO (> 0.80) y prueba de esfericidad de Bartlett; ambos instrumentos mostraron alta fiabilidad ($\alpha > 0.85$; $\omega > 0.87$). El AFE identificó seis dimensiones teóricas, explicando el 78.2 % de la varianza para RSD y el 78.8 % para LinkedIn. Estos hallazgos ofrecen una aproximación inicial a la estructura de la CMP y su potencial diagnóstico en la formación docente, destacando el papel estratégico de LinkedIn como entorno profesional para la divulgación académica. Se subraya la necesidad de incorporar estrategias comunicativas y mediáticas en la formación del profesorado. Como línea futura, se recomienda realizar análisis factorial confirmatorio con muestras más amplias para consolidar los resultados obtenidos.

Palabras clave: competencia mediática; profesores universitarios; redes sociales digitales; evaluación; validez de contenido; confiabilidad.

INTRODUCTION

Within the framework of digital education and open science, the communication of scientific knowledge through digital social networks (DSN) has emerged as a key competence for teachers. Nevertheless, teacher training in this area continues to face both conceptual and methodological challenges. In this context, this study is aligned with the Digital Teaching Competence Framework 2.2 (INTEF, 2022), particularly in its areas 2 (Digital content) and 6 (Developing students' digital competence), because of its focus on the creation, dissemination and critical analysis of digital content. Alongside Ferrés and Piscitelli's (2012) media literacy model, these two frameworks underpin the design and validation of two constructs for assessing teachers' media competence (TMC). The factor structure of the indicators and their applicability in training contexts are analysed.

Teachers' media competence in the context of digital social networks

Digital competence and TMC are complementary and fundamental concepts in the current educational context (Marcelo-Martínez et al., 2025). The former focuses on the pedagogical use of technological tools to improve teaching and learning (Cabero-Almenara et al., 2020), while the latter emphasises critical media literacy and its impact on knowledge construction (Ferrés & Piscitelli, 2012). Both are essential for science outreach in DSN, where teachers act as creators and mediators of academic knowledge (Buzón et al., 2019).

Scientific dissemination in DSN has become a key element in bringing academic knowledge closer to society, broadening the impact of research and fostering interaction with different audiences (Valencia-Oliveros & Martín-Gutiérrez, 2024). Increasing digitisation and the rise of platforms such as Twitter/X (Vásquez-Rocca, 2024), Instagram (Romero Rodríguez et al., 2019), Facebook (Verstappen & Opgenhaffen, 2024) and YouTube (Marchal et al., 2025) have generated new opportunities to communicate science beyond traditional academic circles. However, their use among researchers remains limited, highlighting the need to strengthen training in digital communication (Said-Hung et al., 2024).

In a context of increasing misinformation, these platforms not only facilitate the dissemination of knowledge, but also enable interaction with the public and adaptation to new digital formats (Martin-Neira et al., 2023). Despite their potential in education and social sciences, the challenge lies in balancing the accessibility of the message without losing scientific rigour, especially in networks such as LinkedIn and Instagram, where content must be brief and attractive (Gallego et al., 2023).

LinkedIn has established itself as a key platform for scientific dissemination, differentiating itself from other DSNs by its professional approach and its ability to connect researchers with academic, business and education communities (Knight, 2019). Its algorithm favours the visibility of scientific content, enabling the dissemination of articles, preprints and research findings, as well as interaction in specialised communities (Baruffaldi et al., 2017). Recent studies have identified LinkedIn as the social network most used by academics to share their scientific output, surpassing platforms such as ResearchGate and Academia.edu, due to its ability to amplify the impact of knowledge and facilitate transfer between academia and industry (Huang et al., 2019; Knight, 2019; Ordóñez Castillo et al., 2024; Said-Hung et al., 2024; Tolbert, 2024). However, barriers to their widespread adoption persist, such as the lack

of training in digital communication and the scarce institutional recognition of these activities (Argüello-Gutiérrez & Moreno-López, 2024). Despite these limitations, LinkedIn offers an environment conducive to the construction of a digital academic identity, the dissemination of knowledge and the generation of collaborative networks, which highlights the need to promote institutional strategies that encourage its effective use in scientific dissemination.

Previous research proposes general models of TMC (Gümüş & Kukul, 2023; Jiang & Yu, 2024; Vergili & Kara, 2024) but has not structured it in science outreach in DSN. Although they identify media skills for educational and scientific purposes, they do not establish a specific construct to serve as a basis for teacher training. Therefore, it responds to this gap by means of an integrative proposal that articulates the Spanish Framework for the Digital Competence of Teachers, hereafter referred to by its Spanish acronym, MRCDD 2.2 (INTEF, 2022) and the media literacy model of Ferrés and Piscitelli (2012).

Integration allows for a comprehensive analysis of TMC in science popularisation. In this regard, several recent systematic reviews support such convergence. Claro et al. (2024) stress the importance of consolidating the operationalisation of teachers' digital competence within normative frameworks such as MRCDD 2.2, while Olaya Guerrero et al. (2025) emphasise the fundamental role of institutional strategies in teacher training in digital competence. In the field of media literacy, Fuentes Cancell et al. (2021) highlight that effective use of DSN in higher education requires not only technological skills, but also critical thinking to assess the impact and credibility of media.

In summary, the complementarity between the MRCDD 2.2 and Ferrés and Piscitelli's model justifies their integration to assess TMC in scientific dissemination in DSN, consolidating a robust framework for teacher training in digital contexts. This combination makes it possible to consolidate a solid model for teacher training in digital environments, aligned with the current demands of scholarly communication.

Based on these references, this study addresses the need to strengthen the participation of teachers in the communication and dissemination of knowledge through DSN, contributing to the consolidation of their academic and professional impact (Hidalgo & Hidalgo, 2024; Jiang & Yu, 2024). To this end, the design and validation of two specific constructs is proposed through a theoretical and empirical approach. The findings will allow structuring an evaluation model applicable to both teacher training and future research on media literacy and scientific dissemination in digital environments.

Research problem and objectives

In this context, the study poses the following problem: How is TMC structured for science popularisation in DSN, with a special focus on LinkedIn, and what are the implications for teacher education? It also examines the need to incorporate this competence into initial and in-service teacher education programmes.

To answer this question, the study aims to achieve the following specific objectives: (1) to design and validate two constructs that define the TMC for scientific dissemination in DSN and LinkedIn; (2) to analyse the factorial structure of these constructs; (3) to assess the reliability and validity of the indicators that compose them; and (4) to identify the pedagogical implications derived from the model for teacher training in digital contexts.

METHODS

Research design

This study follows an exploratory approach based on conceptual adequacy and structural validation of constructs (Hernández-Sampieri & Mendoza, 2018). The design is structured in three phases: (1) theoretical validation and expert evaluation of the indicators; (2) analysis of the factorial structure of the constructs; and (3) evaluation of the reliability and validity of the defined indicators.

Instruments

Two instruments were designed to assess teachers' media competence in science communication: one focused on DSN and the other on LinkedIn. Both are based on the MRCDD 2.2 and the media literacy model of Ferrés and Piscitelli (2012).

The DSN instrument covers six dimensions (language, technology, interaction, production and dissemination, ideology and values, and aesthetics) with 34 indicators, while the LinkedIn instrument, with the same structure, includes 28 indicators adapted to dissemination in professional environments. Although they share a conceptual basis, the DSN instrument assesses disclosure in general platforms, while the LinkedIn instrument focuses on communication strategies within a professional network. Both instruments were validated through expert judgement and exploratory factor analysis. Reliability was high in both cases, with internal consistency coefficients above $\alpha = 0.85$ and $\omega = 0.87$ ($\alpha = 0.91$ and $\omega = 0.93$ for DSN; $\alpha = 0.89$ and $\omega = 0.90$ for LinkedIn), ensuring their relevance and stability.

The MRCDD 2.2 is selected for its alignment with DigCompEdu and its contribution to the European Education Area, while Ferrés and Piscitelli's (2012) model is a reference in media literacy and widely cited in the scientific literature. While there are several models of competences and media literacy such as Buckingham (2003), UNESCO (Grizzle et al., 2021), Hobbs (2010) and Area and Guarro (2012), Ferrés and Piscitelli's approach stands out for its comprehensive approach and its impact on academia.

The MRCDD 2.2 structures teachers' digital competence in five areas: professional engagement, teaching and learning, assessment and feedback, student empowerment and development of digital competence. For its part, Ferrés and Piscitelli's (2012) model defines media competence in six dimensions, allowing for a comprehensive analysis of TMC in science communication (Table 1).

The instruments developed - one for dissemination in DSN and another specifically for LinkedIn - respond to complementary phases in the process of modelling teachers' media competence. They do not constitute a single consolidated instrument, but rather differentiated approaches adapted to different contexts of science communication, aimed at validating specific constructs for the future development of a comprehensive model.

Table 1

Correspondence between the Dimensions of Ferrés & Piscitelli (2012) and the MRCDD 2.2

Dimensions of media competence	Areas of the MRCDD 2.2	Relationship and similarities
1. Languages	Area 2: Digital content	Media competence emphasises literacy in multiple languages (audio-visual, digital, iconic, etc.), while digital competence in teaching promotes the creation, selection and use of digital content in teaching.
2. Technology	Area 1: Professional engagement Area 3: Teaching and learning	Media competence considers technology as a tool for communication, whereas the MRCDD emphasises its pedagogical use, its integration in teaching and its professional application.
3. Interaction processes	Area 5: Empowering learners Area 6: Developing learners' digital competence	Both models emphasise interaction in digital environments. The MRCDD highlights collaboration and digital citizenship, while media competence addresses active participation in media and networks.
4. Production and dissemination processes	Area 2: Digital content Area 6: Development of learners' digital competence	The production of own messages and interaction with others is key in both models, linking the creation of digital content with digital literacy.
5. Ideology and values	Area 5: Empowerment of learners Area 6: Developing students' digital competence	Critical analysis of media and understanding the underlying ideology of messages are related to the formation of responsible and critical digital citizens.
6. Aesthetic dimension	Area 2: Digital content	Media competence emphasises the aesthetic value of media, which is linked to teachers' ability to design attractive and effective digital content in the MRCDD.

Source: Prepared by the authors.

Procedure

The theoretical validation of the constructs was carried out by means of a literature review and expert judgement (September 2024 to January 2025), ensuring their alignment with the MRCDD 2.2 and the model of Ferrés and Piscitelli (2012). To this end, a digital questionnaire was designed in which 30 experts assessed the clarity, relevance and coherence of the indicators. The process guaranteed anonymity, confidentiality and compliance with ethical criteria. The selection of experts with extensive experience allowed for a robust validation of the content, the results of which are detailed in the data analysis section.

Data analysis

The sample was selected by purposive sampling, considering the trajectory in media literacy, educational technology or teacher training. Thirty specialists from universities in Spain, Portugal and Latin America participated, 70% with more than 10 years of experience. The inclusion criteria were: relevant publications or projects and

experience in instrument validation. The sample, composed of 60% men and with a mean age of 45.2 years ($\sigma = 7.1$), provided disciplinary and geographical diversity, guaranteeing a comprehensive expert judgement. For validation, a digital questionnaire was applied to assess the clarity, relevance and coherence of the indicators, following ethical and confidentiality criteria.

In terms of professional background, 70% of the participants had more than 10 years of experience in university teaching and teacher training in digital environments. In addition, 20% were specialised in media literacy research, and 10% in the development of technologies applied to education. The experts were selected on the basis of their experience and background in the aforementioned fields, ensuring the representation of different academic and professional profiles. This approach allowed for a multidimensional evaluation of the instrument, guaranteeing its validity from a holistic and transdisciplinary perspective.

The sample of 30 experts used in this study is considered methodologically adequate for content validation, in line with the recommendations of the specialised literature on instrument development. Authors such as DiFazio et al. (2018) and Polit and Beck (2006) argue that between 5 and 30 experts are sufficient for this type of analysis, provided that thematic specialisation and professional diversity of the participants is guaranteed. Instead of opting for large samples with non-specialised participants, as in other similar studies, the quality and relevance of the expert judgement was prioritised by selecting professionals with a consolidated track record in media literacy, educational technology and teacher training. This choice responds to the need to validate highly specialised constructs with a rigorous theoretical approach. Robust statistical procedures such as the Content Validity Index (CVI) and the Fleiss Kappa concordance coefficient (Fleiss et al., 2003), which provide quantitative evidence of the level of agreement between evaluators, even in small but highly qualified samples, were also applied.

Based on expert feedback, adjustments were made to improve the accuracy of the instrument, redundancies were eliminated by merging indicators with conceptual overlap, and items on critical thinking and source verification were incorporated to strengthen media literacy. These adjustments guarantee the theoretical and methodological validity of the constructs, ensuring their alignment with reference models and their applicability in teacher education.

Exploratory Factor Analysis (EFA) using SPSS v. 29.0 allowed us to assess construct validity and identify the underlying structure of the indicators (López-Aguado & Gutiérrez-Provecho, 2019). The Kaiser-Meyer-Olkin sample adequacy criteria and Bartlett's test of sphericity were used. Using the Maximum Likelihood (ML) method with Varimax rotation, factors were identified to ensure their independence. The reliability of the instrument was assessed using Cronbach's Alpha (α) and McDonald's Omega (ω) internal consistency coefficients (Viladrich et al., 2017). Given its exploratory nature, the EFA was applied to a purposive sample of 30 experts. Although the results provide clues about the factor structure of the constructs, this sample does not allow for conclusive validation. Therefore, it will be necessary to replicate the analysis with samples larger than 200 university teachers in future studies.

As a visual complement to the quantitative analysis, heat maps were developed to graphically represent the intensity of the factor loadings between the indicators and the factors extracted in the EFAs. This resource facilitates the identification of empirical groupings and possible cross-loadings, reinforcing the interpretation of the

results. It should be noted that, as this is a study based on EFA, neither structural equations nor latent model diagrams are presented, as these correspond to the confirmatory approach (CFA), which is specific to later phases of empirical validation

RESULTS

Theoretical validation and concordance analysis

The TMC construct for scientific dissemination in DSN is structured according to the dimensions proposed by Ferrés and Piscitelli, (2012), describing its indicators and relationship with the MRCDD 2.2 (Table 2).

Table 2

Dimension and indicators of the TMC construct for scientific dissemination in DSN

Id	Domain	Indicators
<i>Language Dimension</i>		
Id.01	Analysis	Identify the typology of scientific information suitable for publication in different DSN.
Id.02		Assess which information published in DSN is relevant and aligned with professional objectives.
Id.03		Select DSN functionalities according to scientific dissemination objectives.
Id.04	Expression	Synthesise academic content in DSN highlighting objective, scope and main results.
Id.05		Design infographics, posters, animated images and videos to disseminate scientific results.
Id.06		Transform and adapts existing digital resources in DSN for scientific dissemination.
Id.07		Write informative content in DSN using clear, accessible language adapted to different audiences.
<i>Technology Dimension</i>		
Id.08	Analysis	Understand key functionalities of DSN to disseminate research activity.
Id.09		Distinguish professional DSN and manage its basic functionalities.
Id.10		Select specific DSN to publish research results according to defined objectives.
Id.11	Expression	Use digital tools to schedule publications and analyse impact metrics in scientific DSN.
Id.12		Use hypermedia, transmedia and multimodal environments for scientific dissemination.
Id.13		Design digital resources (images, videos, infographics) for scientific results.
Id.14		Publish scientific results in DSN in an effective way.
<i>Interaction Dimension</i>		
Id.15	Analysis	Select appropriate scientific content for dissemination in DSN.
Id.16		Identify attractive digital resources for dissemination in DSN.
Id.17		Perceive motivations and emotions in DSN interactions.
Id.18		Apply ethical and copyright standards in DSN content.
Id.19	Expression	Encourage collaboration with other researchers through DSN.
Id.20		Interact with cultural diversity while respecting particularities in DSN.
Id.21		Design strategies to promote participation in scientific publications in DSN.

Id	Domain	Indicators
<i>Production and dissemination dimension</i>		
Id.22	Analysis	Select sources for publication in DSN (personal, group or institutional accounts).
Id.23		Select appropriate strategies for the dissemination of scientific content on multiple digital platforms.
Id.24	Expression	Collaborate in the dissemination of scientific production using DSN functionalities.
Id.25		Share scientific results in DSN in a comprehensible way.
Id.26		Produce scientific content adapted to the language of DSN.
<i>Ideology and values</i>		
Id.27	Analysis	Evaluate reliability and relevance of scientific information in DSN.
Id.28		Contrast scientific information published in DSN with academic standards.
Id.29		Identify information in DSN for purposes contradictory to scientific standards.
Id.30	Expression	Write content in DSN respecting ethics and scientific principles.
Id.31		Interact from an inclusive and respectful perspective with scientific divergences.
<i>Aesthetic Dimension</i>		
Id.32	Analysis	Adapt scientific content to the particularities of each DSN without duplicating information.
Id.33	Expression	Design engaging content that generates interest in diverse audiences.
Id.34		Produce creative and original messages that promote scientific debate in DSN.

Source: Prepared by the authors.

Once the indicators have been defined, their levels of achievement are established according to the MRCDD 2.2, and examples of performance are presented (Table 3).

Table 3
Levels of progression for TMC in DSN

Stage	Level	Indicators of achievement	Examples of performance
Initial	A1 - Basic	Identifies appropriate DSNs for science outreach, posts basic content, recognises relevant information.	Posts a simple scientific finding on an academic social network; follows popular science accounts.
	A2 - Exploratory	Interacts with scientific publications on different networks, explores options for format and presentation of content.	Comments on or shares scientific publications on Twitter or LinkedIn; identifies scientific trends in networks.
Intermediate	B1 - Intermediate	Designs scientific publications with images, infographics or videos; uses basic metrics to evaluate impact.	Creates a publication with explanatory graphics; reviews analytics of a publication on networks.
	B2 - Upper Intermediate	Participates in scientific debates, adapts his/her content to different platforms, follows digital ethics rules.	Actively comments on a scholarly discussion thread; adapts a Twitter post for ResearchGate.

Stage	Level	Indicators of achievement	Examples of performance
Advanced	C1 - Expert	Plans publication strategies, uses automation tools, encourages scientific interaction and collaboration in networks.	Schedules posts using tools such as Hootsuite or Buffer; create Twitter threads about a scientific study.
	C2 - Senior expert	Leads scientific debates in networks, evaluates the impact of their dissemination and designs advanced engagement strategies.	Organises a Twitter Space on a research topic; uses LinkedIn Analytics to improve scientific dissemination.

Source: Prepared by the authors.

The construct: TMC for scientific dissemination on LinkedIn was adapted to the specificity of LinkedIn, defining 28 indicators and levels of progression. Table 4 shows its 28 indicators.

Table 4

Dimensions and indicators of the TMC construct for science outreach on LinkedIn

Id	Scope	Indicators
<i>Language dimension</i>		
Id.35	Expression	Creates professional publications that highlight scientific achievements and their relevance.
Id.36		Designs short, engaging presentations for publications in LinkedIn Stories format.
Id.37		Writes scientific blog posts on LinkedIn while maintaining an accessible and professional tone.
Id.38	Analysis	Identifies relevant LinkedIn groups and communities for sharing scientific research.
Id.39		Recognises prominent scientific publications within the professional network.
Id.40		Synthesises scientific research into publications that combine professional text and graphics.
<i>Technology Dimension</i>		
Id.41	Expression	Uses multimedia tools to enrich publications (videos, presentations, infographics).
Id.42		Participates in virtual events or webinars organised through LinkedIn Live.
Id.43		Automates interaction tracking through integrations with contact management tools.
Id.44	Analytics	Uses digital tools to schedule posts and analyses impact metrics on LinkedIn.
Id.45		Understands the functionality of LinkedIn Analytics to assess the impact of scientific publications.
Id.46		Identifies professional networking opportunities based on shared research interests.
<i>Interaction Process Dimension</i>		
Id.47	Expression	Interacts with scientific publications through comments that add professional value.
Id.48		Participates in scientific discussions within professional groups on LinkedIn.
Id.49		Collaborates on shared projects using LinkedIn features such as group messaging.
Id.50	Analysis	Devises strategies to encourage participation in scientific publications on LinkedIn.
Id.51		Identifies key profiles of experts and institutions in the research area.
Id.52		Recognises high impact publications to generate scientific discussions on LinkedIn.

Id	Scope	Indicators
<i>Production and Dissemination Process Dimension</i>		
Id.53	Expression	Designs specific publications to highlight scientific results in the LinkedIn feed.
Id.54		Creates content for LinkedIn that meets copyright and confidentiality standards.
Id.55		Selects appropriate strategies for disseminating scientific content on LinkedIn and other professional platforms.
Id.56	Analysis	Selects appropriate publication formats for sharing scientific results (articles, videos, etc.).
<i>Ideology and Values Dimension</i>		
Id.57	Expression	Writes inclusive and ethical scientific content that respects professional diversity.
Id.58		Acts responsibly in sharing research, respecting intellectual property.
Id.59	Analysis	Evaluates the credibility of profiles and sources of information shared on LinkedIn.
Id.60		Identifies relevant content that promotes ethical and scientific values.
<i>Aesthetic Dimension</i>		
Id.61	Expression	Creates original posts that motivate professional interaction and participation on LinkedIn.
Id.62	Analysis	Designs content that projects professionalism and visual appeal (high quality images and graphics).
<i>Source:</i> Prepared by the authors.		

Source: Prepared by the authors.

Once the indicators were defined, their levels of achievement and examples of performance were established (Table 5).

Table 5: Levels of progression of the second construct
Levels of progression of the second construct

Stages of progression	Levels of progression	Indicators of achievement	Examples of performance
Initial	A1 - Basic	Creates professional publications that highlight scientific achievements and their relevance; Recognises outstanding scientific publications; Identifies key profiles of experts and institutions; Understands the basic functionality of LinkedIn.	Publishes a personal scientific achievement on LinkedIn; Identifies a relevant scientific publication and share it.
	A2 - Explorer	Identifies relevant LinkedIn groups and communities; Interacts with scientific publications through comments; Selects appropriate formats; Creates original posts that motivate interaction. Explores LinkedIn Analytics to identify trends in scientific publications.	Joins a LinkedIn group and comment on relevant posts; Chooses to use text or video to share a result.

Intermediate	B1 - Intermediate	Writes scientific blog articles; Synthesises research in publications with text and graphics; Uses multimedia tools; Evaluates credibility of profiles and sources; Designs visually appealing content.	Writes a LinkedIn Blog article; Combines infographics and text to synthesise research; Evaluates a researcher profile before going online.
	B2 - Upper Intermediate	Designs short presentations for LinkedIn Stories; Participates in scientific group discussions; Designs targeted posts to highlight results; Creates ethical and copyright compliant content. Plans a publishing strategy using scheduling tools.	Designs a LinkedIn Story about a scientific breakthrough; comments in a group discussion about an academic publication; creates a publication summarising a scientific result with high-quality graphics.
Advanced	C1 - Expert	Participates in virtual events via LinkedIn Live; Identifies connections based on shared interests; Collaborates on projects using LinkedIn features; Creates inclusive and ethical posts.	Conducts a seminar on LinkedIn Live; Creates an inclusive publication celebrating diversity in scientific research; Sends group messages to collaborators on a joint project.
	C2 - Expert top	Automates interaction tracking; Understands LinkedIn Analytics to assess impact; Designs advanced outreach strategies; Leads scientific debates and generates impact. Implements engagement strategies through questions, polls and interactive content on LinkedIn.	Sets up automation with external tools to manage interactions; uses LinkedIn Analytics to measure the impact of a post; leads a professional group discussion on a high-impact scientific topic.

Source: Prepared by the authors.

Exploratory Factor Analysis

The suitability of the data for factor analysis was confirmed by KMO (0.85 and 0.83) and Bartlett's test of sphericity. All factor loadings exceeded 0.40, ensuring a significant contribution of the items to their respective items. The identified factors explained 78.2 % for the TMC construct in DSN (Table 6) and 78.8 % of the total variance in DSN and LinkedIn (Table 7), respectively.

Table 6

Exploratory Factor Analysis of the first construct

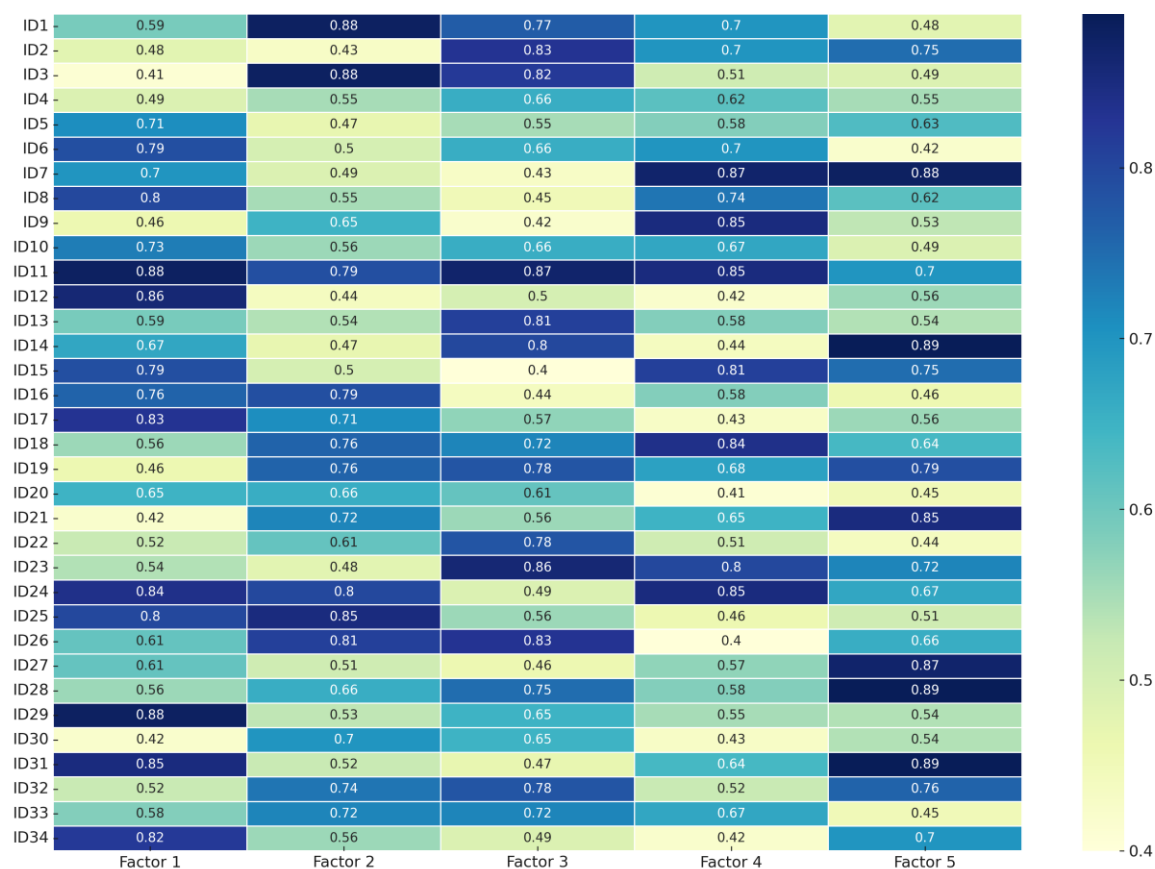
Indicator	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
ID1	0.59	0.88	0.77	0.70	0.48
ID2	0.48	0.43	0.83	0.70	0.75
ID3	0.41	0.88	0.82	0.51	0.49
ID4	0.49	0.55	0.66	0.62	0.55
ID5	0.71	0.47	0.55	0.58	0.63
ID6	0.79	0.50	0.66	0.70	0.42

Indicator	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
ID7	0.70	0.49	0.43	0.87	0.88
ID8	0.80	0.55	0.45	0.74	0.62
ID9	0.46	0.65	0.42	0.85	0.53
ID10	0.73	0.56	0.66	0.67	0.49
ID11	0.88	0.79	0.87	0.85	0.70
ID12	0.86	0.44	0.50	0.42	0.56
ID13	0.59	0.54	0.81	0.58	0.54
ID14	0.67	0.47	0.80	0.44	0.89
ID15	0.79	0.50	0.40	0.81	0.75
ID16	0.76	0.79	0.44	0.58	0.46
ID17	0.83	0.71	0.57	0.43	0.56
ID18	0.56	0.76	0.72	0.84	0.64
ID19	0.46	0.76	0.78	0.68	0.79
ID20	0.65	0.66	0.61	0.41	0.45
ID21	0.42	0.72	0.56	0.65	0.85
ID22	0.52	0.61	0.78	0.51	0.44
ID23	0.54	0.48	0.86	0.80	0.72
ID24	0.84	0.80	0.49	0.85	0.67
ID25	0.80	0.85	0.56	0.46	0.51
ID26	0.61	0.81	0.83	0.40	0.66
ID27	0.61	0.51	0.46	0.57	0.87
ID28	0.56	0.66	0.75	0.58	0.89
ID29	0.88	0.53	0.65	0.55	0.54
ID30	0.42	0.70	0.65	0.43	0.54
ID31	0.85	0.52	0.47	0.64	0.89
ID32	0.52	0.74	0.78	0.52	0.76
ID33	0.58	0.72	0.72	0.67	0.45
ID34	0.82	0.56	0.49	0.42	0.70

Source: Prepared by the authors.

In order to facilitate the visual interpretation of the factor structure resulting from the EFA, a heat map was elaborated representing the intensity of the factor loadings of the 34 indicators on the five empirically identified factors. This representation allows us to observe clustering patterns, identify cross-loaded items and verify the internal consistency of the proposed model. Figure 1 shows the heat map of the factor loadings corresponding to the TMC construct for DSN. This representation allows visualising the intensity of the saturations and the internal consistency of the factors, facilitating the identification of clustering and possible cross-loadings. Its interpretation complements the EFA results and provides graphical evidence of the model's consistency.

Figure 1
Heat map of factor loadings of the first construct



As can be seen in the graph, most of the indicators show high loadings on a single factor, which reinforces the validity of the proposed empirical structure. However, some items show relevant loadings on more than one dimension, suggesting a possible interdependence between the components of the model, especially between the dimensions of interaction, technology and production and diffusion. Table 7, corresponding to the EFA of the second construct, focusing on teachers' media competence for scientific dissemination on LinkedIn, is presented below.

Table 7
Exploratory Factor Analysis of the second construct

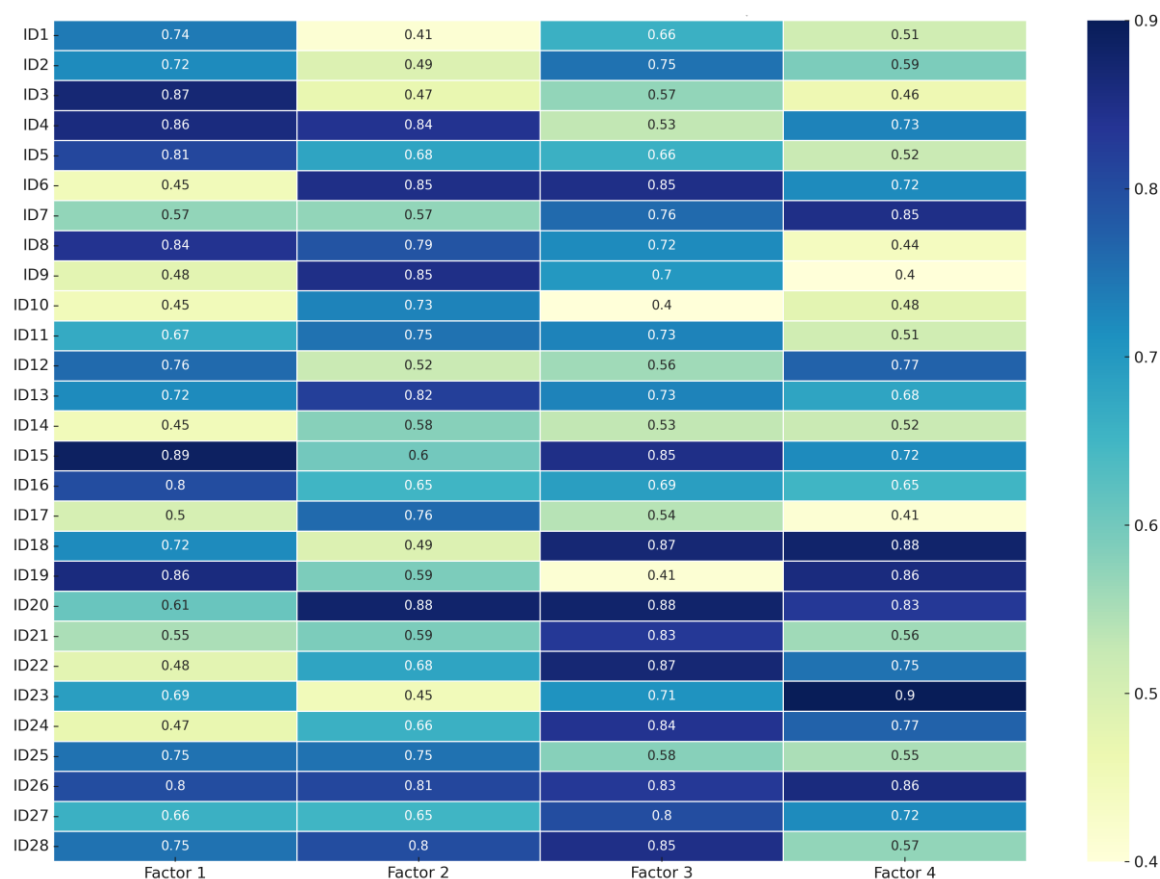
Indicator	Factor 1	Factor 2	Factor 3	Factor 4
ID1	0.74	0.41	0.66	0.51
ID2	0.72	0.49	0.75	0.59
ID3	0.87	0.47	0.57	0.46
ID4	0.86	0.84	0.53	0.73
ID5	0.81	0.68	0.66	0.52
ID6	0.45	0.85	0.85	0.72
ID7	0.57	0.57	0.76	0.85

ID8	0.84	0.79	0.72	0.44
ID9	0.48	0.85	0.70	0.40
ID10	0.45	0.73	0.40	0.48
ID11	0.67	0.75	0.73	0.51
ID12	0.76	0.52	0.56	0.77
ID13	0.72	0.82	0.73	0.68
ID14	0.45	0.58	0.53	0.52
ID15	0.89	0.60	0.85	0.72
ID16	0.80	0.65	0.69	0.65
ID17	0.50	0.76	0.54	0.41
ID18	0.72	0.49	0.87	0.88
ID19	0.86	0.59	0.41	0.86
ID20	0.61	0.88	0.88	0.83
ID21	0.55	0.59	0.83	0.56
ID22	0.48	0.68	0.87	0.75
ID23	0.69	0.45	0.71	0.90
ID24	0.47	0.66	0.84	0.77
ID25	0.75	0.75	0.58	0.55
ID26	0.80	0.81	0.83	0.86
ID27	0.66	0.65	0.80	0.72
ID28	0.75	0.80	0.85	0.57

Source: Prepared by the authors.

Although the TMC theoretical model for science outreach in DSN was designed with six dimensions (Language, Technology, Interaction, Production and Dissemination, Ideology and Values, Aesthetics), the EFA revealed a different empirical structure. This suggests that some theoretical dimensions may have been grouped into the same factor in practice, reflecting the interdependence between them. In particular, dimensions such as Interaction, Technology and Production and Dissemination clustered together, indicating that technological use directly influences the way knowledge is interacted with and disseminated.

To complement the matrix of factor loadings obtained in the EFA of the second construct, a heat map (Figure 2) was produced showing the concentration of high loadings on specific factors. This visual analysis reinforces the preliminary validity of the model and allows observing the empirical consistency of the indicators.

Figure 2*Heat map of factor loadings of the second construct*

The heat map for the second construct reveals, as in the previous case, a clear concentration of high loadings around specific factors, indicating a good empirical definition of the proposed dimensions. The consistency of the results obtained in both exploratory factor analyses supports the preliminary structural validity of the instruments developed. This empirical evidence strengthens the applicability of the model in training contexts and provides a solid basis for future research into the assessment of teachers' media competence in DSN.

Reliability assessment

Reliability analysis confirmed the internal consistency of the instrument, with coefficients of $\alpha \geq 0.70$ and $\omega \geq 0.70$ in all dimensions.

In the first construct (TMC for science outreach on DSN), α values ranged between 0.84 and 0.89, with an overall coefficient of 0.91, while ω ranged between 0.86 and 0.91, reaching an overall value of 0.93. For the second construct (TMC for science outreach on LinkedIn), the coefficients per dimension showed a similar trend, with an overall α of 0.89 and an overall ω of 0.90. These values indicate a high internal consistency across dimensions, which reinforces the stability of the model (Table 8).

The item-total correlation analysis showed coefficients above 0.30 for all indicators, which guarantees their adequate discrimination within each dimension. No

items were identified with critical values suggesting their elimination or modification. This confirms that each item contributes significantly to the measurement of the construct to which it belongs.

Table 8

Results of Cronbach's Alpha and McDonald's Omega analysis

Construct	Dimension	Cronbach's alpha (α)	McDonald's Omega (ω)
TMC for scientific dissemination in DSN	Language	0.86	0.88
	Technology	0.89	0.91
	Interaction	0.85	0.87
	Production and Dissemination	0.88	0.90
	Ideology and Values	0.87	0.89
	Aesthetics	0.84	0.86
	Overall	0.91	0.93
TMC for science outreach on LinkedIn	Language	0.85	0.87
	Technology	0.88	0.89
	Interaction	0.84	0.85
	Production and Dissemination	0.87	0.88
	Ideology and Values	0.86	0.87
	Aesthetics	0.83	0.84
	Overall	0.89	0.90

Source: Prepared by the authors.

The EFA results showed a reduction from the initial six theoretical dimensions to five factors in DSN and four in LinkedIn. This grouping suggests conceptual integrations, particularly between the interaction, technology and production/dissemination dimensions, which tend to converge empirically. This pattern could be explained by the functional convergence that characterises teaching practice in digital environments, where networked interaction, strategic use of technology and content creation develop in an interrelated way. Moreover, the professional and academic nature of platforms such as LinkedIn seems to enhance this integration, blurring the boundaries between dimensions theoretically envisaged. These findings advance the empirical understanding of teachers' media competence, providing a basis for its critical interpretation and possible application in teacher education and scientific dissemination. Theoretical and practical implications are discussed below.

DISCUSSION AND CONCLUSIONS

The results obtained offer an empirical approach to the structure of the TMC for scientific dissemination in DSN, differentiating between the generalist use of digital platforms and scholarly communication in professional environments such as LinkedIn. The reduction of the initial theoretical dimensions and the observed groupings suggest the need to reconsider certain conceptual overlaps, in line with digital teaching practice and the functional integration of interaction, technology and content production. The high reliability indices and preliminary factorial consistency support the theoretical relevance of the instrument, in line with the MRCDD 2.2 and media literacy frameworks (Ferrés & Piscitelli, 2012). These findings are in line with previous studies that highlight the importance of strengthening TMC in academic contexts (Cabero-Almenara et al., 2020; Suelves et al., 2022).

These findings are in line with previous research that has addressed the assessment and development of digital competence in teaching from psychometric and formative perspectives. For example, Cabero-Almenara et al. (2020) validated competence frameworks through expert judgement, highlighting the importance of the analysis of internal consistency and relevance of indicators. Palacios-Rodríguez et al. (2025) conducted a comparative macroassessment between Spain and Portugal based on the DigCompEdu framework, underlining the need for more specific diagnostic approaches. At the conceptual level, studies such as Marcelo-Martínez et al. (2025) have highlighted the gap between the instrumental use of DSN and its pedagogical use in teacher education. This paper extends these contributions by integrating media literacy into the assessment of teaching competence, with a contextual specificity (LinkedIn) that has not been previously addressed.

The findings also reinforce the idea that media and digital literacy is essential in teacher education, in line with research that has assessed TMC and its pedagogical applications (Jiang & Yu, 2024; Simons et al., 2017). In particular, this study provides a differentiation between TMC in DSN and its specificity in LinkedIn, which has not been sufficiently addressed in previous studies.

However, there are discrepancies with previous studies that have addressed TMC and digital as a global construct without differentiating their specific applications in academic outreach (Ho et al., 2024; Kara et al., 2017; Luan et al., 2023; Vergili & Kara, 2024). Previous research (Palacios-Rodríguez et al., 2025) has analysed TMC in more general digital-oriented terms by highlighting media competence in general DSN and specific DSN in professional settings. This distinction brings a novel approach by highlighting that scientific communication requires specific strategies that are not fully covered in general models of teacher media literacy.

This study makes an original contribution to the field of teacher education and media literacy by proposing two distinct and empirically validated constructs for assessing teachers' media competence in science communication. Unlike previous models, the proposal presented here allows for a specific assessment for professional contexts such as LinkedIn, responding to the current demands of academic communication in digital environments. This specificity broadens the field of application of the TMC and provides a useful evaluative framework both for designing training programmes and for future research in the area.

In addition to its training applications, the validated model offers relevant implications for institutional policies in higher education. The inclusion of specific indicators on science outreach in DSN allows universities to design action plans more

aligned with the principles of open science and knowledge transfer. This is particularly relevant in the context of the Sustainable Development Goals (SDGs), which promote inclusive and accessible science communication. Institutionalising the recognition of media competence as an assessable dimension in teaching careers can not only favour the visibility of research, but also a more socially connected academic culture. In this line, the proposal is articulated as a useful tool for diagnosing needs, defining training itineraries and establishing quality indicators in digital academic dissemination, aspects that are increasingly valued by quality assurance agencies and competitive calls for funding.

The results of the EFA show a factorial structure that, although it does not coincide exactly with the theoretical proposal of six dimensions, presents an organisation of the factors that is in line with previous findings in studies on media and digital competences. Several studies (García et al., 2022; Romero-Rodríguez et al., 2019) have found or assessed similar structures, where the dimensions of technology, interaction and production tend to be grouped together due to their interdependence in digital contexts. This pattern suggests that, in practice, media competences may not manifest themselves as strictly independent constructs, but as interrelated dimensions that respond to the dynamics of the digital environment. Future studies could apply a CFA to assess the stability of the model and compare it with factor structures reported in the literature.

The integration of TMC into teacher training and institutional strategies is key to strengthening science outreach in DSN. The results confirm the robustness of the content validation, supported by clear theoretical criteria and expert judgement, and provide theoretical value for future research and training programmes. LinkedIn stands out as an essential tool for the visibility of scientific production and knowledge transfer. At the institutional level, it is necessary to recognise dissemination in DSN as a legitimate academic activity and to align teacher training with the MRCDD 2.2 in order to promote media literacy.

From an applied perspective, the instruments developed offer a starting point for the diagnosis and orientation of training plans and institutional policies aimed at ethical and effective scientific dissemination. However, the factorial evidence obtained is preliminary and needs to be contrasted with large samples before a definitive model can be consolidated. Educational policies should promote the critical use of these platforms, providing resources and training in line with the principles of open science and the Sustainable Development Goals.

Limitations and lines of future work

The main limitations of the study include the small sample size ($n = 30$), which makes it necessary to consider exploratory factor analysis (EFA) as a preliminary approximation. Although it is a purposive sample of experts specialising in media literacy, educational technology and teacher training, the factorial validity of the instrument needs to be contrasted with large and heterogeneous samples.

As a future line, a CFA is proposed to test the proposed structure of the TMC, under a second-order hierarchical model with six dimensions grouped into a general factor. The CFA will be applied to a sample of more than 300 university teachers, using robust maximum likelihood (MLR) with AMOS and assessing the goodness of fit by means of χ^2/df , CFI, TLI, RMSEA and SRMR. The empirical application of the instruments in real training contexts is also proposed in order to contrast their

diagnostic usefulness, carry out cross-validation and develop test-retest reliability studies. We also plan to extend their application to international samples and compare results between different social platforms, which will allow us to reinforce their predictive validity and comparative usefulness.

Acknowledgments

Dieter Reynaldo Fuentes Cancell has been funded under the UVa 2024 predoctoral contracts call, co-funded by Banco Santander.

REFERENCES

- Area, M., & Guarro, A. (2012). La alfabetización informacional y digital: fundamentos pedagógicos para la enseñanza y el aprendizaje competente. *Revista Española de Documentación Científica*, 35(Monográfico), 46-74. <https://doi.org/10.3989/redc.2012.mon0.977>
- Argüello-Gutiérrez, C., & Moreno-López, R. (2024). Attitudes and practices of educational researchers towards the use of social media to disseminate science. *Journal of Information Science*. Advance online publication. <https://doi.org/10.1177/01655515241245958>
- Baruffaldi, S. H., Di Maio, G., & Landoni, P. (2017). Determinants of PhD holders' use of social networking sites: An analysis based on LinkedIn. *Research Policy*, 46(4), 740-750. <https://doi.org/10.1016/j.respol.2017.01.014>
- Buckingham, D. (2003). Media education and the end of the critical consumer. *Harvard Educational Review*, 73(3), 309-327. <https://doi.org/10.17763/haer.73.3.c149w3g81t381p67>
- Buzón, V. M., Pérez, R. G., & Catalán, Á. R. (2019). Explorando factores predictores de la competencia digital en las redes sociales virtuales. *Pixel-Bit. Revista de Medios y Educación*, 56(56), 51-69. <https://doi.org/10.12795/pixelbit.2019.i56.03>
- Cabero-Almenara, J., Romero-Tena, R., & Palacios-Rodríguez, A. (2020). Evaluation of teacher digital competence frameworks through expert judgement: The use of the expert competence coefficient. *Journal of New Approaches in Educational Research*, 9(2), 275-283. <https://doi.org/10.7821/naer.2020.7.578>
- Claro, M., Castro-Grau, C., Ochoa, J. M., Hinostroza, J. E., & Cabello, P. (2024). Systematic review of quantitative research on digital competences of in-service school teachers. *Computers & Education*, 215, 105030. <https://doi.org/10.1016/j.compedu.2024.105030>
- DiFazio, R. L., Strout, T. D., Vessey, J. A., & Luloff, A. (2018). Item generation and content validity of the child-adolescent bullying scale. *Nursing Research*, 67(4), 294-304. <https://doi.org/10.1097/NNR.0000000000000283>
- Ferrés, J., & Piscitelli, A. (2012). La competencia mediática: propuesta articulada de dimensiones e indicadores. *Comunicar*, 19(38), 75-82. <https://doi.org/10.3916/C38-2012-02-08>
- Fleiss, J. L., Levin, B., & Cho Paik, M. (2003). *Statistical Methods for Rates and Proportions* (3rd ed.). John Wiley and Sons. <https://doi.org/10.1002/0471445428>
- Fuentes Cancell, D. R., Estrada Molina, O., & Delgado Yanes, N. (2021). Las redes sociales digitales: una valoración socioeducativa. Revisión sistemática. *Revista Fuentes*, 23(1), 41-52. <https://doi.org/10.12795/revistafuentes.2021.v23.i1.11947>
- Gallego, A. M., Lacerda, J. D. S., & Araujo, A. C. C. (2023). La divulgación científica en Instagram: el reto del discurso audiovisual científico ante los contenidos efímeros. *Revista de Comunicación de la*

- SEECI, 56, 148-175.
<https://doi.org/10.15198/seeci.2023.56.e823>
- García, M. Á. G., Muñoz-Repiso, A. G.-V., & Duarte, M. A. A. (2022). Competencias digitales de los docentes en formación: dimensiones y componentes que promueven su desarrollo. *Civilizar*, 22(42), e20220105.
<https://doi.org/10.22518/jour.ccsb/20220105>
- Grizzle, A., Wilson, C., Tuazon, R., Cheung, C. K., Lau, J., Fischer, R., Gordon, D., Akyempong, K., Singh, J., Carr, P. R., Stewart, K., Tayie, S., Suraj, O., Jaakkola, M., Thésée, G., Gulston, C., Andzongo Menyeng, B. P., & Zibi Fama, P. A. (2021). *Media and information literate citizens: think critically, click wisely!* UNESCO.
<https://unesdoc.unesco.org/ark:/48223/pf0000377068>
- Gümüş, M. M., & Kukul, V. (2023). Developing a digital competence scale for teachers: validity and reliability study. *Education and Information Technologies*, 28(3), 2747-2765.
<https://doi.org/10.1007/s10639-022-11213-2>
- Hernández-Sampieri, R., & Mendoza, C. (2018). *Metodología de la investigación: Las rutas cuantitativa, cualitativa y mixta* (6th ed.). McGraw-Hill.
- Hidalgo, M., & Hidalgo, M. (2024). Análisis del concepto de Competencia Digital Docente: una revisión sistemática de la literatura. *Revista Latinoamericana de Tecnología Educativa - RELATEC*, 23(1), 25-41.
<https://doi.org/10.17398/1695-288X.23.1.25>
- Ho, H. T. N., Dinh, V. H., & Phan, Q. A. (2024). The important role of New Media Literacy (NML) in Vietnam: Adaptation and Validation of the NML Scale. *Communication Reports*, 37(1), 28-44.
<https://doi.org/10.1080/08934215.2023.2272270>
- Hobbs, R. (2010). *Digital and Media Literacy: A Plan of Action*. The Aspen Institute.
<https://www.aspeninstitute.org/publications/digital-media-literacy-plan-action-2>
- Huang, L., Clarke, A., Heldsinger, N., & Tian, W. (2019). The communication role of social media in social marketing: a study of the community sustainability knowledge dissemination on LinkedIn and Twitter. *Journal of Marketing Analytics*, 7(2), 64-75.
<https://doi.org/10.1057/s41270-019-00053-8>
- Instituto Nacional de Tecnologías Educativas y de Formación del Profesorado (INTEF). (2022). *Marco de Referencia de la Competencia Digital Docente 2.2*. Ministerio de Educación y Formación Profesional.
https://intef.es/wp-content/uploads/2022/03/MRCDD_Vo6B_GTTA.pdf
- Jiang, L., & Yu, N. (2024). Developing and validating a Teachers' Digital Competence Model and Self-Assessment Instrument for secondary school teachers in China. *Education and Information Technologies*, 29(7), 8817-8842.
<https://doi.org/10.1007/s10639-023-12182-w>
- Kara, M., Caner, S., Günay Gökben, A., Cengiz, C., İlgör Şimşek, E., & Yıldırım, S. (2017). Validation of an instrument for preservice teachers and an investigation of their new media literacy. *Journal of Educational Computing Research*, 56(7), 1005-1029.
<https://doi.org/10.1177/0735633117731380>
- Knight, M. (2019). Teaching responsible social media practices in business and professional communication: the importance of LinkedIn. *Business and Professional Communication Quarterly*, 82(4), 399-400.
<https://doi.org/10.1177/2329490619884740>
- López-Aguado, M., & Gutiérrez-Provecho, L. (2019). Cómo realizar e interpretar un análisis factorial exploratorio utilizando SPSS. *REIRE Revista d'Innovació i Recerca en Educació*, 12(2), 1-14.
<https://doi.org/10.1344/reire2019.12.227057>
- Luan, L., Liang, J. C., Chai, C. S., Lin, T. B., & Dong, Y. (2023). Development of the new media literacy scale for EFL learners in China: a validation study. *Interactive Learning Environments*, 31(1), 244-257.
<https://doi.org/10.1080/10494820.2020.1774396>

- Marcelo-Martínez, P., Yot-Domínguez, C., & Yanes Cabrera, C. (2025). Conectados fuera, desconectados dentro. Las redes sociales en la formación inicial docente. *RIED-Revista Iberoamericana de Educación a Distancia*, 28(1), 83-106. <https://doi.org/10.5944/ried.28.1.41343>
- Marchal, N., Hoes, E., Klüser, K. J., Hamborg, F., Alizadeh, M., Kubli, M., & Katzenbach, C. (2025). How Negative Media Coverage Impacts Platform Governance: Evidence from Facebook, Twitter, and YouTube. *Political Communication*, 42(2), 215-233. <https://doi.org/10.1080/10584609.2024.2377992>
- Martin-Neira, J. I., Trillo-Domínguez, M., & Olvera-Lobo, M. D. (2023). El periodismo científico ante la desinformación: decálogo de buenas prácticas en el entorno digital y transmedia. *Revista ICONO 14. Revista científica de Comunicación y Tecnologías emergentes*, 21(1), 1-28. <https://doi.org/10.7195/ri14.v21i1.1949>
- Olaya Guerrero, J. C., Contreras Contreras, F., & Salinas Ponce, Á. F. B. (2025). Competencias digitales en los docentes universitarios: una revisión sistemática. *Revista InveCom*, 5(1). <https://doi.org/10.5281/zenodo.12659838>
- Ordóñez Castillo, G. A., Montesdeoca Estrada, S. J., Henríquez Mendoza, E. F., Santín Picoita, F. G., & Granda Cruz, C. A. (2024). Divulgación científica y plataformas digitales: Caso de los profesores de comunicación en tres universidades ecuatorianas. *Estudios y Perspectivas Revista Científica y Académica*, 4(3), 3147-3166. <https://doi.org/10.61384/r.c.a.v4i3.602>
- Palacios-Rodríguez, A., Llorente-Cejudo, C., Lucas, M., & Bem-haja, P. (2025). Macroevaluación de la competencia digital docente. Estudio DigCompEdu en España y Portugal. *RIED-Revista Iberoamericana de Educación a Distancia*, 28(1), 177-196. <https://doi.org/10.5944/ried.28.1.41379>
- Polit, D. F., & Beck, C. T. (2006). The content validity index: Are you sure you know what's being reported? Critique and recommendations. *Research in Nursing & Health*, 29(5), 489-497. <https://doi.org/10.1002/nur.20147>
- Romero Rodríguez, J. M., Campos Soto, M. N., & Gómez García, G. (2019). Follow me y dame like: Hábitos de uso de Instagram de los futuros maestros. *Revista Interuniversitaria de Formación del Profesorado*, 33(1), 83-96. <https://doi.org/10.47553/rifop.v33i1.72046>
- Romero-Rodríguez, L. M., Contreras-Pulido, P., & Pérez-Rodríguez, M. A. (2019). Media competencies of university professors and students. Comparison of levels in Spain, Portugal, Brazil and Venezuela. *Cultura y Educación*, 31(2), 326-368. <https://doi.org/10.1080/11356405.2019.1597564>
- Said-Hung, E., Martín-Gutiérrez, Á., & Marcano, B. (2024). A study of social media use for scientific communication and dissemination among Spanish education researchers. *Knowledge Management & E-Learning: An International Journal*, 16(2), 237-258. <https://doi.org/10.34105/j.kmel.2024.16.012>
- Simons, M., Meeus, W., & T'Sas, J. (2017). Measuring media literacy for media education: development of a questionnaire for teachers' competencies. *Journal of Media Literacy Education*, 9(1), 99-115. <https://doi.org/10.23860/JMLE-2017-9-1-7>
- Suelves, D. M., Méndez, V. G., & Mas, J. A. R. L. (2022). Análisis de la competencia digital en el futuro profesorado a través de un diseño mixto. *Revista de Educación a Distancia (RED)*, 22(70), 1-17. <https://doi.org/10.6018/red.523071>
- Tolbert, C. L. (2024). The leadership identity presented on LinkedIn. *The Journal of Values-Based Leadership*, 18(1), 11. <https://doi.org/10.22543/1948-0733.1539>
- Valencia-Oliveros, N. Y., & Martín-Gutiérrez, Á. (2024). Divulgación en RRSS de las profesoras universitarias del campo educativo en España. *Profesorado, Revista de Currículum y Formación del Profesorado*, 28(3), 311-331.

<https://doi.org/10.30827/profesorado.v28i3.29601>

Vásquez-Rocca, L. (2024). Comunicación científica en Twitter. Autoría, función y estructuras prototípicas de académicos en Chile. *Galáxia (São Paulo)*, 49, e66109. <https://doi.org/10.1590/1982-2553202466109>

Vergili, M., & Kara, M. (2024). An investigation of students and teachers' new media literacy: the contributing characteristics with the moderator role of gender. *Research and Practice in Technology Enhanced Learning*, 19, 029.

<https://doi.org/10.58459/rptel.2024.19029>

Verstappen, M., & Opgenhaffen, M. (2024). Making it fit: how science news gets remediated for Facebook and Instagram. *Journalism Studies*, 25(9), 1010-1028. <https://doi.org/10.1080/1461670X.2023.2263799>

Viladrich, C., Angulo-Brunet, A., & Doval, E. (2017). Un viaje alrededor de alfa y omega para estimar la fiabilidad de consistencia interna. *Anales de Psicología / Annals of Psychology*, 33(3), 755-782. <https://doi.org/10.6018/analesps.33.3.268401>

Date of reception: 1 June 2025

Date of acceptance: 31 July 2025

Date of approval for layout: 4 November 2025

Date of publication in OnlineFirst: 11 October 2025

Date of publication: 1 January 2026

Immersive technologies in higher education: faculty profiles and barriers to integration

Tecnologías inmersivas en la universidad: perfiles del profesorado y obstáculos para su integración



 Julio Cabero-Almenara - *Universidad de Sevilla, US (Spain)*

 Antonio Palacios-Rodríguez - *Universidad de Sevilla, US (Spain)*

 Julio Barroso-Osuna - *Universidad de Sevilla, US (Spain)*

 Carmen Siles-Rojas - *Universidad de Sevilla, US (Spain)*

ABSTRACT

Introduction: The study analyses the barriers perceived by university faculties to integrate immersive technologies (AR, VR and MRI) in higher education. Despite their pedagogical potential, these technologies face technical, pedagogical, economic, institutional and ethical/social barriers. **Methodology:** A quantitative, non-experimental approach was used by means of an online survey of 775 teachers from Spanish and Latin American universities with experience in XR technologies. The instrument included socio-demographic variables and 23 items on perceived difficulties, analysed by Multiple Correspondence Analysis (MCA). **Results:** The MCA identified four teacher profiles: (1) critical and experienced, (2) average or transitional, (3) technopositive or innovative, and (4) selective critical. Economic difficulties were the most prominent, followed by pedagogical and technical difficulties. Perceptions varied according to age, gender, subject area and institutional context. **Discussion:** Barriers are not homogeneous and respond to structural and cultural factors. Technical and economic difficulties affect older teachers or those from institutions with fewer resources. Pedagogical and ethical barriers are of particular concern to teachers in the humanities and social sciences. Institutional resistance to change also emerges as a key obstacle. **Conclusions:** The study evidences the need for differentiated training strategies and institutional support. It is recommended to move towards longitudinal and qualitative research that delves into the evolution of these perceptions and the impact of educational innovation policies.

Keywords: immersive technologies; higher education; virtual and augmented reality; difficulties; multiple correspondence analysis.

RESUMEN

Introducción: El estudio analiza las barreras percibidas por el profesorado universitario para integrar tecnologías inmersivas (RA, RV y RM) en la educación superior. A pesar de su potencial pedagógico, estas tecnologías enfrentan obstáculos técnicos, pedagógicos, económicos, institucionales y éticos/sociales. **Metodología:** Se empleó un enfoque cuantitativo, no experimental, mediante encuesta online a 775 docentes de universidades españolas e iberoamericanas con experiencia en tecnologías XR. El instrumento incluyó variables sociodemográficas y 23 ítems sobre dificultades percibidas, analizados mediante Análisis de Correspondencias Múltiples (ACM). **Resultados:** El ACM permitió identificar cuatro perfiles docentes: (1) crítico y experimentado, (2) promedio o de transición, (3) tecnopositivo o innovador, y (4) crítico selectivo. Las dificultades económicas fueron las más destacadas, seguidas de las pedagógicas y técnicas. Las percepciones variaron según edad, género, área disciplinar y contexto institucional. **Discusión:** Las barreras no son homogéneas y responden a factores estructurales y culturales. Las dificultades técnicas y económicas afectan más a docentes mayores o de instituciones con menos recursos. Las pedagógicas y éticas preocupan especialmente a docentes de Humanidades y Ciencias Sociales. La resistencia institucional al cambio también emerge como un obstáculo clave. **Conclusiones:** El estudio evidencia la necesidad de estrategias diferenciadas de formación y apoyo institucional. Se recomienda avanzar hacia investigaciones longitudinales y cualitativas que profundicen en la evolución de estas percepciones y en el impacto de las políticas de innovación educativa.

Palabras clave: tecnologías inmersivas; educación superior; realidad virtual y aumentada; dificultades; análisis de correspondencias múltiples.

INTRODUCTION

The significance that Augmented Reality (AR), Virtual Reality (VR) and Mixed Reality (MR) are acquiring in the field of university education can be seen directly from the volume of research that has been carried out on them in recent times. The result of this research is the increase in meta-analyses and systematic reviews on different aspects of them, such as: the acquisition of learning by students (Victoria Maldonado et al., 2024; Yu & Xu, 2022), the pedagogical theories and approaches used for their design (Marougkas et al., 2024; Mohsen & Alangari, 2024), the ethical and psychological aspects involved in its use (Mogrovejo-Zambrano et al., 2024), the possibilities of its use in desktop and immersive versions (Hamilton et al., 2021; Liu et al., 2024) or the cognitive load involved in its use (Luna et al., 2023).

To clarify the terms to be used, it can be pointed out that there is a strong difference between AR and VR, "since, in the latter, virtual data replaces physical data, creating a new reality. On the contrary, in augmented reality, the two realities are superimposed in different layers of information in different formats (computer-generated images, video sequences, animations, etc.) to configure a new reality with which the person interacts" (Cabero-Almenara et al., 2022, 12). For its part, when analysing MRI, it is important to note that it is a technology that combines elements of AR and VR. Thanks to this integration, it is possible to generate virtual objects that allow users to interact with three-dimensional environments, both through total immersion in virtual spaces, characteristic of VR, and through the superimposition of digital content on the real world, characteristic of AR.

These meta-analyses also address the focus of this research: what are the difficulties perceived by teachers and school managers in incorporating them into training (Obeidallah et al., 2023; Sales et al., 2023).

Both research and systematic reviews suggest that teachers, students and school principals perceive several difficulties in incorporating these emerging technologies into university education. The first of these difficulties are financial (Angulo et al., 2023; Familoni & Onyebuchi, 2024; Menjivar et al., 2021; Mulders et al., 2020; Sales et al., 2023; Toala-Palma et al., 2020). This dimension is not limited exclusively to the initial costs associated with the implementation of these technologies, such as the acquisition of devices, the costs of placing the objects produced on specific digital platforms, the investment in specific software for their production, or the costs derived from licences for their use. It also encompasses the costs of the ongoing maintenance of these resources and their inevitable periodic updating.

In addition, the speed at which the technology used in these devices is advancing means that constant investment is required to update them and bring them into line with new standards as they emerge. Therefore, this is not just a one-off investment, but a constant economic investment, which is a clear difficulty for those institutions with limited economic resources.

On the other hand, it should not be forgotten that these economic difficulties include the difficulty of connectivity existing in certain educational institutions, since, for correct use and to achieve an acceptable sense of immersion and movement through the object, a stable internet connection and sufficient bandwidth are required (Sales et al., 2023; Subirats-Blanco & Conde del Rio, 2024).

A number of authors (Obeidallah et al., 2023; Mogrovejo-Zambrano et al., 2024) point to the lack of equity in the use of these technologies, both for institutions

and students, as another difficulty, which to some extent derives from the previous one. Therefore, we could speak of the existence of a digital divide that affects both students and educational institutions themselves (computers, virtual reality glasses, tablets and the latest generation of smartphones). This divide not only reflects differences in access to technology but also shows profound inequalities in the real possibilities of participating in learning environments produced by these technologies and, consequently, not being able to use the possibilities they offer (Cabero-Almenara et al., 2025).

To a certain extent, these aspects create another difficulty, which refers to the equity of use of these technologies, both for institutions and students (Mogrovejo-Zambrano et al., 2024). Therefore, one could speak of the existence of a digital divide between students and institutions, as not all students and institutions have access to devices to fully participate in AR, VR or MR-based activities, and this may result in their exclusion from these innovations, widening the digital divide between institutions.

Another large group of difficulties pointed out by different authors focus on the lack of teacher training in the use of AR, VR and MRI in educational contexts and the resistance to change that they show not only to incorporate digital technologies in teaching, but above all the emerging technologies that are appearing (AlGerafi et al., 2023; Marín-Díaz et al., 2022; Perifanou et al., 2023; Sales et al., 2023). This is compounded, on the one hand, by the difficulty of integrating them into learning scenarios (Mulders et al., 2020), and on the other hand, by the lack of relationship with learning theories (Caballero-Garriazo et al., 2023). All of this is a consequence of the lack of existing research on how to use and design these learning objects (Radianti et al., 2020; Sales et al., 2023) and the scarcity of these objects in open format (Upadhyay et al., 2024).

On the other hand, it should not be forgotten that in the research that has found significant results, the effect sizes are not very large (Coban et al., 2022). Although some research suggests that without proper integration of learning strategies in AR scenarios, the effectiveness of learning outcomes may be limited, or even less effective than traditional technology-supported learning methods (Jingru et al., 2025).

Another type of difficulty relates to the existence of different psychological impacts on their use, such as anxiety, fatigue and feelings of disorientation caused by certain types of use such as immersive use (Cevikbas et al., 2023; Mogrovejo-Zambrano et al., 2024; Mulders et al., 2020; Obeidallah et al., 2023) and the cognitive load that requires the learner to invest if objects are not properly designed (Angulo et al., 2023; Bermejo et al., 2023; Bautista et al., 2025; Cevikbas et al., 2023).

The results pointed out by the meta-analyses allow us to identify some major blocks of difficulties that could be determining factors in the application of AR, VR and MR technologies: technical, pedagogical, economic, ethical, social and psychological, as well as resistance to change on the part of institutions and teachers.

In any case, with a view to the integration of the technologies discussed, it is also important to consider technology analysis models such as the TAM (Technology Acceptance Model) and UTAUT (Unified Theory of Technology Adoption and Use of Technology) technology acceptance models, and the TPACK (Technological Pedagogical Content Knowledge) and SAMR (Substitution and Augmentation, Modification and Redefinition) technology integration models. Models formulated at the time by Davis (1989), Venkatesh (2000), Mishra and Koehler (2006) and Puentedura (2006), respectively.

Of all these models, TAM and UTAUT have been the ones that have sparked the most research with respect to AR, VR and RE technologies (Cabero-Almenara & Pérez Díez, 2018; Cabero-Almenara et al., 2021). They point to a series of conclusions that are fundamentally aimed at the fact that they are technologies that arouse a high degree of acceptance that leads them to show potential users a high intention to use them. With regard to the other models, they incorporate two ideas to consider: on the one hand, that a correct incorporation of these technologies requires thinking about the fact that the technology does not position itself alone, but depends on the content it transmits and the methodologies applied with it, and on the other hand, that the incorporation of any technology is not direct in teaching, but implies a slow process of incorporation.

METHOD

The research presented here is non-experimental and of the survey type (Tourrón, 2023) and aims to achieve the following objectives:

1. To identify the main difficulties perceived in the implementation of AR, VR and MR in the context of higher education, considering technical (TD), pedagogical (PD), economic (ED), ethical, social and psychological (DES) aspects, as well as resistance to change on the part of institutions and faculty (DRC).
2. To explore the association between perceived difficulties and teachers' socio-demographic characteristics, to determine differentiated typologies of teachers in perceived barriers to VR use.
3. To determine clusters or profiles of difficulty based on perceived difficulties and socio-demographic characteristics of teachers, to provide a comprehensive characterisation of barriers to the integration of VR in higher education.

Instrument

For the present research, an instrument was designed to collect information specifically adapted to the objectives of the study. This questionnaire consists of two sections. The first section is aimed at collecting socio-demographic data on the participant, such as gender, age, among other personal aspects. The second section, structured using a Likert-type scale, includes a total of 23 items designed to evaluate, on a scale of 1 to 10, the degree of perception regarding the influence of various difficulties.

These items are distributed according to the categories of difficulty presented above in the objectives to be achieved: technical difficulties (4 items), pedagogical (9 items), economic (3 items), related to resistance to change on the part of institutions and teaching staff (2 items), and those of an ethical, social and psychological nature (5 items).

For the validation of the questionnaire, a pilot test was applied to some members of the "Didactic Research Group" from the University of Seville and the "Educational Technology Research Group" from the University of Malaga, and to researchers of the MEREVIA project who had published articles and communications on the application of AR, VR and MR. Subsequently, the statistical analyses of Cronbach's alpha and McDonald's omega were applied to obtain the reliability index.

This research has all the necessary ethical safeguards to ensure the protection of the rights and privacy of the participants. Firstly, all data collected were treated strictly anonymously, ensuring that no individual can be identified from the information obtained. Furthermore, the participants were duly informed about the objectives and procedures of the study, and gave their informed consent freely and voluntarily, also accepting a specific clause on the transfer of data for exclusively scientific purposes. The protocol followed in this research complies with the ethical principles set out in the Declaration of Helsinki, ensuring the respect, dignity and protection of the individuals involved. Furthermore, Regulation (EU) 2016/679 of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation - GDPR) has been strictly followed, thus ensuring responsible, secure and transparent use of the information collected.

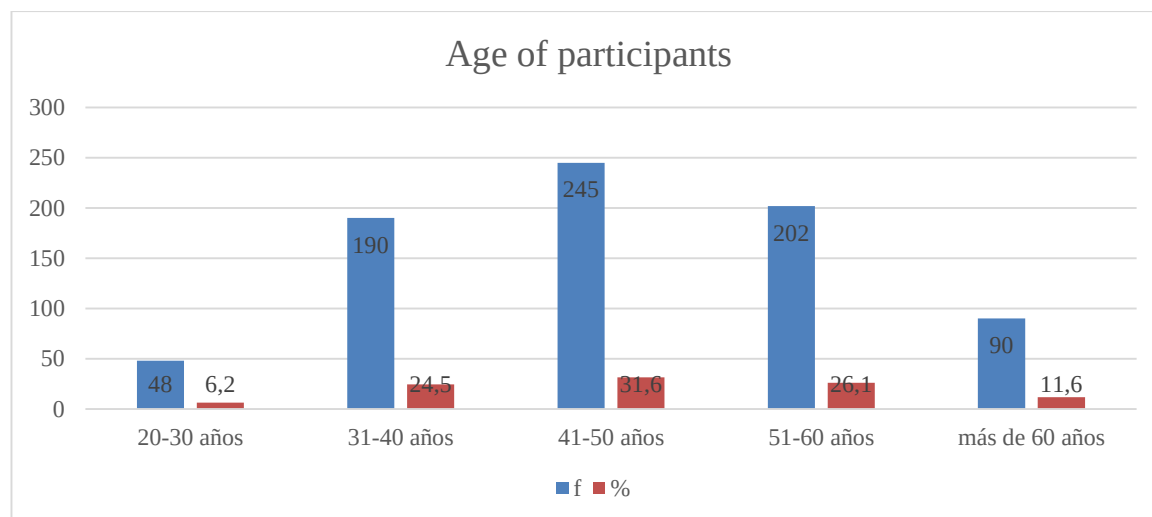
Sample

The study sample consisted of a total of 775 participants. Of these, 389 (50.2%) identified themselves as men, 378 (48.8%) as women, while 8 people (1.0%) chose not to answer the question on gender.

Figure 1 shows the frequencies and percentages of participants according to their age.

Figure 1

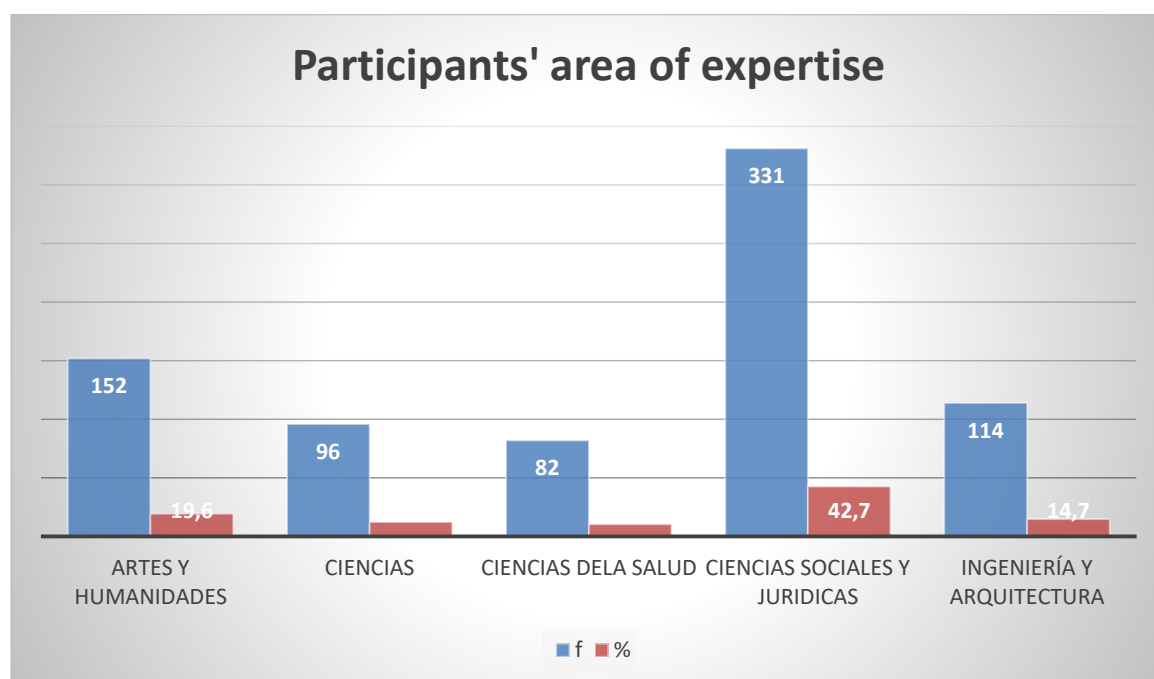
Frequency and percentage of participants' age



As can be seen, 82% of the distribution was between 31-60 years of age. Those between 41-50 years of age stood out.

Of those surveyed, 395 (51.0%) worked in Spanish universities and 380 (49.0%) in Latin American universities.

With regard to the area of knowledge to which the participants belonged, Figure 2 shows the frequencies and percentages to which they belonged.

Figure 2*Frequency and percentage of participants' age*

As can be seen, the highest percentage of participants belonged to the "Social and Legal Sciences" area of knowledge, which could be explained by the procedure followed for the selection of participants, as they tend to be one of the groups most interested in the incorporation of these technologies into educational practice.

The selection of participants was carried out by means of several strategic procedures. Firstly, we included people who had published, between 2020 and February 2025, an article related to AR, VR and RE in specialised Spanish-language journals. Secondly, those who presented papers on these technologies at the Edutec 2024 Congress were considered. Finally, teachers belonging to different research groups from Spanish and Latin American universities were also included, from whom information was available that showed evidence of their work in the incorporation of these technologies in educational contexts.

Procedure

Prior to the main analysis, a reliability study was carried out on the scales used to measure the different difficulties perceived in the use of AR, ER and MR in higher education. Cronbach's Alpha and McDonald's Omega internal reliability indices were calculated for each of the dimensions included in the questionnaire: technical difficulties (TD), pedagogical difficulties (PD), economic difficulties (ED), resistance to institutional and faculty change (DRC), and ethical, social and psychological difficulties (ESD). These coefficients allow us to estimate the internal consistency of the items that make up each dimension, and are considered robust indicators for assessing the reliability of psychometric scales (Dunn et al., 2014). The values obtained were within acceptable ranges in educational research, indicating adequate reliability for each subscale.

In addition, means and standard deviations were calculated for each dimension of perceived difficulty in order to provide an initial description of the behaviour of the data and to explore the degree of difficulty that participants attribute to each of the categories assessed. This preliminary statistical description provides a solid basis for the subsequent interpretation of the MCA results by contextualising the level of difficulty perceived by teachers in relation to technical, pedagogical, economic, institutional and ethical aspects.

For the analysis of teacher profiles and typologies, we applied the MCA statistical test, a statistic that is used when more than two categorical variables are to be analysed simultaneously and which allows both the categories of the variables and the individuals (cases or subjects) to be represented graphically in a reduced space of two or more dimensions, with the aim of visualising associations, similarities and latent structures in the data (Bond & Michailides, 1997). It has many qualitative variables (nominal or ordinal variables treated as nominal), and aims to identify characteristic profiles of individuals or groups. As Cornejo (1988, p. 97) points out, the aim is to "delve into the relationships of dependence established between two qualitative variables observed in the same population".

This approach makes it possible to identify latent patterns of association between categorical variables and to construct a factor space that graphically represents the configuration of these perceptions, thus facilitating the segmentation and interpretation of teacher profiles (Greenacre, 2007; Le Roux & Rouanet, 2010).

The use of MCA in this context is justified by the categorical nature of the data, derived from ordinal rating scales that capture perceived levels of difficulty in each dimension. The technique makes it possible to reduce the complexity of the information and to represent in a two-dimensional space the most significant relationships between categories, preserving a considerable proportion of the original variance (Abdi & Valentin, 2007; Escofier & Pagès, 2008). The analysis not only allows us to observe how responses are clustered, but also how they interact with possible socio-demographic or institutional factors, offering a more holistic view of barriers to the adoption of immersive technologies (Le Roux & Rouanet, 2010).

In social research it has been used for the analysis of different topics such as: "locus control" (Algañaraz, 2016), the image of teachers in the press (Cabero-Almenara & Loscertales-Abril, 1998), characteristics of Nobel Prize winners or mathematical problem solving as a methodology.

RESULTS

The overall reliability index of the instrument analysed through Cronbach's alpha and McDonald's omega were 0.915 and 0.909, respectively. As for the different dimensions identified, the values achieved are shown in Table 1.

Table 1
Reliability index of the instrument

Reliability index		
Variable	Cronbach's α	McDonald's Ω
Technical difficulties	.783	.778
Pedagogical difficulties	.847	.830
Economic difficulties	.789	.783

Difficulties due to resistance to change in institutions	.736	.734
Ethical, social and psychological difficulties	.799	.792

These values, according to Roco-Videla et al. (2024), indicate a high level of reliability of the instrument.

In line with the objectives set out, and with the aim of obtaining an initial descriptive view of the degree of difficulty perceived in the use of AR, VR and RM in higher education, we proceeded to analyse the mean scores and standard deviations corresponding to each of the dimensions assessed. Table 2 shows the mean scores, and standard deviations achieved for each of the difficulties mentioned and for the instrument.

Table 2

Mean and standard deviations of the different dimensions that make up the instrument

	Mean	Standard deviation
Technical difficulties	7.14	1.86
Pedagogical difficulties	7.40	1.47
Economic difficulties	8.04	1.71
Difficulties due to resistance to institutional change	7.20	2.11
Ethical, social and psychological difficulties	7.13	1.71
Total instrument	7.36	1.32

The analysis of the dimensions considering the socio-demographic variables also reveals information to be highlighted. In terms of gender, women report higher levels of difficulties in all the dimensions evaluated, standing out especially in economic difficulties (8.14) and ethical, social and psychological difficulties (7.37). By branches of knowledge, Engineering and Architecture professors had the highest mean scores in the instrument (7.62), with special emphasis on economic difficulties (8.55), while Natural Sciences reported the lowest (7.3). In relation to the type of university, Latin American teachers report more difficulties in all dimensions, with an overall average of 7.67 compared to 7.13 in the Spanish case. By age, an increasing trend is observed: those over 61 years of age report the highest scores in almost all dimensions, reaching an overall average of 7.75, especially in institutional resistance to change (8.06).

After the initial description of the mean scores in each dimension of difficulty, an MCA was carried out in order to explore in more depth the relational structure between the different perceived difficulties and the socio-demographic variables of the teachers, such as gender and age. This technique is particularly suited to address one of the central objectives of the study: to identify differential perception profiles of barriers to the integration of AR, VR and MR technologies based on personal or institutional characteristics. Since the data used comes from categorised scales (treated as qualitative), the MCA allows us to graphically visualise latent associations between categories, revealing patterns of clustering and proximity that might not be evident through traditional statistical analysis. In this sense, the MCA provides a powerful analytical framework to interpret in an integrated way the different dimensions of the phenomenon studied and to generate a meaningful segmentation of the teaching staff

according to their perception of the difficulties in adopting immersive technologies in university environments (Cardona et al., 2021).

To facilitate the interpretation of the data in the multiple correspondence analysis, the scores obtained on the difficulty scales (originally measured on a scale of 1 to 10) were transformed into qualitative categories. This recording allowed the values to be grouped into three levels: Low (scores 1 to 4), Medium (score 5 to 8) and High (score 9 to 10). This categorisation strategy responds to the need to adapt the data to the requirements of the applied factor analysis, given that the MCA works with qualitative variables, and contributes to a more intuitive reading of the profiles and associations between levels of difficulty and socio-demographic variables, without losing the informative richness of the original instrument.

As part of the analysis procedure, the ACM was subjected to an iterative optimisation process until an acceptable convergence criterion was reached. This process ensures the stability and validity of the factorial solution obtained, making it possible to adequately represent the relationships between the categories included in the analysis. Table 3 shows the history of iterations carried out during model fitting.

Table 3
Iteration history

Iteration number	Variance accounted for		
	Total	Increase	Loss
29	38.699544	.000009	24.300456

The estimation algorithm achieved stable convergence at iteration 29, with a total variance explained of 38.6% and a final marginal increase of only 0.000009, indicating that an optimal solution was reached (Greenacre, 2007). This value is common in Multiple Correspondence Analysis (MCA) when working with categorical variables related to subjective perceptions and experiences, such as the difficulties faced by teachers when integrating emerging technologies (Le Roux & Rouanet, 2010; Abdi & Valentin, 2007). The residual loss was 24.3, reflecting unexplained variability, which is to be expected in ACM and does not compromise the validity of the model (Escofier & Pagès, 2008). Overall, the results confirm the stability and reliability of the solution, validating the interpretive use of two-dimensional space to analyse the relationships between types of perceived difficulties and teacher characteristics (Greenacre, 2007). This evidence supports the design of more precise strategies for training and institutional support in the adoption of VR/AR/RM in university teaching (Le Roux & Rouanet, 2010).

To assess the internal consistency of the instrument and the quality of the factorial solution obtained through the ACM, a summary of the model is presented below (Table 4).

Table 4
Summary of the model

Dimension	Cronbach's alpha	Total (eigenvalue)	Inertia	% of variance
1	.930	10.192	.309	40.884
2	.888	7.207	.218	21.840

The analysis reveals a robust and explanatory factor structure, with two main dimensions that capture a significant part of the variability in teachers' perceptions of difficulties in integrating emerging technologies. The first dimension, with a Cronbach's alpha of 0.930 and an eigenvalue of 10.192, explains 40.884% of the variance, reflecting high internal consistency between the grouped categories (Greenacre, 2007; Le Roux & Rouanet, 2010).

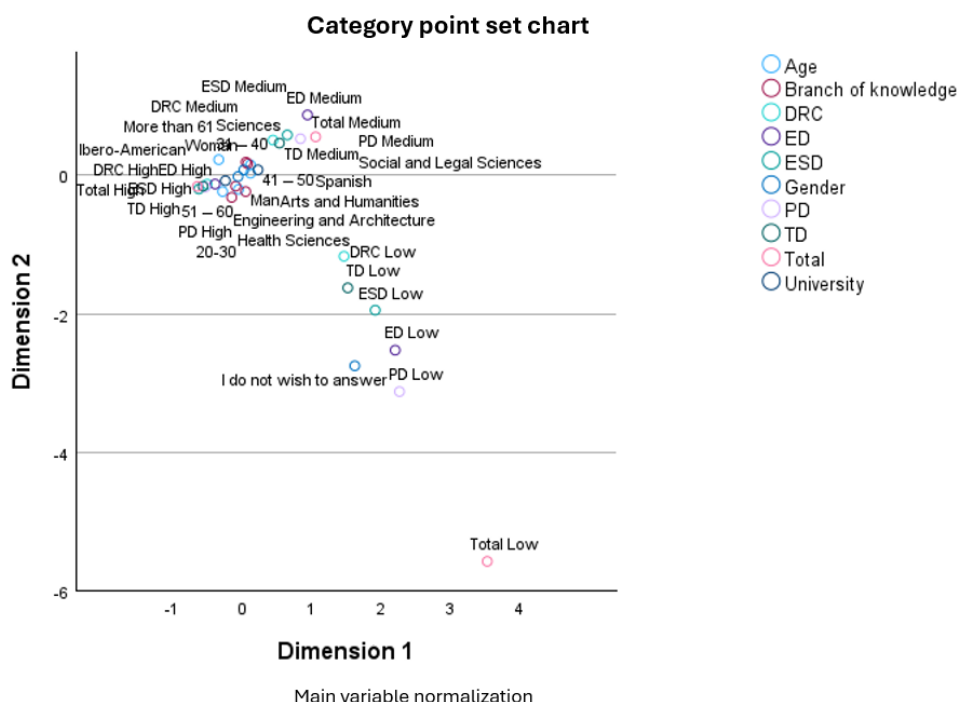
The total explained inertia reaches 76.4%, an unusually high value in the context of the MCA, given the complexity of categorical data (Abdi & Valentin, 2007; Escofier & Pagès, 2008). Furthermore, the average Cronbach's alpha (0.913) and eigenvalue (35.058) reinforce the psychometric reliability of the model, indicating that the dimensions are methodologically and conceptually sound (Greenacre, 2007).

This configuration makes it possible to identify logical groupings of barriers - technical, pedagogical, economic, institutional or ethical/social - and to profile different types of teachers according to their experience of these difficulties. The robustness of the model provides a reliable interpretative basis for analysing how variables such as type of university, age, gender or subject area influence the perception of barriers, offering valuable inputs for designing more contextualised and effective training policies, technological investment and support strategies.

Below is the joint graph of category points (Figure 3), the result of the MCA, which allows us to visually observe the associations between the different categories of perceived difficulty and the socio-demographic and academic variables analysed (age, gender, branch of knowledge, and location of the university in which they work).

Figure 3

Joint graph of category points (note: ED = Economic Difficulties, ESD = Ethical, Social and Psychological Difficulties; PD = Pedagogical Difficulties; DRC = Difficulties of Resistance to Change; TD = Technical Difficulties)



In this two-dimensional space, proximities between points indicate relationships or similarities in response profiles, while distances suggest differentiation between categories. In the following section, the most salient results of the analysis are discussed, considering the different types of difficulties considered in the study: technical, pedagogical, economic, institutional resistance to change, ethical, social and psychological.

Difficulties

The multiple correspondence analysis reveals a clear polarisation in the perception of university teaching staff regarding the different barriers to the integration of immersive technologies (AR, VR, RE). Technical (TD) and economic (ED) difficulties are associated with older, male, Latin American university lecturers, while pedagogical (PD), ethical/social/psychological (ESD) and institutional resistance to change (IRC) difficulties present a more nuanced but consistent segmentation. In general, those who perceive high levels of difficulty do so cumulatively across several dimensions, while profiles with low perceptions tend to coincide in minimising multiple barriers, indicating consistent patterns of response. These differences suggest the need for differentiated strategies for training, technological and institutional support, addressing both critical and technopositive profiles.

Total difficulties

The Total variable acts as an integrating axis summarising the overall perception of difficulties. "High Total" is grouped with high scores on all dimensions, representing critical teachers who perceive significant structural barriers to the adoption of XR technologies. "Medium Total", located near the centre of the graph, reflects a more balanced position and is representative of the majority of teachers. On the other hand, "Low Total" is positioned in a separate quadrant, next to all the other low categories, indicating a subgroup with a technopositive view, probably with more experience or institutional support. This consistency between levels reveals a latent structure in teachers' perceptions and allows for the identification of extreme and transitional profiles.

Socio-demographic variables

Socio-demographic variables show differential effects on perceptions. Age does not seem to exert a determining influence, with a balanced distribution between the different age groups. In terms of gender, men and women present similar perceptions, but the category "I do not wish to answer" stands out, which is associated with low levels of difficulty and differentiated response patterns. The branch of knowledge introduces greater diversity: disciplines such as Engineering and Architecture tend to minimise barriers, while Social Sciences and Humanities present greater ethical and pedagogical concerns. Finally, institutional location (Spain vs. Latin America) does not generate a strong differentiation, although nuances are detected that could be explored in future studies. These findings allow us to delineate useful teacher profiles for designing contextualised training interventions.

The profiles identified are described below, developing in depth the characteristics that define them, their positioning in relation to emerging technologies and the implications that these findings have for educational innovation at the university.

Profiles

Profile 1: Critical and Experienced Teacher is made up of teachers over 61 years of age, mainly in Latin America and in disciplines such as Social Sciences and Law. They present a high perception of technical, pedagogical, economic, ethical and institutional resistance barriers, adopting a critical stance towards the hasty incorporation of XR technologies without pedagogical or ethical reflection (Aquino Negrin et al., 2021; Tapia Cortes, 2020).

Profile 2: Average or Transition Teacher includes mostly women between 41 and 50 years of age in the Spanish university system. Their perception of difficulties is average in all dimensions, with an attitude open to change if they are provided with training and adequate institutional support (Cajas Bravo et al., 2023). This group is key to facilitating the technological transition from a moderate stance.

Profile 3: Technopositive or Innovative Teacher is composed of young teachers (31-40 years old), some of whom do not declare gender, and come from areas such as engineering, science or education. Their perception of barriers is low, and they adopt an active and enthusiastic attitude towards innovation, although they may underestimate ethical or social risks without critical reflection (Arancibia Muñoz et al., 2017; UNESCO, 2023).

Finally, Profile 4: Selective Critical Teacher groups together mostly women between 51 and 60 years of age, from Spain and humanistic disciplines. Although they do not reject XR technologies, they are particularly sensitive to the ethical and social dimensions, and recognise certain institutional and economic obstacles. Their humanistic and educational justice approach allows them to lead from a critical, ethical and transformative perspective (UNESCO, 2023; Angulo Salazar et al., 2025).

DISCUSSION AND CONCLUSIONS

The results obtained in this research allow us to advance in the understanding of the barriers faced by university professors in the integration of immersive technologies such as AR, VR and MRI in their teaching practices. Using the ACM, clearly differentiated profiles of teachers have been identified, whose perceptions of the difficulties in implementing these emerging technologies are influenced by socio-demographic variables such as age, gender, area of knowledge or location of the university. With regard to technical difficulties (TD), the study reveals that part of the teaching staff -especially the most senior or those linked to less technological areas- perceive significant barriers related to the use, configuration and maintenance of the technologies mentioned, as has also been pointed out in a variety of research studies (Sales et al., 2023; Subirats-Blanco & Conde del Rio, 2024). This dimension is particularly influenced by age, subject area and previous familiarity with digital environments, and is more accentuated in institutional contexts with less technological investment. Despite advances in equipment, technical gaps remain a limiting factor in certain profiles, underlining the need to reinforce technical support and practical training.

Regarding pedagogical difficulties (PD), the results indicate that many teachers face challenges when it comes to pedagogically integrating these technologies within coherent didactic frameworks, as suggested by different research, with which the results obtained here coincide (Mulders et al., 2020; Caballero-Garriazo et al., 2023). These difficulties do not depend solely on technical mastery, but on the methodological understanding of how to apply AR, VR and MRI tools to enrich teaching-learning processes. Teachers in transition and with experience in educational innovation show more receptive attitudes, but demand support to redesign their practices. This dimension reflects the urgent need to articulate training proposals focused on pedagogical transformation and not only on operational instruction.

Economic difficulties (ED) emerge as a transversal barrier, although this is particularly accentuated among young teachers or those from institutions with limited resources. The perception of lack of funding, shortage of equipment or lack of equitable access to immersive technologies generates a sense of practical impossibility to incorporate these tools. This finding coincides with the findings of a variety of research (Mulders et al., 2020; Angulo et al., 2023; Familoni & Onyebuchi, 2024; Obeidallah et al., 2023, Sales et al., 2023) regarding the significance of this type of difficulty in incorporating these technologies in university education. This dimension is closely linked to structural conditions, so overcoming it depends more on institutional decisions and public policies than on the individual action of teachers. Investing in technological infrastructure and guaranteeing its maintenance is therefore a basic condition for progress towards inclusive digitisation.

With regard to ethical, social and psychological difficulties (ESD), a teaching profile was identified that is particularly sensitive to the implications that the use of AR, VR and MRI technologies may have on student subjectivity, educational equity, privacy or the dehumanisation of pedagogical relationships, in short, a problem of equity (Mogrovejo-Zambrano et al., 2024). These concerns are especially frequent in areas such as the humanities and social sciences, and among teachers with a humanistic or critical orientation. This dimension, often forgotten in technocentric discourses, should be seriously considered in the design of innovation policies, through the inclusion of ethical frameworks and protocols of use that guarantee a responsible and person-centred implementation.

Finally, difficulties due to resistance to institutional change (DRC) constitute one of the most relevant barriers, even pointed out by teachers with a high technological disposition. The perception of rigid structures, lack of leadership in innovation, lack of recognition for teaching and excessive bureaucracy hinder motivation and the possibility of implementing the technologies considered in this study. This dimension shows that it is not enough to train teaching staff, but that it is essential to transform the organisational culture of universities, generating environments that value pedagogical experimentation, promote collaborative networks and make their structures more flexible to adapt to change.

Taken together, the five dimensions analysed allow us to understand the complexity of the phenomenon and reinforce the need to adopt a systemic and integrated approach in the promotion of technological innovation in higher education. Although the teaching profiles identified make it possible to differentiate specific attitudes and needs, the barriers cannot be addressed in isolation or individually, as many of them respond to structural and cultural conditions of the university system, on the one hand, in Spain and, on the other, in Latin America.

Another of the main findings of the study is the evidence of a solid and explanatory factor structure, which made it possible to segment teachers according to their levels of perceived difficulty -technical, pedagogical, economic, institutional and ethical/social-. This segmentation revealed not only the existence of extreme profiles, such as the technopositive or the critical and experienced teacher, but also intermediate or hybrid typologies reflecting transitional or selective critical stances. The identification of these profiles represents a significant step forward, as it provides an empirical basis for designing training, support and resource provision policies that are better adapted to the specific realities and needs of teachers.

The study also shows that, although there are differences between subject areas - for example, in the perception of technical difficulties in Engineering versus ethical ones in the Humanities - some barriers are transversal to the entire university system, such as those related to the lack of infrastructure, the scarce specific training of teaching staff or the disconnection with consolidated pedagogical frameworks. These barriers, far from being merely individual, reflect structural limitations of the university environment that need to be addressed from a comprehensive institutional perspective. The analysis also showed that, although age and professional experience are variables that influence the perception of certain difficulties, they do not act as unique or deterministic predictors, but interact with cultural, organisational and attitudinal factors.

In this sense, the use of the MCA has proven to be a valuable methodological tool to reveal the complexity of the phenomenon under study. Unlike other more unidimensional approaches, this method has made it possible to graphically visualise latent relationships between multiple variables, favouring the identification of patterns of perception that would not be easily detectable through traditional analyses. This approach is particularly relevant in the field of education, where the diversity of teacher profiles and institutional contexts demands rich and multifactorial interpretative approaches.

The practical implications derived from this study are highly relevant for the design and implementation of strategies that facilitate the effective integration of immersive technologies - AR, VR and MRI - in university teaching. Firstly, the clear identification of differentiated teaching profiles, reflecting different perceptions of technical, pedagogical, economic, ethical and organisational barriers, underlines the need to develop specific training adapted to the characteristics and needs of each group. This implies designing training programmes that not only address technical operational aspects, but also strengthen methodological skills for coherent pedagogical integration, as well as raise awareness of the ethical and social dimensions of these technologies. Furthermore, the recognition that technical difficulties predominate among teachers with less digital familiarity or in less technological areas calls for a reinforcement of continuous technical support, as well as the simplification and standardisation of tools to facilitate their use. On the other hand, economic and structural limitations, closely linked to institutional conditions, highlight the need to promote sustained investment policies in technological infrastructure that guarantee equitable access and adequate maintenance of resources, especially in universities with less financial capacity. In addition, institutional barriers related to resistance to change, and organisational rigidity require a profound cultural transformation that promotes leadership in innovation, formal recognition of teaching and the creation of collaborative spaces that encourage experimentation and the exchange of good practices. Finally, the ethical and social concerns detected demand the inclusion of

clear regulatory frameworks and protocols that guide a responsible use focused on the well-being of students, ensuring that the incorporation of immersive technologies contributes to a more inclusive, equitable and humane education.

Despite the contributions of this work, it is necessary to recognise some limitations that should be considered when interpreting the results. Firstly, the study is based on a purposive sample of teachers with experience or interest in the use of immersive technologies, which may imply a positive selection bias towards this type of tool. This means that the profiles identified could differ if the sample were extended to teachers with no previous contact with these technologies. However, this would also imply another limitation, namely the teachers' lack of knowledge of these technologies.

Secondly, the self-reported nature of the questionnaire and its online application imply certain inherent methodological limitations, such as the possibility of socially desirable responses or the difficulty in controlling homogeneous understanding of the items. Although the questionnaire showed high internal reliability, it would be convenient to contrast the results with observational studies or in-depth interviews.

Furthermore, the categorisation of the ordinal variables into levels (Low, Medium, High) responds to statistical and graphical interpretation criteria, but this recoding may have softened important nuances in the individual responses. Finally, the MCA, although robust, implies a reduction of the original information to a two-dimensional space, so that a part of the variability inevitably remains unrepresented in the interpretative planes.

Future lines of research should move towards a more longitudinal and mixed approach. It would be relevant to develop studies that not only capture the perception of barriers at a given point in time but also analyse the evolution of these perceptions after teachers' participation in training actions or practical experiences with XR technologies. This would make it possible to assess the real impact of training and infrastructure on the transformation of teachers' attitudes, beliefs and practices.

On the other hand, it would be interesting to replicate the research after a couple of years, and therefore to hold two Edutec conferences and publish new articles in scientific journals, to analyse which difficulties have increased and which have been reduced and maintained. And to carry out a comparative study between the two points in time.

It is also proposed to complement the quantitative analyses with qualitative studies that delve deeper into the subjective experience of the perceived difficulties, especially in the extreme profiles detected (for example, the technopositive teacher or the selective critic). The integration of interviews, focus groups or case analyses would enrich the understanding of the meanings that teachers attribute to these technologies, as well as the emotional, ethical or ideological factors that mediate their acceptance or rejection.

It is important to assume, for the correct interpretation of the results, that the procedure followed to obtain the participants, although it is usual in studies of this type, and the criteria used justified their selection. They undoubtedly incorporate a limitation. At the same time, the decompensation by areas of knowledge may imply a limitation, which is why it would be convenient to replicate the research with stratified sample selection procedures, with the application of the "coefficient of expert competence" (Cabero-Almenara & Barroso-Osuna, 2013) and to consider in the sample the existing percentage of teaching staff for each area of knowledge; however, the latter

poses a limitation, as we have not worked with university teaching staff, but with those who publish on the incorporation of AR, VR and MR.

Finally, it would be pertinent to extend the sample to other international university contexts, as well as to explore the relationships between the barriers identified and variables such as institutional culture, predominant pedagogical models or the type of academic leadership. This broader and more comparative view would make it possible to generate transferable interpretative frameworks to guide the design of global and local policies that are more coherent with the real needs of university teaching staff in the era of digital transformation. In this sense, it is essential to structure specific training programmes that respond to the different levels of technological mastery and to the teaching profiles identified, so that training is truly contextualised and effective, addressing both those who are in the initial stages of familiarisation with immersive technologies and those with greater experience or a critical orientation. As a future line, it is also proposed to incorporate complementary statistical analyses that make it possible to contrast significant differences between teaching profiles and socio-demographic variables, to further refine training and support strategies.

Funding

The research is funded through the State Programme to Promote Scientific and Technological Research and its Transfer, within the framework of the State Plan for Scientific, Technical and Innovation Research 2021-2023. Ministry of Science and Innovation (MEREVIA (PID2022-136430OB7-IOO)).

REFERENCES

- Abdi, H., & Valentin, D. (2007). Multiple correspondence analysis. In N. J. Salkind (Ed.), *Encyclopedia of measurement and statistics* (pp. 651-657). SAGE Publications.
- Algañaraz, V. H. (2016). El “Análisis de Correspondencias Múltiples” como herramienta metodológica de síntesis teórica y empírica. Su aporte al estudio del locus universitario privado argentino (1955-1983). *Revista Latinoamericana de Metodología de las Ciencias Sociales*, 6(1), e003.
- AlGerafi, M. A. M., Zhou, Y., Oubibi, M., & Wijaya, T. (2023). Unlocking the potential: A comprehensive evaluation of augmented reality and virtual reality in education. *Electronics*, 12(18), 3953. <https://doi.org/10.3390/electronics12183953>
- Angulo, G., Lewis, F., Plante, P., & Brassard, C. (2023). Estado del arte sobre el uso de la realidad virtual, la realidad aumentada y el video 360° en educación superior. *Edutec. Revista Electrónica de Tecnología Educativa*, 84, 35-51. <https://doi.org/10.21556/edutec.2023.84.2769>
- Angulo Salazar, D. B., Andy Shiguango, N. P., Tipan Lugmania, N. D., & Vera Tigreiro, M. del P. (2025). Las TIC en el perfil docente: Análisis de los conocimientos y uso de la tecnología. *Estudios y Perspectivas Revista Científica y Académica*, 5(1), 1495-1514.
- Aquino Negrin, O., & Hernández Romero, G. (2021). Tecnologías de la Información y Comunicación como herramientas de trabajo en docentes mayores de 60 años: Uso de las TIC en docentes mayores de 60 años. *Revista Científica Sinapsis*, 1(19). <https://doi.org/10.37117/s.v19i1.429>
- Arancibia Muñoz, M. L., Valdivia Zamorano, I., Araneda Riveros, S. M., & Cabero-Almenara, J. (2017). Tipologías para la innovación tecnológica en docentes de educación superior a partir de un análisis de conglomerados: Un estudio

- exploratorio. *Revista de Educación a Distancia (RED)*, 17(55). <https://doi.org/10.6018/red/55/5>
- Bautista, L. E., Maradei, F., & Pedraza, G. (2025). Análisis de la disposición espacial de contenido en entornos de realidad aumentada y su efecto en la carga cognitiva de los usuarios. *Pixel-Bit. Revista de Medios y Educación*, 72, 39-69.
- Bermejo, B., Cortes, D., Oskam, J., Moilanen, T., Loijas, J., Govender, P., Hussey, J., Schmidt, A. L., Burbach, R., King, D., O'Connor, C., & Dunlea, D. (2023). AR/VR teaching-learning experiences in higher education institutions (HEI): A systematic literature review. *Informatics*, 10(2), 45. <https://doi.org/10.3390/informatics10020045>
- Bond, J., & Michailides, G. (1997). Interactive correspondence analysis in a dynamic object-oriented environment. *Journal of Statistical Software*, 2(8), 1-30. <https://doi.org/10.18637/jss.v002.i08>
- Caballero-Garriazo, J., Rojas-Huacanca, J., Sánchez-Castro, A., & Lázaro-Aguirre, A. (2023). Revisión sistemática sobre la aplicación de la realidad virtual en la educación universitaria. *Revista Electrónica Educare*, 27(3), 1-18. <https://doi.org/10.15359/ree.27-3.17271>
- Cabero-Almenara, J., & Barroso-Osuna, J. (2013). La utilización del juicio de experto para la evaluación de TIC: el coeficiente de competencia experta. *Bordón*, 65(2), 25-38. <https://doi.org/10.13042/brp.2013.65202>
- Cabero-Almenara, J., Barroso-Osuna, J., & Martínez-Roig, R. (2021). Mixed, augmented and virtual reality applied to the teaching of mathematics for architects. *Applied Sciences*, 11(15), 7125. <https://doi.org/10.3390/app11157125>
- Cabero-Almenara, J., & Loscertales-Abril, F. (Eds.). (1998). *¿Cómo nos ven los demás? La imagen del profesor y la enseñanza en los medios de comunicación social*. Secretariado de Publicaciones de la Universidad de Sevilla.
- Cabero-Almenara, J., & Pérez Díez, J. L. (2018). Validación del modelo TAM de adopción de la realidad aumentada mediante ecuaciones estructurales. *Estudios sobre Educación*, 34, 129-153. <https://doi.org/10.15581/004.34.129-153>
- Cabero-Almenara, J., Rodríguez Gallego, M., & Llorente-Cejudo, C. (2025). Realidad mixta, virtual y aumentada: Tecnologías para el aprendizaje. *Texto libre. Linguagem e Tecnologia*, 18, 1-19.
- Cabero-Almenara, J., Valencia, R., & Llorente, C. (2022). Ecosystem of emerging technologies: Augmented, virtual and mixed reality. *Tecnología, Ciencia y Educación*, 23, 7-22. <https://doi.org/10.51302/tce.2022.1148>
- Cajas Bravo, T. V., Silva Infantes, M., & Dávila Morán, R. C. (2023). Percepción docente sobre el uso de la tecnología en la educación superior. *Revista Conrado*, 19(90), 326-335.
- Cardona, C., Padilla, L., & Gutierrez, J. (2021). *Análisis de correspondencia múltiple* (ACM). <https://rpubs.com/ocamilocardona/813536>
- Cevikbas, M., Bulut, N., & Kaiser, G. (2023). Exploring the benefits and drawbacks of AR and VR technologies for learners of mathematics: Recent developments. *Systems*, 11(5), 244. <https://doi.org/10.3390/systems11050244>
- Coban, M., Bolat, Y., & Goksu, I. (2022). The potential of immersive virtual reality to enhance learning: A meta-analysis. *Educational Research Review*, 36, 100452. <https://doi.org/10.1016/j.edurev.2022.100452>
- Cornejo, J. M. (1988). *Técnicas de investigación social. El análisis de correspondencia*. PPU.
- Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13(3), 319-340. <https://doi.org/10.2307/249008>
- Dunn, T. J., Baguley, T., & Brunnsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British*

- Journal of Psychology*, 105(3), 399-412.
<https://doi.org/10.1111/bjop.12046>
- Escofier, B., & Pagès, J. (2008). *Analyses factorielles simples et multiples: Objectifs, méthodes et interprétation* (4th ed.). Dunod.
- Familoni, B. T., & Onyebuchi, N. C. (2024). Augmented and virtual reality in US education: A review: Analyzing the impact, effectiveness, and future prospects of AR/VR tools in enhancing learning experiences. *International Journal of Applied Research in Social Sciences*, 6(4), 642-663.
<https://doi.org/10.51594/ijarss.v6i4.1043>
- Greenacre, M. (2007). *Correspondence analysis in practice* (2nd ed.). Chapman and Hall/CRC.
- Hamilton, D., McKechnie, J., Edgerton, E., & Wilson, C. (2021). Immersive virtual reality as a pedagogical tool in education: A systematic literature review of quantitative learning outcomes and experimental design. *Journal of Computer Assisted Learning*, 37(5), 1-32.
<https://doi.org/10.1007/s40692-020-00169-2>
- Jingru, Z., Yahaya, W., & Sanmugam, M. (2025). An educational mobile augmented reality system to enhance undergraduate. *Interactive Learning Environments*.
- Le Roux, B., & Rouanet, H. (2010). *Multiple correspondence analysis*. SAGE Publications.
<https://doi.org/10.4135/9781412993906>
- Liu, J., Liu, Q., Yu, S., Ma, J., Liu, M., & Wu, L. (2024). How do autonomy and learner characteristics combine to influence learners' learning outcomes and cognitive load in virtual reality learning environments? A fuzzy-set qualitative comparative analysis approach. *Education and Information Technologies*, 29(1), 77-101.
<https://doi.org/10.1007/s10639-023-12262-x>
- Luna, A., Ortiz Colón, A. M., & Rodríguez Moreno, J. (2023). Tecnologías inmersivas en el aprendizaje autorregulado: Revisión sistemática de literatura científica. *Digital Education Review*, 44, 105-113.
<https://doi.org/10.1344/der.2023.44.105-113>
- Marín-Díaz, V., Sampedro-Requena, B. E., Vega-Gea, E. M., & Ruiz-Palmero, J. (2022). Conocimientos tecnopedagógicos del profesorado de secundaria sobre la realidad mixta. *Revista Tecnología, Ciencia y Educación*, (23), 23-48.
<https://doi.org/10.51302/tce.2022.1984>
- Marougkas, A., Troussas, C., Krouska, A., & Sgouropoulou, C. (2024). How personalized and effective is immersive virtual reality in education? A systematic literature review for the last decade. *Multimedia Tools and Applications*, 83, 18185-18233.
<https://doi.org/10.1007/s11042-023-15986-7>
- Menjivar, E., Sanchezs, E., Ruiz Palmero, J., & Valenzuela, T. (2021). Revisión de la producción científica sobre la realidad virtual entre 2016 y 2020 a través de Scopus y WoS. *EDMETIC, Revista de Educación Mediática y TIC*, 10(2), 26-55.
<https://doi.org/10.21071/edmetic.v10i2.13422>
- Mishra, P., & Koehler, M. J. (2006). Technological pedagogical content knowledge: A framework for teacher knowledge. *Teachers College Record*, 108(6), 1017-1054.
<https://doi.org/10.1111/j.1467-9620.2006.00684.x>
- Mogrovejo-Zambrano, J. N., Montalván-Vélez, C. L., Barragan-Espinoza, G. M., & Cabrera-Davila, M. A. (2024). Fenomenología de la realidad virtual: Explorando la experiencia humana en entornos digitales inmersivos. *Journal of Economic and Social Science Research*, 4(1), 149-159.
<https://doi.org/10.55813/gaea/jessr/v4/n1/88>
- Mohsen, M. A., & Alangari, T. S. (2024). Analyzing two decades of immersive technology research in education: Trends, clusters, and future directions. *Education and Information Technologies*, 29(3), 3571-3587.
<https://doi.org/10.1007/s10639-023-11968-2>
- Mulders, M., Buchner, J., & Kerres, M. (2020). A framework for the use of immersive virtual reality in learning

- environments. *International Journal of Emerging Technologies in Learning (iJET)*, 15(24), 208–224. <https://doi.org/10.3991/ijet.v15i24.16615>
- Obeidallah, R., Al Ahmad, A., & Qutishat, D. (2023). Challenges of extended reality technology in higher education: A review. *Journal of Emerging Technologies in Learning (iJET)*, 18(14), 39–50. <https://doi.org/10.3991/ijet.v18i14.39871>
- Perifanou, M., Economides, A. A., & Nikou, S. A. (2023). Teachers' views on integrating augmented reality in education: Needs, opportunities, challenges and recommendations. *Future Internet*, 15(1), 20. <https://doi.org/10.3390/fi15010020>
- Puentedura, R. (2006). *Transformation, technology, and education*. Hippasus. <http://hippasus.com/resources/tte/>
- Radianti, J., Majchrzak, T. A., Fromm, J., & Wohlgenannt, I. (2020). A systematic review of immersive virtual reality applications for higher education: Design elements, lessons learned, and research agenda. *Computers & Education*, 147, 103778. <https://doi.org/10.1016/j.compedu.2019.103778>
- Roco-Videla, Á., Aguilera-Eguía, R., & Olguin-Barraza, M. (2024). Ventajas del uso del coeficiente de omega de McDonald frente al alfa de Cronbach. *Nutrición Hospitalaria*, 41(1), 262–263. <https://doi.org/10.20960/nh.04879>
- Sales, S., Macédo, R., De Oliveira, V., Souza, W., Gama, L., De Brito, S., & Winkler, I. (2023). Augmented reality in education for industry 4.0: What are the barriers to adoption? *Cuadernos de Educación y Desarrollo*, 15(1), 481–505. <https://doi.org/10.55905/cuadv15n1-025>
- Subirats-Blanco, R., & Conde del Rio, M. A. (2024). Integración de la realidad aumentada en la enseñanza del inglés en educación primaria: Mejorando el aprendizaje lingüístico y la adquisición de vocabulario. *EDMETIC*, 13(2), art.7. <https://doi.org/10.21071/edmetic.v13i2.17072>
- Tapia Cortes, C. (2020). Tipologías de uso educativo de las Tecnologías de la Información y Comunicación: Una revisión sistemática de la literatura. *EduTec. Revista Electrónica De Tecnología Educativa*, 7(1), 16–34. <https://doi.org/10.21556/edutec.2020.71.1489>
- Toala-Palma, J., Arteaga-Mera, J., Quintana, J., & Santana, I. (2020). Virtual reality as a tool for educational innovation. *Episteme Koinonia*, 3(5). <https://fundacionkoinonia.com.ve/ojs/index.php/epistemekoinonia/article/view/835>
- Tourrón, J. (2023). *Análisis de datos y medidas en educación* (Vol. II). Unir.
- UNESCO. (2023). *Informe GEM 2023: Tecnología en la educación. Una herramienta en los términos de los docentes*. UNESCO.
- Upadhyay, B., Chalil Madathil, K., Hegde, S., Anderson, D., Wooldridge, E., Presley, D., Perez, L., & Reid, B. (2024). Barriers toward the implementation of extended reality (XR) technologies to support education and training in workforce development programs. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 68(1), 265–269. <https://doi.org/10.1177/10711813241275080>
- Venkatesh, V. (2000). Determinants of perceived ease of use: Integrating control, intrinsic motivation, and emotion into the technology acceptance model. *Information Systems Research*, 11(4), 342–365. <https://doi.org/10.1287/isre.11.4.342.11872>
- Victoria Maldonado, J. J., Fuentes-Cabrera, A., Fernández-Cerero, J., & Sadio-Ramos, F. J. (2024). Influencia de la realidad virtual en el rendimiento académico en educación secundaria a través de un meta-análisis. *Pixel-Bit. Revista de Medios y Educación*, 71, 107–121. <https://doi.org/10.12795/pixelbit.104279>
- Yu, Z., & Xu, W. (2022). A meta-analysis and systematic review of the effect of virtual reality technology on users' learning outcomes. *Computer Applications in Engineering Education*, 30(6), 1470–1484. <https://doi.org/10.1002/cae.22532>

Date of reception: 1 June 2025
Date of acceptance: 31 July 2025
Date of approval for layout: 5 September 2025
Date of publication in OnlineFirst: 12 October 2025
Date of publication: 1 January 2026

Development and validation of the AI-Enhanced Self-Regulated Learning (AI-SRL) scale

Desarrollo y validación de la escala de Aprendizaje Autorregulado con IA (AI-SRL)



-  Mehmet Ali Ayaz - *Giresun University, GRÜ (Turkey)*
 Fatih Karataş - *Nevşehir Hacı Bektaş Veli University, NEVÜ (Turkey)*
 Erkan Yüce - *Aksaray University, ASÜ (Turkey)*

ABSTRACT

The integration of Artificial Intelligence (AI) into educational settings offers new opportunities for enhancing self-regulated learning (SRL) among students. However, current tools lack precision in measuring how AI influences learners' self-regulatory capabilities. This study introduces the AI-Enhanced Self-Regulated Learning (AI-SRL) Scale, designed to assess SRL in AI-supported educational environments. Through a rigorous scale development process involving a literature review, expert consultation, and systematic validation, we developed a comprehensive measurement instrument. The development employed a two-phase validation approach with separate samples for exploratory and confirmatory factor analyses. The final validated scale comprises 22 items organized into five dimensions: AI Competence, Learning Awareness, Learning Strategies, Engagement and Efficiency, and Ethical Collaboration. These factors capture the multifaceted nature of how students regulate their learning when using AI tools, from technical proficiency to ethical considerations. The scale demonstrated strong psychometric properties, with excellent internal consistency and robust construct validity. This validated instrument has practical applications for educators aiming to optimize AI integration in their classrooms, researchers investigating the intersection of AI and SRL, and institutions developing AI-enhanced curricula. The AI-SRL Scale provides a reliable framework for assessing how effectively students leverage AI tools while maintaining their self-regulatory capabilities, thereby contributing to more effective and responsible AI implementation.

Keywords: AI; AIED; self-regulated learning; AI-SRL; AI-SRL scale.

RESUMEN

La integración de la Inteligencia Artificial (IA) en entornos educativos brinda nuevas oportunidades para mejorar el aprendizaje autorregulado (SRL) en los estudiantes. Sin embargo, las herramientas actuales carecen de precisión para evaluar cómo la IA influye en las capacidades autorregulatorias de los estudiantes. Este estudio presenta la Escala de Aprendizaje Autorregulado Mejorada por IA (AI-SRL), diseñada para evaluar el aprendizaje autorregulado en entornos educativos asistidos por IA. A través de un riguroso proceso de desarrollo de escala, que incluyó revisión de literatura, consulta con expertos y validación sistemática, se construyó un instrumento de medición integral. En el desarrollo se empleó un enfoque de validación en dos fases con muestras separadas para análisis factoriales exploratorios y confirmatorios. La escala final validada consta de 22 ítems organizados en cinco dimensiones: Competencia en IA, Conciencia de Aprendizaje, Estrategias de Aprendizaje, Implicación y Eficiencia, y Colaboración Ética. Estos factores capturan la naturaleza multifacética de cómo los estudiantes regulan su aprendizaje al usar herramientas de IA, desde competencia técnica hasta consideraciones éticas. La escala demostró sólidas propiedades psicométricas, con excelente consistencia interna y robusta validez de constructo. Este instrumento validado tiene aplicaciones prácticas para docentes que buscan optimizar la integración de IA en sus aulas, investigadores que investigan la intersección de IA y aprendizaje autorregulado, e instituciones desarrollando currículos mejorados con IA. La Escala AI-SRL proporciona un marco confiable para evaluar en qué medida los estudiantes aprovechan las herramientas de IA mientras mantienen sus capacidades autorregulatorias, contribuyendo así a una implementación de IA más efectiva y responsable.

Palabras clave: IA; AIED; aprendizaje autorregulado; AI-SRL; escala AI-SRL.

INTRODUCTION

Self-regulated learning (SRL), a critical component of success in digital and lifelong learning, involves learners actively setting goals, monitoring progress, and adjusting behaviors to achieve academic success (Hilpert et al., 2023; Lai, 2024). In the digital age, SRL has become increasingly crucial for lifelong learning, enabling individuals to adapt to changing educational needs while promoting problem-solving abilities (Adams et al., 2024; Hilpert et al., 2023; Kong & Yang, 2024). As educational practices evolve, integrating advanced technologies like AI into learning environments offers opportunities to enhance SRL while addressing associated challenges.

The theoretical foundations of SRL are primarily based on two comprehensive frameworks that have shaped educational psychology for over three decades. Zimmerman's cyclical phase model, grounded in social cognitive theory, conceptualizes SRL as structured into three recursive phases: forethought, performance, and self-reflection (Panadero, 2017; Schunk & Zimmerman, 2012; Zimmerman, 1989). The forethought phase encompasses task analysis, including goal setting and strategic planning, alongside self-motivational beliefs such as self-efficacy, outcome expectations, and intrinsic interest. The performance phase involves self-control strategies such as focused attention, task strategies, and self-instruction, combined with self-observation through metacognitive monitoring of one's progress and strategy effectiveness. The self-reflection phase includes self-judgment, where learners evaluate their performance against standards, and self-reaction processes involving satisfaction responses and adaptive or defensive reactions that influence future learning efforts (Panadero, 2017; Tinajero et al., 2024).

Complementing this temporal view, Pintrich's (2000) framework offers a taxonomy with four phases—forethought/planning/activation, monitoring, control, and reaction/reflection—applied across four regulatory areas: cognition, motivation/affect, behavior, and context (Tinajero et al., 2024; Yot-Domínguez & Marcelo, 2017). This framework is particularly valuable for its explicit treatment of motivational regulation and contextual factors, offering a 16-cell matrix of specific regulatory processes that provides exhaustive coverage of SRL components. Pintrich's model emphasizes that SRL is an active, constructive process whereby learners set goals for their learning and then attempt to monitor, regulate, and control their cognition, motivation, and behavior, guided and constrained by their goals and the contextual features in the environment (Panadero, 2017). These foundational models, while developed for conventional learning formats, provide the essential theoretical architecture for investigating how AI transforms established SRL processes.

The integration of AI in educational settings is transforming the structure of learning, bringing both opportunities and challenges in fostering SRL (Chang et al., 2023; Grüneke et al., 2024). AI technologies offer innovative ways to enhance traditional SRL strategies, providing personalized real-time feedback, flexible support, and real-time observing that supports learners in managing their learning processes effectively (Huang et al., 2023; Molenaar et al., 2023; Ng et al., 2024). Research demonstrates that AI applications effectively support metacognitive, cognitive, and behavioral regulation across Pintrich's four phases, with learning analytics dashboards significantly enhancing metacognitive awareness through real-time progress tracking and performance visualization (Gkintoni et al., 2025; Seo et al., 2021; Wang & Lin, 2023). Additionally, AI tools support metacognitive strategies through goal-setting support and strategic planning assistance, aligning with Zimmerman's forethought

phase (Anggoro & Pratiwi, 2023; Liu et al., 2024). These systems can trace learning behaviors and provide insights into SRL processes, facilitating solutions that support the development of SRL (Chang et al., 2023; Jin et al., 2023; Molenaar et al., 2023).

Though AI tools have potential to support metacognitive skills and providing real-time monitoring (Dahri et al., 2024; Jin et al., 2023), researchers express concerns about overreliance on AI hindering essential skill development (Boguslawski et al., 2025). The literature highlights disagreement between AI's capacity to provide personalized support and the risk of diminishing learner autonomy through cognitive offloading effects, where frequent AI usage is negatively associated with critical thinking abilities (Chiu, 2024; Gerlich, 2025; Kong & Yang, 2024). Studies reveal that while learners perceive AI as useful for metacognitive, cognitive, and behavioral regulation, they show a strong preference for human support when addressing motivational needs, suggesting that AI applications consistently struggle to support motivational regulation—a critical component of Pintrich's framework (Lan & Zhou, 2025; Seo et al., 2021). Another critical aspect emphasized in studies is that studies emphasize the importance of maintaining a balance between AI support and human guidance in educational settings (Kong & Yang, 2024; Lai, 2024). While AI can provide detailed oversight and feedback mechanisms, educators play a crucial role in designing learning activities and providing emotional support (Chiu et al., 2023). Studies suggest that effective implementation of AI in SRL requires careful consideration of how these tools can complement rather than replace human interaction, particularly in supporting students' metacognitive development and motivation (Wang & Lin, 2023; Xia et al., 2023).

However, the potential of AI to fully enhance SRL has not yet been fully realized due to several barriers. While some studies highlight AI's potential to alleviate learning anxiety and enhance motivation (Biju et al., 2024), others raise concerns about the need for balanced integration to maintain active student participation (Grüneke et al., 2024). Furthermore, significant ethical concerns have arisen in multiple domains, including data privacy and surveillance issues, algorithmic bias that may embed educational inequalities, and academic integrity challenges as AI-generated content complicates the assessment of authentic learning (Gkintoni et al., 2025; Walter, 2024). To address these challenges, researchers emphasize the critical need for developing reliable measurement instruments to evaluate AI's impact on SRL processes (Baker, 2023; Grüneke et al., 2024; Jin et al., 2023; Molenaar et al., 2023).

Building on these theoretical foundations, the present study conceptualizes AI-enhanced SRL through five dimensions, extending and operationalizing Zimmerman's and Pintrich's frameworks for contemporary technological contexts. The AI Competence factor aligns with the forethought phase of both models, encompassing self-efficacy beliefs specifically related to AI tool usage and strategic planning for their deployment. Learning Awareness corresponds to metacognitive monitoring central to Pintrich's monitoring phase and Zimmerman's self-observation subprocess, capturing learners' awareness of their understanding and progress while interacting with AI. Learning Strategies and Engagement and Efficiency factors map onto the performance and control phases of both frameworks, reflecting cognitive strategy selection, adaptation, and the behavioral and motivational regulation necessary to persist in learning tasks. Critically, the Ethical Collaboration factor represents a necessary extension to classical models, operationalizing modern context regulation (Pintrich's fourth regulatory area) that addresses algorithmic bias, data privacy, and academic

integrity concerns-regulatory demands unique to the AI era not anticipated in traditional SRL frameworks (Gkintoni et al., 2025; Walter, 2024).

The development of the AI-SRL Scale has important implications for evaluating diverse educational technology environments. In platforms utilizing real-time formative assessment tools like Socrative, the scale's Learning Awareness and Learning Strategies dimensions can measure whether instant feedback effectively prompts metacognitive monitoring and strategy adjustment (Bauer et al., 2025). Within Learning Management Systems such as Moodle featuring AI assistants like Corolair or CodeRunner Agent, the AI Competence and Engagement and Efficiency subscales can assess whether these 24/7 AI tutors enhance students' self-regulatory capabilities outside direct instructor supervision. For Intelligent Tutoring Systems providing adaptive scaffolding and personalized learning paths, the complete scale offers holistic evaluation of the entire regulatory cycle, from planning and strategy use to ethical considerations, offering valuable feedback for designing systems that are both intelligent and pedagogically sound (Wu & Chiu, 2025). As students increasingly engage with generative AI tools for academic work, the scale provides structured methods to investigate their regulation of inquiry, critical evaluation, and ethical AI use, addressing key gaps in understanding modern digital learning ecosystems (Li et al., 2025; Mohamed et al., 2025). Therefore, this study addresses the following research questions:

- RQ1: How can AI-enhanced self-regulated learning be comprehensively measured through a scale that operationalizes and extends the established theoretical frameworks of Zimmerman and Pintrich for contemporary technological contexts?*
- RQ2: What is the factorial structure and psychometric properties of AI-SRL Scale when validated through exploratory and confirmatory factor analysis procedures?*
- RQ3: To what extent does the AI-SRL Scale function as a reliable and valid instrument for capturing the unique dimensions of SRL in AI-supported educational environments?*

Literature review

Recent advancements in AI have fundamentally transformed the landscape of SRL in educational settings, creating both unprecedented opportunities and complex challenges that demand careful theoretical and empirical examination. The literature reveals a nuanced picture where AI technologies demonstrate strong potential for supporting cognitive and metacognitive dimensions of SRL (Seo et al., 2021; Wang & Lin, 2023), while simultaneously presenting significant challenges for motivational regulation and raising critical ethical concerns about learner agency and digital well-being (Lan & Zhou, 2025).

The theoretical foundations of SRL, particularly Zimmerman's cyclical model and Pintrich's comprehensive framework, have proven remarkably robust when applied to AI-enhanced learning contexts. Zimmerman's three-phase cyclical structure—forethought, performance, and self-reflection—maps effectively onto AI applications (Chiu, 2024), with AI systems providing differential support across these phases. Research by Lan and Zhou (2025) found that while 50% of AI applications primarily support the performance phase, comprehensive support across all three

phases remains rare (only 21% of applications), with the self-reflection phase receiving the least AI support. This suggests a continued need for human guidance in deeper reflective processes. Similarly, Pintrich's framework, with its four areas of regulation (cognitive, motivational/affective, behavioral, and contextual), finds clear parallels in AI-mediated learning environments (Yot-Domínguez & Marcelo, 2017), though motivational regulation remains particularly problematic (Jin et al., 2023; Seo et al., 2021).

AI technologies have demonstrated significant support for core SRL components across disciplines, particularly in cognitive and metacognitive domains. Recent empirical studies reveal significant relationships between AI self-efficacy and academic performance, with practical application of AI tools serving as the strongest predictor of self-efficacy (Bećirović et al., 2025). Research by Liang et al. (2023) found that student-AI interaction significantly improved learning achievement through enhanced cognitive engagement ($\beta = 0.046$), while AI-powered feedback systems enhance students' metacognitive awareness through real-time progress tracking and performance visualization (Gkintoni et al., 2025; Molenaar et al., 2023). The implementation of AI in SRL reveals complex support-hindrance dynamics that require careful consideration. While AI tools like ChatGPT have enhanced students' research skills and supported self-directed learning through autonomy and engagement (Li et al., 2025), studies reveal concerning negative correlations ($r = -0.39$) between over-reliance on AI tools and critical thinking abilities, particularly among younger learners aged 17-25 (Gerlich, 2025). This "cognitive offloading" effect suggests that frequent AI usage may prevent the development of independent metacognitive skills, with learners becoming overly reliant on AI systems rather than developing autonomous self-regulation capabilities (Gerlich, 2025). Research by Glick et al. (2024) demonstrated that while AI-powered training modules significantly increased use of planning tools and goal-setting activities, the balance between AI support and learner agency remains a critical concern.

Perhaps the most significant finding across the literature is that AI applications consistently struggle to support motivational regulation, a critical component of effective SRL (Seo et al., 2021). Multiple systematic reviews confirm that while learners perceive AI as useful for metacognitive, cognitive, and behavioral regulation, they show a strong preference for human support when addressing motivational needs (Jin et al., 2023; Lan & Zhou, 2025). From a Self-Determination Theory perspective, while AI tools can support competence through personalized feedback and autonomy through learning pathway control (Chiu, 2024; Li et al., 2025), relatedness remains the weakest aspect of AI support.

Recent empirical evidence from 2020-2025 increasingly points toward hybrid human-AI regulatory systems as optimal approaches for supporting SRL. Research distinguishes between human-centered self-regulation, where AI serves as a facilitator providing tools and data for learner control, and AI-centered self-regulation, where AI generates SRL cycles based on data analysis (Lan & Zhou, 2025). The most effective implementations appear to involve human-AI interaction where AI engages with human-centered SRL cycles across behavioral, emotional, and cognitive dimensions while maintaining human agency (Wu & Chiu, 2025).

Overall, while AI presents transformative potential for SRL, the evidence suggests that effective implementation requires a carefully balanced approach that harnesses technological capabilities while preserving fundamental human elements of SRL. The development of new measurement tools like the AI-SRL Scale becomes

crucial for providing deeper insights into the effectiveness and impact of AI tools, addressing the complex interplay between cognitive enhancement and potential motivational deficits, and guiding future integrations that maintain learner agency while leveraging AI's strengths in personalization and cognitive load management.

METHOD AND FINDINGS

In this study, we aimed to develop a scale to measure participants' SRL levels in the context of AI use. Despite established guidelines, researchers often face limitations in scale development, including sample characteristics, methodological issues, and psychometric challenges. To improve future practices, researchers should carefully consider these potential limitations and make informed methodological choices throughout the scale development process (Morgado et al., 2017). Therefore, acknowledging these limitations, we employed a robust methodology for the development of a reliable and valid scale in this study.

Participants

Two separate samples were used for the exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). The EFA sample consisted of 504 participants, while the CFA sample included 424 participants. Participants were recruited online through convenience and purposeful sampling (Creswell, 2014). Convenience sampling was used in a way that enabled us to reach participants more easily. We found participants from the universities that we somehow had connections with. Our sampling was purposeful, because we selected participants who could readily understand the scale items. Most of them were from the foreign language teaching or literature departments of universities. For those who were from other departments, we made sure they could easily understand the items on our scale. We preferred to collect data mainly from these departments to ensure clarity in item comprehension, which were in English. This might have affected the generalizability of the responses to some extent and may have influenced factor interpretation across some constructs. Future studies to validate the scale could collect data across diverse academic populations to enhance generalizability and statistics. The demographics of our participants are summarized in the table below.

Table 1
Demographics

Gender	n	Age	n	Class	n	Department	n
Male	334	18 or below	195	English prep class	212	English Language Teaching	288
Female	594	19	202	Freshman/1 st year	184	English Language and Literature	262
Total	928	20	228	Sophomore/2 nd year	260	Pharmacy	149
		21	193	Junior/3 rd year	180	International Relations	31
		22	195	Senior/4 th year+	92	Psychology	28
		23 +	202	Total	928	Law	25
		Total	928			Tourism and Guidance	23
						Engineering	24
						Others	98
						Total	928

As Table 1 indicates, we had a total of 928 participants for the first and second steps (EFA and CFA) of the study. 594 participants (64.0%) were female, while 334 (36.0%) were male. Among departments, 288 participants (31.0%) were from English Language Teaching, 262 (28.2%) from English Language and Literature, and 149 (16.1%) from Pharmacy. Class distribution included 212 participants (22.8%) from preparatory English classes, 184 (19.8%) freshmen, 260 (28.0%) sophomores, 180 (19.4%) juniors, and 92 (9.9%) seniors. We collected some other information related to the participants' use of technology and AI in learning. Table 2 summarizes this information.

Table 2
Participants' use of technology and AI in learning

Level of Using Technology in Learning	n	Experience with AI	n	Frequency of using AI for Learning	n
Very low	8	1 year	543	Every day	104
Low	46	2 years	239	Several times a week	212
Moderate	395	3 years	81	Once a week	141
High	338	4 years	29	Several times a month	180
Very high	141	5 years	13	Once a month	189
Total	928	6 +	23	Not regularly/on need	102
		Total	928	Total	928

As seen from Table 2, the participants indicated their level of using technology in learning, with a vast majority reporting moderate (395, 42.5%), high (338, 36.4%) and very high levels (141, 14%). The ones whose level was low and very low were 5.8% of the participants. The participants also provided information on how much experience they had with AI. 543 of them (58.5%) had one year of experience, 239 had 2 years (25.7%), 81 had 3 years (8.7%), and the others had 4, 5, 6 or more years of experience with AI. Lastly, we gathered input from the participants on how often they used AI for learning. The number of participants who reported to be using it every day was 104 (11.2%), 212 (almost 22.8%) used it several times a week, 141 (15.2%) once a week, 180 (19.3%) several times a month, and 189 (20.3%) once a month. Also, 102 (almost 11%) participants reported using AI for learning when they needed it.

Item generation

Scale development is a critical process in social and behavioral research, typically involving three main phases: item generation, scale construction, and scale evaluation (Boateng et al., 2018). The item generation phase begins with clearly defining the construct to be measured and creating an initial item pool (DeVellis, 1991). In this study, the initial pool of items was generated through a comprehensive review of the literature on SRL and AI. Content validity is assessed, often through expert review (Teeluckdhar et al., 2021). The items in our scale were designed to capture key dimensions of SRL in the context of AI, ensuring alignment with existing theoretical frameworks. The initial item pool underwent a rigorous content validation process by a panel of seven experts. These experts were selected based on specific criteria: (1) possession of a Ph.D. in Educational Technology, Educational Psychology, or a related field; (2) a minimum of five years of research experience; and (3) at least

two peer-reviewed publications on either SRL or the application of AI in education. This purposive selection ensured substantial theoretical and practical expertise. The panel provided feedback on item clarity, relevance, and representativeness. Content validity was assessed using the Content Validity Index (CVI). Items achieving $CVI \geq .80$ from at least 6 of 7 experts were retained. The Scale-level CVI average (S-CVI/Ave) was .92, exceeding the recommended threshold of .90. Items were revised or removed based on this feedback, yielding a preliminary scale for pilot testing. The literature suggests various sample sizes for pilot studies. Johanson and Brooks (2009) reported that Isaac and Michael (1995) and Hill (1998) suggested 10 to 30 samples for pilot studies, and they also stated that Mooney and Duval (1993) offered 30-50, and Hertzog (2008) suggested 25-40 for scale development. In our case, the pilot test was conducted with a sample of 33 participants representing similar characteristics of the target study groups. The pilot study aimed to identify any issues with item wording, response options, or overall scale length. Based on the pilot test results, minor revisions were made to enhance item clarity and reliability.

Exploratory Factor Analysis (EFA)

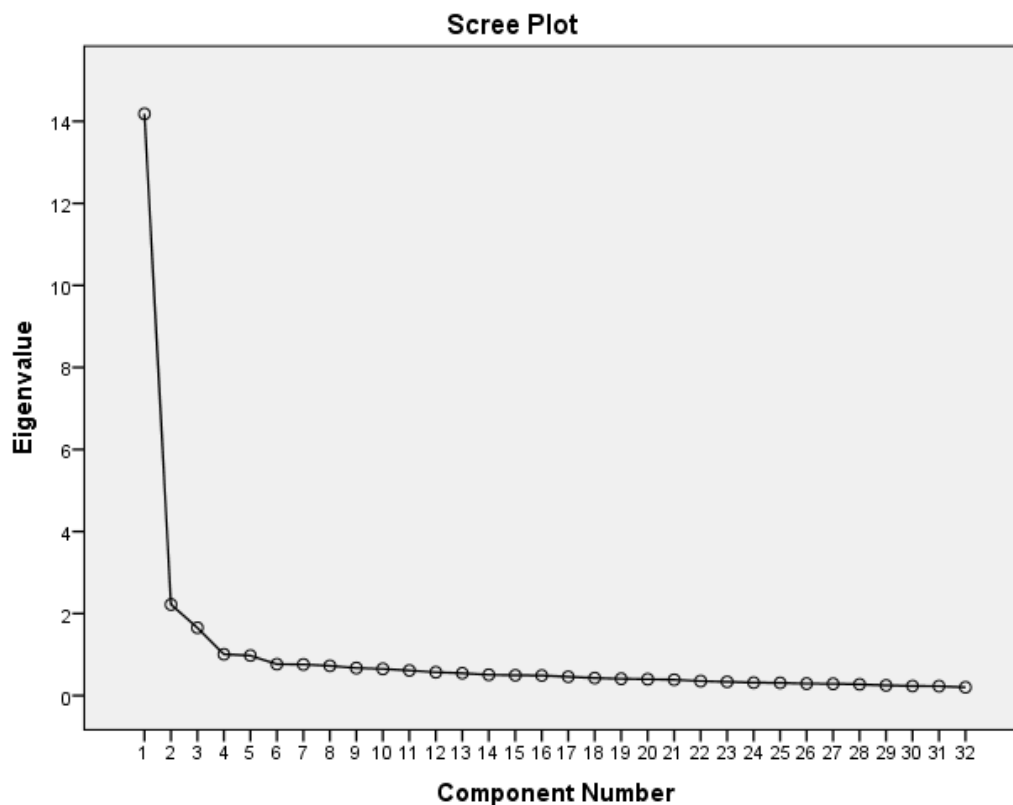
The scale construction phase includes pre-testing, survey administration, item reduction, and factor extraction (Boateng et al., 2018). After administering the scale to the first study group, we proceeded to the next step. The scale evaluation phase involves testing dimensionality, reliability, and validity (Teeluckdharry et al., 2021). The suitability of the data from the first sample ($n = 504$) for factor analysis was assessed. The Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy was .96, indicating marvelous suitability for factor analysis (Kaiser & Rice, 1974). Additionally, Bartlett's Test of Sphericity was significant, $\chi^2 = 9846.80$, $p < .001$, confirming that the correlation matrix was not an identity matrix and was appropriate for factor analysis. Following this, Exploratory Factor Analysis (EFA) was carried out using SPSS 22 to examine and confirm the factor structure of the AI-SRL Scale. EFA was conducted to reveal the underlying factor structure and to identify clusters of related items, thereby establishing the construct validity of the scale by reducing the data into a smaller number of meaningful factors (Field, 2018). Principal Component Analysis (PCA) was employed as the extraction method, and promax rotation - one of the oblique rotation methods - was used to allow for correlations between factors (Kline, 2014; Muthén & Muthén, 2017). While Principal Axis Factoring (PAF) and Maximum Likelihood (ML) were tested, they did not yield stable or interpretable factor solutions. Given our aim of exploring the underlying dimensions and reducing the item pool into coherent components, PCA was selected for its robustness and interpretability in this context. This approach aligns with prior recommendations for early-stage scale development (Fabrigar & Wegener, 2012; Jolliffe, 2014). To further ensure validity, we used a separate dataset to conduct a Confirmatory Factor Analysis (CFA) via AMOS, which confirmed the PCA-derived factor structure with satisfactory model fit indices. The EFA most often yielded four factors, and in some cases five, based on eigenvalues larger than one. After trying a very large number of models with EFA and CFA, we chose the five-factor solution, because it yielded the best model fit. At the EFA stage, items with factor loadings of 0.4 or higher were retained, as this threshold was considered significant (Floyd & Widaman, 1995; Hair et al., 2010). During the EFA stage, three items were deleted due to cross-loadings below the 0.10 difference threshold (Büyüköztürk, 2018): *"I do not share my personal information when using*

AI tools,” “I check the originality of AI-generated work before using it,” and “I ensure that my peers also use AI tools ethically.” At the CFA stage, nine additional items were removed due to low or marginal standardized loadings, redundancy with higher-loading items, or problematic cross-factor modification indices. These included items related to AI Competence (e.g., “I can quickly learn to use new AI tools”), reflective and evaluative Learning Awareness (e.g., “I evaluate whether AI tools give me accurate information”), and Ethical Collaboration (e.g., “I discuss AI-related ethical issues with peers”). Removal decisions were based on both statistical and conceptual considerations to ensure the final scale retained psychometric rigor and theoretical coherence. The items removed and the rationale for their exclusion are presented in Appendix B. The EFA revealed a five-factor structure, accounting for 62.42% of the total variance. Items loaded strongly on their respective factors, with factor loadings ranging from 0.44 to 0.96.

Table 3
Eigenvalues of parallel analysis

Factor	Initial eigenvalues	% of variance	Cumulative %
Factor 1	14.183	43.99	43.99
Factor 2	2.220	6.95	50.95
Factor 3	1.652	5.25	56.20
Factor 4	1.006	3.13	59.34
Factor 5	.996	3.080	62.42

As seen from Table 3, the eigenvalues for the first four factors are higher than one, and the fifth one is 0.99, just below 1.00. Because the fifth factor was supported by the previous literature and the upcoming analyses, we decided that it stay as a factor in the final scale. Parallel analysis was conducted using 1000 randomly generated data matrices with the same number of variables (31) and participants (504) as the actual dataset. The first three factors clearly exceeded the 95th percentile of random eigenvalues, strongly supporting their retention. Factors 4 and 5 had eigenvalues slightly below the 95th percentile threshold but were retained based on (a) theoretical alignment with established SRL frameworks, (b) adequate factor loadings (all > .40), and (c) improved model fit in subsequent CFA. This decision aligns with recommendations to consider multiple criteria beyond eigenvalues alone when determining factor retention (Fabrigar & Wegener, 2012). The five-factor solution was also supported by a scree plot, which illustrates the potential factors within the scale (Eisinga et al., 2013).

Figure 1*Scree plot for the initial Exploratory Factor Analysis*

The scree plot also supported a five-factor solution, as the slope of the plot flattened considerably after the fifth factor. Appendix C presents the factor loadings for each item from the initial model, outlining the sub-dimensional nature of the scale. Following this analysis, items 21, 24, and 29 were removed. The remaining 31 items formed a five-factor structure, which was then subjected to CFA using data from the second study group. This subsequent analysis resulted in the final 22-item scale.

Confirmatory Factor Analysis (CFA)

After the initial EFA model was confirmed, we collected new data based on the latest EFA model. The new data was used for CFA. Sun (2005) indicates three main purposes for conducting a CFA: Construct validity evaluation, response pattern comparison and competing model comparison. We used it for evaluating construct validity and model comparison. The CFA, conducted with a separate sample ($n = 424$), was used to test the five-factor structure identified in the EFA. The initial model was refined based on standardized loadings and modification indices. Items with low loadings or those creating significant cross-factor covariance were iteratively removed to improve model fit while maintaining theoretical coherence. This process resulted in the final 22-item model. Table 2 summarizes the main fit indices from the CFA analyses.

Table 4
CFA results

Key indices	Hu and Bentler (1999)	Our results
CFI	> .95 excellent, > .90 good, > .80 sometimes acceptable	0.96
RMSEA	< .05 good, .10 moderate, > .10 not good	0.04
CMIN/DF	< 3 good, < 5 acceptable	2.82
GFI	> .95 excellent, > .90 good	0.94
AGFI	> .80	0.92
SRMR	< .09	0.04
TLI	> .95 excellent, > .90 acceptable	.95

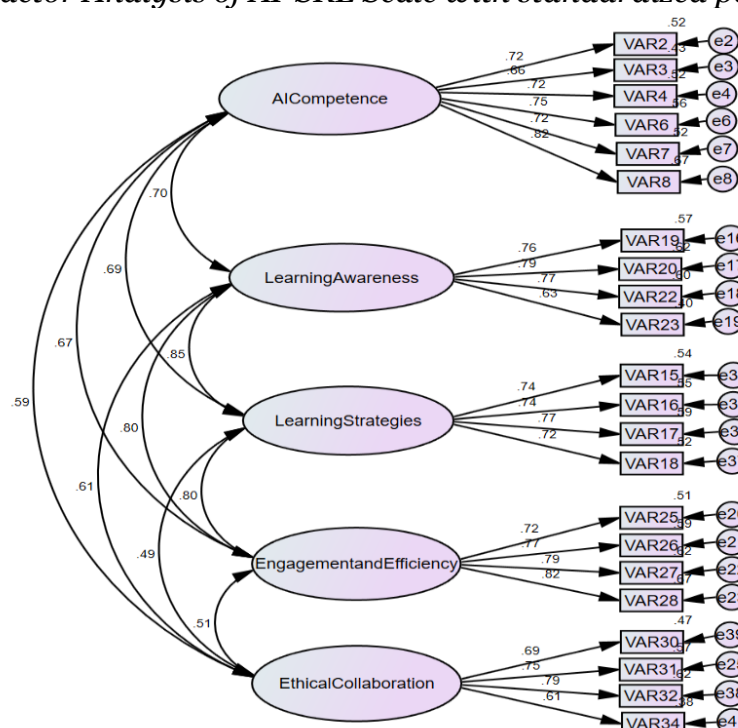
As Table 4 indicates, the final model demonstrated a good fit to the data, as assessed against conventional fit criteria (Hu & Bentler, 1999). The fit indices were: χ^2/df (CMIN/DF) = 2.82, Comparative Fit Index (CFI) = .97, Tucker-Lewis Index (TLI) = .95, Root Mean Square Error of Approximation (RMSEA) = .04 (90% CI [.035, .045]), and Standardized Root Mean Square Residual (SRMR) = .04. As the CFI and TLI were above .95 and the RMSEA and SRMR were below .05, the five-factor model was confirmed as a valid representation of the data.

Confirmatory Factor Analysis (CFA) Model

The figure below illustrates the CFA model and the coefficients between variables and items.

Figure 2

Confirmatory Factor Analysis of AI-SRL Scale with standardized path coefficients



The figure confirms the final model that consists of 5 factors and 22 items. The five factors confirmed through CFA represent distinct dimensions of AI-enhanced self-regulated learning with clear pedagogical implications. *AI Competence* (Factor 1, 6 items) captures learners' self-efficacy and technical proficiency in utilizing AI tools for educational purposes, reflecting their confidence in adapting to various AI interfaces, understanding AI-generated instructions, and effectively interpreting AI-produced information for learning. *Learning Awareness* (Factor 2, 4 items) encompasses metacognitive monitoring processes where learners consciously track their understanding, progress, and learning effectiveness while interacting with AI systems, including critical reflection on their dependency on these tools. *Learning Strategies* (Factor 3, 4 items) involves the selection, adaptation, and implementation of AI-supported cognitive strategies, including how learners use AI for finding effective study methods, improving problem-solving skills, and enhancing critical thinking through multiple perspective analysis. *Engagement and Efficiency* (Factor 4, 4 items) measures learners' behavioral and motivational regulation in maintaining participation, achieving permanent learning outcomes, and sustaining engagement when using AI tools adapted to their learning style. *Ethical Collaboration* (Factor 5, 4 items) represents a novel dimension addressing learners' awareness of responsible AI use, including considerations of personal responsibility, privacy concerns, potential biases, and information security in AI-assisted learning contexts. These factors collectively operationalize how contemporary learners navigate the complex landscape of AI-enhanced education while maintaining their self-regulatory capabilities. The complete validated 22-item AI-SRL Scale with all items organized by these five factors is provided in Appendix A. In the upcoming sections, it has been elaborated on the reliability and validity of the AI-SRL scale.

Reliability analysis

We performed several tests to ensure the AI-SRL scale is a reliable scale. These tests are explained below.

Internal consistency

Cronbach's alpha was calculated to assess the internal consistency of the overall scale and each factor, which is used to indicate the overall degree of homogeneity of test items. Cronbach's alpha values above .90 are considered to reflect excellent consistency, while values between .80 and .90 suggest good consistency. Values between .70 and .80 indicate acceptable internal consistency, values between .60 - .70 are questionable, .50 to .60 are poor, and values between .50 are unacceptable (George & Mallery, 2003). The overall scale demonstrated excellent internal consistency ($\alpha = .94$, 95% CI [.93, .95]). The individual sub-factors also showed good to excellent reliability: AI Competence ($\alpha = .91$), Learning Awareness ($\alpha = .88$), Learning Strategies ($\alpha = .84$), Engagement and Efficiency ($\alpha = .86$), and Ethical Collaboration ($\alpha = .87$). Detailed statistics, including item-factor correlations, are presented in Appendix D.

Split-half reliability

Split-half reliability analysis was performed on the AI-SRL scale to further assess the consistency of the test items. Table 5 shows the results of this analysis.

Table 5
Split-half Reliability Analysis Results

	Number of items	Max – Min Means	Means	Cronbach's Alpha	Correlation between forms	Guttman split-half coefficient	Spearman-Brown coefficient
Part 1	11 ^a	3.56 ± 3.16	3.39	0.90	0.80	0.89	0.89
Part 2	11 ^b	3.59 ± 3.19	3.41	0.89			
Total	22	3.59 ± 3.16	3.40	0.94			

Notes: ^a The items are: 1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21.

^b The items are: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22.

As Table 5 indicates, the Spearman-Brown coefficient was .89 (95% CI [.87, .91]), and the Guttman Split-Half coefficient was .89 (95% CI [.87, .91]), indicating excellent split-half reliability. The correlation between forms was .80 (95% CI [.77, .83]).

Validity

According to Cheung et al. (2024), convergent validity requires CR ≥ 0.70, standardized loadings ≥ 0.50, and AVE ≥ 0.50-all of which are met in our model. Table 6 shows the scores of AI-SRL scales.

Table 6
CR, AVE, MSV, and Factor Correlations

	CR	AVE	MSV	Ethical collab.	AI comp.	Learning awareness	Engagemt and effic.	Learning strategies
Ethical collaboration	0.80	0.51	0.37	0.71				
AI competence	0.87	0.54	0.49	0.59	0.73			
Learning awareness	0.83	0.55	0.73	0.61	0.70	0.74		
Engagement and efficiency	0.85	0.59	0.65	0.51	0.67	0.80	0.77	
Learning strategies	0.83	0.55	0.73	0.49	0.68	0.85	0.80	0.74

Note: 95% confidence intervals for correlations are available in supplementary materials. All correlations significant at $p < .001$. CR = Composite Reliability; AVE = Average Variance Extracted; MSV = Maximum Shared Variance.

Discriminant validity was tested using multiple criteria: (a) no item cross-loads across constructs, (b) AVE for each construct exceeded squared inter-construct correlations, and (c) inter-factor correlations fell below the recommended values -none exceeded 0.85- (Cheung et al., 2024; Hair et al., 2010). While some high correlations were found among Learning Awareness, Engagement and Efficiency, and Learning Strategies, these constructs maintained theoretical distinctiveness within the AI-SRL framework. Thus, their retention was considered both psychometrically justifiable and conceptually meaningful.

Statistical analyses were performed using IBM SPSS Statistics 22 and AMOS 22. All confidence intervals were calculated at the 95% level using bias-corrected accelerated (BCa) bootstrap with 5000 resamples. Parallel analysis employed O'Connor's (2000) SPSS syntax, generating 1000 random correlation matrices for eigenvalue comparison. Significance levels were set at $\alpha = .05$ for all tests, with Bonferroni corrections applied for multiple comparisons where appropriate. The validated scale offers a reliable and valid measure of SRL in the context of AI, with potential applications in educational research and practice. Future research could explore the scale's use in diverse populations and settings.

DISCUSSION

This study successfully developed and validated the AI-SRL Scale, providing the psychometrically robust instrument specifically designed to measure SRL capabilities in AI-supported educational environments. For the preliminary analyses, a tool of 34 items was developed based on the literature and expert views. Factor analysis utilizing KMO revealed an excellent suitability (.96), and Bartlett's Test of Sphericity proved the multivariate normal distribution of the dataset. After EFA and PCA to establish meaningful factors and reveal correlations among these factors, 12 items were removed from the scale to preserve the scale's theoretical coherence and psychometric integrity. Based on the responses from 928 participants, EFA and CFA analyses were performed. The content validity, construct validity, internal consistency and split-half reliability results of the study proved that AI-SRL is a psychometrically sound scale in assessing the SRL in AI in educational environments. The finalized version of this scale involved 22 items under 5 factors, which are AI Competence, Learning Awareness, Learning Strategies, Engagement and Efficiency, and Ethical Collaboration.

This five-factor structure shows that AI use in educational environments deepens and contributes to the traditional construct of self-regulation by elaborating on competence, learning awareness, learning strategies, engagement and efficiency, and ethical collaboration. This new five-factor structure complies with the recent literature by highlighting how AI integration into educational settings advances the conventional SRL theories proposed by Zimmerman (1989) and Pintrich (1999). Furthermore, the AI Competence factor in the current study contributes to the previous literature. For example, Bećirović et al. (2025) explored a significant correlation between academic performance and AI self-efficacy. Also, AI Competence factor can be associated to cognitive regulation strategies of Pintrich (1999) since the current study highlighted that AI tools can assist elaboration strategies by providing individualized explanations and suggestions. Moreover, the Learning Awareness factor echoes Zimmerman's self-observation process of performance phase (Panadero, 2017; Zimmerman, 1989) by indicating AI's potentials in enhancing metacognitive monitoring and learning awareness. Gkintoni et al. (2025) revealed that feedback from AI tools, enabling simultaneous performance measuring, promote learners' metacognitive awareness. This awareness supports students in following their learning processes more consciously. In addition, the factor of Ethical Collaboration contributes to the previous literature. For example, the issues such as data privacy, surveillance, algorithmic bias, and academic integrity were highlighted as ethical concerns that may directly affect SRL in educational contexts in terms of AI applications (Gkintoni et al., 2025; Walter, 2024). Thus, this factor underpins students' ethical responsibilities in using AI tools and highlights the challenges that may emerge in AI-based educational

environments. The Ethical Collaboration factor represents a paradigmatic extension beyond classical SRL frameworks, introducing dimensions that neither Zimmerman nor Pintrich anticipated. This factor operationalizes what Pintrich (1999) termed "context regulation" but extends it to encompass algorithmic transparency, data sovereignty, and collaborative ethics in human-AI interactions. The emergence of this factor as statistically and conceptually distinct suggests that ethical considerations in AI-mediated learning constitute not merely an add-on to existing frameworks but a fundamental regulatory domain requiring dedicated theoretical attention. This finding aligns with recent calls for "algorithmic literacy" as a core competency in digital education (Walter, 2024), suggesting that self-regulated learners must now navigate not only cognitive and motivational challenges but also ethical complexities inherent in AI-assisted learning ecosystems. Regarding the factor of Engagement and Efficiency, AI technologies can present invaluable potentials to support cognitive and metacognitive dimensions of SRL while they may fall short in motivational regulation (Lan & Zhou, 2025; Seo et al., 2021; Wang & Lin, 2023). Considering Zimmerman's (1989) and Pintrich (1999) theories, motivational regulation is closely associated with trust, affinity, and individualized affective support which are provided by AI supported educational tools. Thus, Engagement and Efficiency as a factor in the current study emerges as a critical area for new research to elaborately investigate the motivational impacts of AI implementations. The Learning Strategies factor aligns with both Pintrich's (1999) cognitive control phase and Zimmerman's (1989) performance phase, capturing how learners select, adapt, and implement AI-supported cognitive strategies for learning. This factor reflects the literature's findings that AI tools effectively support strategy adaptation and selection, with students using AI to find effective study methods, improve problem-solving skills, and enhance critical thinking through multiple perspective analysis (Liu et al., 2024; Seo et al., 2021). The inclusion of this factor acknowledges that contemporary learners actively integrate AI recommendations into their strategic repertoire, representing a shift from traditional strategy use to AI-augmented strategic learning approaches that require both technological fluency and metacognitive judgment about when and how to deploy AI-suggested strategies.

CONCLUSION AND FUTURE DIRECTIONS

Rapid advancements in AI tools have made it necessary for educational environments to adapt them into educational processes appropriately. Along with other psychological constructs, it has become evident that developing a measurement tool regarding SRL in using AI in educational environments to employ in research studies for various purposes seems essential. For this purpose, the current study aimed to conduct the validity and reliability study of AI-SRL scale to address this need in the literature. The current study proposed the AI-SRL scale, and it can be used as a psychometrically sound scale to measure students' SRL skills in AI use in educational environments. The AI-SRL scale can make an original contribution to the literature, and it has potential as an effective instrument for implementations in AI-mediated educational contexts. For future research, cross-cultural validation studies of this instrument can be conducted to analyze its applicability in different cultural settings. Also, the data of the current study were collected in higher education contexts, and the future data can be collected in secondary education contexts or professional contexts to see whether it yields different results.

The AI-SRL scale represents a pioneering contribution to educational measurement by being among the first validated instruments specifically designed to capture the multidimensional nature of SRL in AI-enhanced environments. Unlike generic SRL measures or technology acceptance scales, this instrument uniquely bridges theoretical foundations from educational psychology with contemporary realities of AI integration, offering researchers and practitioners a psychometrically robust tool for understanding how learners navigate the complexities of AI-mediated education. The scale's potential extends beyond individual assessment to inform institutional policy and pedagogical innovation. Educational institutions implementing AI-powered learning systems can utilize the scale to benchmark student readiness, identify support needs, and evaluate the effectiveness of AI integration initiatives. The five-factor structure provides actionable insights for curriculum designers, suggesting that AI literacy programs should address not only technical competence but also metacognitive awareness, strategic adaptation, motivational engagement, and ethical reasoning.

For future research, cross-cultural validation studies of this instrument can be conducted to analyze its applicability in different cultural settings. Also, the data of the current study were collected in higher education contexts, and the future data can be collected in secondary education contexts or professional contexts to see whether it yields different results. Future validation studies should prioritize cross-cultural investigations, particularly in non-Western educational contexts where collectivist learning orientations might interact differently with AI tools. The scale's applicability in K-12 settings warrants investigation, as younger learners may exhibit distinct patterns of AI-enhanced self-regulation. Professional training contexts, including corporate learning environments and continuing education programs, represent another frontier where the AI-SRL scale could illuminate how working professionals integrate AI tools into their lifelong learning practices. Longitudinal studies tracking changes in AI-SRL profiles over time would provide valuable insights into developmental trajectories and the impact of educational interventions.

PRACTICAL IMPLICATIONS AND LIMITATIONS

The AI-SRL scale can be implemented as an effective reliable tool from planning to assessment in educational settings. It can be applied to measure students' AI-SRL levels, learning analytics can be designed or observed based on the reports from the scale, and it can provide insights for instructional designs. Also, the scale can be employed in face to face, online, and hybrid education. The AI-SRL scale offers concrete applications for enhancing teaching practices across diverse educational modalities. In diagnostic assessment, educators can administer the scale at course onset to identify students who may struggle with AI-enhanced learning components, enabling targeted support interventions. For learning analytics applications, scale scores can be integrated with digital trace data to create comprehensive learner profiles, revealing patterns between self-reported AI-SRL capabilities and actual tool usage behaviors. This integration becomes particularly valuable in massive open online courses (MOOCs) where personalized support at scale remains challenging. In instructional design, the five-factor structure guides the development of scaffolding strategies tailored to different aspects of AI-enhanced learning. For instance, students scoring low on AI Competence might benefit from technical orientation modules, while those with weak Ethical Collaboration scores could engage with case studies on

responsible AI use. In hybrid learning environments, the scale can inform decisions about when to deploy AI support versus human instruction, optimizing the blend based on students' self-regulatory profiles. Distance education programs can use the scale to assess whether students possess the requisite AI-SRL skills for success in autonomous learning environments, potentially incorporating scale-based prerequisites or preparatory courses.

However, the current study involved several limitations. First, most of the participants were chosen from foreign language teaching and English language literature departments of universities in Turkey. Involving most of the participants from two departments may limit generalizability of the findings to other higher education contexts. Second, convenience and purposeful sampling methods were used to reach the participants in this study. Non-random sampling may hinder the generalizability of the findings to other populations. Third, the items of the scale can be interpreted differently by individuals from different linguistic backgrounds. Last, disciplinary biases may influence the effective implementation of the instrument negatively. However, several limitations warrant consideration. The scale's development within Turkish higher education contexts, with participants predominantly from language-related disciplines, may have limited cross-cultural and cross-disciplinary generalizability. The English-language items may have introduced response biases among non-native speakers, potentially affecting measurement invariance across linguistic groups. Additionally, the rapidly evolving nature of AI technologies means the scale will require periodic updates to maintain relevance.

Ethical approval

This research received ethical approval from Social Sciences University of Ankara, as per decision number 98038.

Ethical use of AI tools

In accordance with transparency principles, we disclose the use of AI tools in manuscript preparation. We utilized Claude Opus primarily for linguistic refinement—specifically to improve grammatical structures, enhance sentence flow, and ensure consistency in academic terminology across sections. This tool served exclusively as editorial assistants for language polishing, not for generating research content, data analysis, or theoretical contributions. All substantive content, including research design, data collection, statistical analyses, and interpretation of findings, represents original work by the authors. We maintained full editorial control throughout the writing process, critically reviewing and modifying all AI-suggested revisions. The authors assume complete responsibility for the accuracy, originality, and scholarly integrity of the final manuscript.

Data availability

The data and materials used in this study are available upon request from the corresponding author.

Funding

The authors report no funding.

REFERENCES

- Adams, D., Chuah, K.-M., Devadason, E., & Azzis, M. S. A. (2024). From novice to navigator: Students' academic help-seeking behaviour, readiness, and perceived usefulness of ChatGPT in learning. *Education and Information Technologies*, 29(11), 13617-13634. <https://doi.org/10.1007/s10639-023-12427-8>
- Anggoro, K. J., & Pratiwi, D. I. (2023). Fostering self-assessment in English learning with a generative AI platform: A case of Quizizz AI. *Studies in Self-Access Learning Journal*, 14(4), 489-501. <https://doi.org/10.37237/140406>
- Baker, R. S. (2023). AI and self-regulated learning theory: What could be on the horizon? *Computers in Human Behavior*, 147, 107849. <https://doi.org/10.1016/j.chb.2023.107849>
- Bauer, E., Greiff, S., Graesser, A. C., Scheiter, K., & Sailer, M. (2025). Looking beyond the hype: Understanding the effects of AI on learning. *Educational Psychology Review*, 37(2). <https://doi.org/10.1007/s10648-025-10020-8>
- Bećirović, S., Polz, E., & Tinkel, I. (2025). Exploring students' AI literacy and its effects on their AI output quality, self-efficacy, and academic performance. *Smart Learning Environments*, 12, 29. <https://doi.org/10.1186/s40561-025-00384-3>
- Biju, N., Abdelrasheed, N. S. G., Bakiyeva, K., Prasad, K. D. V., & Jember, B. (2024). Which one? AI-assisted language assessment or paper format: An exploration of the impacts on foreign language anxiety, learning attitudes, motivation, and writing performance. *Language Testing in Asia*, 14, 45. <https://doi.org/10.1186/s40468-024-00322-z>
- Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quinonez, H. R., & Young, S. L. (2018). Best practices for developing and validating scales for health, social, and behavioral research: A primer. *Frontiers in Public Health*, 6. <https://doi.org/10.3389/fpubh.2018.00149>
- Boguslawski, S., Deer, R., & Dawson, M. G. (2025). Programming education and learner motivation in the age of generative AI: Student and educator perspectives. *Information and Learning Sciences*, 126(1/2), 91-109. <https://doi.org/10.1108/ILS-10-2023-0163>
- Büyüköztürk, Ş. (2018). *Sosyal bilimler için veri analizi el kitabı*. Pegem Akademi Yayıncılık. <https://doi.org/10.14527/9789756802748>
- Chang, D. H., Lin, M. P.-C., Hajian, S., & Wang, Q. Q. (2023). Educational design principles of using AI chatbot that supports self-regulated learning in education: Goal setting, feedback, and personalization. *Sustainability*, 15(17), 12921. <https://doi.org/10.3390/su151712921>
- Cheung, G. W., Cooper-Thomas, H. D., Lau, R. S., & Wang, L. C. (2024). Reporting reliability, convergent and discriminant validity with structural equation modeling: A review and best-practice recommendations. *Asia Pacific Journal of Management*, 41(2), 745-783. <https://doi.org/10.1007/s10490-023-09871-y>
- Chiu, T. K. F. (2024). A classification tool to foster self-regulated learning with generative artificial intelligence by

- applying self-determination theory: A case of ChatGPT. *Educational Technology Research and Development*, 72(4), 2401-2416. <https://doi.org/10.1007/s11423-024-10366-w>
- Chiu, T. K. F., Moorhouse, B. L., Chai, C. S., & Ismailov, M. (2023). Teacher support and student motivation to learn with artificial intelligence (AI) based chatbot. *Interactive Learning Environments*, 1-17. <https://doi.org/10.1080/10494820.2023.2172044>
- Creswell, J. W. (2014). *Research design: Qualitative, quantitative and mixed methods approaches* (4th ed.). SAGE.
- Dahri, N. A., Yahaya, N., Al-Rahmi, W. M., Aldraiweesh, A., Alturki, U., Almutairy, S., Shutaleva, A., & Soomro, R. B. (2024). Extended TAM based acceptance of AI-powered ChatGPT for supporting metacognitive self-regulated learning in education: A mixed-methods study. *Heliyon*, 10(8), e29317. <https://doi.org/10.1016/j.heliyon.2024.e29317>
- DeVellis, R. F. (1991). *Scale development: Theory and applications*. SAGE.
- Eisinga, R., te Grotenhuis, M., & Pelzer, B. (2013). The reliability of a two-item scale: Pearson, Cronbach, or Spearman-Brown? *International Journal of Public Health*, 58(4), 637-642. <https://doi.org/10.1007/s00038-012-0416-3>
- Fabrigar, L. R., & Wegener, D. T. (2012). *Exploratory factor analysis*. Oxford University Press. <https://doi.org/10.1093/acprof:osobl/9780199734177.001.0001>
- Field, A. P. (2018). *Discovering statistics using IBM SPSS statistics* (5th ed.). SAGE.
- Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment*, 7(3), 286-299. <https://doi.org/10.1037/1040-3590.7.3.286>
- George, D., & Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference, 11.0 update* (4th ed.). Allyn & Bacon.
- Gerlich, M. (2025). AI tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies*, 15(1), 6. <https://doi.org/10.3390/soc15010006>
- Gkintoni, E., Antonopoulou, H., Sortwell, A., & Halkiopoulos, C. (2025). Challenging cognitive load theory: The role of educational neuroscience and artificial intelligence in redefining learning efficacy. *Brain Sciences*, 15(2), 203. <https://doi.org/10.3390/brainsci15020203>
- Glick, D., Miedijensky, S., & Zhang, H. (2024). Examining the effect of AI-powered virtual-human training on STEM majors' self-regulated learning behavior. *Frontiers in Education*, 9, 1465207. <https://doi.org/10.3389/feduc.2024.1465207>
- Grüneke, T., Guggenberger, T., Hofmeister, S., & Stoetzer, J. C. (2024). AI-enabled self-regulated learning: A multi-layer taxonomy development. In *Proceedings of the 32nd European Conference on Information Systems (ECIS 2024)*, Paphos, Cyprus. (ISBN 978-1-958200-10-0). https://aisel.aisnet.org/ecis2024/track13_learning_teach/track13_learning_teach/4
- Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis* (7th ed.). Pearson.
- Hertzog, M. A. (2008). Considerations in determining sample size for pilot studies. *Research in Nursing & Health*, 31, 180-191. <https://doi.org/10.1002/nur.20247>
- Hill, R. (1998). What sample size is 'enough' in Internet survey research? *Interpersonal Computing and Technology: An Electronic Journal for the 21st Century*, 6, 1-10.
- Hilpert, J. C., Greene, J. A., & Bernacki, M. (2023). Leveraging complexity

- frameworks to refine theories of engagement: Advancing self-regulated learning in the age of artificial intelligence. *British Journal of Educational Technology*, 54(5), 1204-1221. <https://doi.org/10.1111/bjet.13340>
- Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1-55. <https://doi.org/10.1080/10705519909540118>
- Huang, A. Y. Q., Lu, O. H. T., & Yang, S. J. H. (2023). Effects of artificial intelligence-enabled personalized recommendations on learners' learning engagement, motivation, and outcomes in a flipped classroom. *Computers & Education*, 194, 104684. <https://doi.org/10.1016/j.compedu.2022.104684>
- Isaac, S., & Michael, W. B. (1995). *Handbook in research and evaluation: A collection of principles, methods, and strategies useful in the planning, design, and evaluation of studies in education and the behavioral sciences* (3rd ed.). EdITS Publishers.
- Jin, S.-H., Im, K., Yoo, M., Roll, I., & Seo, K. (2023). Supporting students' self-regulated learning in online learning using artificial intelligence applications. *International Journal of Educational Technology in Higher Education*, 20, 37. <https://doi.org/10.1186/s41239-023-00406-5>
- Johanson, G. A., & Brooks, G. P. (2009). Initial scale development: Sample size for pilot studies. *Educational and Psychological Measurement*, 70(3), 394-400. <https://doi.org/10.1177/0013164409355692>
- Jolliffe, I. (2014). Principal component analysis. In N. Balakrishnan, T. Colton, B. Everitt, & W. W. Piegorsch (Eds.), *Wiley StatsRef: Statistics Reference Online*. John Wiley & Sons. <https://doi.org/10.1002/9781118445112.stat06472>
- Kaiser, H. F., & Rice, J. (1974). Little jiffy, mark IV. *Educational and Psychological Measurement*, 34(1), 111-117. <https://doi.org/10.1177/001316447403400115>
- Kline, P. (2014). *An easy guide to factor analysis*. Routledge. <https://doi.org/10.4324/9781315788135>
- Kong, S. C., & Yang, Y. (2024). A human-centered learning and teaching framework using generative artificial intelligence for self-regulated learning development through domain knowledge learning in K-12 settings. *IEEE Transactions on Learning Technologies*, 17, 1562-1573. <https://doi.org/10.1109/TLT.2024.3392830>
- Lai, J. W. (2024). Adapting self-regulated learning in an age of generative artificial intelligence chatbots. *Future Internet*, 16(6), 218. <https://doi.org/10.3390/fi16060218>
- Lan, M., & Zhou, X. (2025). A qualitative systematic review on AI empowered self-regulated learning in higher education. *npj Science of Learning*, 10(1). <https://doi.org/10.1038/s41539-025-00319-0>
- Li, Y., Sadiq, G., Qambar, G., & Zheng, P. (2025). The impact of students' use of ChatGPT on their research skills: The mediating effects of autonomous motivation, engagement, and self-directed learning. *Education and Information Technologies*, 30(4), 4185-4216. <https://doi.org/10.1007/s10639-024-12981-9>
- Liang, J., Wang, L., Luo, J., Yan, Y., & Fan, C. (2023). The relationship between student interaction with generative artificial intelligence and learning achievement: Serial mediating roles of self-efficacy and cognitive engagement. *Frontiers in Psychology*, 14, 1285392. <https://doi.org/10.3389/fpsyg.2023.1285392>

- Liu, Z. M., Hwang, G. J., Chen, C. Q., Chen, X. D., & Ye, X. D. (2024). Integrating large language models into EFL writing instruction: Effects on performance, self-regulated learning strategies, and motivation. *Computer Assisted Language Learning*, 1-25. <https://doi.org/10.1080/09588221.2024.2389923>
- Mohamed, A. M., Shaaban, T. S., Bakry, S. H., Guillén-Gámez, F. D., & Strzelecki, A. (2025). Empowering the faculty of education students: Applying AI's potential for motivating and enhancing learning. *Innovative Higher Education*, 50(2), 587-609. <https://doi.org/10.1007/s10755-024-09747-z>
- Molenaar, I., Mooij, S. de, Azevedo, R., Bannert, M., Järvelä, S., & Gašević, D. (2023). Measuring self-regulated learning and the role of AI: Five years of research using multimodal multichannel data. *Computers in Human Behavior*, 139, 107540. <https://doi.org/10.1016/j.chb.2022.107540>
- Mooney, C. Z., & Duval, R. D. (1993). *Bootstrapping: A nonparametric approach to statistical inference*. SAGE. <https://doi.org/10.4135/9781412983532>
- Morgado, F. F. R., Meireles, J. F. F., Neves, C. M., Amaral, A. C. S., & Ferreira, M. E. C. (2017). Scale development: Ten main limitations and recommendations to improve future research practices. *Psicologia: Reflexão e Crítica*, 30(1). <https://doi.org/10.1186/s41155-016-0057-1>
- Muthén, L. K., & Muthén, B. O. (2017). *Mplus: Statistical analysis with latent variables: User's guide* (Version 8). Muthén & Muthén.
- Ng, D. T. K., Tan, C. W., & Leung, J. K. L. (2024). Empowering student self-regulated learning and science education through ChatGPT: A pioneering pilot study. *British Journal of Educational Technology*, 55(4), 1328-1353. <https://doi.org/10.1111/bjet.13454>
- O'Connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instruments, & Computers*, 32(3), 396-402. <https://doi.org/10.3758/BF03200807>
- Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology*, 8, 422. <https://doi.org/10.3389/fpsyg.2017.00422>
- Pintrich, P. R. (1999). The role of motivation in promoting and sustaining self-regulated learning. *International Journal of Educational Research*, 31(6), 459-470. [https://doi.org/10.1016/S0883-0355\(99\)00015-4](https://doi.org/10.1016/S0883-0355(99)00015-4)
- Pintrich, P. R. (2000). The role of goal orientation in self-regulated learning. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 451-502). Academic Press. <https://doi.org/10.1016/B978-012109890-2/50043-3>
- Schunk, D. H., & Zimmerman, B. J. (2012). Self-regulation and learning. In I. B. Weiner, G. E. Miller, & W. M. Reynolds (Eds.), *Handbook of psychology: Vol. 7. Educational psychology* (2nd ed., pp. 59-78). John Wiley & Sons. <https://doi.org/10.1002/9781118133880.hop207003>
- Seo, K., Tang, J., Roll, I., Fels, S., & Yoon, D. (2021). The impact of artificial intelligence on learner-instructor interaction in online learning. *International Journal of Educational Technology in Higher Education*, 18(1), 54. <https://doi.org/10.1186/s41239-021-00292-9>
- Sun, J. (2005). Assessing goodness of fit in confirmatory factor analysis. *Measurement and Evaluation in Counseling and Development*, 37(4), 240-256. <https://doi.org/10.1080/07481756.2005.11909764>

- Teeluckdharry, N., Teeroovengadam, V., & Seebaluck, A. (2021). Scale development in marketing research. In R. Nunkoo, V. Teeroovengadam, & C. M. Ringle (Eds.), *Handbook of research methods for marketing management* (pp. 17-50). Edward Elgar. <https://doi.org/10.4337/9781788976954.00007>
- Tinajero, C., Mayo, M. E., Villar, E., & Martínez-López, Z. (2024). Classic and modern models of self-regulated learning: Integrative and componential analysis. *Frontiers in Psychology*, 15, 1307574. <https://doi.org/10.3389/fpsyg.2024.1307574>
- Walter, Y. (2024). Embracing the future of artificial intelligence in the classroom: The relevance of AI literacy, prompt engineering, and critical thinking in modern education. *International Journal of Educational Technology in Higher Education*, 21(1), 15. <https://doi.org/10.1186/s41239-024-00448-3>
- Wang, C. Y., & Lin, J. J. H. (2023). Utilizing artificial intelligence to support analyzing self-regulated learning: A preliminary mixed-methods evaluation from a human-centered perspective. *Computers in Human Behavior*, 144, 107721. <https://doi.org/10.1016/j.chb.2023.107721>
- Wu, X. Y., & Chiu, T. K. F. (2025). Integrating learner characteristics and generative AI affordances to enhance self-regulated learning: A configurational analysis. *Journal of New Approaches in Educational Research*, 14(1). <https://doi.org/10.1007/s44322-025-00028-x>
- Xia, Q., Chiu, T. K. F., Chai, C. S., & Xie, K. (2023). The mediating effects of needs satisfaction on the relationships between prior knowledge and self-regulated learning through artificial intelligence chatbot. *British Journal of Educational Technology*, 54(4), 967–986. <https://doi.org/10.1111/bjet.13305>
- Yot-Domínguez, C., & Marcelo, C. (2017). University students' self-regulated learning using digital technologies. *International Journal of Educational Technology in Higher Education*, 14, 38. <https://doi.org/10.1186/s41239-017-0076-8>
- Zimmerman, B. J. (1989). Models of self-regulated learning and academic achievement. In B. J. Zimmerman & D. H. Schunk (Eds.), *Self-regulated learning and academic achievement: Theory, research, and practice* (pp. 1-25). Springer. https://doi.org/10.1007/978-1-4612-3618-4_1

APPENDICES

Appendix A: Items from the AI-Enhanced Self-Regulated Learning (AI-SRL) Scale

Factors and items	Completely disagree		Completely agree		
	1	2	3	4	5
AI Competence					
1. I can easily use AI tools with different interfaces.	☐	☐	☐	☐	☐
2. Active use of AI tools improves my ability to learn.	☐	☐	☐	☐	☐
3. I can efficiently utilize AI tools to find information relevant to my studies.	☐	☐	☐	☐	☐
4. I feel confident in my ability to learn via new AI tools.	☐	☐	☐	☐	☐
5. The instructions from AI tools are easy to understand.	☐	☐	☐	☐	☐
6. I can easily make sense of AI-generated information in my learning process.	☐	☐	☐	☐	☐
Learning Awareness	1	2	3	4	5
7. Feedback from AI tools helps me understand my learning strengths.	☐	☐	☐	☐	☐
8. AI-provided feedback helps me identify areas for improvement.	☐	☐	☐	☐	☐
9. I assess the effectiveness of AI tools in helping me achieve my learning objectives.	☐	☐	☐	☐	☐
10. I think critically about my dependency on AI tools for learning.	☐	☐	☐	☐	☐
Learning Strategies	1	2	3	4	5
11. I adapt my study methods according to recommendations provided by AI tools.	☐	☐	☐	☐	☐
12. I use AI tools to find the most effective ways to learn new subjects.	☐	☐	☐	☐	☐
13. I utilize AI-generated information to improve my problem-solving skills.	☐	☐	☐	☐	☐
14. By using AI tools to analyze different perspectives, I can enhance my critical thinking skills.	☐	☐	☐	☐	☐

Engagement and Efficiency	1	2	3	4	5
15. My participation in learning activities is higher with AI use.	☐	☐	☐	☐	☐
16. AI tools help me learn more permanently.	☐	☐	☐	☐	☐
17. I think I learn better when I use AI tools.	☐	☐	☐	☐	☐
18. I think AI tools that adapt to my learning style enhance my engagement in learning.	☐	☐	☐	☐	☐
Ethical Collaboration	1	2	3	4	5
19. I know I have to take all the responsibility when I use AI tools for learning.	☐	☐	☐	☐	☐
20. I am conscious of privacy concerns in AI-tools.	☐	☐	☐	☐	☐
21. I am aware of potential biases in AI tools.	☐	☐	☐	☐	☐
22. I am conscious about the security of the information that I provide to AI tools.	☐	☐	☐	☐	☐

Appendix B: Items removed from final model with empirical and conceptual rationale

Stage	Item (full wording)	Original factor	Loading	Reason for removal
EFA	I use AI tools to check how effectively I'm learning.	Probably Learning Awareness	.463/.439	Cross-loading < 0.10 difference; with loadings .463 and .439 for two factors
EFA	Thinking deeply about the reasons for using AI tools helps me apply them better in my studies.	Probably Learning Strategies	.285/.463	Cross-loading on two items; similar to other academic integrity items, also decreased the AVE when tried in CFA
EFA	I know how to use AI generated information ethically when learning with AI.	Probably Ethical Collaboration	.329/-.260/.491	Cross-loading < 0.10 difference; with loadings .329, -.260 and .491 for three factors
CFA	I find it easy to adapt to the changing features in AI tools.	AI Competence	.66	Decreased the AVE; concept covered by stronger AI Competence items
CFA	I am skilled at using specific functions of AI tools (like data analysis or content recommendation) for learning.	AI Competence	.64	Decreased discriminant validity; probably vague

CFA	AI tools help me prioritize my learning goals.	Learning Awareness	.64	Low loading, decreased AVE and discriminant validity
CFA	AI tools help me access a wide range of learning resources.	Learning Awareness	.62	Decreased AVE; probably vague
CFA	AI tools make me more autonomous (independent) in my learning process.	Learning Awareness	.59	Low loading; weak representation of construct
CFA	AI tools help me manage my learning time better.	Learning Awareness	.63	Low loading; decreased AVE
CFA	AI tools help me develop strategies for learning tasks.	Learning Awareness	.65	Low loading; concept covered by other self-monitoring items
CFA	AI tools help me understand complex topics more easily.	Learning Awareness	.64	Decreased discriminant validity; better represented by other learning-monitoring items
CFA	I am skeptical about the capability of AI tools in providing information accuracy.	Ethical Collaboration	.65	Marginal loading; probably vague

Appendix C: Factor loadings for the initial EFA model

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Item1	.860				
Item2	.853				
Item3	.555				
Item4	.698				
Item5	.693				
Item6	.690				
Item7	.672	.			
Item8	.628				
Item9		.665			
Item10		.464			
Item11		.676			
Item12		.776			
Item13		.489			
Item14		.449			
Item15				.725	
Item16				.751	

Item17		.835	
Item18		.589	
Item19	.684		
Item20	.571		
Item22	.498		
Item23	.669		
Item25			.636
Item26			.916
Item27			.861
Item28			.669
Item30	.625		
Item31	.867		
Item32	.789		
Item33	.729		
Item34	.850		

Note: All loadings significant at $p < .001$. Standard errors ranged from .028 to .045.

Appendix D: Cronbach's Alpha for each factor and item-factor correlations

Factors/item description		Corrected item-total correlation	Cronbach's alpha if item deleted	Cronbach's alpha
Factor 1. AI Competence				
1.	I can easily use AI tools with different interfaces.	.60	.94	.88
2.	Active use of AI tools improves my ability to learn.	.60	.94	
3.	I can efficiently utilize AI tools to find information relevant to my studies.	.66	.93	
4.	I feel confident in my ability to learn via new AI tools.	.66	.93	
5.	The instructions from AI tools are easy to understand.	.64	.93	
6.	I can easily make sense of AI-generated information in my learning process.	.72	.93	
Factor 2. Learning Awareness				
7.	Feedback from AI tools helps me understand my learning strengths.	.61	.94	.83
8.	AI-provided feedback helps me identify areas for improvement.	.62	.93	
9.	I assess the effectiveness of AI tools in helping me achieve my learning objectives.	.65	.93	
10.	I think critically about my dependency on AI tools for learning.	.67	.93	
Factor 3. Learning Strategies				
11.	I adapt my study methods according to recommendations provided by AI tools.	.66	.93	.84

Factors/item description	Corrected item-total correlation	Cronbach's alpha if item deleted	Cronbach's alpha
12. I use AI tools to find the most effective ways to learn new subjects.	.71	.93	
13. I utilize AI-generated information to improve my problem-solving skills.	.70	.93	
14. By using AI tools to analyze different perspectives, I can enhance my critical thinking skills.	.56	.94	
Factor 4. Engagement and Efficiency			
15. My participation in learning activities is higher with AI use.	.64	.93	
16. AI tools help me learn more permanently.	.60	.94	
17. I think I learn better when I use AI tools.	.63	.93	.86
18. I think AI tools that adapt to my learning style enhance my engagement in learning.	.74	.93	
Factor 5. Ethical Collaboration			
19. I know I have to take all the responsibility when I use AI tools for learning.	.58	.94	
20. I am conscious of privacy concerns in AI-tools.	.54	.94	
21. I am aware of potential biases in AI tools.	.60	.94	.84
22. I am conscious about the security of the information that I provide to AI tools.	.55	.94	
Whole scale			.94
Note: 95% CIs calculated using 5000 bootstrap samples. All item-total correlations significant at $p < .001$.			

Date of reception: 1 June 2025

Date of acceptance: 27 August 2025

Date of approval for layout: 25 September 2025

Date of publication in OnlineFirst: 16 October 2025

Date of publication: 1 January 2026

The influence of learning motivation and procrastination on ChatGPT dependence

Influencia de la motivación hacia el aprendizaje y la procrastinación en la dependencia a ChatGPT

 Héctor Galindo-Domínguez - Universidad del País Vasco, UPV/EHU (Spain)

 Martín Sainz-de-la-Maza - Universidad del País Vasco, UPV/EHU (Spain)

 Lucía Campo - Universidad del País Vasco, UPV/EHU (Spain)

 Daniel Losada-Iglesias - Universidad del País Vasco, UPV/EHU (Spain)



ABSTRACT

In a short period of time, generative artificial intelligences such as ChatGPT have become a widely used support tool for students' learning processes, altering how many approach academic tasks. Their immediacy and accessibility make them attractive alternatives to traditional study resources, to the extent that some students may have even developed a certain degree of dependence on these tools. The present study investigates the impact of learning motivation and procrastination on ChatGPT dependence. A total of 467 university students from the field of education participated by completing a series of validated scales measuring different types of motivation defined in Self-Determination Theory (extrinsic motivation, intrinsic motivation, and amotivation), procrastination, and ChatGPT dependence. The indirect effects of the mediation analyses revealed that students with lower levels of intrinsic motivation ($\beta = -.076$; $LL = -.121$; $UL = -.037$) and higher levels of amotivation ($\beta = .090$; $LL = .041$; $UL = .144$) were more likely to procrastinate frequently, with procrastination emerging as a significant factor contributing to greater ChatGPT dependence. Similarly, results indicated that students with high extrinsic motivation (without procrastination serving as a mediator) were more prone to develop greater dependence on ChatGPT ($\beta = .122$; $p = .022$). These findings highlight the importance of implementing strategies that foster intrinsic motivation and self-regulation, helping students use generative AI-based tools appropriately while developing essential competencies that could be at risk from excessive use of these tools, such as critical thinking and problem-solving.

Keywords: motivation; dependence; ChatGPT; procrastination; artificial intelligence; higher education.

RESUMEN

En poco tiempo, las inteligencias artificiales generativas como ChatGPT se han convertido en herramientas de apoyo para el aprendizaje, transformando la forma en que muchos estudiantes afrontan las tareas académicas. Su inmediatez y accesibilidad las hacen especialmente atractivas frente a los recursos tradicionales, hasta el punto de que algunos estudiantes pueden desarrollar cierta dependencia a ellas. Este trabajo investiga el impacto de la motivación hacia el aprendizaje y la procrastinación en la dependencia a ChatGPT. 467 estudiantes universitarios del área de la Educación participaron completando una serie de escalas validadas para evaluar diferentes tipos de motivación definidas en la *Teoría de la Autodeterminación* (motivación extrínseca, intrínseca y amotivación), procrastinación y dependencia a ChatGPT. Los análisis de mediación mostraron que los estudiantes con menor motivación intrínseca ($\beta = -.076$; $LI = -.121$; $LS = -.037$) y mayor amotivación ($\beta = .090$; $LI = .041$; $LS = .144$) tendían a procrastinar con mayor frecuencia, siendo la procrastinación un factor clave que aumentaba la dependencia a *ChatGPT*. Además, el alumnado con alta motivación extrínseca (incluso sin la mediación de la procrastinación) resultó ser más propenso a desarrollar una mayor dependencia a *ChatGPT* ($\beta = .122$; $p = .022$). Estos hallazgos destacan la importancia de implementar estrategias que fomenten la motivación intrínseca y la autorregulación, ayudando a los estudiantes a utilizar adecuadamente las herramientas basadas en la IA generativa mientras desarrollan competencias esenciales que podrían estar en riesgo por el uso excesivo de estas herramientas, como el pensamiento crítico y la resolución de problemas.

Palabras clave: motivación; dependencia; ChatGPT; procrastinación; inteligencia artificial; educación superior.

How to cite: Galindo-Domínguez, H., Sainz-de-la-Maza, M., Campo, L., & Losada-Iglesias, D. (2026). The influence of learning motivation and procrastination on ChatGPT dependence [Influencia de la motivación hacia el aprendizaje y la procrastinación en la dependencia a ChatGPT]. *RIED-Revista Iberoamericana de Educación a Distancia*, 29(1), 217-240.

<https://doi.org/10.5944/ried.45497>

INTRODUCTION

In this study, the theoretical framework is designed to test a mediation model in which motivation serves as the independent variable, procrastination as the mediating variable, and dependence on ChatGPT as the dependent variable. To provide a clear roadmap for the reader, the literature review is structured logically. Firstly, it explores the foundational theories of motivation and procrastination, examining the established relationship between these two constructs. Secondly, dependence on artificial intelligence within the context of higher education is defined, focusing on an excessive reliance on these systems for academic tasks. Finally, the framework integrates these three concepts by linking motivation to dependence on ChatGPT and, subsequently, procrastination to this dependence, aligning the theoretical flow with the proposed model.

Motivation and procrastination in learning processes

Motivation constitutes a complex construct within human psychology and behavior, exerting an influence on the way people decide to allocate their time, the level of effort they put into a task, their thoughts and emotions about the activity, and their perseverance in completing it (Filgona et al., 2020).

Although numerous influential theories have been proposed to understand individual motivation, including *Attribution Theory*, *Expectancy-Value Theory*, the *ARCS model*, *Social Cognitive Theory*, and *Goal-Oriented Theory*, among others, the *Self-Determination Theory* (SDT henceforth) has recently gained increased prominence in explaining shifts in attitudes toward Artificial Intelligence (e.g., Chai et al., 2023; Chiu, 2024; Chiu & Chai, 2020; Xia et al., 2022). The taxonomy associated with this theory explains how different forms of motivation vary depending on the degree of self-determination, ranging from the absence of motivation, through motivations controlled by external factors, to more autonomous motivations linked to personal interests (Ryan & Deci, 2020). Indeed, amotivation refers to the absence of intention to act, often resulting from a perceived lack of competence or value in the activity. Extrinsic motivation, in contrast, is driven by external rewards or pressures, yet it varies in its level of autonomy depending on the degree of internalization. The highest form of autonomy in motivation is intrinsic, emerging from the inherent enjoyment and satisfaction derived from the activity itself, which promotes both engagement and persistence (Ryan & Deci, 2020).

Likewise, SDT posits that the different forms of motivation (intrinsic, extrinsic, and amotivation) are closely linked to the extent to which three basic psychological needs are fulfilled: Competence, Autonomy and Relatedness (Ryan & Deci, 2020).

Firstly, competence can be understood as the feeling of being effective and capable in one's interactions with the environment. It involves mastering tasks and achieving desired outcomes, fostering a sense of accomplishment and growth. Feeling competent strengthens intrinsic motivation as well as the more self-determined forms of extrinsic motivation, while a lack of competence may lead to amotivation. A variety of studies have shown how strong competence, often reflected in high performance in task execution, is a predictor of increased intrinsic motivation among students (e.g., Garon-Carrier et al., 2016; Van Bergen et al., 2022).

Secondly, autonomy can be defined as the sense of being in control of one's own actions and decisions. It allows individuals to act volitionally and pursue goals in a self-

endorsed way. Intrinsic motivation thrives when individuals perceive a sense of freedom and choice, whereas a lack of autonomy may contribute to amotivation or controlled forms of extrinsic motivation. In this regard, previous studies have revealed that students from autonomy-supportive teaching contexts have higher intrinsic motivation, greater classroom engagement and skill development, as well as a lower likelihood of dropout (e.g., Cheon et al., 2020; Ljubin-Golub et al., 2020; Reeve & Cheon, 2021).

Thirdly, the concept of relatedness concerns the fundamental need to establish connections with others, to experience belonging, and to develop meaningful interpersonal bonds. It encompasses the perception of being cared for, supported, and understood within social contexts. When this need is satisfied, individuals are more inclined to display autonomous forms of motivation, engaging more fully in activities that nurture their sense of connection. Previous evidence has shown that fostering relatedness in the classroom partially affects the increase in students' intrinsic motivation (e.g., Ganotice et al., 2024; Guay et al., 2019)

As previously observed, from the perspective of SDT, the main basic psychological needs have a significant impact on students' motivation. Furthermore, according to previous evidence, having this motivation is key to reducing certain constructs such as procrastination, a construct that is analyzed in more detail in this paper.

Procrastination is commonly described as an irrational tendency to delay essential tasks or responsibilities, despite the negative consequences this postponement may have for both individuals and organizations (Klingsieck, 2013). Procrastination represents a self-regulatory deficit associated with both personal dispositions and environmental factors (Hen & Goroshit, 2018). Specifically, studies indicate that task-related factors (such as unclear instructions, timing of rewards and punishments, and task aversiveness), personality traits (e.g., the five-factor model, motivation, and cognition), and environmental influences (e.g., temptation, incentives, and accountability) are key contributors to procrastination (Johnson & Bloom, 1995; Wypych et al., 2018). Procrastination may act as a barrier to achievement, exerting detrimental effects on emotional well-being and contributing to heightened levels of anxiety, depressive symptoms, and diminished self-esteem (Duru & Balkis, 2017). Moreover, chronic procrastinators frequently demonstrate reduced performance outcomes, such as lower academic results, postponed career advancements, and adverse health consequences (Legood et al., 2018; Bolden & Fillauer, 2020).

Focusing on the relationship between motivation and procrastination, this has been a widely explored topic in the scientific literature, showing that the relationship between motivation and procrastination is complex and appears to be primarily unidirectional, with motivation influencing procrastination rather than the other way around (e.g., Grund & Fries, 2018; Li et al., 2021; Wu & Fan, 2016). Specifically, previous evidence has revealed that lower levels of attainment value (Wypych et al., 2018; Wu & Fan, 2016), self-determination (Grund & Fries, 2018), perseverance (Wypych et al., 2018), and intrinsic motivation (Kok, 2016), as well as higher levels of cost (Wu & Fan, 2016), delay discounting (Wypych et al., 2018), amotivation and external motivation (Kok, 2016) lead to higher academic procrastination levels.

Dependence on artificial intelligence in higher education

Artificial Intelligence (AI) is defined as the ability of machines or computational systems to emulate and perform functions traditionally attributed to human intelligence, including reasoning, learning, and problem-solving (Morandín-Ahuerma, 2022). While the integration of AI into educational contexts remains a relatively recent phenomenon, existing research underscores several potential advantages, including support in drafting written assignments, the provision of automated feedback, and assistance with data and information analysis (Chen et al., 2020; Delgado et al., 2024; Wang et al., 2018). Nonetheless, studies also highlight notable limitations, such as reduced human interaction in the learning process, limited contextual understanding, and overdependence on AI, which may diminish essential skills like critical thinking and problem-solving (Ahmad et al., 2021; Jara & Ochoa, 2020; Korteling et al., 2021). This study centers on the latter concern: AI dependence.

In the clinical domain, dependence has historically been linked to the use of substances such as alcohol and drugs, typically identified through indicators like tolerance and withdrawal (American Psychiatric Association, 2013). Yet, with the advent of the artificial intelligence era, this understanding has expanded to encompass behavioral forms of addiction, including pathological gambling and internet addiction disorder, which exhibit comparable neurobiological and behavioral patterns. (e.g., Bucur et al., 2021; Siste et al., 2019). In an academic context, dependence on AI is defined as an excessive reliance on automated systems for tasks such as decision-making or task execution (Morales-García et al., 2024), a phenomenon that manifests as a lack of confidence in one's own academic abilities and a preference for using AI as the primary, if not sole, source of assistance (Del Cisne et al., 2024). In fact, recent evidence suggests that simply knowing that advice comes from an AI can lead students to overrely on it, even when it contradicts available information or their own judgment, potentially producing suboptimal outcomes for themselves and others (Klingbeil et al., 2024). One of the main reasons, as noted by Baird and Maruping (2021), is that AI systems and chatbots increasingly replicate human interactions in a more natural manner, leading students to place greater trust in these types of systems.

This phenomenon can be interpreted through the lens of the *Unified Theory of Acceptance and Use of Technology* (Venkatesh et al., 2003), which offers a framework for analyzing dependence on AI by identifying the primary factors affecting its adoption and use. The theory emphasizes that constructs such as “performance expectancy, effort expectancy, social influence, and facilitating conditions” are pivotal in shaping user behavior. Dependence on AI may emerge when users hold high performance expectancy, perceiving that AI substantially improves task efficiency or accuracy, and low effort expectancy, as AI simplifies otherwise complex processes. Furthermore, social influence, including peer pressure to adopt and use such tools, can amplify dependence on these systems. Lastly, facilitating conditions, such as accessible infrastructure and resources, create an environment conducive to frequent AI use, potentially leading to overdependence.

Currently, there are some studies that reveal how the prevalence of dependence on artificial intelligence tools such as ChatGPT is moderately low, ranging from 16% to 25% among university students (Hong & Chen, 2024; Huang et al., 2024; Stojanov, 2024). Some authors even suggest that this should not yet be a cause for excessive societal alarm (Huang et al., 2024). Although the rapid advancements of artificial intelligence in education, together with its wide range of functionalities, have created

new opportunities, the risk of excessive dependence calls for institutional regulation (Liu et al., 2023; Stojanov, 2024). In this regard, universities could play a crucial role by guiding students toward a balanced use of these tools, encouraging their integration in a critical manner while ensuring that learners continue to strengthen their own skills (Lee et al., 2024). This aligns with models of open and distance education, such as the *Community of Inquiry framework*, which emphasize the need for cognitive, social, and teaching presence to support self-regulated learning and prevent overreliance on technological tools (Carroll et al., 2025).

Impact of motivation and procrastination on ChatGPT dependence

ChatGPT (*Chat Generative Pre-Trained Transformer*), created by OpenAI, is a sophisticated AI-powered language model designed to handle complex linguistic tasks. Renowned for its ability to create human-like text, it facilitates a diversity of applications, including automated text generation, question answering, summarization, and image creation. By delivering responses that closely emulate human communication, ChatGPT has revolutionized interactions with chatbots and virtual assistants, achieving substantial global popularity and widespread use since its launch in November 2022 (Yu et al., 2024).

Although the existing literature analyzing the relationship between motivation, procrastination, and dependence on ChatGPT is extremely limited due to the novelty of the topic, certain studies provide some valuable insights.

First, regarding the relationship between motivation and dependence on ChatGPT, several studies have revealed that high levels of lazy thinking (which may reflect amotivation toward tasks), high levels of emotional stress (which may reflect exhaustion toward tasks), low levels of academic self-efficacy, and low levels of decision-making skills were associated with an increase in dependence on Generative AI (Ahmad et al., 2023; Estrada-Araoz et al., 2025; Naseer et al., 2025; Ye et al., 2024). Similarly, previous research highlights the effect of different types of motivation on dependence on AI. Specifically, low levels of achievement motivation (Hong & Chen, 2024) and Mastery Approach orientation (Stojanov, 2024), as well as high levels of escape motivation, social motivation (Huang et al., 2024), mastery avoidance orientation, and performance approach orientation (Stojanov, 2024), have been associated with increased levels of dependence on ChatGPT. As stated by Ye et al. (2024), individuals who have had a positive experience using ChatGPT may be more motivated to avoid learning tasks and rely on the responses provided by ChatGPT.

Secondly, regarding the relationship between procrastination and dependence on ChatGPT, previous studies, such as the one by Feng et al. (2023), hypothesize how procrastination may influence dependence on ChatGPT. This hypothesis has been partially tested. For example, Stojanov (2024) observed that procrastinating students, who represented nearly a quarter of the university student population, were highly dependent on ChatGPT for completing assigned tasks. Similarly, Swargiary (2024), through a pre-post design with a control group, conducted an intervention using ChatGPT for task completion with the intervention group. The results showed that the intervention group exhibited increased levels of procrastination and reduced academic performance compared to the comparison group. Finally, studies with qualitative approaches (e.g., Ayele, 2024; Paraso et al., 2024) revealed that university students reported that AI was extremely helpful, especially for those who tended to leave tasks until the last moment.

Based on these studies, procrastination might mediate the relationship between different types of motivation and AI dependence. This suggests that students who lack internal motivation or feel disengaged may increasingly postpone tasks, turning to AI tools as a compensatory strategy, a dynamic particularly relevant in certain environments, like open and distance education environments, where self-regulation is crucial.

The current study

The impact of Artificial Intelligence on education has garnered growing attention, prompting a proliferation of opinion articles. However, empirical research on this subject remains limited (Ye et al., 2024). One of the primary concerns for educators and researchers is the potential negative consequences of ChatGPT, particularly due to its problematic use (Ye et al., 2024). Despite this, as several studies point out, there is a severe lack of comprehensive analyses on inappropriate uses, such as AI dependence, which warrants further exploration (Yu et al., 2024). As a result, some scholars stress the importance of investigating the psychological factors that contribute to the development of this dependence (Ahmad et al., 2023; Ye et al., 2024), due to the fact that the intensive use of these tools might negatively impact students' self-regulation, critical thinking and problem-solving skills (Ali et al., 2024; Kamalov et al., 2023; Zhai et al., 2024).

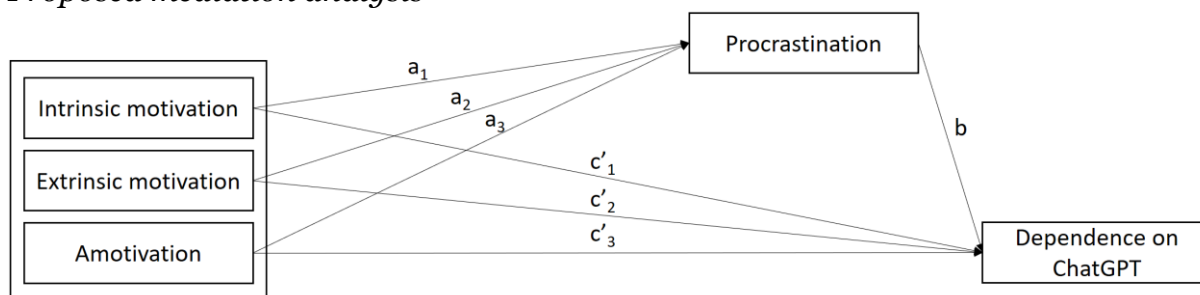
Similarly, as previously observed, there is some evidence linking motivation and procrastination-related factors with dependence on ChatGPT. However, to the authors' knowledge, the current evidence employing more complex analyses, such as structural equation modeling, mediation analysis, or moderation analysis, is extremely limited. This scarcity hinders a comprehensive understanding of the relationship between various constructs of interest, such as motivation or procrastination. Additionally, this is compounded by the lack of empirical studies on Artificial Intelligence dependence that are approached through robust frameworks, such as SDT, which is gaining significant traction in the field of AI (e.g., Chai et al., 2023; Chiu, 2024; Chiu & Chai, 2020; Xia et al., 2022). In addition, the past few years have witnessed a growth in scholarly articles examining procrastination in university students, clearly indicating growing scientific interest in the topic (Yan & Zhang, 2022).

Accordingly, this study seeks to address three distinct research questions:

- RQ₁: To what extent do different types of motivation (amotivation, intrinsic and extrinsic) impact dependence on ChatGPT?
- RQ₂: To what extent does procrastination impact dependence on ChatGPT?
- RQ₃: To what degree does procrastination mediate the relationship between various forms of motivation and dependence on ChatGPT?

Figure 1 presents the diagram of the model that is intended to be tested in the subsequent analyses.

Figure 1
Proposed mediation analysis



METHODOLOGY

Design and participants

The study employed a quantitative, cross-sectional, non-experimental design, following a descriptive–correlational approach. Mediation analyses were conducted to examine the indirect effects of learning motivation and procrastination on ChatGPT dependence.

In this design, 467 Spanish university students from the field of education at the University of the Basque Country participated ($M = 20.44$; $SD = 2.83$). Among them, 292 identified as male, 173 as female, and 2 as another gender. Regarding their academic programs, 40 students were enrolled in the Bachelor's Degree in Early Childhood Education, 379 in the Bachelor's Degree in Primary Education, and 48 in the master's degree in Secondary Education Teaching. In terms of academic progression, 274 participants were in their second year, 103 in their third year, 45 in their fourth year, and 45 in the Master's program (equivalent to a fifth year).

Participants were selected through non-probabilistic sampling based on proximity to the research team, as all students had been taught by one or more team members. Notably, previous studies suggest that non-probabilistic sampling methods often produce results similar to those obtained through randomized sampling techniques (Coppock et al., 2018). Likewise, an a priori power analysis was conducted using G*Power 3.1 for a linear multiple regression model with four predictors (three independent variables and one mediator). Assuming a medium effect size ($f^2 = 0.15$) and $\alpha = 0.05$, the analysis indicated that a minimum sample of 55 participants was required to achieve a power of 0.80, a commonly used value in the context of correlational studies in education and psychology (Cohen, 1988). With the current sample of 467 participants, the study provides a robust level of statistical power to detect the anticipated effects.

Instruments

• Learning motivation

Intrinsic Motivation: The dimensions of Intrinsic Motivation to Know and Intrinsic Motivation toward Accomplishment were used from the short version of the Academic Motivation Scale (Kotera et al., 2023). These two dimensions comprise four 5-point Likert-type items each, capturing the respondent's internal intention to study

for the purpose of expanding knowledge (e.g., “I study because my studies allow me to continue to learn about many things that interest me”) and improving as a student (e.g., “I study for the pleasure that I experience while I am surpassing myself in one of my personal accomplishments”). According to Kotera et al. (2023), these dimensions demonstrated reliability with $\alpha = .84$ and $.79$, respectively. *Extrinsic Motivation*: To assess extrinsic motivation, the External Regulation dimension of the same scale was employed. This dimension consists of two 5-point Likert-type items that measure the respondent’s intention to study for external rewards (e.g., “I study in order to obtain a more prestigious job later on”). The reliability of this dimension was reported as $\alpha = .82$ (Kotera et al., 2023). *Amotivation*: Amotivation was measured using the Amotivation dimension of the same scale. This dimension includes two 5-point Likert-type items that capture the respondent’s lack of interest in studying (e.g., “I can’t see why I go to university and frankly, I couldn’t care less”). According to Kotera et al. (2023), the reliability for this dimension was $\alpha = .79$.

- **Procrastination**

To assess the level of procrastination, the Short Form of the Academic Procrastination Scale (Yockey, 2016) was utilized. This scale consists of five 5-point Likert-type items grouped into a single factor, designed to measure the extent to which students delay and postpone completing their tasks until the last moment (e.g., “When given an assignment, I usually put it away and forget about it until it is almost due”). According to Yockey (2016), the reliability of this scale is reported as $\alpha = .87$.

- **Dependence on ChatGPT**

To evaluate the degree of dependence on ChatGPT, the ChatGPT Dependence Scale (Ye et al., 2024) was employed. This scale comprises a single dimension with five 5-point Likert-type items, measuring the extent to which individuals feel the need to use ChatGPT for academic tasks (e.g., “I would feel uncomfortable without ChatGPT”). According to Ye et al. (2024), the reliability of this scale is reported as $\alpha = .71$.

Procedure

The procedure commenced with a review of various databases to identify appropriate scales for the study, which were then digitized using Google Forms. Subsequently, the sample was defined, including all students taught by the researchers involved in the study and enrolled in education-related programs. As noted earlier, although this was a non-probabilistic sample, prior research indicates that convenience samples often produce results comparable to those obtained from randomized samples (Coppock et al., 2018). Participants completed the Google Forms questionnaire during class hours using individual computers, after consenting to the study’s participation conditions. These conditions complied with the principles outlined in the Declaration of Helsinki, ensuring anonymity and privacy of responses, the right to withdraw before or during participation, and appropriate data management and disposal during and after the research. Upon study completion, participants received a report summarizing the main findings and their implications.

Data analysis

First, to assess the validity of the data as much as possible, the goodness of fit of the model was examined through the X^2/df , CFI, RMSEA, and AIC indices using SPSS AMOS 25 software, expecting the indices to meet the following thresholds: $X^2/df < 5$; CFI $> .90$; RMSEA $< .080$; AIC = lowest possible value (Hooper et al., 2008). Subsequently, in SPSS Statistics 27, a descriptive analysis was conducted by studying the arithmetic mean and standard deviation.

This was followed by other analyses, such as the correlational analysis using Pearson's r between the main dimensions, as well as the reliability analysis using Cronbach's alpha. Finally, also in SPSS Statistics 27, the PROCESS macro was used to perform a mediation analysis, as shown in Figure 1. This analysis was carried out employing a total of 10,000 samples of bootstrapping. In this mediation analysis, particular attention was given to the direct and indirect effects obtained. Specifically, to answer RQ₁ and RQ₂, the direct effects were examined. For these effects, the R^2 values were observed to determine the effect size, indicating the proportion of the variance in the dependent variable explained by the independent variable. For RQ₃, the indirect effects were studied. To evaluate their effect size, the Proportion of Mediation (PM) was calculated. This metric indicates the proportion of the total effect of the independent variable on the dependent variable that is accounted for by the mediator. All parametric analyses were justified by the normal distribution of all measured variables, assessed through skewness and kurtosis, with all values falling within the skewness range of ± 2 and the kurtosis range of ± 7 (Hair et al., 2010). Multicollinearity was also tested, with all values falling within the acceptable range of VIF < 10 and Tolerance $> .10$ (Field, 2013). Independence of errors was assessed using the Durbin–Watson statistic, which yielded a value of 1.636; this falls within the 1.5–2.5 range, indicating a slight positive autocorrelation but generally considered acceptable (Field, 2013).

RESULTS

First, to determine whether any preliminary adjustments were necessary, the goodness of fit of the proposed model was assessed using the indices X^2/df , CFI, RMSEA, and AIC. The resulting model demonstrated good fit ($X^2/df = 2.328$; CFI = .955; RMSEA = .053; AIC = 382.995), as the values fell within the acceptable thresholds established for evaluating model fit in this study of $X^2/df < 5$; CFI $> .90$; RMSEA $< .80$; AIC = lowest value possible (Hooper et al., 2008).

Subsequently, descriptive statistics, correlations, and reliability indices for the main dimensions were analyzed. As shown in Table 1, participants generally exhibited high levels of intrinsic ($M = 3.48$; $SD = .843$) and extrinsic motivation ($M = 3.95$; $SD = .849$), alongside low levels of amotivation ($M = 1.63$; $SD = .884$). Additionally, moderate levels of procrastination ($M = 2.57$; $SD = .947$) and dependence on ChatGPT ($M = 2.37$; $SD = .996$) were observed. Likewise, correlation analyses indicated that intrinsic motivation was inversely related to procrastination ($r = -.287$, $p < .001$), while amotivation demonstrated a positive association with procrastination ($r = .403$, $p < .001$). Furthermore, the only type of motivation significantly associated with dependence on ChatGPT was amotivation ($r = .162$; $p < .001$), which also showed a positive correlation. Procrastination similarly demonstrated a positive correlation with ChatGPT dependence ($r = .231$; $p < .001$). Finally, all reliability indices exceeded the

threshold of $\alpha = .70$ (Tavakol & Dennick, 2011), indicating good internal consistency for the dimensions used in the study.

Table 1

Descriptives, correlations and reliability indices of the main dimensions

	M	SD	1	2	3	4	5
IM	3.48	.843	(.865)	.176***	-.200***	-.287***	-.084
EM	3.95	.849		(.785)	-.046	-.091	.082
AM	1.63	.884			(.813)	.403***	.162***
PRC	2.57	.947				(.858)	.231***
GPT	2.37	.996					(.858)

Note: * $p < .05$; ** $p < .01$; *** $p < .001$; IM, Intrinsic Motivation; EM, Extrinsic Motivation; AM, Amotivation; PRC, Procrastination; GPT, Dependence on ChatGPT. Cronbach's alpha reliability indices are presented on the main diagonal in parentheses.

To address the research questions posed in this study, a mediation analysis was conducted. Mediation analysis is employed to assess whether, and to what extent, the effect of an independent variable on a dependent variable operates indirectly through a mediator variable. In this model, the independent variables were the different types of motivation, the dependent variable was dependence on ChatGPT, and the mediating variable was procrastination.

The main results from the direct effects, presented in Table 2, indicate that both intrinsic motivation ($R^2 = .082$; $\beta = -.322$; $p < .001$) and amotivation ($R^2 = .162$; $\beta = .432$; $p < .001$) are significantly associated with procrastination. These R^2 values indicate that intrinsic motivation alone explains about 8% of the variance in procrastination, whereas amotivation accounts for 16%. Thus, lower intrinsic motivation and higher amotivation are associated with more pronounced procrastination, suggesting that the absence of self-determined reasons for engaging in academic tasks may be particularly detrimental. Additionally, and concerning RQ₁ and RQ₂, procrastination was found to be significantly associated with dependence on ChatGPT ($R^2 = .053$; $\beta = .237$; $p < .001$) indicating that procrastination explains around 5% of the variance in dependence on ChatGPT. Among the three types of motivation examined, the only one that was directly associated with dependence on ChatGPT was extrinsic motivation ($R^2 = .026$; $\beta = .122$; $p = .022$), indicating that although the effect was statistically significant, the proportion of variance explained was relatively low, suggesting that other factors likely contribute more substantially to dependence on ChatGPT.

Table 2

Results from the mediation analysis

	Path	R ²	PM	β	SE	LL	UL	p
Direct effects								
IM $\rightarrow$ PRC	a ₁	.082	-	-.322	.050	-.420	-.224	< .001
EM $\rightarrow$ PRC	a ₂	.008	-	-.101	.052	-.203	.000	.054
AM $\rightarrow$ PRC	a ₃	.162	-	.432	.045	.342	.521	< .001

	Path	R ²	PM	β	SE	LL	UL	p
PRC → GPT	b	.053	-	.237	.050	.140	.334	< .001
IM → GPT	c' ₁	.007	-	-.023	.056	-.133	.086	.677
EM → GPT	c' ₂	.007	-	.122	.053	.018	.226	.022
AM → GPT	c' ₃	.026	-	.093	.055	-.016	.202	.096
Indirect effects								
IM → PRC → GPT	a ₁ * b	-	.764	-.076	.022	-.121	-.037	***
EM → PRC → GPT	a ₂ * b	-	-.270	-.026	.015	-.058	.001	(ns)
AM → PRC → GPT	a ₃ * b	-	.493	.090	.026	.041	.144	***

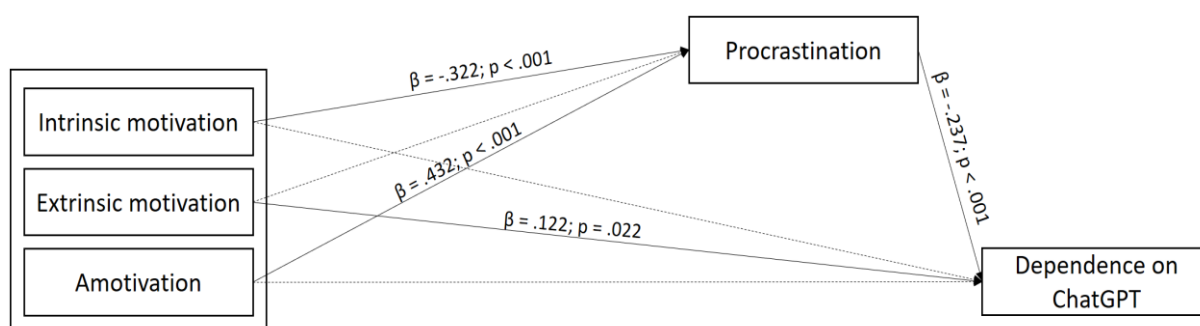
Note: IM, Intrinsic Motivation; EM, Extrinsic Motivation; AM, Amotivation; PRC, Procrastination; GPT, Dependence on ChatGPT.

Table 2 also displays the indirect effects, showing whether the independent variable influences the dependent variable via the mediator. Significance is established when the confidence interval excludes zero (MacKinnon et al., 2002). In this case, and answering RQ3, procrastination was found to mediate the relationships from intrinsic motivation ($\beta = -.076$; $LL = -.121$; $UL = -.037$) and amotivation ($\beta = .090$; $LL = .041$; $UL = .144$) to dependence on ChatGPT. In other words, high intrinsic motivation and minimal amotivation are associated with reduced procrastination, which subsequently serves to decrease dependence on ChatGPT. A representation of the main results can be observed in Figure 2.

Likewise, the proportion of mediation (PM) values were also calculated in Table 2 to determine the effect size of these indirect relationships. For the relationship from intrinsic motivation to dependence on ChatGPT mediated by procrastination, the PM was .764, indicating that a significant 76.4% of the total effect is mediated by procrastination. This suggests a nearly complete mediation, with a high relevance of procrastination. Similarly, for the path from amotivation to dependence on ChatGPT through procrastination, the PM was .493, showing that procrastination mediates almost half (49.3%) of the total effect. In both cases, procrastination was observed to be pivotal in understanding the relationship between intrinsic motivation, amotivation, and ChatGPT dependence.

Figure 2

Figure presenting the outcomes of the mediation analysis



DISCUSSION

The objective of the present study was threefold. First, it aimed to analyze the impact of motivation on dependence on ChatGPT. Second, it sought to examine the impact of procrastination on dependence on ChatGPT. Finally, it aimed to determine whether the association between different types of motivation and dependence on ChatGPT is mediated by levels of procrastination.

The results revealed that the only type of motivation analyzed that had a direct effect on dependence on ChatGPT was extrinsic motivation. However, the proportion of variance explained was relatively low, suggesting that other factors may have a stronger influence on ChatGPT dependence. Regarding intrinsic motivation and amotivation, it was observed that their relationship with dependence on ChatGPT is significantly mediated by students' procrastination. Specifically, high levels of intrinsic motivation and low levels of amotivation are associated with a reduction in students' procrastination levels, which in turn reduces dependence on ChatGPT.

Although these findings are novel, as to the authors' knowledge no previous mediation analyses have linked these variables, they are consistent with existing literature. First, evidence from multiple studies indicates that higher intrinsic motivation toward learning is associated with reduced dependence on ChatGPT. (e.g., Hong & Chen, 2024; Stojanov, 2024). Conversely, high levels of motivation driven by external factors (e.g., Huang et al., 2024; Stojanov, 2024), lack of motivation, or interest in learning (e.g., Ahmad et al., 2023; Ye et al., 2024) are associated with an increase in dependence on ChatGPT. Similarly, previous quantitative and qualitative studies indicate that students with a propensity to procrastinate are more likely to rely on ChatGPT for their academic tasks (e.g., Ayele, 2024; Feng et al., 2023; Paraso et al., 2024; Stojanov, 2024; Swargiary, 2024). This study integrates these findings through mediation analysis, revealing that strong intrinsic motivation and low amotivation are key to reducing students' procrastination, which, in turn, is a critical factor in decreasing dependence on ChatGPT. This can be explained by the SDT, which proposes a continuum of internalization (Ryan & Deci, 2020). Extrinsic motivation operates directly; students use ChatGPT as a quick fix to meet external demands with minimal effort. In contrast, intrinsic motivation and amotivation affect students' perceived cognitive cost of tasks, which triggers a coping mechanism: procrastination. Our findings suggest that this procrastination acts as the true mediator, facilitating the use of ChatGPT to avoid the aversive task. Thus, intrinsic motivation influences dependence indirectly through procrastination, while extrinsic motivation appears to lead to direct use.

Next, the present study has different implications that may be of interest to various stakeholders. First, these findings may be valuable for researchers as they contribute to a better understanding and consolidation of a theoretical framework as recent as the impact of emotional factors on dependence on generative AI, such as ChatGPT.

Likewise, these findings may assist university faculty by demonstrating that increasing intrinsic motivation and reducing amotivation and procrastination can significantly reduce university students' dependence on ChatGPT. As discussed in the theoretical framework of this paper, according to SDT, intrinsic motivation is influenced by the basic psychological needs of competence, autonomy, and relatedness (Ryan & Deci, 2020). Based on the results showing that procrastination mediates the relationships from both intrinsic motivation and amotivation to ChatGPT dependence,

there is a wide range of literature that contributes to improving each of these needs, with consequent effects on students' intrinsic motivation.

First, some strategies that teachers could apply to foster students' basic psychological needs of competence, autonomy and relatedness might include promoting positive and constructive feedback (García et al., 2019; Koike & Pearson, 2005), using scaffolding or gradual support techniques (e.g., Mulatsih, 2011; Neitzel & Stright, 2003), encouraging the use of executive functions such as short- and long-term goal planning (Benavides-Nieto et al., 2017; Ganesalingam et al., 2011; Roebbers et al., 2012), fostering the use of metacognitive skills, such as self-reflection (Roebbers et al., 2012; Jho & Chae, 2014) promoting the usage of cooperative learning techniques, like 1-2-4 or Jigsaw (Fernández-Río et al., 2017; Namaziandost et al., 2019), or using other learning strategies to organize information and link it with existing knowledge such as distributed learning, retrieval in any of its forms (practice testing), interleaved practice, summarizing, doing concept maps, or highlighting the relevant information (Donoghue & Hattie, 2021; Maldonado-Sánchez et al., 2019). These strategies would be particularly relevant for e-learning and distance education settings, as they can be easily adapted for online platforms. For instance, online learning systems can integrate features that provide automated and instant feedback; course modules can be designed with a scaffolded structure to promote autonomy; and virtual group projects can facilitate cooperative learning, which can help mitigate the isolation that can foster amotivation and dependence on external tools. In this context, through digital learning strategies and the promotion of self-regulation, educators would be able to guide students to perceive AI as a supportive aid rather than as a replacement for their own abilities (Darvishi et al., 2024).

Similarly, authors such as Kachgal et al. (2001) propose a series of effective strategies to reduce procrastination, including fostering self-awareness of how students manage their procrastination, encouraging reflection on the benefits they have gained from task avoidance, conducting formal assessments that allow students to understand their levels of procrastination (e.g., using the PASS scale), promoting self-regulation strategies that have a significant impact on time management (e.g., preparing calendars that allow students to track their progress), or teaching stress management strategies. These recommendations directly address the finding that procrastination is a key mediator in the relationship between amotivation and dependence on ChatGPT, suggesting that tackling procrastination might be essential for reducing this form of dependence. For e-learning, these strategies can be implemented through digital tools like time management apps, built-in quizzes for self-assessment, and integrated forums for peer-to-peer discussions on managing academic stress. These online resources might provide students with the structure and support needed to combat procrastination, thereby decreasing their dependence on generative AI.

Finally, this study has several limitations that should be considered when interpreting the results. First, the sample is entirely drawn from the social sciences field, with the majority of participants enrolled in the Primary Education program, while a smaller proportion are from the Early Childhood Education program or the Master's in Secondary Education Teaching. This imbalance, combined with the use of a convenience sampling method, may have partially biased the results and limits the generalizability of the findings. Future studies could conduct comparative analyses across different degree programs and universities, which would allow examining potential variations in educational AI use. Replicating these findings in students from

other disciplines, such as experimental or health sciences, would be valuable and could confirm the robustness of these results across different educational contexts. Additionally, the cross-sectional design restricts causal inferences, highlighting the need for longitudinal studies to strengthen the conclusions. Designs combining longitudinal and mixed-methods approaches could further explore the evolution of educational AI use over time. Data were also collected through self-report measures, which may introduce biases such as common method variance and social desirability, potentially affecting the observed relationships. Finally, while the current study tested a single theoretical mediation model, future research could provide more precise information about the conditions under which the proposed model operates most effectively by incorporating potential moderating variables. For instance, a lack of confidence in one's abilities may cause students with low academic self-efficacy to perceive AI as a resource that provides immediate answers, thereby reducing the need to confront challenges independently. This behavior, especially when students face academic pressure or stress, might reinforce their dependence on technology (Rodríguez et al., 2024). Consequently, variables like academic self-efficacy and workload could serve as control or moderating factors in future models, helping to explain the nuanced conditions under which students are more likely to develop dependence on AI. Despite these limitations, this work provides a foundation for future research exploring the impact of emotional variables on dependence on generative AI, exemplified by ChatGPT.

CONCLUSION

This study demonstrates that high intrinsic motivation and low amotivation reduce students' procrastination, which in turn decreases dependence on ChatGPT. These findings suggest that fostering students' basic psychological needs and implementing strategies to enhance motivation and self-regulation may help mitigate dependence on generative AI in educational contexts. Future research should replicate these results across different disciplines, employing longitudinal designs, replacing self-report measures, and exploring additional moderating or mediating factors influencing students' dependence on AI tools.

REFERENCES

- Ahmad, S. F., Han, H., Alam, M. M., Rehmat, M. K., Irshad, M., Arraño-Muñoz, M., & Ariza-Montes, A. (2023). Impact of artificial intelligence on human loss in decision making, laziness and safety in education. *Humanities & Social Sciences Communications*, 10(1), 311. <https://doi.org/10.1057/s41599-023-01842-4>
- Ahmad, S. F., Rahmat, M. K., Mubarik, M. S., Alam, M. M., & Hyder, S. I. (2021). Artificial intelligence and its role in education. *Sustainability*, 13(22), 12902. <https://doi.org/10.3390/su132212902>
- Ali, K., García, A., & Vadsariya, A. (2024). Impact of the AI dependency revolution on both physical and mental health. *Journal of Strategic Innovation and Sustainability*, 19(2), 70-81. <https://doi.org/10.33423/jsis.v19i2.7006>
- American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). American Psychiatric Publishing.
- Ayele, N. (2024). *The usage of new AI technologies by college students and its influence on learning and dependence* [Manuscrito, Caldwell University Archives]. JSTOR Community Collections. <https://www.jstor.org/stable/community.37312365>

- Baird, A., & Maruping, L. M. (2021). The next generation of research on IS use: A theoretical framework of delegation to and from agentic IS artifacts. *MIS Quarterly*, 45(1), 315-341. <https://doi.org/10.25300/MISQ/2021/15882>
- Benavides-Nieto, A., Romero-López, M., Quesada-Conde, A. B., & Alba, G. (2017). Basic executive functions in early childhood education and their relationship with social competence. *Procedia - Social and Behavioral Sciences*, 237, 471-478. <https://doi.org/10.1016/j.sbspro.2017.02.092>
- Bolden, J., & Fillauer, J. P. (2020). "Tomorrow is the busiest day of the week": Executive functions mediate the relation between procrastination and attention problems. *Journal of American College Health*, 68(8), 854-863. <https://doi.org/10.1080/07448481.2019.1626399>
- Bucur, A. M., Cosma, A., & Dinu, L. P. (2021). *Early risk detection of pathological gambling, self-harm and depression using BERT* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2106.16175>
- Carroll, N., Lang, M., & Connolly, C. (2025). An extended community of inquiry framework supporting students in online and digital education. *Innovations in Education and Teaching International*, 62(2), 369-385. <https://doi.org/10.1080/14703297.2024.2326658>
- Chai, C. S., Chiu, T. K. F., Wang, X., Jiang, F., & Lin, X.-F. (2023). Modeling Chinese secondary school students' behavioral intentions to learn artificial intelligence with the theory of planned behavior and self-determination theory. *Sustainability*, 15, 605. <https://doi.org/10.3390/su15010605>
- Chen, N.-S., Yin, C., Isaias, P., & Psotka, J. (2020). Educational big data: Extracting meaning from data for smart education. *Interactive Learning Environments*, 28(2), 142-147. <https://doi.org/10.1080/10494820.2019.1635395>
- Cheon, S. H., Reeve, J., & Vansteenkiste, M. (2020). When teachers learn how to provide classroom structure in an autonomy-supportive way: Benefits to teachers and their students. *Teaching and Teacher Education*, 90, 103004. <https://doi.org/10.1016/j.tate.2019.103004>
- Chiu, T. K. F. (2024). A classification tool to foster self-regulated learning with generative artificial intelligence by applying self-determination theory: A case of ChatGPT. *Educational Technology Research & Development*. <https://doi.org/10.1007/s11423-024-10366-w>
- Chiu, T. K. F., & Chai, C.-S. (2020). Sustainable curriculum planning for artificial intelligence education: A self-determination theory perspective. *Sustainability*, 12(14), 5568. <https://doi.org/10.3390/su12145568>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.
- Coppock, A., Leeper, T. J., & Mullinix, K. J. (2018). Generalizability of heterogeneous treatment effect estimates across samples. *Proceedings of the National Academy of Sciences*, 115(49), 12441-12446. <https://doi.org/10.1073/pnas.1808083115>
- Darvishi, A., Khosravi, H., Sadiq, S., Gašević, D., & Siemens, G. (2024). Impact of AI assistance on student agency. *Computers & Education*, 210, 104967. <https://doi.org/10.1016/j.compedu.2023.104967>
- Del Cisne, M., Romero, J., Sancho, D., & Romero, A. (2024). Consecuencias de la dependencia a la inteligencia artificial en habilidades críticas y aprendizaje autónomo en los estudiantes. *Ciencia Latina Revista Científica Multidisciplinar*, 8(2), 2368-2382. https://doi.org/10.37811/cl_rcm.v8i2.10678
- Delgado, N., Campo Carrasco, L., Sainz de la Maza, M., & Etxabe-Urbieta, J. M. (2024). Application of artificial intelligence (AI) in education: Benefits and limitations of AI as perceived by primary, secondary, and higher education teachers. *Revista Electrónica*

- Interuniversitaria de Formación del Profesorado*, 27(1), 207-224.
<https://doi.org/10.6018/reifop.577211>
- Donoghue, G. M., & Hattie, J. A. C. (2021). A meta-analysis of ten learning techniques. *Frontiers in Education*, 6, 581216.
<https://doi.org/10.3389/feduc.2021.581216>
- Duru, E., & Balkis, M. (2017). Procrastination, self-esteem, academic performance, and well-being: A moderated mediation model. *International Journal of Educational Psychology*, 6(2), 97-119.
<https://doi.org/10.17583/ijep.2017.2584>
- Estrada-Araoz, E. G., Mamani-Roque, M., Quispe-Aquise, J., Manrique-Jaramillo, Y. V., & Cruz-Laricano, E. O. (2025). Academic self-efficacy and dependence on artificial intelligence in a sample of university students. *Sapienza: International Journal of Interdisciplinary Studies*, 6(1), e25008.
<https://doi.org/10.51798/sijis.v6i1.916>
- Feng, M., Xia, A., & Xia, X. (2023). The association between ChatGPT usage and college students' online learning burnout: The mediating role of self-control. In X. Chen & Q. Liu (Eds.), *Proceedings of the International Conference of Educational Innovation through Technology* (pp. 209-212). IEEE.
<https://doi.org/10.1109/EITT61659.2023.00046>
- Fernández-Río, J., Sanz, N., Fernández-Cando, J., & Santos, L. (2017). Impact of a sustained cooperative learning intervention on student motivation. *Physical Education and Sport Pedagogy*, 22(1), 89-105.
<https://doi.org/10.1080/17408989.2015.1123238>
- Field, A. (2013). *Discovering statistics using IBM SPSS statistics* (4th ed.). Sage.
- Filgona, J., Sakiyo, J., Gwany, D. M., & Okoronka, A. U. (2020). Motivation in learning. *Asian Journal of Education and Social Studies*, 10(4), 16-37.
<https://doi.org/10.9734/ajess/2020/v10i430273>
- Ganesalingam, K., Yeates, K. O., Taylor, H. G., Chertkoff, N., Stancin, T., & Wade, S. (2011). Executive functions and social competence in young children 6 months following traumatic brain injury. *Neuropsychology*, 25(4), 466-476.
<https://doi.org/10.1037/a0022768>
- Ganotice, F. A., Mendoza, N. B., Dizon, J. I. W. T., Shen, X., Lee, J. C.-Y., Chan, E., Luk, P., Manio, M. M., He, Q., Khoo, U. S., Lam, M. P. S., Chan, S. C., Chow, A. Y. M., Wang, N., & Tipoe, G. L. (2024). Students' motivation and engagement in interprofessional education: The mediating role of peer relatedness. *Medical Education Online*, 29(1), 2430593.
<https://doi.org/10.1080/10872981.2024.2430593>
- García, J. A., Carcedo, R. J., & Castaño, J. L. (2019). The influence of feedback on competence, motivation, vitality, and performance in a throwing task. *Research Quarterly for Exercise and Sport*, 90(2), 172-179.
<https://doi.org/10.1080/02701367.2019.1571677>
- Garon-Carrier, G., Boivin, M., Guay, F., Kovas, Y., Dionne, G., Lemelin, J. P., Séguin, J. R., Vitaro, F., & Tremblay, R. E. (2016). Intrinsic motivation and achievement in mathematics in elementary school: A longitudinal investigation of their association. *Child Development*, 87(1), 165-175.
<https://doi.org/10.1111/cdev.12458>
- Grund, A., & Fries, S. (2018). Understanding procrastination: A motivational approach. *Personality and Individual Differences*, 121, 120-130.
<https://doi.org/10.1016/j.paid.2017.09.035>
- Guay, F., Stupnisky, R., Boivin, M., Japel, C., & Dionne, G. (2019). Teachers' relatedness with students as a predictor of students' intrinsic motivation, self-concept, and reading achievement. *Early Childhood Research Quarterly*, 48(3), 215-225.
<https://doi.org/10.1016/j.ecresq.2019.03.005>
- Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis* (7th ed.). Pearson Education International.
- Hen, M., & Goroshit, M. (2018). General and life-domain procrastination in highly

- educated adults in Israel. *Frontiers in Psychology*, 9, 1173. <https://doi.org/10.3389/fpsyg.2018.01173>
- Hong, J.-C., & Chen, N. (2024). *Effects of achievement motivation and ChatGPT self-efficacy on ChatGPT addiction and ChatGPT continuance intention* [Preprint]. SSRN. <https://doi.org/10.2139/ssrn.5064583>
- Hooper, D., Coughlan, J., & Mullen, M. R. (2008). Structural equation modeling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods*, 6(1), 53-60.
- Huang, S., Lai, X., Ke, L., Li, Y., Wang, H., Zhao, X., Dai, X., & Wang, Y. (2024). AI technology panic—Is AI dependence bad for mental health? A cross-lagged panel model and the mediating roles of motivations for AI use among adolescents. *Psychology Research and Behavior Management*, 17, 1087-1102. <https://doi.org/10.2147/PRBM.S440889>
- Jara, I., & Ochoa, J. M. (2020). Usos y efectos de la inteligencia artificial en educación. *Banco Interamericano de Desarrollo*. <https://doi.org/10.18235/0002380>
- Jho, M. Y., & Chae, M. O. (2014). Impact of self-directed learning ability and metacognition on clinical competence among nursing students. *Journal of Korean Academic Society of Nursing Education*, 20(4), 513-522. <https://doi.org/10.5977/jkasne.2014.20.4.513>
- Johnson, J. L., & Bloom, A. M. (1995). An analysis of the contribution of the five factors of personality to variance in academic procrastination. *Personality and Individual Differences*, 18(1), 127-133. [https://doi.org/10.1016/0191-8869\(94\)00109-6](https://doi.org/10.1016/0191-8869(94)00109-6)
- Kachgal, M. M., Hansen, L. S., & Nutter, K. J. (2001). Academic procrastination prevention/intervention: Strategies and recommendations. *Journal of Developmental Education*, 25(1), 14-24. <https://www.jstor.org/stable/42775842>
- Kamalov, F., Santandreu Calonge, D., & Gurrib, I. (2023). New era of artificial intelligence in education: Towards a sustainable multifaceted revolution. *Sustainability*, 15(16), 12451. <https://doi.org/10.3390/su151612451>
- Klingbeil, A., Grützner, C., & Schreck, P. (2024). Trust and reliance on AI—An experimental study on the extent and costs of overreliance on AI. *Computers in Human Behavior*, 160, 108352. <https://doi.org/10.1016/j.chb.2024.108352>
- Klingsieck, K. B. (2013). Procrastination: When good things don't come to those who wait. *European Psychologist*, 18, 24-34. <https://doi.org/10.1027/1016-9040/a000138>
- Koike, D. A., & Pearson, L. (2005). The effect of instruction and feedback in the development of pragmatic competence. *System*, 33(3), 481-501. <https://doi.org/10.1016/j.system.2005.06.008>
- Kok, J. K. (2016). The relationships between procrastination and motivational aspects of self-regulation. *Jurnal Psikologi Malaysia*, 30(1), 30-39.
- Korteling, J. E., van de Boer-Visschedijk, G. C., Blankendaal, R. A. M., Boonekamp, R. C., & Eikelboom, A. R. (2021). Human-versus artificial intelligence. *Frontiers in Artificial Intelligence*, 4, 622364. <https://doi.org/10.3389/frai.2021.622364>
- Kotera, Y., Conway, E., & Green, P. (2023). Construction and factorial validation of a short version of the Academic Motivation Scale. *British Journal of Guidance & Counselling*, 51(2), 274-283. <https://doi.org/10.1080/03069885.2021.1903387>
- Lee, D., Arnold, M., Srivastava, A., Plastow, K., Strelan, P., Ploeckl, F., Lekkas, D., & Palmer, E. (2024). The impact of generative AI on higher education learning and teaching: A study of educators' perspectives. *Computers and Education: Artificial Intelligence*, 6, 100221. <https://doi.org/10.1016/j.caeai.2024.100221>
- Legood, A., Lee, A., Schwarz, G., & Newman, A. (2018). From self-defeating to other defeating: Examining the effects of leader procrastination on follower work outcomes. *Journal of Occupational and*

- Organizational Psychology*, 91, 430-439.
<https://doi.org/10.1111/joop.12205>
- Li, Y., Zhang, L., Zhang, R., Xu, T., & Feng, T. (2021). The neural basis linking achievement motivation with procrastination: Left precuneus connectivity with right anterior cingulate cortex. *Personality and Social Psychology Bulletin*, 48, 1382-1392.
<https://doi.org/10.1177/01461672211040677>
- Liu, M., Ren, Y., Nyagoga, L. M., Stonier, F., Wu, Z., & Yu, L. (2023). Future of education in the era of generative artificial intelligence: Consensus among Chinese scholars on applications of ChatGPT in schools. *Future in Educational Research*, 1(1), 72-101.
<https://doi.org/10.1002/fer3.10>
- Ljubin-Golub, T., Rijavec, M., & Olčar, D. (2020). Student flow and burnout: The role of teacher autonomy support and student autonomous motivation. *Psychological Studies*, 65, 145-156.
<https://doi.org/10.1007/s12646-019-00539-6>
- MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods*, 7(1), 83-104.
<https://doi.org/10.1037/1082-989X.7.1.83>
- Maldonado-Sánchez, M., Aguinaga-Villegas, D., Nieto-Gamboa, J., Fonseca-Arellano, F., Shardin-Flores, L., & Cadenillas-Albornoz, V. (2019). Learning strategies for the development of the autonomy of secondary school students. *Propósitos y Representaciones*, 7(2), 415-439.
<https://doi.org/10.20511/pyr2019.v7n2.290>
- Morales-García, W. C., Sairitupa-Sánchez, L. Z., Morales-García, S. B., & Morales-García, M. (2024). Development and validation of a scale for dependence on artificial intelligence in university students. *Frontiers in Education*, 9, 1323898.
<https://doi.org/10.3389/feduc.2024.1323898>
- Morandín-Ahuerma, F. (2022). What is artificial intelligence? *International Journal of Research Publication and Reviews*, 3(12), 1947-1951.
<https://doi.org/10.55248/gengpi.2022.31261>
- Mulatsih, S. (2011). The use of scaffolding technique to improve the students' competence in writing genre-based texts. *Parole: Journal of Linguistics and Education*, 2(1), 101-109.
- Namaziandost, E., Shatalebi, V., & Nasri, M. (2019). The impact of cooperative learning on developing speaking ability and motivation toward learning English. *Journal of Language and Education*, 5(3), 83-101.
<https://doi.org/10.17323/jle.2019.9809>
- Naseer, A., Rafaqat, N., & Amjad, M. (2025). Psychological impacts of AI dependence: Assessing the cognitive and emotional costs of intelligent systems in daily life. *Review of Applied Management and Social Sciences*, 8(1), 291-307.
<https://doi.org/10.47067/ramss.v8i1.458>
- Neitzel, C., & Stright, A. D. (2003). Mothers' scaffolding of children's problem solving: Establishing a foundation of academic self-regulatory competence. *Journal of Family Psychology*, 17(1), 147-159.
<https://doi.org/10.1037/0893-3200.17.1.147>
- Paraso, Y. M., Sedon, S. B., & Mahilum, D. Z. E. (2024). Student's artificial intelligence (AI) dependency: The lived experience of STEM students at Tongantongan National High School. *International Journal of All Research Writings*, 6(4), 34-43.
- Reeve, J., & Cheon, S. H. (2021). Autonomy-supportive teaching: Its malleability, benefits, and potential to improve educational practice. *Educational Psychologist*, 56(1), 54-77.
<https://doi.org/10.1080/00461520.2020.1862657>
- Rodríguez, J., Marín, I., & Espejo, R. (2024). Is artificial intelligence use related to self-control, self-esteem, and self-efficacy among university students? *Education and Information Technologies*, 30, 2507-2524.
<https://doi.org/10.1007/s10639-024-12906-6>

- Roebers, C. M., Cimeli, P., Röthlisberger, M., & Neunswander, R. (2012). Executive functioning, metacognition, and self-perceived competence in elementary school children: An explorative study on their interrelations and their role for school achievement. *Metacognition and Learning*, 7, 151-173. <https://doi.org/10.1007/s11409-012-9089-9>
- Ryan, R. M., & Deci, E. L. (2020). Intrinsic and extrinsic motivation from a self-determination theory perspective: Definitions, theory, practices, and future directions. *Contemporary Educational Psychology*, 61, 101860. <https://doi.org/10.1016/j.cedpsych.2020.101860>
- Siste, K., Assandi, P., Irawati, R., Setiawan, M. W., & Wiguna, T. (2019). Internet addiction: A new addiction? *Medical Journal of Indonesia*, 28(1), 82-91. <https://doi.org/10.13181/mji.v28i1.2752>
- Stojanov, A. (2024). Achievement goal motivation and reliance on ChatGPT for learning. In C. Pracana & M. Wang (Eds.), *Psychological Applications and Trends* (pp. 547–551). inScience Press. <https://doi.org/10.36315/2024inpact127>
- Swargiary, K. (2024). *The impact of ChatGPT on student learning outcomes: A comparative study of cognitive engagement, procrastination, and academic performance* [Preprint]. SSRN. <https://doi.org/10.2139/ssrn.4914743>
- Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, 2, 53-55. <https://doi.org/10.5116/ijme.4dfb.8dfd>
- Van Bergen, E., Hart, S. A., Latvala, A., Vuoksima, E., Tolvanen, A., & Torppa, M. (2022). Literacy skills seem to fuel literacy enjoyment, rather than vice versa. *Developmental Science*, 26(3), e13325. <https://doi.org/10.1111/desc.13325>
- Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, 27(3), 425-478. <https://doi.org/10.2307/30036540>
- Wang, Z., Liu, J., & Dong, R. (2018). Intelligent auto-grading system. In *Proceedings of the 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS)* (pp. 430-435). IEEE. <https://doi.org/10.1109/CCIS.2018.8691244>
- Wu, F., & Fan, W. (2016). Academic procrastination in linking motivation and achievement-related behaviours: A perspective of expectancy-value theory. *Educational Psychology*, 37(6), 1-17. <https://doi.org/10.1080/01443410.2016.1202901>
- Wypych, M., Matuszewski, J., & Dragan, W. Ł. (2018). Roles of impulsivity, motivation, and emotion regulation in procrastination—Path analysis and comparison between students and non-students. *Frontiers in Psychology*, 9, 891. <https://doi.org/10.3389/fpsyg.2018.00891>
- Xia, Q., Chiu, T. K. F., Lee, M., Sanusi, I. T., Dai, Y., & Chai, C. S. (2022). A self-determination theory (SDT) design approach for inclusive and diverse artificial intelligence (AI) education. *Computers & Education*, 189, 104582. <https://doi.org/10.1016/j.compedu.2022.104582>
- Yan, B., & Zhang, X. (2022). What research has been conducted on procrastination? Evidence from a systematic bibliometric analysis. *Frontiers in Psychology*, 13, 809044. <https://doi.org/10.3389/fpsyg.2022.809044>
- Ye, J. H., Zhang, M., Nong, W., Wang, L., & Yang, X. (2024). The relationship between inert thinking and ChatGPT dependence: An I-PACE model perspective. *Education and Information Technologies*. <https://doi.org/10.1007/s10639-024-12966-8>
- Yockey, R. D. (2016). Validation of the short form of the Academic Procrastination Scale. *Psychological Reports*, 118(1), 171-179. <https://doi.org/10.1177/0033294115626825>
- Yu, S.-C., Chen, H.-R., & Yang, Y.-W. (2024). Development and validation the Problematic ChatGPT Use Scale: A preliminary report. *Current Psychology*, 43, 26080–26092.

<https://doi.org/10.1007/s12144-024-06259-z>

Zhai, C., Wibowo, S., & Li, L. D. (2024). The effects of over-reliance on AI dialogue systems on students' cognitive abilities: A

systematic review. *Smart Learning Environments*, 11, 28.

<https://doi.org/10.1186/s40561-024-00316-7>

Date of reception: 1 June 2025

Date of acceptance: 9 September 2025

Date of approval for layout: 2 October 2025

Date of publication in OnlineFirst: 17 October 2025

Date of publication: 1 January 2026

Feedback on learning with generative artificial intelligence in university students

Retroalimentación de aprendizajes con inteligencia artificial generativa en estudiantes universitarios



-  María Verónica Leiva-Guerrero - *Pontificia Universidad Católica de Valparaíso, PUCV (Chile)*
 Ignacio Araya Zamorano - *Pontificia Universidad Católica de Valparaíso, PUCV (Chile)*
 Rafael Escobar Collins - *Pontificia Universidad Católica de Valparaíso, PUCV (Chile)*
 Francisca Silva Castro - *Pontificia Universidad Católica de Valparaíso, PUCV (Chile)*

ABSTRACT

Assessment for learning has become increasingly important in university teaching, particularly regarding the feedback process. However, there is still a perception of student dissatisfaction with the quality of feedback provided by faculty, highlighting the need to innovate in feedback strategies. This study aimed to explore the pedagogical and technological relevance of integrating Wilson's Feedback Ladder with generative artificial intelligence, specifically GPT-4o, to strengthen formative feedback in university students. The study was conducted using a qualitative and exploratory approach in two phases. First, a prompt was designed and validated using the Delphi method with the participation of eight experts in assessment and artificial intelligence, applying it to seven state-of-the-art language models. In the second phase, the validated prompt was implemented in two university courses of different nature, Assessment for Learning and Data Structures, integrating automatic feedback into the Moodle platform. The results showed that the experts agreed on the suitability of AI-mediated Wilson's Ladder and highlighted the superior performance of GPT-4o. At the classroom level, students valued the clarity, usefulness, and immediacy of the feedback, although they identified limitations in the tool's lack of contextualization and impersonal tone. It is concluded that the integration of Wilson's Ladder with generative artificial intelligence represents a promising innovation, but one that requires disciplinary adjustments, teacher supervision, and careful attention to the human dimension of feedback in e-learning contexts.

Keywords: feedback; formative evaluation; generative artificial intelligence; ChatGPT; Ladder of Feedback; university students.

RESUMEN

La evaluación para el aprendizaje ha adquirido creciente relevancia en la docencia universitaria, especialmente el proceso de retroalimentación. Sin embargo, persiste una percepción de insatisfacción en estudiantes sobre la calidad de retroalimentación brindada por el profesorado, lo que evidencia la necesidad de innovar en sus estrategias. Esta investigación exploró la pertinencia pedagógica y tecnológica de integrar la Escalera de Retroalimentación de Wilson con inteligencia artificial generativa, específicamente GPT-4o, para fortalecer la retroalimentación formativa en estudiantes universitarios. El estudio se desarrolló bajo un enfoque cualitativo y exploratorio en dos fases. En primer lugar, se diseñó y validó un prompt mediante el método Delphi con la participación de ocho expertos en evaluación e inteligencia artificial, aplicándolo a siete modelos de lenguaje de última generación. En la segunda fase, el prompt validado se implementó en dos cursos universitarios de distinta naturaleza, Evaluación para el Aprendizaje y Estructura de Datos, integrando la retroalimentación automática en la plataforma Moodle. Los resultados mostraron que los expertos coincidieron en la idoneidad de la Escalera de Wilson mediada por IA y destacaron el desempeño superior de GPT-4o. A nivel de aula, los estudiantes valoraron la claridad, utilidad e inmediatez de la retroalimentación, aunque identificaron limitaciones en la falta de contextualización y tono impersonal de la herramienta. Se concluye que la integración de la Escalera de Wilson con inteligencia artificial generativa representa una innovación prometedora, pero que requiere ajustes disciplinares, supervisión docente y resguardo de la dimensión humana en los procesos de retroalimentación en entornos abiertos y a distancia.

Palabras clave: retroalimentación; evaluación formativa; inteligencia artificial generativa; ChatGPT; Escalera de Wilson; estudiantes universitarios.

INTRODUCTION

Assessment for learning with a formative approach has gained increasing relevance at a global level, both in school and university contexts, by centering on continuous support and improvement of the educational process (Black & Wiliam, 2018; Andrade & Brookhart, 2019). Within this framework, feedback is situated as a key component, as it uses evidence of student performance to guide learning improvement (Carless & Boud, 2018).

However, various studies report a persistent dissatisfaction of university students regarding the quality of feedback provided by faculty, especially related to immediacy, specificity and utility (Quezada & Salinas, 2021; Galindo-Domínguez et al., 2023). This situation has been evidenced in research carried out in countries such as Australia, United Kingdom, Mexico and Chile, which suggests this is an issue with international scope.

In several educational systems, feedback practices have not had a significant evolution, which limits its formative impact. Strategies such as Wilson's Feedback Ladder (2013) or the Feedback Panel (Booth et al., 2008) have been proposed as effective alternatives to improve assessment processes, although its incorporation in teaching practices is still limited.

In the face of this scenario, it becomes necessary to renew university assessment strategies, incorporating emerging digital technologies that can support formative feedback (Puertas & Cano, 2024). In particular, Generative Artificial Intelligence (IAGen, for its name in Spanish) offers relevant opportunities to transform feedback in writing and complex thinking tasks (Dai et al., 2023; Al-Azawei et al., 2023).

Among the tools based on this technology we can mention ChatGPT (OpenAI, 2024), which integrates models such as GPT-4o, GPT-4.1 or GPT-4-5 and generates coherent and contextualized texts from prompts designed with pedagogical criteria. Although its incorporation to the educational field is recent, preliminary studies show an increasing use in teaching, learning and assessment, which opens new research lines about its applicability and efficacy (Galindo-Domínguez et al., 2023; García-Peñalvo, 2023).

Within this framework, the present study proposes three main questions: Is Wilson's Feedback Ladder adequate for its application through artificial intelligence? What IAGen tool, programmed with prompts according to these criteria, turns out to be more pertinent to provide feedback in university learning? And how do students perceive feedback received through this strategy applied with GPT-4o?

Artificial Intelligence in education

The integration of artificial Intelligence (AI) in the educational field has redefined the traditional teaching framework, becoming a fundamental core of contemporary pedagogical innovation. The role of the teacher, historically centered on content transmission, has experienced a substantial transition towards a role of learning facilitator, thanks to the potentialities of AI to automatize processes, analyze data and customize formative trajectories (Bonales-Daimiel et al., 2025). This reconfiguration allows educators to focus on tasks with greater pedagogical value such as qualitative feedback and differentiated support (García Peñalvo et al., 2024).

In addition, AI has demonstrated to be a strategic ally for the improvement of the educational experience, both in face-to-face and virtual contexts, through the

personalization of resources, automatized evaluation, immediate feedback and real-time monitoring of student performance (López Regalado et al., 2024; Romero Alonso et al., 2025). The implementation of adaptative systems and educational chatbots has allowed not only to optimize teacher efficiency, but also enrich learning processes with student-centered approaches.

Generative Artificial Intelligence

Generative Artificial Intelligence (IAGen) covers techniques and models designed to create new content (text, images, audio, video) that imitates the statistic distribution of large data groups. At the core of many IAGen solutions are the transformers, architectures of attention that, after massive pre-training, learn to preview the net unit (word, pixel, notation) and then generate coherent and contextualized outputs (Vaswani et al., 2017; OpenAI, 2023).

Thanks to fine-tuned algorithms and the incorporation of multiple input and output modalities, these systems can adapt to various tasks – from creative writing to conversational simulations – maintaining fluidity and sense in its answers (Gan et al., 2023).

Since 2024, the LLM ecosystem has diversified with business proposals - GPT-4o y GPT-4.5 (OpenAI, 2024), Claude 3.7 Sonnet (Anthropic, 2024) and Gemini 2.5 (Google DeepMind, 2024)— and from open code —LLaMa 4 (LLaMa, 2024), Mistral (Mistral AI, 2024).

In addition, some models such as DeepSeek R1 (DeepSeek, 2024), stand out by explaining their internal reasoning in the face of complex problems, and xAI Grok 3 (xAI, 2024) dedicates deep “reflection” phases for advanced mathematics and logic.

In education, these tools not only are useful to automatize content, but they help to customize teaching and feedback, adapting to each student and relieving repetitive tasks (work corrections), as long as ethical and methodological aspects are considered. Its responsible integration can significantly strengthen learning (Gan et al., 2023).

Prompts and prompting techniques

In the context of generative artificial intelligence, a prompt is the textual or structure input that is given to a language model to guide their answer. Its design is not trivial: the form, content and format of prompt directly impact quality, precision and utility of the generated output. As a result, the prompt engineering field emerges, dedicated to the creation of specific strategies to build effective prompts according to the task or the educational context (Liu et al., 2023; Reynolds & McDonell, 2021).

Between the main prompting techniques, we can find zero-shot prompting, where the model responds only to a direct instruction with no examples; and few-shot prompting, which incorporates representative examples before the task in order to model the type of expected answer (Brown et al., 2020). These techniques allow the modulation of the model’s behavior and better anticipate the type of content that will be created.

Another important technique is chain-of-thought, which guides the model to reason, step by step before offering a final answer, improving its performance in complex logic or structured assessment (Wei et al., 2022). In addition, there is role prompting, that assigns a role explicitly to the model – like an “expert evaluator” or “university professor” – to control the tone, depth and style of feedback (Sahoo et al.,

2024). Finally, structured prompting specifies the required output format, such as a dictionary or a table, which is especially useful when the information generated must be automatically processed by external systems, such as educational platforms (Schulhoff et al., 2025).

In education, prompting techniques allow automatized feedback to be structured in an accurate and pedagogically aligned way. The combined use of role prompting, chain-of-thought and structure prompting guides the model to reason as an expert evaluator, generating organized and coherent responses. This facilitates the integration to educational platforms and fosters a more useful, traceable and learning-centered feedback.

Formative feedback and its mediation with artificial intelligence

Feedback is a pillar of formative learning, conceived not only as error correction, but as an interactive process that promotes reflection, adjusting strategies and active building of knowledge. Winstone et al. (2022) describes it as a dialogic interaction between teacher and student, or between peers, that favors a deep understanding, as long as it is timely, specific and with clear goals. Along the same line, Carless and Boud (2018) underscore that it must be understood as a process of mutual influence and student agency, While Molloy et al. (2020) highlight the need for students to develop competences to interpret and apply it in autonomous contexts.

However, in practice, there are some structural barriers. In Chile, the Ministry of Education (2019) notes that, despite the theoretical recognition of feedback, evaluative practices continue to be mainly unidirectional and corrective, with little emphasis on metacognition and self-regulation. To address this gap, structured approaches are required to strengthen the formative dimension. Within this framework, Wilson's Feedback Ladder (Goodrich, 2011) offers a clear guide with four progressive levels (Clarify: Make questions that point towards solving ambiguities or aspects omitted in the student's work, promoting conceptual understanding; Value: highlight achievements and progress, acknowledging positive aspects of performance in an authentic way; State concerns: identify observed difficulties with a respectful and constructive tone, avoiding judgmental remarks and Suggesting improvements: offering specific orientations that will guide the student towards a deeper understanding and viable improvement actions) aimed at critical thinking and self-regulation. Its value lies on transforming feedback in an empathetic and participative experience that can adapt to different educational contexts, whether they are online or face-to-face.

In this study, Wilson's Ladder is adopted as a conceptual basis to design automatized feedback mediated by artificial intelligence. From this perspective, AI is proposed as a resource to customize and broaden coverage, in line with Wiliam (2011) and Shute and Rahimi (2017) regarding the potential of emerging technologies in evaluation for learning.

Artificial Intelligence for educational feedback

The incorporation of AI in educational feedback processes responds to the need to transform traditional practices into more interactive, personalized and effective models. Various research has proposed to innovate in this dimension through the use of AI, highlighting its potential to improve the quality of learning and reduce teacher

assessment workload (Puertas & Cano, 2024; Ossa & Willatt, 2023). The Teaching with AI – Assessment, Feedback and Personalization report, prepared by the European Commission (2023), delves into how AI can positively intervene in assessment and feedback processes from four complementary levels: social, institutional, teachers and students.

In this context, AI offers tools that are capable of adapting feedback to the individual needs of students through the automatized generation of comments, real-time assessment and academic progress follow-up (Ayeni et al., 2024; Holmes et al., 2019). This type of support allows teachers to identify with greater precision the specific difficulties of each student and offer a more focalized and differentiated teaching (Luckin & Holmes, 2016). Likewise, immediate feedback provided by intelligent systems has a positive impact on student motivation, since it offers timely answers that facilitate error comprehension and strengthen key learnings.

In particular, chat tools based on IAGen models have gained relevance as efficient pedagogical assistants. These systems allow the automatization of correction and feedback from previously established criteria, increasing the efficiency and coherence of the assessment process (Carless & Winstone, 2023). For teachers, this means a reduction in the operational workload; for students, an opportunity to receive immediate formative orientation, promoting self-assessment and critical thinking.

The efficiency of these tools has been documented in various empirical studies. Zhang et al. (2024) evaluated the use of ChatGPT to create feedback in programming tasks with students at an introductory Computer Science course. The majority of participants gave a positive value to the clarity and utility of the AI-generated comments, although some expressed they would have preferred feedback with more specific examples. For their part, Jauhiainen and Garagorry Guerra (2024) analyzed the ChatGPT-4 app in its assessment of open answers written by university students. Using a rubric with five criteria – relevance, accuracy, thoroughness, coherence and linguistic correction – they observed that the model offered detailed and consistent evaluations, although they insisted in the need for teacher supervision to guarantee its final accuracy.

Additionally, Baral et al. (2024) compared ChatGPT-4 with other models (like LLaMa and SBERT-Canberra) in the correction of secondary level mathematics answers. From 500 evaluated answers with a common rubric, it was found that ChatGPT-4 reached 92% of agreement with human teacher grading and also created high-quality feedback, evaluated through linguistic metrics and expert judgement. Nonetheless, the authors note that, although AI can approximate to human criteria, its use must be framed within a system that considers expert review and pedagogical reflection.

METHODOLOGY

Design

The study adopted a qualitative approach of exploratory nature, oriented towards designing, validating and implementing a formative feedback prompt automatized through generative artificial Intelligence (IAGen), supported by Wilson's Feedback Ladder (2013). The process was carried out in two phases.

In Phase 1, a prompt was created from a documental review on feedback strategies, and it was validated through the Delphi method (Landeta, 1999). The

procedure included two successive rounds of expert judgement. The first one evaluated the adaptation of the prompt and the four steps of the Ladder (clarify, value, state concerns, suggest) applied to seven representative language models (LLM). In the second phase, a synthesis of results was returned, and re-judgement was requested for agreement. A Likert 0-3 scale anchored in performance was used, and an interquartile ≤ 1 deviation was used as consensus criteria. In addition, Kendall's W coefficient was calculated as a measure of global consistency (acceptability criteria $\geq 0,70$).

In Phase 2, the validated prompt was integrated to the Moodle¹ platform in two university courses. Student open answers were processed automatically with GPT-4o to create immediate feedback structured according to Wilson's Ladder. Finally, student perceptions were collected through semi-structured interviews to supplement technical data with student experience.

Procedure

In the first phase, the goal was to ensure the designed prompt generated structured and coherent feedback with the four levels of Wilson's Ladder (2013) before being applied to students. This process included the language model selection, the iterative prompt creation and validation. Seven representative language models were selected for their technical diversity and accessibility: GPT-4o², GPT-4o-mini³ (OpenAI, 2024), Microsoft Copilot, Google Gemini (2024), Claude (Anthropic, 2024), Perplexity AI and LLaMA (2024), accessed from their official channels.

The prompt was designed as a structured template with: (i) theoretical context, (ii) question statement, (iii) evaluation criteria and/or expected response, (iv) student answer and (v) explicit instructions to guide feedback following Wilson's steps. Two activities were used during the pilot assessment: an open question based on a case study and a rubric creation task. The answers for each model were analyzed and then evaluated by a panel of eight experts – three assessment experts, three AI experts and two university professors – using a 0 to 3 scale and qualitative comments, in three rounds according to the Delphi approach (Steurer, 2011).

The purpose of the second phase was to implement the validated prompt in two undergraduate courses: Evaluation for learning and Data Structure. The Assessment (written test) was carried out in Moodle through the Coderunner⁴ plugin, automatizing the flow between student answers and GPT-4o. Each student answered open questions that, once sent, were processed in real-time generating immediate feedback, structured according to Wilson's Ladder, along to a preliminary score of 0 to 10 based on rubrics.

Students were informed about the experimental nature of the system and the option of an appeal was offered, reviewed manually by faculty, for the sake of equality. In addition, a qualitative analysis was later carried out through semi-structured interviews to volunteer students, enquiring on clarity, utility and reliability of the automatized feedback. Interviews were analyzed using deductive categories (based on Wilson's steps) and emerging subcategories by Mayring (2000). ChatGPT assisted in the analysis systematization, supporting thematic coding and extraction of significant patterns in student testimonies.

Instruments

The following instruments and techniques were used:

Leiva-Guerrero, M. V., Araya Zamorano, I., Escobar Collins, R., & Silva Castro, F. (2026). Feedback on learning with generative artificial intelligence in university students [Retroalimentación de aprendizajes con inteligencia artificial generativa en estudiantes universitarios]. *RIED-Revista Iberoamericana de Educación a Distancia*, 29(1), 241-265.
<https://doi.org/10.5944/ried.45547>

1. *Prompt Validation Protocol*

A protocol was designed to validate the performance of the prompt in seven Language models (LLM). This contained feedback generated by each model, organized according to the four steps in Wilson's Ladder (2013): Clarify, Value, State concerns and Suggest improvements. The expert panel assessed each step with an ordinal scale (0-3) and made qualitative observations when the score was equal or less than 1.

Validation took place through the Delphi method with an interdisciplinary panel of eight experts. The process included two consecutive rounds, using a 0-3 Likert scale and agreement criteria based on an interquartile (≤ 1) deviation and Kendall's W coefficient. This dynamic allowed the adjustment of the instrument before its application, confirming its coherence with Wilson's Ladder steps and supporting the selection of GPT-4o as a preferred model for the pilot phase.

2. *Written Test for Students (Summative Evaluation)*

The applied evaluations in both courses included open questions designed to activate higher-order cognitive abilities. In "Data Structure", explanations, algorithmic structures and technical justifications were addressed. In "Evaluation from and for learning", questions demanded answers with arguments about pedagogical concepts from observed classes. Questions were accompanied by specific rubrics, which allowed the prompt to be applied coherently.

3. *Structured Prompt*

The central instrument of the study was a prompt designed according to prompt engineering principles (Brown et al., 2020; Wei et al., 2022; Hao et al., 2022; Liu et al., 2023; Schulhoff et al., 2025), comprised of six components: (i) Theoretical context: synthesizes the conceptual and methodological foundations that frame the task, allowing to situate the evaluation within a solid disciplinary and pedagogical framework. (ii) Question statement: lays out the evaluative task with clarity, ensuring coherence with the learning objectives. (iii) Expected evaluation/answer criteria establishes the quality parameters and pertinence of answers, described through an analytical rubric with all their performance levels. (iv) Student Answer contains the production to be evaluated. (v) Instructions for feedback guide the formulation of judgements and comments following the steps in Wilson's Ladder (value, clarifications, concerns, suggestions), with extension limits and constructive tone. (vi) Output format in Python dictionary: organizes the results in the following fields: analysis and score by each evaluation criteria, in addition to the four steps in Wilson's Ladder. It was written in Spanish, using a markdown format to favor the legibility and the role of "expert evaluator" was assigned to ensure technical aspects. After its validation, it was decided to include in point (iii) the complete description of all levels of the analytic rubric and add the point (vi) to facilitate automatization of the answer in the Moodle platform. The validated version is attached as an annex for its replicability. It is important to highlight that points (v) and (vi) can stay unmodified, while the previous sections can be adapted to the corresponding evaluation.

4. Automatized Feedback System

An integrated system was developed in Moodle through the Coderunner plugin, that connects data with the GPT-4o model. Each time an answer was sent by a student, a structured prompt was generated and synchronically processed, providing immediate feedback. The output was automatically formatted in Moodle along with the preliminary score. In addition, data was stored for its later analysis. The model was put together with a temperature of 0.1, prioritizing more deterministic answers, consistent with the established evaluation rubric.

Technique

Semi-structured interview

With the goal of exploring student perception of the system, an interview was designed with open questions organized around six dimensions: Wilson's four steps, usability of the system and perceived impact of learning. Before its application, each participant signed an informed consent, protecting the principles of confidentiality, anonymity and voluntary nature. Interviews were recorded on audio with student authorization and completely transcribed for its analysis. The guarantee was that, at all times, the information would only be used for academic purposes, in agreement with current institutional ethical regulations.

The analysis was carried out using Mayring's (2000) qualitative content approach, combining deductive categories (derived from the steps in Wilson's Ladder and the complementary dimensions of usability, impact on learning and critical judgement with recommendations) and inductive, resulting from the information itself. Two coders worked independently in the reading, segmentations and coding of interviews, applying the previously agreed matrix. Then, there was a joint reviewing of results in order to solve discrepancies and consolidate definitive coding, which ensured the reliability of the intercoder and provide more strength to the findings.

The categories were enriched with emerging sub-categories based on participants' experiences, making the distinction between favorable assessments (utility, clarity, motivation, formative guide) and problematic ones (rigidity, de-contextualization, automatism, lack of empathy). The process also had the support of ChatGPT as a supplementary methodological tool, used to organize and contrast textual fragments, explore thematic patterns, and verify the coherence of the category system. Its use was limited to facilitating operational tasks, without replacing the analytical judgement of researchers, and contributed to strengthening the traceability and consistency of the process. Table 1 sums up the categories and subcategories used in the analysis.

Table 1
Categories and subcategories used for the analysis

Category	Favorable Assessment	Problematic Assessment
I. General Experience	- Evaluation of Innovation - Speed and precision of system	- Oddness at automatized evaluation - Absence of human connection

Category	Favorable Assessment	Problematic Assessment
II. Clarification (Ladder)	- Points to useful omissions - Clarifying examples	- Unnecessary corrections - Confusing suggestions
III. Value (Ladder)	- Positive reinforcement generates trust - Stimulates learning	- Emotional impact limited by being AI - Forced or mechanic tone
IV. Concerns (Ladder)	- Constructive and clear criticism - Encouraging improvement	- Redundance or irrelevance in certain criticism
V. Suggestions (Ladder)	- Clarity and applicability - Future improvements	- Suggestions outside of the covered content - Little contextual pertinence
VI. Usability and trust	- Trust in stable technical criteria	- Lack of assessment transparency - Doubts about pedagogical sensitivity
VII. Impact on learning	- Comprehension improvement - Feedback that is useful for study	- Demands perceived as excessive
VIII. Critical judgement and Recommendations	- Willingness to continue to use AI - Contribution to system development	- Need for personalization - Criticism to level adjustment
IX. Perceived autonomy and control (emergent)	- Clarity about expectations - Feedback as guide - AI Strategic adjustments (without explicit conflict)	- Rigid evaluation - Penalization by personal style or creativity - Low motivation due to lack of acknowledgement of alternative valid answers.

Source: Prepared by the authors.

RESULTS

Next, the main results regarding research questions are presented: Is Wilson's Feedback Ladder adequate to be applied through artificial intelligence tools?; Which IAGen tools, with prompts, designed according to Wilson's Ladder, result in being more adequate to provide feedback on learning in university students?; and how do university students perceive the feedback offered through ChatGPT-4o?

Wilson's Feedback Ladder with AI

One hundred percent of the assessments from the expert panel agreed that the strategy of Wilson's Ladder to provide feedback on learning with the use of AI was appropriate, since it offers clarity, value, reflection and suggestions to encourage the learning of students and strengthening the teacher-student relationship when the feedback steps are clearly defined. To this respect, the experts stated:

It provided a level of description and detail that not only centers on pointing out positive and negative aspects, but goes further, offering more possibilities for reflection, correction and learning improvement through the clarify, value, state concerns and offer suggestions steps (Expert 6).

Wilson's Feedback Ladder strategy is interesting and appropriate and provides a framework for feedback to be empathetic, descriptive and with a focus on contributing to the building of learning, instead of criticism (Expert 2).

More appropriate AI tools according to Wilson's Feedback Ladder

Table 2 presents the averages of assessments provided by experts about feedback generated by different AI tools, evaluated on a scale of 0 (very low) to 3 (very high) according to the four steps in Wilson's Ladder. Each tool was evaluated in terms of clarity, assessment, state concerns, and suggestions, also including its median and standard deviation. According to these data, GPT-4o stands out as the tool with the best general performance, by obtaining consistently higher scores in all criteria and a median of 2.87, along with a low standard deviation (0.18), which reflects both effectiveness and stability in feedback quality.

Table 2

Assessment Averages of Experts on Feedback with AI

AI Tool	Average Feedback Wilson's Ladder Examples 1 and 2					
	Clarify	Value	State Concerns	Offer Suggestions	Mean	Standard Deviation
GPT-4o	2.9	2.75	2.9	2.95	2.87	0.18
GPT-4o mini	2.2	2.1	2.6	2.35	2.30	0.35
Microsoft Copilot	1.55	1.6	1.75	1.95	1.70	0.49
Gemini	1.65	1.45	1.7	1.95	1.67	0.46
Claude	1.6	1.5	1.1	1.35	1.36	0.62
Perplexity	2.45	1.95	2.3	2.5	2.28	0.25
LlaMa 3.2	2	1.9	2	2.3	2.04	0.27

Source: Prepared by the authors.

The agreement between the eight judges when evaluating the different seven tools (GPT-4, Free GPT, Copilot, Gemini, Claude, Perplexity and LlaMa 3.2) was through Kendall's W coefficient, that reached a value of 0.43, which indicates a moderate level according to the judges. The associated meaning test rendered a value of $\chi^2 = 20.67$ with 6 degrees of freedom and a $p < 0.01$, which allows one to conclude that the observed agreement is statistically significant, indicating that the judges showed consistent agreement and statistical validity regarding tool classification according to Wilson's Feedback criteria.

Application of the tool in the Data Structure course

The tool was applied to a total of 107 second-year students. During the test, 846 open answers were processed, each one automatically evaluated through the GPT-4o model, using the previously validated structured prompt.

A voluntary appeal channel was set up through the institutional e-mail, allowing students to request a review of feedback or the assigned score. Thirty-four appeals were received, 27 of those were considered valid, which represents 3.19% of the total answers, a relatively low disconformity rate with the authorized system.

Table 3 presents an analysis of the main patterns of error of the GPT-4o model, on the basis of the total number of processed answers ($N = 846$). Each identified pattern includes a representative example of generated feedback, an interpretative analysis of the type of error, and the frequency (Freq.) with which it was observed, expressed both in the number of cases and the percentage of the total.

Table 3

Patterns of error observed in responses to the model

Pattern of error	GPT-4o feedback example	Error analysis	Cases	Freq.
Feedback of the model suggests including details that were already present in the student's answer.	"It would be useful to consider a clearer description of the operations that take place on the list, such as the addition of elements at the end and the elimination of the first element".	Feedback suggests that there are still missing details about the operations of adding and eliminating elements on the list. However, the student had already mentioned them in their answer.	31	3.7 %
Feedback of the model asks for details or aspects that were not requested in the question.	"It would be helpful to consider the importance of the hash function and how it affects speech efficiency".	Feedback suggests including additional information (such as the importance of the hash function) that was not requested in the original question. This can confuse the student since it extends the scope of the question and makes them feel their answer was incomplete despite actually answering appropriately to what was asked.	127	15 %
Feedback is unclear or ambiguous, limiting guidance for student improvement.	"Your response does neither address the specific search range nor how to collect elements within this range".	Feedback from AI suggests that the student did not explain about element collection, when in fact they did, but not with the most adequate approach. This can disorient the student since it is not clear if what they wrote was wrong or if they simply had to improve on it.	47	5.6 %

Source: Prepared by the authors.

Application of the tool in the Evaluation for Learning course

The tool was applied to a total of 37 students. During the test, 592 open answers were processed, each one automatically evaluated through the GPT-4o model, using the same structured prompt previously validated.

As in the previous experience, a voluntary appeal channel was activated through the institutional e-mail, so students could request a review of the feedback or the assigned score. In this case, no appeals were received, which suggests a high acceptance of the automatically generated assessments.

Table 4 presents an analysis of the main patterns of error detected in this second experiment, on the basis of the total of processed answers ($N = 592$).

Table 4

Patterns of error observed in responses to the model

Pattern of error	GPT-4o Feedback example	Error Analysis	Cases	Freq.
Feedback of the model asks for details or aspects that were not requested in the question.	“What specific examples of formative evaluation did you observe during class?”	Feedback does not render pertinent neither regarding the student’s answer nor the case stated in the question, which does not reflect an explicit application of the principles of formative evaluation.	3	0.5 %
Feedback has conceptual errors or incorrect statements.	The answer correctly highlights the pedagogical function of assessment and its formative approach, which is fundamental in a context of language learning. In addition, it adequately identifies the use of hetero evaluation, which shows an understanding of the roles of assessing agents in the class.	Feedback suggests that the student’s answer is correct and adequate; however, it does not address all aspects required by the question. This inconsistency can generate confusion, especially if the maximum score is not assigned and feedback does not clearly specify which elements should be improved.	4	0.7 %
Feedback uses negative or rigid concepts that do not agree with constructive feedback.	The answer lacks a discussion on how standardized tests could have an impact...	The word “lack” was replaced for not being appropriate enough in the context of constructive feedback, since it could result in being rigid or negative. In its place, a more formative and improvement-oriented statement was used, such as “maybe you could incorporate...” or “it is still to be addressed...”.	26	4.4 %

Feedback is unclear or ambiguous, limiting the student's orientation to improvement.	The answer correctly highlights the formative nature of the classroom assessment.	A double evaluation when using terms such as "highlights" and "correctly"; however, it does not provide constructive feedback from the perspective of the achieved goal. The answer does not specify which aspect of the formative assessment is valued in the classroom context, which limits its usefulness to guide improvements.	12	2.0 %

Source: Prepared by the authors

Student Perception on Received Feedback

From this category system, the systematic analysis of the interviews was performed, coding student answers in relation to defined categories and subcategories. This coding allowed for the detection of patterns of sense, thematic recurrences and contrast of individual values, as well as relevant nuances in the way in which participants interpreted automatized feedback. The analysis was carried out combining an inductive reading of the discourse with the structured classification by dimensions, which facilitated a holistic and nuanced understanding of the student experience. Each testimony was interpreted in the light of its internal coherence, its relationship with the categories of the analytical framework and its connection with the rest of the cases.

In general terms, students assign positive value to the experience of receiving automatized feedback, especially highlighting its immediacy, clarity and innovation in comparison to teacher traditional feedback. The surprise due to the use of artificial intelligence is recurring: "it is surprising that it reviews your assessment, since as a student, I am used to having this feedback come from the teacher (student 3). This innovation is seen as a significant change in the evaluative practice, perceived as new and promising. However, this technical assessment coexists with emotional nuances: some describe the experience as "a little stressful" (students 1 and 8), because "it makes you think that you have to be more prepared and be more accurate" (students 1,4,6). Together, students recognize the functional value of the system, but they also notice the absence of the "human factor" (students 1,2,5,7 and 8) and a certain affective distance in the feedback style.

Regarding the clarification step, participants appreciate that AI is capable of identifying omissions and ambiguities that they hadn't noticed themselves. It is valued, for example, when the AI indicated that "there are still details of the complete process of pointer adjustments that need to be listed so that the new node integrates correctly to the list" (student 7). These observations are perceived as useful and formative, but they are not exempt from criticism. Some students question the fact that "they correct me and explain something that I already wrote down in my answer" (student 10), which can generate confusion and a sensation of unfair assessment. This tension between structure and subjectivity is reinforced when the "AI becomes more critical when one freely provides an answer" (student 9) which questions the interpretative flexibility of the system in the face of non-literal but valid answers.

In relation to the value step, many students acknowledge the positive reinforcement as an incentive to continue learning. Affirmations such as “highlight what you did well and invites you to continue to study” (student 6) or “confirms that what I know is not so wrong” (student 2) show that recognition, even when automatized, can take a meaningful pedagogical function. However, its emotional impact is limited to some: “if the comments came from a professor, I would probably feel more acknowledged” (student 4). This ambivalence reveals that, although reinforcement validates knowledge and guides the student, it lacks the emotional charge and the empathy that a human teacher normally contributes, which poses ethical considerations about the quality of the pedagogical connection and interaction mediated by AI.

The concerns step shows that AI criticism is, in general, well received thanks to its clear and respectful tone. Student 1 points out that “this criticism is constructive, which reaffirms its mistakes”, while others comment that “I felt well criticized and to the point, straightforward” (student 3). Nonetheless, there are questions that arise when AI demands additional explanations that were not requested by the statement. For example, “although I received all the points, I believe that the part where it mentions including examples is a little unnecessary” (student 9). These observations suggest that feedback would be more effective if the level of demands was adjusted to the content that was actually worked with and the explicit evaluation criteria.

Regarding the suggestions step, students consider that AI offers recommendations that are “clear, concise and easy to apply”, which contributes to their autonomous learning. “It helped me to consider it for future occasions” (student 10), one student comments, while another highlights that “it emphasizes the writing order when explaining my algorithm” (student 3). However, the utility of these suggestions depends on the degree of alignment with the context of the course. Some criticize that “we hadn’t covered temporal complexity, and this confused the feedback” (student 5), showing that the system could lose efficacy if operating disconnected from the scheduled topics. In spite of this, the majority values structure and clarity of the suggestions as a valuable resource for progressive improvement.

Finally, in relation to the impact on learning, students agree that the use of artificial intelligence has strengthened their own formative process, significantly contributing to reflection about their own learning. Immediate feedback was valued for its ability to activate previous knowledge, “remember details I used to forget” (student 6) and allowing for timely review: “feedback the moment” (students 2); in addition, it was described as “useful to reflect about my teaching-learning process” (student 4).

Some participants recognize the contribution of this tool as a strategic resource: “it made me improve in topics I had not gone into depth about” (student 8) and “it helped me to identify important elements I had not considered before” (student 1). These perceptions reflect an active appropriation of automatized feedback as support for the development of reflective and metacognitive competences.

On the other hand, regarding critical judgement and recommendations, we can observe a proactive and analytical attitude. Students not only value the use of AI, but also expressed concrete proposals to improve its functioning. Among them, reorganizing the steps of Wilson’s Ladder, placing value before clarifying, incorporating a discussion section in order to explore contents further, and previously train AI with course materials, and avoid unjustified penalties, not deduct points due to the lack of examples if they are not required or are part of the evaluation criteria.

These proposals show that students do not think of themselves as a passive receptor of the system, but as a reflective agent, committed to perfecting and pedagogical instrument that they consider useful, but that they want to become more contextualized, fair and aligned with the formative goals of the course.

DISCUSSION AND CONCLUSIONS

The results of this study show the pedagogical potential of integrating automatized feedback mediated by generative artificial intelligence, in particular through GPT-4o structured through Wilson's Ladder. This proposal is aligned with the need to transform feedback practices in university teaching, which students tend to perceive as late, superficial and non-specific (Carless & Boud, 2018; Winstone et al., 2022; Quezada & Salinas, 2021). In open and distant contexts, where the immediacy and clarity are critical factors, the IAGen demonstrates that it is capable to reduce response times and deliver structured and focussed comments, strengthening the formative dimension of evaluation (Puertas & Cano, 2024; Holmes et al., 2019).

Wilson's Ladder offered a useful scaffolding to organize feedback in the four steps (clarify, value, state concerns and suggest), which allowed the generation of more understandable, specific and improvement-oriented observations, even in an automatized format. Students highlighted this experience because of its immediacy, clarity and innovation, recognizing it as a tool capable of supplementing the teaching role. Nonetheless, they also mentioned tensions associated to feelings of stress, greater demand on the preparation for answers and the absence of the human factor which offers closeness and warmth. These nuances strengthen the idea that technology cannot replace the teacher but must be conceived as a supplementary support to pedagogical interaction. This interpretation coincides with Valenzuela and Pérez (2025) who underscore the irreplaceability of teachers, in terms of human interaction and personalized support, remembering that artificial intelligence can build valuable resources, but always subordinate to the relational and affective dimension that is typical of teaching.

A relevant finding was the student observation regarding the clarify step. In current practice, the tool does not allow the student to contribute additional information before receiving comments, which reduces the formative reach of this step. To include an interactive mechanism, such as a chat of clarifying window, could enrich the personalization of feedback, aligning comments with real needs of students (Luckin & Holmes, 2016; López Regalado et al., 2024). This possibility gains greater importance in massive contexts or distance education, where interaction spaces tend to be limited.

In the technological dimension, validation by experts and the positive perception of students agree with what is proposed by Zhang et al. (2024) and Baral et al. (2024), who highlight that GPT can generate coherent and functional feedback as long as prompts are carefully designed. This aspect is critical in scenarios that require standardization and scalability (Carless & Winstone, 2023). However, as stated by Jauhiainen and Garagorry Guerra (2024), limitations linked to empathy, contextualization and affective sensibility persist, which reasserts the need for teacher supervision to ensure pedagogical pertinence and formative impact.

The study also shows that students from different fields and academic levels (second and fourth year) valued automatized feedback in a similar way, which suggests that Wilson's Ladder, applied with generative AI, becomes an adaptable and replicable

framework in various contexts. However, its implementation requires incorporating ethical criteria, teacher reviews and rigorous prompt design that will adequately guide the AI production.

Finally, there are challenges that emerge that invite further exploration in future research, in particular regarding the analysis of algorithmic biases, evaluation of ethical risks connected to privacy and equality, and also the need to guarantee the technical sustainability of these tools. It is also critical to explore its adaptation to various academic disciplines, ensuring pedagogical relevance and its alignment with higher education quality standards. Along this line, it is recommended that progress is made towards a responsible and contextualized integration of artificial intelligence in university evaluation processes, in such a way that its technical benefits supplement teacher pedagogical mediation. It is pertinent for the future to expand the reach of the studies to different disciplines and explore new dimensions of its impact in education, considering variables such as argumentative quality, problem resolution, critical thinking, student autonomy, and the transformation of evaluation practices. Likewise, it is suggested that longitudinal research is developed in order to analyze the sustained effect of IAGen use in learning and teaching practice. Simultaneously, the need emerges to strengthen teacher training in prompt design, ethical reflection and critical evaluation of the use of these technologies, with the purpose of ensuring a more conscious and formative implementation.

NOTES

1. Moodle is an open code learning management platform (LMS, because of its name in English) widely used in educational contexts for creation, delivery and follow up of online courses.
2. Version gpt-4o-2024-08-06
3. Version gpt-4o-mini-2024-07-18
4. CodeRunner is a Moodle plugin that allows automatic evaluation of information coding. It integrates with an execution engine (for example, Python) that processes answers in real time, creating feedback and immediate scores according to preset evaluation criteria.

REFERENCES

- Al-Azawei, A., Abdullah, A. A., Mohammed, M. K., & Abod, Z. A. (2023). Predicting online learning success based on learners' perceptions: The integration of the information system success model and the security triangle framework. *International Review of Research in Open and Distributed Learning*, 24(2), 7295. <https://doi.org/10.19173/irrodl.v24i2.6895>
- Andrade, H. L., & Brookhart, S. M. (2019). Classroom assessment as the co-regulation of learning. *Assessment in Education: Principles, Policy & Practice*, 26(1), 103-117. <https://doi.org/10.1080/0969594X.2019.1571992>
- Anthropic. (2024). *Claude 3.7* [Large language model]. <https://www.anthropic.com>
- Ayeni, O. O., Al Hamad, N. M., Chisom, O. N., Osawaru, B., & Adewusi, O. E. (2024). AI in education: A review of personalized learning and educational technology. *GSC Advanced Research and Reviews*, 18(2), 261-271. <https://doi.org/10.30574/gscarr.2024.18.2.0062>
- Baral, S., Worden, E., Lim, W.-C., Luo, Z., Santorelli, C., Gurung, A., & Heffernan, N. (2024). *Automated feedback in math education: A comparative analysis of LLMs for open-ended responses*. arXiv. <https://doi.org/10.48550/arXiv.2411.08910>

- Black, P., & Wiliam, D. (2018). Classroom assessment and pedagogy. *Assessment in Education: Principles, Policy & Practice*, 25(6), 551-575. <https://doi.org/10.1080/0969594X.2018.1441807>
- Bonales-Daimiel, G., Martínez-Estrella, E. C., & Sierra-Sánchez, J. (2025). Evolución del perfil docente y surgimiento de nuevos roles profesionales en la era de la inteligencia artificial (IA). *Pixel-Bit. Revista de Medios y Educación*, 73, Article 3. <https://doi.org/10.12795/pixelbit.109085>
- Booth, W. C., Colomb, G. G., & Williams, J. M. (2008). *The craft of research* (3rd ed.). University of Chicago Press. <https://doi.org/10.7208/chicago/9780226062648.001.0001>
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). *Language models are few-shot learners*. arXiv. <https://doi.org/10.48550/arXiv.2005.14165>
- Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. *Assessment & Evaluation in Higher Education*, 43(8), 1315-1325. <https://doi.org/10.1080/02602938.2018.1463354>
- Carless, D., & Winstone, N. (2023). Teacher feedback literacy and its interplay with student feedback literacy. *Teaching in Higher Education*, 28(1), 150-163. <https://doi.org/10.1080/13562517.2020.1782372>
- Dai, W., Lin, J., Jin, F., Li, T., Tsai, Y. S., Gašević, D., & Chen, G. (2023). *Can large language models provide feedback to students? A case study on ChatGPT*. EdArXiv. <https://doi.org/10.35542/osf.io/hcgzi>
- DeepSeek. (2024). *DeepSeek R1* [Generative AI model]. <https://www.deepseek.com/>
- European Commission. (2023). *Teaching with AI – Assessment, feedback and personalisation*. Briefing report No. 7 (European Digital Education Hub). Erasmus+ Programme. https://resitve.sio.si/wp-content/uploads/sites/7/2023/11/AI-squad-output_briefing-report-7.pdf
- Galindo-Domínguez, H., Delgado, N., Losada, D., & Etxabe, J. M. (2023). An analysis of the use of artificial intelligence in education in Spain: The in-service teacher's perspective. *Journal of Digital Learning in Teacher Education*, 40(1), 41-56. <https://doi.org/10.1080/21532974.2023.2284726>
- García-Peñalvo, F. J. (2023). La percepción de la inteligencia artificial en contextos educativos tras el lanzamiento de ChatGPT: ¿Disrupción o pánico? *Education in the Knowledge Society (EKS)*, 24, e31279-e31279. <https://doi.org/10.14201/eks.31279>
- García Peñalvo, F. J., Llorens-Largo, F., & Vidal, J. (2024). La nueva realidad de la educación ante los avances de la inteligencia artificial generativa. *RIED-Revista Iberoamericana de Educación a Distancia*, 27(1), 9-39. <https://doi.org/10.5944/ried.27.1.37716>
- Goodrich, H. (2011). Ladder of Feedback [Adaptation based on Ron Berger's Ladder of Feedback]. In A. Goodrich (Ed.), *Protocols in the classroom*. Harvard Project Zero.
- Google DeepMind. (2024). *Gemini 2.5* [Multimodal AI model]. <https://deepmind.google/>
- Hao, Y., Sun, Y., Dong, L., Han, Z., Gu, Y., & Wei, F. (2022). *Structured prompting: Scaling in-context learning to 1,000 examples*. arXiv. <https://doi.org/10.48550/arXiv.2212.06713>
- Holmes, W., Bialik, M., & Fadel, C. (2019). *Artificial intelligence in education: Promises and implications for teaching and learning*. Center for Curriculum Redesign.
- Jauhainen, J. S., & Garagorry Guerra, A. (2024). Generative AI in education: ChatGPT-4 in evaluating students' written responses. *Innovations in Education and Teaching International*, 1-18.

- <https://doi.org/10.1080/14703297.2024.2422337>
- Landeta, J. (1999). *El método Delphi: Una técnica de previsión para la incertidumbre*. Ariel.
- Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9), 1-35. <https://doi.org/10.1145/3560815>
- LLaMa. (2024). LLaMA4 [Large language model]. <https://www.llama.com/>
- López Regalado, O., Núñez-Rojas, N., López Gil, O. R., & Sánchez-Rodríguez, J. (2024). El análisis del uso de la inteligencia artificial en la educación universitaria: Una revisión sistemática. *Pixel-Bit. Revista de Medios y Educación*, 70, 97-122. <https://doi.org/10.12795/pixelbit.106336>
- Luckin, R., & Holmes, W. (2016). *Intelligence unleashed: An argument for AI in education*. Pearson. <https://www.pearson.com/content/dam/one-dot-com/one-dot-com/global/Files/about-pearson/innovation/open-ideas/IntelligenceUnleashedSPANISH.pdf>
- Mayring, P. (2000). Qualitative content analysis. *Forum: Qualitative Social Research*, 1(2), 1-10. <https://doi.org/10.17169/fqs-1.2.1089>
- Ministry of Education of Chile. (2019). *Orientaciones para la implementación del Decreto 67/2018 de evaluación, calificación y promoción*. MINEDUC. <https://bibliotecadigital.mineduc.cl/bitstream/handle/20.500.12365/14279/orientaciones%20decreto%2067.pdf>
- Mistral AI. (2024). Mistral [Large language model]. <https://mistral.ai/>
- Molloy, E., Boud, D., & Henderson, M. (2020). Developing a learning-centred framework for feedback literacy. *Assessment & Evaluation in Higher Education*, 45(4), 527-540. <https://doi.org/10.1080/02602938.2019.1667955>
- OpenAI. (2023). *Best practices for prompt engineering with the OpenAI API*. <https://help.openai.com/en/articles/6654000-best-practices-for-prompt-engineering-with-the-openai-api>
- OpenAI. (2024). GPT-4o, GPT-4.5 model cards. <https://openai.com>
- Ossa, C., & Willatt, C. (2023). Uso de inteligencia artificial generativa para retroalimentar escritura académica en procesos de formación inicial docente. *European Journal of Education and Psychology*, 16(2), 1-16. <https://doi.org/10.32457/ejep.v16i2.2412>
- Puertas, E., & Cano, E. (2024). ¿Puede la inteligencia artificial proporcionar un feedback más sostenible? *Digital Education Review*, 45(1), 50-58.
- Quezada, S., & Salinas, C. (2021). Modelo de retroalimentación para el aprendizaje: Una propuesta basada en la revisión de literatura. *Revista Mexicana de Investigación Educativa*, 26(88), 225-251.
- Reynolds, L., & McDonell, K. (2021, May). Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1-7). <https://doi.org/10.1145/3411763.3451760>
- Romero Alonso, R., Araya Carvajal, K., & Reyes Acevedo, N. (2025). Rol de la inteligencia artificial en la personalización de la educación a distancia: Una revisión sistemática. *RIED-Revista Iberoamericana de Educación a Distancia*, 28(1), 9-36. <https://doi.org/10.5944/ried.28.1.41538>
- Sahoo, S. S., Plasek, J. M., Xu, H., Uzuner, Ö., Cohen, T., Yetisgen, M., & Wang, Y. (2024). Large language models for biomedicine: Foundations, opportunities, challenges, and best practices. *Journal of the American Medical Informatics Association*, 31(9), 2114-2124. <https://doi.org/10.1093/jamia/ocae074>
- Schulhoff, S., Ilie, M., Balepur, N., Kahadze, K., Liu, A., Si, C., Li, Y., Gupta, A., Han, H., Schulhoff, S., Dulepet, P. S., Vidyadhara, S., Ki, D., Agrawal, S., Pham, C., Kroiz, G., Li, F., Tao, H., Srivastava, A., ... Resnik, P. (2025). *The prompt report: A systematic survey of prompt*

- engineering techniques. arXiv. <https://doi.org/10.48550/arXiv.2406.06608>
- Shute, V. J., & Rahimi, S. (2017). Review of computer-based assessment for learning in elementary and secondary education. *Journal of Computer Assisted Learning*, 33(1), 1-19. <https://doi.org/10.1111/jcal.12172>
- Steurer, J. (2011). The Delphi method: An efficient procedure to generate knowledge. *Skeletal Radiology*, 40, 959-961. <https://doi.org/10.1007/s00256-011-1145-z>
- Valenzuela Caico, R., & Pérez Carvajal, A. (2025). Inteligencia artificial en educación superior: ¿Un reemplazo para los profesores o una herramienta de apoyo? *Revista Iberoamericana de Investigación en Educación*, (9). <https://doi.org/10.58663/riied.vi9.221>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2023). *Chain-of-thought prompting elicits reasoning in large language models*. arXiv. <https://doi.org/10.48550/arXiv.2201.11903>
- Weng, X., Xia, Q., Gu, M., Rajaram, K., & Chiu, T. K. (2024). Assessment and learning outcomes for generative AI in higher education: A scoping review on current research status and trends. *Australasian Journal of Educational Technology*, 40(6), 37-55. <https://doi.org/10.14742/ajet.9540>
- William, D. (2011). What is assessment for learning? *Studies in Educational Evaluation*, 37(1), 3-14. <https://doi.org/10.1016/j.stueduc.2011.03.001>
- Wilson, D. (2013). *Ladder of feedback* [Working paper]. Project Zero, Harvard Graduate School of Education. <https://pz.harvard.edu/resources/ladder-of-feedback>
- Winstone, N. E., Boud, D., Dawson, P., & Heron, M. (2022). From feedback-as-information to feedback-as-process: A linguistic analysis of the feedback literature. *Assessment & Evaluation in Higher Education*, 47(2), 213-230. <https://doi.org/10.1080/02602938.2021.1902467>
- xAI. (2024). *Grok 3* [Large language model]. <https://grok.com/>
- Zhang, Z., Dong, Z., Shi, Y., Price, T., Matsuda, N., & Xu, D. (2024). Students' perceptions and preferences of generative artificial intelligence feedback for programming. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(21), 23250-23258. <https://doi.org/10.1609/aaai.v38i21.30372>

Date of reception: 1 June 2025

Date of acceptance: 2 September 2025

Date of approval for layout: 1 October 2025

Date of publication in OnlineFirst: 17 October 2025

Date of publication: 1 January 2026

PathRAG application in adaptive learning with generative AI for inclusive and sustainable education

Aplicación de PathRAG en aprendizaje adaptativo con IA generativa para una educación inclusiva y sostenible



 Rubén Juárez Cádiz - *Universidad Alfonso X El Sabio, UAX (Spain)*

ABSTRACT

This study presents the implementation of the PathRAG model within an adaptive, hybrid, and inclusive learning environment, supported by generative artificial intelligence. Aligned with the Sustainable Development Goals (SDGs), the proposal aims to personalize university-level learning through dynamic and equitable educational pathways. The objective is to address student diversity while reducing access and participation gaps through advanced educational technology. A quasi-experimental design was applied to a sample of 52 students enrolled in a Master's program in Inclusive Education at a Spanish university. The intervention was developed in a hybrid format, combining the PathRAG algorithm with generative AI tools (GPT-3.5 turbo). Key indicators such as active participation, competence development, perceived inclusion and equity, and overall student satisfaction were assessed. Findings show significant improvements in active engagement, skill acquisition, and inclusive perception, especially among students with special educational needs or limited technological access. Overall satisfaction was high, particularly regarding the usefulness of personalized learning paths. The study concludes that PathRAG fosters more equitable and adaptive learning processes. Nevertheless, limitations such as the absence of a control group, short duration, and lack of validated instruments are acknowledged. Future research should involve controlled designs, broader samples, and longitudinal approaches. This work highlights the transformative potential of generative AI in promoting sustainable and inclusive educational models.

Keywords: generative artificial intelligence; adaptive learning; educational inclusion; sustainability; hybrid education; personalized learning.

RESUMEN

Este estudio presenta la aplicación del modelo PathRAG en un entorno de aprendizaje adaptativo, híbrido e inclusivo, mediante el uso de inteligencia artificial generativa. Enmarcado en los Objetivos de Desarrollo Sostenible (ODS), el trabajo se orienta a personalizar el aprendizaje en contextos universitarios a través de rutas educativas dinámicas y equitativas. La propuesta busca responder a la diversidad del alumnado y reducir la brecha de acceso y participación mediante tecnología educativa avanzada. Se implementó un diseño cuasi experimental con una muestra de 52 estudiantes del Máster en Educación Inclusiva, perteneciente a una universidad española. La intervención se desarrolló en un entorno híbrido, integrando el modelo PathRAG con herramientas basadas en IA generativa (GPT-3.5 turbo). Se evaluaron indicadores de participación, desarrollo competencial, percepción de equidad e inclusión, así como la satisfacción general del alumnado. Los resultados indican mejoras significativas en la participación activa, el logro competencial y la percepción de inclusión, especialmente entre estudiantes con necesidades educativas específicas o dificultades de acceso tecnológico. La satisfacción alcanzó niveles elevados, destacando la utilidad de las rutas personalizadas. Se concluye que el modelo PathRAG potencia un aprendizaje más equitativo y adaptativo, aunque se reconoce la necesidad de futuros estudios con diseños más robustos, instrumentos validados y muestras más amplias. Este trabajo evidencia el potencial transformador de la IA generativa aplicada a contextos educativos sostenibles e inclusivos.

Palabras clave: inteligencia artificial generativa; aprendizaje adaptativo; inclusión educativa; sostenibilidad; educación híbrida; personalización del aprendizaje.

INTRODUCTION

The digital transformation in higher education has led to the development of pedagogical models that promote personalized learning and educational inclusion. In this context, adaptive learning has been established as an effective strategy for addressing student diversity by designing educational pathways tailored to their characteristics, pace, and needs. However, its implementation in real-world environments still faces significant challenges, including a lack of accessible technical tools, limited empirical validation of adaptive models, and the insufficient integration of inclusive and sustainable approaches.

In parallel, the recent incorporation of generative artificial intelligence (AI-G) models—such as GPT-3.5 turbo—offers new opportunities to facilitate the design of personalized, accessible, and contextualized pathways. However, it also brings associated risks, such as algorithmic bias in content generation or technological dependence on proprietary platforms, which requires a critical and pedagogically supervised implementation (Birhane, 2023; Bender et al., 2021).

This paper presents the application of the PathRAG model, based on a semantic graph of educational resources, in a hybrid university environment. The design was quasi-experimental with no control group, using an intentional and limited sample of students. The intervention is framed within the Sustainable Development Goals (SDGs), particularly target 4.5, which aims to eliminate inequalities in access to education and promote quality, inclusive, and equitable education (UNESCO, 2022).

THEORETICAL FRAMEWORK

The integration of advanced technology into educational contexts has intensified the debate on learning personalization, digital equity, and the role of artificial intelligence in building sustainable educational pathways. This theoretical framework is structured around three main pillars: adaptive learning, the potential and limitations of generative AI, and the principles of educational inclusion in hybrid environments, incorporating recent findings.

Adaptive learning and graph-based models

Adaptive learning is a pedagogical approach that dynamically adjusts content, pace, and difficulty levels based on a student's progress (Conati & Kardan, 2013; Ifenthaler & Yau, 2020). In recent years, knowledge graphs have been recognized as effective structures for representing concepts and semantic relationships in personalized educational systems (Zhou & Wang, 2020; Zhang et al., 2021; Ruipérez-Valiente et al., 2022).

Recent models like PathRAG leverage the structure of directed graphs to prioritize relevant learning paths based on a student's initial query, thereby promoting learning coherence and progression (Chen et al., 2024). This approach has proven effective in technical disciplines and hybrid environments, where automated AI support allows for adjusting recommendations to diverse profiles (Lu et al., 2024; Zhang & Wang, 2025).

Generative artificial intelligence: opportunities and challenges

Generative AI, particularly through language models like GPT-3.5-turbo, has had a substantial impact on higher education. It facilitates automated tutoring, real-time feedback, and the generation of adaptive content (Kasneji et al., 2023; van de Sande et al., 2023). Its integration into educational platforms has shown improvements in student efficiency and motivation, especially when it is used to support teachers rather than replace them (Susnjak, 2022; Luckin et al., 2022).

However, this technology is not without its ethical and pedagogical challenges. Recent research warns about the reproduction of algorithmic biases present in training data, which can reinforce pre-existing inequalities (Bender et al., 2021; Birhane, 2023; Selwyn et al., 2023). Furthermore, a reliance on closed technological providers limits transparency, traceability, and pedagogical control, compromising educational autonomy and creating risks of structural dependency (Zawacki-Richter et al., 2019; Chen et al., 2024).

In this regard, authors such as Floridi (2019) and Holmes et al. (2021) emphasize the need for applying robust ethical frameworks for implementing AI in educational contexts. Algorithmic governance, semantic auditing, and advanced digital literacy training for teachers are key elements to minimize system opacity and ensure AI is focused on student well-being.

Mathematical foundation of the path pruning algorithm

The PathRAG algorithm is based on a resource propagation model over a directed graph $G = (V, E)$, where V represents the set of nodes (educational entities such as competencies, concepts, or resources) and E is the set of edges (semantic relationships between these nodes, such as prerequisites or thematic dependencies). The algorithm's objective is to identify the most relevant relational paths from an initial node v_{start} , which is related to the student's query, to a set of target nodes V_q previously retrieved by a search or semantic selection system.

Initialization and resource propagation

Each node $v_i \in V$ receives an accumulated resource value, $S(v_i)$, which is initialized as follows:

$$S(v_{start}) = 1; S(v_i) = 0 \quad \forall i \neq start$$

The resource propagation follows an iterative dynamic, inspired by diffusion models. The resource value in each node v_i is updated as:

$$S(v_i) = \sum_{v_j \in N_{in}(v_i)} [\alpha \cdot S(v_j) / |N_{out}(v_j)|]$$

where:

- $\alpha \in (0,1]$ is a decay factor that penalizes longer paths.
- $N_{in}(v_i)$ is the set of nodes with incoming edges to v_i .
- $N_{out}(v_j)$ is the set of nodes to which v_j has outgoing edges.

This propagation scheme ensures that the closest and most traversed nodes in the graph receive a greater accumulation of resources, reflecting their structural relevance.

Reliability calculation and pruning

For each path $P = (v_0 \rightarrow \dots \rightarrow v_n)$, the reliability is defined as the average of the accumulated resource in its nodes:

$$S(P) = (1 / |E_P|) \sum_{v_i \in V_P} S(v_i)$$

where E_P is the number of edges in the path and V_P is the set of nodes that make it up. This metric allows for the identification of the most semantically robust routes—those through which more resources flow and, therefore, provide more explanatory value for the generative model.

Paths that do not exceed a minimum reliability threshold, θ , are discarded early:

$$\text{Prune if } S(v_i) / |N_{\text{out}}(v_i)| < \theta$$

This flow-based pruning operation drastically reduces the search space and focuses attention on the most promising paths, avoiding redundant or uninformative combinations.

Final selection

Finally, the top K paths with the highest reliability are selected to be transformed into structured inputs that feed the language model. Each path is converted into a textual string composed of the visited nodes, their relationships, and context. This guides the generation of explanations or adaptive content based on the student's profile.

This selection process, based on dynamic relational structures, offers significant advantages over flat retrieval methods. In educational settings, it allows for the construction of personalized itineraries, improves the coherence of AI-generated content, and ensures semantic alignment with defined learning objectives.

Inclusive education, equity, and sustainability

Inclusive education means ensuring that all students, regardless of their personal, social, or technological conditions, have equal opportunities for quality learning processes (Ainscow, 2020). Equity refers not only to access but also to effective personalization and meaningful participation, especially in hybrid or digital contexts (Anderson et al., 2020).

In this regard, adaptive educational technology—when well-designed—can serve as a tool to compensate for inequalities (García-Peñalvo, 2022). Models like PathRAG, by incorporating generative AI and semantic representation, allow for the adjustment of learning paths to a student's technological limitations, learning styles, and specific needs, thereby fostering a fairer and more sustainable learning environment.

Inclusive education, understood as the ability to equitably address student diversity, has been strengthened by the use of adaptive technologies that personalize resources and reduce access barriers (UNESCO, 2023; Cobo et al., 2022). Recent studies highlight that intelligent personalization, when implemented with principles of universal accessibility, can improve outcomes for students with specific needs and promote more sustainable trajectories (Pedro et al., 2021; Liu et al., 2023).

Furthermore, platforms that integrate graph-based augmented retrieval and content generation offer significant opportunities for designing contextualized learning paths, respecting individual paces, styles, and abilities. These strategies align with SDG 4 of the 2030 Agenda, fostering quality, inclusive, and equitable education.

METHODOLOGY

This study adopts a mixed-methods approach of a quantitative, quasi-experimental design without a control group, supported by qualitative, interpretive techniques. This strategy was chosen for its suitability to comprehensively evaluate both the effectiveness of the PathRAG system—based on generative artificial intelligence and graph-based augmented retrieval—and the subjective experience of students in personalized virtual learning environments. The combination of quantitative and qualitative methods aimed to provide interpretive robustness to the results through data triangulation and the convergence of perspectives.

The overall objective was to analyze the impact of implementing PathRAG in distance higher education, focusing on three key dimensions: (1) adaptive learning personalization, (2) improvement of the experience of students with diverse educational needs, and (3) alignment with principles of educational sustainability and digital equity.

The design was structured in three sequential phases:

- Phase 1 (Pre-test): Administration of a diagnostic test and an initial questionnaire to assess prior knowledge, autonomy, and motivation toward virtual learning.
- Phase 2 (Intervention): Development of three practical exercises in the virtual classroom, assisted by PathRAG. The system generated personalized learning paths and adapted content according to each student's cognitive profile.
- Phase 3 (Post-test): Evaluation of the evolution in conceptual understanding, specific skills, and perception of inclusion, supplemented with open-ended questions about the clarity, usefulness, and relevance of the generative AI intervention.

The quasi-experimental design without a control group was chosen due to the specific characteristics of the institutional context and access to a single consolidated group. Although this design limits direct causal inference, it allows for an approximation of the system's effects under real-world ecological conditions, providing relevant preliminary evidence for future research.

Participants

The sample consisted of 350 students from the Online Master's in Cybersecurity, delivered 100% virtually in an asynchronous format with online tutoring. This inter-

university, professional-focused program offers advanced training in cybersecurity within a distributed, multicultural, and technologically diverse environment.

The students had a heterogeneous profile, with ages ranging from 22 to 54, varying levels of work experience, and originating from over 10 Spanish-speaking countries (primarily Spain and Latin America). Twelve percent (12%) of the sample reported specific educational needs, broken down into three categories:

- Learning disorders: dyslexia, ADHD, or other (6%)
- Mild sensory disabilities: visual or auditory (4%)
- Technological limitations: poor connectivity or lack of equipment (2%)

The sampling was non-probabilistic and based on convenience, utilizing an already established, accessible group that aligned with the study's exploratory objectives. While this strategy limits the generalizability of the findings, it was justified by the feasibility of access, the diversity of profiles, and the thematic relevance of the program. For future studies, it is recommended to implement controlled designs with probabilistic sampling and comparative groups to enhance the external validity of the results and make more robust causal inferences.

Implementation environment and variables

The intervention was developed in a virtual learning environment (VLE) based on Moodle 4.1, which was integrated with a Python microservice (FastAPI) connected to a generative model (GPT-3.5-turbo). PathRAG was supported by a knowledge graph with over 12,000 nodes and 19,000 edges, built from the master's course materials (PDFs, transcribed videos, SCORM files, glossaries, etc.) and processed using NLP techniques (tokenization, entity extraction, semantic analysis).

- The intervention consisted of three practical activities:
- Vulnerability analysis in critical systems.
- OSINT attack simulation.
- Design of adaptive defenses.

Each exercise was preceded by a pre-test and followed by a post-test. PathRAG generated personalized learning paths and adaptive content (summaries, feedback, guided exercises) that were delivered dynamically through an individualized dashboard, based on student errors, interactions, and navigation patterns.

Integration of mixed methods

This study is based on a mixed-methods design, featuring a quantitative, quasi-experimental approach without a control group, complemented by a qualitative, interpretive analysis. The goal was to triangulate the results and enrich the understanding of PathRAG's impact from multiple learning dimensions.

Justification for the mixed approach

The combination of methods is justified by the complex nature of the educational phenomenon being evaluated: algorithmic personalization in virtual

environments. While quantitative data allowed for the objective measurement of indicators like participation, comprehension, and skills through standardized tests and activity logs, qualitative data provided access to the students' lived experience. This captured subjective aspects like their perception of inclusion, the clarity of personalized paths, and cognitive barriers.

Qualitative phase: instruments and procedure

During the third phase of the study (post-test), open-ended items were included in the evaluation instruments, designed to collect student feedback on:

- The usefulness of the personalized paths.
- The clarity and relevance of the generated content.
- Satisfaction with the AI-mediated learning experience.
- The perception of equity, cognitive effort, and accessibility.

These responses were coded using an inductive thematic analysis (Braun & Clarke, 2006), following this procedure:

- Open Coding: Two researchers independently analyzed the corpus of open-ended responses.
- Consensus Categorization: The units of meaning were grouped into common emergent categories.
- Triangulation with Quantitative Results: The categories were cross-referenced with statistical results to identify correspondences or contradictions.
- The inter-coder agreement index was over 90%, ensuring reliability in the coding.

Triangulation and integrative interpretation

The qualitative findings provided an explanation for the numerical patterns observed in the quantitative analyses. For example:

- The increase in active participation, detected in log records, was reinforced by comments that highlighted the usefulness of personalized paths and the motivating nature of automated feedback.
- Improvements in conceptual understanding coincided with positive perceptions regarding the logical organization of the generated content.
- The subgroups with technological or sensory limitations (12% of the sample) expressed feeling better attended to thanks to the progressive and tailored delivery of content, which aligned with their superior improvements in inclusion and equity indicators.

A representative example was the statement from a student with dyslexia: “For the first time, the system gave me just what was needed, without overwhelming me. It was as if it knew what I struggled with the most.” This type of qualitative evidence reinforces the conclusion that PathRAG helped to reduce cognitive overload and improve educational equity.

Strengthening the model

This methodological integration allowed us to do more than just validate statistical results from a more human and contextualized perspective. It also helped us identify areas for improvement and suggestions for future implementations. Thus, the qualitative data were not merely illustrative; they were complementary and explanatory, fully adhering to the principles of mixed-methods approaches aimed at improving education.

Instruments and indicators

Data collection was carried out using a set of instruments designed specifically for this study or adapted from existing scales. These instruments were validated by a panel of six experts in educational technology, accessibility, and pedagogical evaluation. They were then refined through a pilot test with 30 students from a related master's program, which helped ensure the clarity, relevance, and coherence of the items.

The instruments used were validated by an expert panel and adjusted in a pilot test. Table 1 summarizes the variables and instruments applied:

Table 1
Data collection instruments by variable

Variable	Instrument	Type
Active Participation	Moodle log records (clicks, logins, time on task)	Quantitative
Conceptual Understanding	Pre/post-test of specific knowledge	Quantitative
Skill Development	Teacher's rubric + self-assessment (based on DigCompEdu)	Quantitative
Inclusion and Perception of Equity	7-item Likert questionnaire + 2 open-ended items	Mixed
Satisfaction	Validated 10-item scale ($\alpha = 0.89$, test-retest = 0.81)	Quantitative

The open-ended answers were analyzed with inductive thematic analysis (Braun & Clarke, 2006), using independent double coding (intercoder agreement index > 90 %).

Student classification

To avoid overly broad groupings under the term “specific educational needs,” participants who reported any condition were classified into three distinct groups, following international categorization criteria (UNESCO, 2020):

- **Learning Disorders (6%):** Includes dyslexia, ADHD, or other conditions that affect the cognitive processing of formal learning.
- **Sensory Disabilities (4%):** Includes mild, non-disabling auditory or visual limitations that required format adjustments.

- Technological Limitations (2%): Includes a lack of stable connectivity or exclusive access to mobile devices that were not compatible with the virtual environment.

This classification allowed for a more precise analysis of the effect of adaptive personalization on diverse profiles and an evaluation of the system's inclusive capacity in real time.

Active participation

To measure active participation, the Moodle activity log was used, including the total number of logins, clicks on key resources, forum participation, and time spent in the VLE. "Full engagement" was defined as having accessed all three practical exercises, completed at least 80% of the proposed tasks, and participated in both the pre- and post-test questionnaires.

The data were anonymized and processed in an aggregate manner to preserve participant privacy, following the criteria of the GDPR.

Conceptual understanding

Conceptual understanding was assessed using a specific 15-item multiple-choice test, administered before and after each exercise. These items were developed by the teaching team and reviewed by external experts to ensure their content validity. The internal reliability index obtained was $\alpha = 0.81$.

Skill development

A teacher's rubric based on the DigCompEdu framework (European Framework for the Digital Competence of Educators) was used. This rubric covers six dimensions: digital literacy, instructional design, implementation, evaluation, empowerment, and professional development. Each dimension was rated on five levels (A1–C2). The rubric was applied by two evaluators with teaching experience in digital technologies and was supplemented by a student self-assessment using the same scale.

The inter-rater consistency was over 92%, and the internal reliability index of the scale was $\alpha = 0.86$.

Inclusion and perception of equity

To evaluate perceived inclusion, a mixed questionnaire was designed, consisting of:

- 7 Likert-type items (1–5 scale) on accessibility, perception of algorithmic fairness, language appropriateness, and adaptation to specific needs.
- 2 open-ended items: "To what extent did you feel the system adapted to your characteristics?" and "What would you improve in terms of inclusion?"

The questionnaire was validated by five specialists in digital inclusion and accessible education. Its internal consistency was $\alpha = 0.83$. The qualitative responses were coded using an inductive thematic analysis (Braun & Clarke, 2006), with independent double coding (inter-coder agreement index > 90%).

Satisfaction

A validated satisfaction scale for educational AI systems was used, consisting of 10 Likert-type items (1–5) that address aspects such as clarity, usefulness, suitability, reliability, and generated motivation, among others. The scale was adapted from Conati et al. (2021), with adjustments made for terminology and digital context.

The reliability analysis yielded an $\alpha = 0.89$ and a test-retest of 0.81, which guarantees the stability and consistency of the results.

Data analysis

To provide a holistic understanding of the effects of the PathRAG system in virtual learning environments, a mixed-methods analysis strategy was applied, combining quantitative statistical inference techniques with a qualitative thematic analysis.

The quantitative analysis included:

- Descriptive statistics: means, standard deviation, and frequencies.
- Inferential statistics: Student's t-test for related samples and repeated measures ANOVA.
- Effect sizes: Cohen's d and partial η^2 .
- Correlations between variables: Pearson's r coefficient.

The following assumptions were verified: Shapiro-Wilk (normality), Levene (homogeneity), and Mauchly (sphericity).

The qualitative analysis helped identify patterns related to accessibility, system clarity, cognitive effort, and the perception of educational fairness. The triangulation of these results strengthened the integrated interpretation of the findings.

Quantitative Analysis

Descriptive and inferential analyses were employed, prioritizing methodological rigor through the verification of assumptions and the calculation of effect size measures and confidence intervals, as detailed below:

Descriptive statistics:

- Calculation of arithmetic means, standard deviations (SD), absolute frequencies (n), and relative frequencies (%).

Inferential contrasts:

- Student's t-test for related samples (pre-test–post-test), for the three sequential exercises.
- Repeated measures ANOVA to evaluate the evolution of continuous variables (comprehension, skills, satisfaction) throughout the exercises.

Effect sizes:

- Cohen's d for comparisons of two moments (pre/post).
- Partial η^2 for multivariate analyses (ANOVA).
- Values were interpreted according to Cohen (1988): small (0.2), medium (0.5), and large (0.8).

Bivariate correlations:

Calculation of the Pearson correlation coefficient (r) between key variables (active participation, conceptual understanding, satisfaction) to explore significant linear relationships ($p < .05$).

Verification of statistical assumptions

To ensure the validity of the inferential analyses, the following assumptions were verified:

- Normality: Shapiro–Wilk test, with $p > .05$ as the acceptance criterion.
- Homogeneity of variances: Levene's test, particularly when comparing across exercises.
- Sphericity: Verified using Mauchly's test; in case of a violation, the Greenhouse–Geisser correction was applied.

Quantitative Analysis

Table 2 summarizes the results of the repeated measures ANOVA on conceptual understanding:

Table 2

Repeated measures ANOVA: conceptual understanding (n = 350)

Exercise	Mean	SD	F	df	η^2	Partial
Pretest	5.28	1.12				
Post 1	6.42	1.03				
Post 2	6.89	0.98	45.21	2, 698	< .001	0.179

The results show a significant increase in conceptual understanding over the course of the three exercises ($p < .001$), with a considerable effect size ($\eta^2 = .179$).

Additionally, a moderate positive correlation was observed between active participation and satisfaction ($r = .47$, $p < .01$), suggesting that more engaged students perceived greater usefulness from the system.

Qualitative analysis

The qualitative analysis was conducted on the open-ended questionnaire items and comments from the exercises using an inductive thematic analysis (Braun & Clarke, 2006). The process included:

- Open coding performed independently by two researchers.
- Categorization by consensus and triangulation with quantitative results.
- The identification of emerging themes related to:
 - Clarity of personalized paths.
 - Reduction of perceived cognitive effort.
 - Improvements in content delivery accessibility and equity.
 - Critical evaluation of AI use in training.

The inter-coder agreement index was over 90%, ensuring reliability in the categorization. The integration of qualitative and quantitative data allowed for the explanation of phenomena, such as the system's greater impact on students with technological or sensory difficulties.

Implementation environment

Construction of the educational graph

To develop the system's semantic foundation, an educational knowledge graph was built using the content from the training program. This graph was generated using Natural Language Processing (NLP) techniques applied to the master's structured and instructional texts (PDFs, SCORM files, glossaries, questionnaires, and transcribed videos). Conceptual entities (nodes) were extracted using the spaCy model, combined with customized heuristics based on lexical, grammatical, and pedagogical patterns. The relationships between entities (edges) were identified through frequent co-occurrences, syntactic dependencies, and explicit links such as "is a prerequisite for," "expands on," or "is solved with."

The final graph incorporated approximately 12,000 nodes and 19,400 semantic edges, representing a structured map of competencies, key concepts, and logical curriculum connections. This served as the core for the tasks of augmented retrieval and the inference of personalized learning paths.

Extraction of adaptive paths

The PathRAG engine used this graph to calculate the most relevant semantic paths for each student in real-time, based on their interaction within the VLE. Behavior patterns (e.g., resources consulted, frequent errors, omissions of key materials) were analyzed, and personalized implicit queries were generated. The algorithm was executed with the following parameters:

- Number of retrieved nodes: $N = 40$
- Number of generated paths: $K = 15$
- Decay rate: $\alpha = 0.8$
- Pruning threshold: $\theta = 0.05$

This procedure made it possible to identify highly reliable and didactically relevant paths, prioritizing fundamental concepts and learning trajectories consistent with the course objectives and student profile. The result was a knowledge graph consisting of over 12,000 nodes and approximately 19,000 relational edges, which reflected the pedagogical interdependencies between the program's concepts. This

graph was used by PathRAG to guide the generation of adaptive paths, providing each student with a personalized sequence of content and activities, generated in real-time based on their previous interactions, common errors, and areas of lower performance.

The integration of this system into the VLE allowed for the traceability of student actions, facilitated the automatic generation of educational resources coherent with their level and trajectory, and evaluated the impact of generative AI on personalized learning—all within a controlled, replicable environment linked to the program's curricular objectives.

Adaptive content generation

The selected paths were transformed into structured prompts, which served as input for the GPT-3.5-turbo generative model (OpenAI). This model generated adaptive content such as:

- Personalized explanatory summaries
- Directed practical exercises
- Automated feedback based on performance

These resources were delivered to the student through an individualized dashboard integrated into the Moodle platform, dynamically updating based on their performance, resource consultation, and evolution during each activity.

Real-time personalization

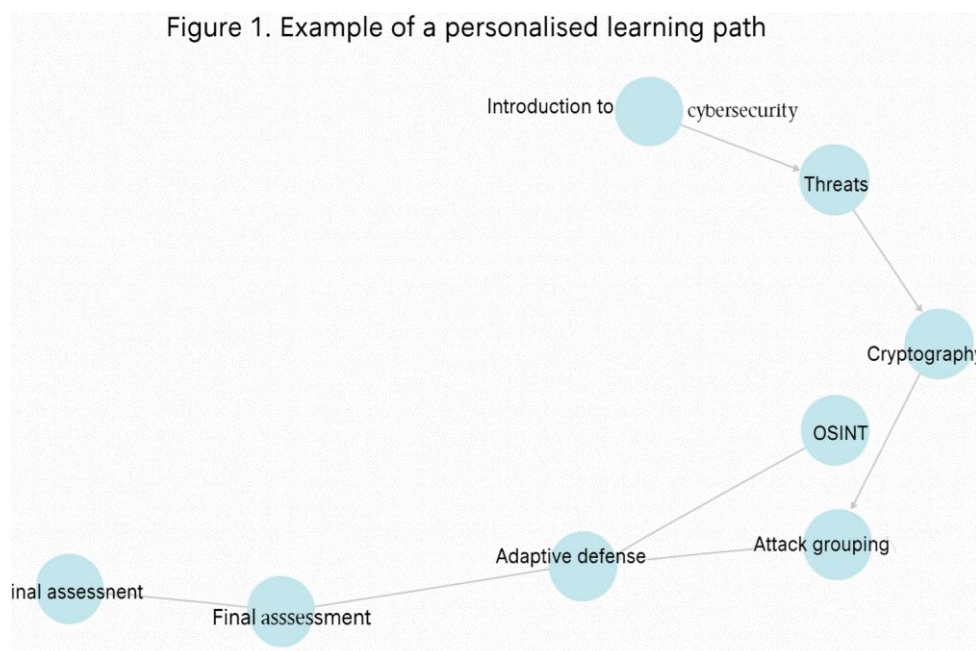
Each selected path was converted into a sequence of personalized learning activities—summaries, examples, practical tasks, and formative assessments—generated in real-time and presented on an individual dashboard. The content generation took into account:

- Areas of low performance detected (based on recurring errors and low time on task).
- Unachieved competencies from previous exercises.
- Format preferences (visual, textual, practical).
- Depth level required based on accumulated progress.

This approach enabled dynamic semantic personalization aimed at maximizing conceptual understanding, reducing unnecessary cognitive load, and promoting autonomous progression in virtual environments, as demonstrated in Figure 1.

Figure 1

Example of a personalized path in the knowledge graph



Ethical considerations

The entire procedure was developed within the framework of the General Data Protection Regulation (GDPR, EU 2016/679) and current Spanish educational legislation. The complete anonymization of data, voluntary participation, and access to a digital informed consent form, previously accepted by the students, were all guaranteed. The research team was careful to maintain the traceability of the generated content, the explainability of the automated processes, and human supervision throughout the intervention, in line with the principles of the ethical use of AI in educational contexts.

Artificial intelligence was applied at two key moments:

- **Augmented Retrieval with PathRAG:** This acted as a semantic filtering system to build the most relevant paths from the knowledge graph. It applied accumulated flow pruning to ensure quality and informational efficiency.
- **Adaptive Content Generation:** This used the selected paths as the basis for textual production via generative AI. The model was adjusted to limit excessive creative generation and ensure accuracy, brevity, and alignment with the SDGs and the curricular framework.

The use of AI was supervised by the research team, ensuring the application of principles of transparency, explainability, and algorithmic equity. Specific measures were taken to ensure data anonymization, participants' informed consent, and the traceability of the generated content. These procedures sought to safeguard the integrity of the educational process, avoid biases, and ensure the ethical and pedagogical coherence of the intervention.

Study limitations

This study has several limitations that should be considered when interpreting the findings:

- The quasi-experimental design without a control group prevents the establishment of conclusive causal relationships.
- The intentional sampling focused on a single master's program restricts the representativeness of the sample and makes it difficult to generalize the findings to other educational levels or disciplines.
- The limited duration of the intervention (September–December 2024) restricts the evaluation of long-term impacts.
- Some of the instruments used were developed ad hoc. While they were validated by expert judgment and a pilot test, they were not subjected to standardized validation processes.
- No a priori statistical power analysis was conducted, nor were expected effect sizes estimated, which affects the precision of the inferential conclusions.

These methodological limitations primarily affect the study's internal validity (the possible influence of uncontrolled variables) and external validity (limited generalizability). However, the exploratory nature of the study, the use of methodological triangulation, and the robustness of the instructional design partially compensate for these weaknesses and lay the groundwork for more robust future research.

RESULTS

The results obtained from implementing the PathRAG system with generative AI support in the online Cybersecurity Master's program show statistically significant improvements across all dimensions evaluated. Five key variables were analyzed: participation, conceptual understanding, skill development, perceived inclusion, and overall satisfaction. The analyses included pre-test/post-test comparisons, longitudinal evolution, and a breakdown by subgroups. The findings were also triangulated with qualitative evidence from open-ended responses.

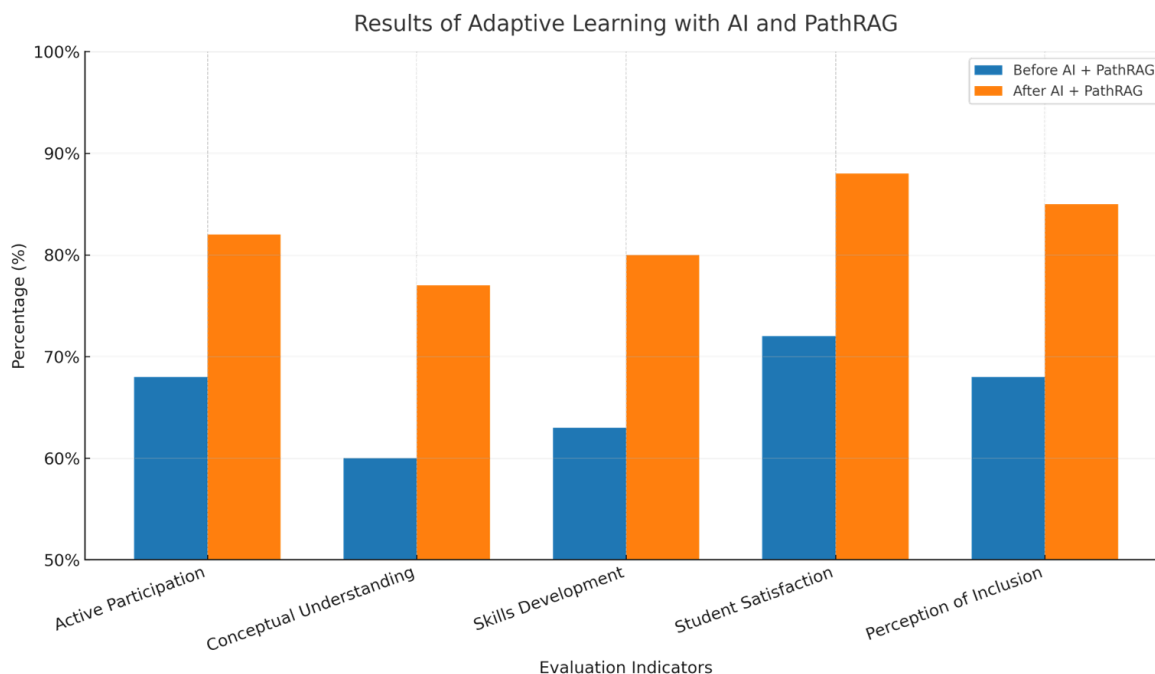
Overall improvement and learning progression

The main indicators showed a statistically significant increase following the intervention (see Figure 2). The observed improvements ranged from 15% to 20%, with moderate to large effect sizes (Cohen's $d = 0.63$ – 0.71), indicating a relevant impact on key learning indicators. These improvements were consistent across all three practical exercises, as confirmed by the repeated measures ANOVA test: conceptual understanding ($F(2,698) = 19.6, p < .001, \eta^2 = .22$), skills ($F(2,698) = 18.1, p < .001, \eta^2 = .19$), and participation ($F(2,698) = 15.7, p < .001, \eta^2 = .17$).

This evolution suggests that the adaptive generation of content, supported by personalized paths from the educational graph, promotes progressive learning tailored to the students' pace. These findings align with recent studies such as those by Chen et al. (2024) and van de Sande et al. (2023), which highlight the potential of generative AI in personalized tutoring and improving motivation in hybrid contexts.

The results indicate an overall improvement of between 15% and 20%, with moderate to large effect sizes. Figure 2 visually summarizes this evolution.

Figure 2
Results of adaptive learning with AI and PathRAG



The study represents five key indicators:

Active participation

Active participation, measured through the number of logins, time spent on the platform, and task completion, showed a significant increase. The percentage of students with full engagement rose from 65% to 82%, with statistically significant differences ($t(349) = 11.78, p < .001$). Figure 2 shows the evolution of participation before and after each exercise.

Conceptual understanding

The knowledge test administered for each exercise reflected an average improvement from 60% to 78%. The largest increase was observed in the second exercise (+18%), where the personalized content generated by the AI included contextualized practical simulations. Table 2 presents the evolution by exercise and indicator.

Skill development

Based on the teacher rubrics and self-assessment, an increase was identified in key skills such as critical analysis, problem-solving, and systemic thinking, with a global average improvement from 63% to 80%. 87% of students improved in at least

two dimensions of the rubric between the first and third exercises. Figure 3 shows the cumulative progression.

Student satisfaction

The subgroup of students with specific educational needs showed substantial improvements. Among them, participation increased from 55% to 75%, comprehension from 50% to 70%, and satisfaction rose from 60% to 82%.

The data from the validated satisfaction scale ($\alpha = 0.89$) showed a medium-high level of acceptance for the PathRAG + generative AI system. 89% of students gave a positive rating to the integration of personalized content, while 92% indicated that the system helped them better understand complex topics. Items related to usefulness, clarity, contextualization, and degree of personalization received scores higher than 4.2 on a 5-point scale.

Perception of inclusion

89% of students positively rated the use of artificial intelligence for personalizing resources, indicating that it improved their motivation and comprehension. The overall perception of the system's usefulness (mean = 4.3/5) was higher among those who received more generated prompts. Total satisfaction increased from 72% to 88% after the intervention.

Results by practical exercise

Table 3 shows a detailed breakdown of the evolution per exercise. A consistent pattern of improvement is observed, with significant increases in each indicator across all three exercises. For example, in conceptual understanding, the average went from 58% (Exercise 1) to 78% (Exercise 3); in skills, from 60% to 80%; and in participation, from 63% to 84%.

The repeated measures ANOVA test confirmed significant differences at all three stages for comprehension ($F(2,698) = 19.6, p < .001, \eta^2 = .22$), skills ($F(2,698) = 18.1, p < .001, \eta^2 = .19$), and participation ($F(2,698) = 15.7, p < .001, \eta^2 = .17$).

Table 3
Results per exercise

Exercise	Indicator	Before	After
Exercise 1	Comprehension (%)	58	76
Exercise 1	Skills (%)	60	78
Exercise 1	Participation (%)	63	80
Exercise 2	Comprehension (%)	61	77
Exercise 2	Skills (%)	62	79
Exercise 2	Participation (%)	66	83
Exercise 3	Comprehension (%)	62	78
Exercise 3	Skills (%)	64	80
Exercise 3	Participation (%)	67	84

These improvements reflect a pedagogical progression and effective content adaptation throughout the training process.

Impact on students with specific educational needs

Table 4 shows the results for the subgroup of students who reported access limitations or specific educational needs. This group started with lower levels (e.g., initial comprehension of 50% and participation of 55%), but after the intervention, they reached values of up to 75% and 70%, respectively.

Table 4

Impact on students with educational needs

Indicator	Before	After
Participation (%)	55	75
Comprehension (%)	50	70
Skills (%)	52	72
Satisfaction (%)	60	82

This finding underscores the inclusive value of the PathRAG approach, especially in offering personalized paths and adaptive content that are automatically generated based on the student's profile and learning trajectory.

Impact on students with specific educational needs

The subgroup of students with specific educational needs showed a particularly noteworthy evolution, improving from lower initial levels (50-55%) to values comparable to the rest of the sample (70-75%). This behavior suggests that the system helps to compensate for initial inequalities, aligning with the principles of educational equity under SDG 4 and the findings of Cobo et al. (2022) on inclusive AI.

Qualitative evidence reinforces this analysis. Statements such as "for the first time, I felt the content was adapting to my pace, not the other way around" (Student 42, with dyslexia) or "the explanations were clearer when they were adjusted for my errors" (Student 117, with limited connectivity) illustrate the positive subjective effect of algorithmic personalization, which has also been documented by Selwyn et al. (2023).

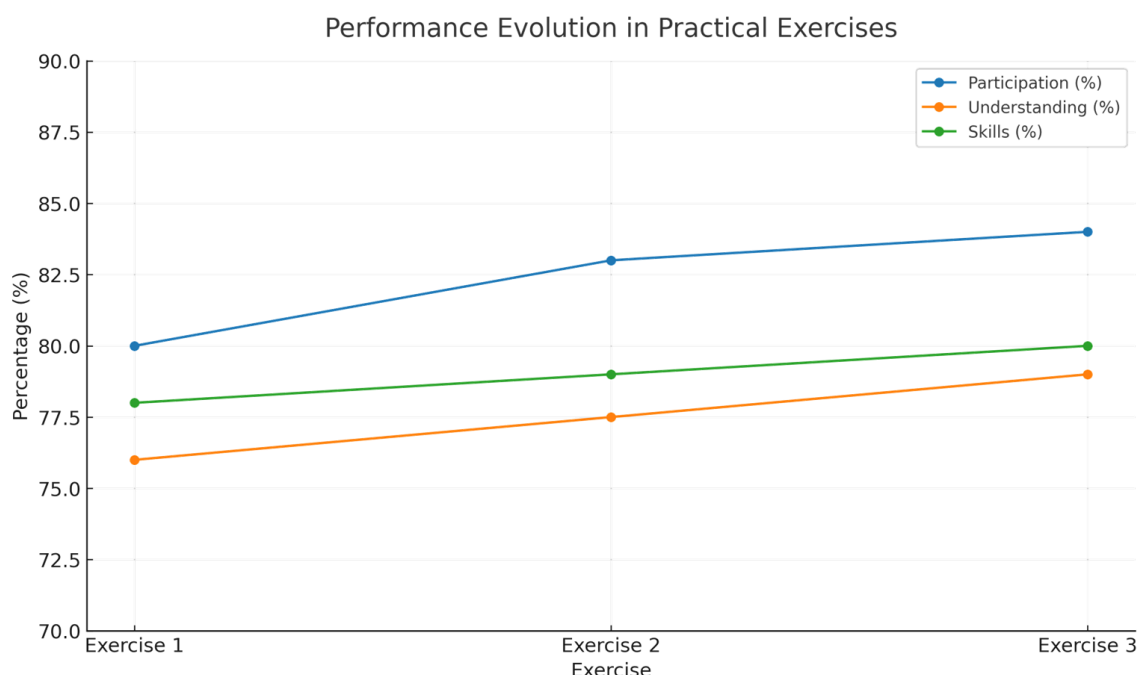
Qualitative responses highlighted that adaptive personalization was key to their progress. Here are representative examples from the thematic coding:

- "For the first time, I felt the content was adapting to my pace, not the other way around" (Student 42, with dyslexia).
- "The explanations were clearer when they were adjusted for my errors" (Student 117, limited access).

Progressive evolution and longitudinal analysis

Figure 3 presents the continuous evolution of the three main indicators, showing progressive improvement with each exercise. Comprehension increased from 58% to 78%, skills from 60% to 80%, and participation from 63% to 84%. These increases demonstrate a sustained learning curve.

Figure 3
Performance evolution in practical exercises



The consistency of the improvement supports the system's utility as an adaptive learning tool, extending beyond a one-time intervention.

Perception of inclusion and satisfaction

Overall satisfaction increased from 72% to 88%. The Likert questionnaires ($\alpha = 0.89$, test-retest = 0.81) reflect a high level of acceptance for the PathRAG system, with ratings over 4.2/5 for usefulness, clarity, contextualization, and personalization. 92% of students stated that the generative AI helped them better understand complex content, which aligns with the findings of Luckin et al. (2022) and Pedro et al. (2021) on the effectiveness of AI-mediated adaptive systems.

The thematic analysis identified patterns such as:

- Clarity in personalized explanations
- Reduction of cognitive load
- Increased motivation and autonomy

Geographical distribution and multicultural context

The geographical and professional diversity of the sample ($n = 350$, representing more than 10 countries) provides some external validity, but the limitation to a single technical master's program, non-probabilistic convenience sampling, and the absence of a control group prevent generalizing the results without reservations. The applicability to non-university levels or other disciplines should be verified empirically.

However, the technical framework of PathRAG—based on semantic graphs and directed prompts—could be easily adapted to other curricular domains, provided there is a well-structured documentary base and expert teacher supervision.

Table 5 shows a diverse geographical distribution, with a predominance of students from Spain (34%), Mexico (17%), and Colombia (14%). This heterogeneity reinforces the contextual validity of the study and its potential for application in global and sustainable educational environments.

Table 5
Geographical distribution of the sample

Country	Number of students
Spain	120
México	60
Colombia	50
Argentina	40
Perú	30
Ecuador	25
Chile	15
Others	10

This diversity reinforces the applicability of the PathRAG model in global, sustainable, and multicultural educational contexts, a characteristic that is especially relevant for distance learning programs.

Potential risks: algorithmic bias and technological dependence

However, some risks have also been identified that limit the model's generalizability and transferability. First, the structure of the educational graph, although manually curated, could incorporate epistemological biases derived from biases in the teaching materials or from unaudited human decisions. This phenomenon has been highlighted by Birhane (2023) and Holmes et al. (2021) as a critical source of algorithmic inequality in educational settings.

Furthermore, the dependence on OpenAI's GPT-3.5-turbo model, while functional in this research, reduces replicability and transparency, as it is a closed-source solution that is not fully explainable (Bender et al., 2021; Liu et al., 2023). These technical limitations compromise educational autonomy and could affect the system's scalability in contexts with stricter technological access or regulatory restrictions.

DISCUSSION

The results suggest that the PathRAG system, which combines augmented retrieval based on knowledge graphs with AI content generation, has a positive impact on the personalization and equity of learning in online higher education. The improvements observed in participation, comprehension, skills, and inclusion, both globally and within vulnerable subgroups, point to the pedagogical effectiveness of the

model. However, the quasi-experimental design used does not allow for direct causal relationships to be established.

Unlike traditional approaches centered on static content searches, PathRAG generates personalized semantic paths in real-time. These paths are adapted to the student's profile and optimized through pruning based on accumulated flow. This approach, supported by recent studies (Liang et al., 2023; Luckin et al., 2022), suggests an improvement in autonomous learning progression and a reduction of cognitive overload—two key factors in diverse and large-scale virtual environments.

A particularly significant finding was the performance increase among students with learning difficulties, sensory disabilities, or technological barriers. This subgroup started from lower levels but achieved improvements comparable to the rest of the sample, which indicates a democratizing effect of the system. These observations are aligned with recent research (UNESCO, 2023; Cobo et al., 2022) that recognizes the role of well-designed AI in reducing learning gaps.

From a contextual standpoint, the analyzed master's program—with students from over ten countries and heterogeneous digital competence levels—offered a natural laboratory to evaluate the system's intercultural adaptability. The analysis showed no significant differences by nationality or age, which reinforces the potential transferability of PathRAG to international, multicultural, and distributed environments.

However, several limitations must be acknowledged. The dependence on OpenAI's GPT-3.5-turbo as the generative engine poses significant risks: algorithmic biases, model opacity, and a lack of pedagogical control over responses (Bender et al., 2021; OECD, 2023). Although filters and adjustments were implemented, there remains a need to use auditable models trained on validated educational corpora with systematic human supervision.

There is also a possible source of bias in the construction of the educational graph itself, as it is based on existing teaching materials that may reflect curriculum limitations or imbalances. The partial manual validation and intensive instructional curation, while rigorous, require technical and human resources that are difficult to replicate in other contexts with less digital maturity.

Methodologically, the absence of a control group, non-probabilistic sampling, and the limited duration of the intervention affect internal validity and the ability to generalize the findings. However, the triangulation between quantitative results and qualitative evidence allowed for the interpretation of patterns beyond simple averages. For example, students with low initial scores reported how AI-generated personalized summaries "allowed them to restart from the basics without getting discouraged," while advanced profiles valued the ability to delve into complex content.

These findings are part of a growing body of research on adaptive and inclusive AI in education. Recent works by Jivet et al. (2021) and Zawacki-Richter et al. (2019) reinforce the idea that algorithmic personalization can foster learning self-regulation and promote fairer trajectories when designed with pedagogical, ethical, and digital equity principles in mind.

Future research directions

Based on the findings and limitations identified, several lines of research are proposed for future work:

- Longitudinal studies that analyze the sustained impact of the system over a complete academic year.
- Experimental designs with a control group to isolate effects and better estimate causality.
- Comparison between different generative models (e.g., GPT-4, Claude, LLaMA) to evaluate the quality, bias, and efficiency of the generated content.
- Semantic audits of the educational graph, focused on detecting cognitive, cultural, or gender biases.
- Applications in other disciplinary or educational contexts, such as secondary education, vocational training, or programs for adults with low digital literacy.

CONCLUSIONS AND FINAL CONSIDERATIONS

This study suggests that the combination of generative artificial intelligence and graph-based semantic retrieval, implemented through PathRAG, represents an innovative methodological approach with the potential to enhance adaptive learning in virtual, hybrid, and inclusive environments. By generating personalized paths based on a structured curriculum, the system facilitates coherent learning progression, adjusted to a student's errors, needs, and trajectory.

The evidence collected must be interpreted within the limitations of the methodological design. The absence of a control group, convenience sampling, the brief duration of the intervention, and the use of partially ad hoc instruments limit the internal and external validity of the results. As such, the findings should be considered exploratory and promising, rather than definitive proof of effectiveness.

Despite these limitations, the results suggest that the system particularly benefits students with differentiated educational needs, such as learning difficulties, sensory disabilities, or technological barriers, by promoting more equitable learning. This reinforces its alignment with the Sustainable Development Goals (SDGs), specifically SDG 4 (quality education) and SDG 10 (reduced inequalities), and underscores its potential to contribute to more inclusive and sustainable educational models.

From a technical standpoint, the system's success largely depends on the quality and coherence of the knowledge graph. Its construction requires demanding processes of curation, instructional design, and the application of natural language processing techniques. Additionally, the model currently relies on the exclusive use of GPT-3.5-turbo, which introduces risks of technological dependency and algorithmic bias, as well as limitations in transparency and pedagogical control.

In terms of future applicability, the PathRAG approach is scalable and adaptable to other educational levels (such as secondary or vocational training) and different disciplinary areas (humanities, social sciences, or health), provided there is an adequate documentary base and a teaching team trained in algorithmic literacy. It can also be applied in continuous training programs, multicultural contexts, and environments with high student diversity.

On an ethical level, the study was developed under the principles of the GDPR and teaching supervision, without automating evaluative decisions or compromising teacher autonomy. However, the move toward AI in education requires not only technical tools but also clear frameworks for governance, equity audits, and specialized, critical teacher training.

In conclusion, PathRAG should not be seen as a closed solution or a substitute for the teacher's role, but rather as a strategic support tool for personalized pedagogical design. Its value lies in increasing educational systems' responsiveness to diversity, facilitating tailored learning paths, and contributing to a more ethical, just, and resilient education model. When used critically and with supervision, generative artificial intelligence can become a key engine for democratizing access to knowledge and building more inclusive and sustainable educational ecosystems.

REFERENCES

- Ainscow, M. (2020). *Promoting inclusion and equity in education: Lessons from international experiences*. UNESCO. <https://doi.org/10.1080/20020317.2020.1729587>
- Anderson, T., Borup, J., & West, R. E. (2020). Learning analytics in higher education: A review of the literature. *Journal of Research on Technology in Education*, 52(4), 433-451. <https://doi.org/10.1080/15391523.2020.1739846>
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610-623). <https://doi.org/10.1145/3442188.3445922>
- Birhane, A. (2023). Algorithmic injustice: A relational ethics approach. *Patterns*, 4(2), 100693. <https://doi.org/10.1016/j.patter.2023.100693>
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77-101. <https://doi.org/10.1191/1478088706qpo630a>
- Chen, B., Guo, Z., Yang, Z., Chen, Y., Chen, J., Liu, Z., Shi, C., & Yang, C. (2024). PathRAG: Pruning graph-based retrieval augmented generation with relational paths. *arXiv*. <https://doi.org/10.48550/arXiv.2502.14902>
- Chen, X., Wang, L., & Li, Z. (2024). Generative AI in higher education: Ethical and pedagogical considerations for ChatGPT. *AI & Society*. <https://doi.org/10.1007/s00146-024-01720-9>
- Cobo, C., Morales, A., & Navarro, C. (2022). Artificial intelligence in education: Equity and ethics in Latin America. *International Journal of Educational Technology in Higher Education*, 19, 32. <https://doi.org/10.1186/s41239-022-00312-3>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.
- Conati, C., Holstein, K., & Roll, I. (2021). The future of intelligent tutoring systems: A survey. *IEEE Transactions on Learning Technologies*, 14(4), 464-484. <https://doi.org/10.1109/TLT.2021.3071740>
- Conati, C., & Kardan, S. (2013). User modeling in intelligent tutoring systems: A survey. In P. Brusilovsky, G. D. Verbert, & P. De Bra (Eds.), *Adaptive and adaptable learning* (pp. 45-63). Springer. https://doi.org/10.1007/978-3-642-40866-0_3
- Floridi, L. (2019). *The ethics of artificial intelligence*. Oxford University Press.
- García-Peñalvo, F. J. (2022). *Tecnología y Educación: Hacia una visión transformadora*. Ediciones Universidad de Salamanca.
- Holmes, W., Porayska-Pomsta, K., Holstein, K., & Rodrigo, M. M. T. (2021). Ethics of AI in education: Towards a community-wide framework. *European Journal of Education*, 56(4), 541-558. <https://doi.org/10.1111/ejed.12481>
- Ifenthaler, D., & Yau, J. Y.-K. (2020). *Utilising learning analytics to support study success in higher education*. Springer. <https://doi.org/10.1007/978-3-319-64792-0>

- Jivet, I., Scheffel, M., Drachsler, H., & Specht, M. (2021). Awareness is not enough: Pitfalls of learning analytics dashboards in the educational practice. *British Journal of Educational Technology*, 52(6), 1864-1880. <https://doi.org/10.1111/bjet.13168>
- Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., ... Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 102274. <https://doi.org/10.1016/j.lindif.2023.102274>
- Liang, W., Tadesse, G. A., Ho, D., Li, F., Liu, S., Yu, H., & Sun, J. (2023). Personalized education in the era of big data. *IEEE Transactions on Knowledge and Data Engineering*, 35(3), 2245-2261. <https://doi.org/10.1109/TKDE.2021.3117424>
- Liu, Y., Tamkin, A., Geng, X., & Ganguli, D. (2023). Lost in the middle: How LLMs use long contexts. *Transactions of the Association for Computational Linguistics*, 11, 1200-1218. https://doi.org/10.1162/tacl_a_00606
- Lu, X., van de Sande, B., & Marsh, J. A. (2024). Hybrid learning environments and adaptive support. *Computers & Education*, 210, 104975. <https://doi.org/10.1016/j.compedu.2023.104975>
- Luckin, R., Holmes, W., Griffiths, M., & Forcier, L. B. (2022). *Artificial intelligence and human learning: Partners in progress*. OECD Publishing.
- OECD. (2023). *The OECD AI principles: An overview*. OECD Publishing.
- Pedro, F., Subosa, M., Rivas, A., & Valverde, P. (2021). *Artificial intelligence in education: Challenges and opportunities for sustainable development*. UNESCO.
- Ruipérez-Valiente, J. A., Cobos, R., & Muñoz-Merino, P. J. (2022). Knowledge graphs in education: A systematic literature review. *IEEE Transactions on Learning Technologies*, 15(4), 485-500. <https://doi.org/10.1109/TLT.2022.3190858>
- Selwyn, N., Macgilchrist, F., & Williamson, B. (2023). EdTech and equity: Critical perspectives on the promises of digital education. *Learning, Media and Technology*, 48(1), 1-11. <https://doi.org/10.1080/17439884.2022.2156041>
- Susnjak, T. (2022). ChatGPT: The end of online exam integrity? *arXiv*. <https://doi.org/10.48550/arXiv.2212.09292>
- UNESCO. (2020). *Global education monitoring report 2020: Inclusion and education: All means all*. UNESCO. <https://doi.org/10.54676/JJNK6989>
- UNESCO. (2022). *Reimagining our futures together: A new social contract for education*. UNESCO.
- UNESCO. (2023). *AI and education: Guidance for policy-makers*. <https://unesdoc.unesco.org/ark:/48223/pf0000384803>
- van de Sande, B., Marsh, J. A., & Lu, X. (2023). Large language models in education: Opportunities and challenges. *Computers & Education: Artificial Intelligence*, 4, 100132. <https://doi.org/10.1016/j.caeai.2023.100132>
- Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education—Where are the educators? *International Journal of Educational Technology in Higher Education*, 16, 39. <https://doi.org/10.1186/s41239-019-0171-0>
- Zhang, D., & Wang, X. (2025). Adaptive learning systems: A review and future directions. *Journal of Educational Technology & Society*, 28(1), 1-15.
- Zhang, D., Wang, X., & Chen, X. (2021). Adaptive learning technologies in higher education: A meta-analysis of effectiveness. *Educational Research Review*, 33, 100390. <https://doi.org/10.1016/j.edurev.2021.100390>
- Zhou, M., & Wang, D. (2020). Ethical considerations in AI-supported education: A perspective from responsible

innovation. *AI & Society*, 35(3), 601-611.

<https://doi.org/10.1007/s00146-019-0910-1>

Date of reception: 1 June 2025

Date of acceptance: 11 August 2025

Date of approval for layout: 15 September 2025

Date of publication in OnlineFirst: 12 October 2025

Date of publication: 1 January 2026

Smart leveling: an AI-driven adaptive learning strategy in higher education

Nivelación inteligente: estrategia de aprendizaje adaptativo con IA en educación superior



Patricia Aldape-Valdes - *Tecnológico de Monterrey, ITESM (Mexico)*

Elvira G. Rincon-Flores - *Tecnológico de Monterrey, ITESM (Mexico)*

Leticia Castano - *Tecnológico de Monterrey, ITESM (Mexico)*

Sadie Guerrero - *Tecnológico de Monterrey, ITESM (Mexico)*

ABSTRACT

Academic leveling is one of the primary challenges in higher education, particularly in STEM disciplines—one that has intensified after the academic lockdowns of COVID-19 and for which traditional remedial courses have not yielded the expected results. This study evaluated the impact of an Adaptive Learning Strategy (ALS) designed to reinforce prior knowledge in 1,309 first-year students enrolled in the courses of Computational Thinking, Fundamental Mathematical Modeling, Mathematical Reasoning, and Mathematical Thinking at a private university in Mexico. Unlike conventional remedial approaches, the ALS offers a flexible, student-centered learning experience through brief adaptive modules integrated into regular courses. These modules, supported by an artificial intelligence platform, generate personalized learning pathways, adapt content, produce analytics, and facilitate data-driven instructional decision-making. The research adopted a mixed-methods approach (QUAN > QUAL) and a quasi-experimental design with matched samples to ensure group comparability. Results revealed statistically significant differences in academic performance in favor of the experimental groups compared to the control groups. Additionally, both students and instructors evaluated the ALS positively in terms of usefulness and learning experience. Findings suggest that an adaptive, integrated, and technology-mediated strategy is a promising alternative to address academic leveling in introductory university courses, offering substantial advantages over traditional remedial methods.

Keywords: adaptive learning; artificial intelligence; academic leveling; higher education.

RESUMEN

La nivelación académica es uno de los principales retos de la educación superior, particularmente en las disciplinas STEM, un desafío que se ha intensificado tras el confinamiento académico y ante el cual los cursos remediales tradicionales no han mostrado los resultados esperados. Este estudio evaluó el impacto de una estrategia de aprendizaje adaptativo (EAA) orientada a reforzar conocimientos previos en 1.309 estudiantes de primer año inscritos en las unidades formativas de Pensamiento Computacional, Modelación Matemática Fundamental, Razonamiento Matemático y Pensamiento Matemático de una universidad privada en México. A diferencia de los enfoques remediales convencionales, la EAA ofrece una experiencia de aprendizaje flexible y centrada en el estudiante mediante breves módulos adaptativos integrados en los cursos regulares. Estos módulos, respaldados por una plataforma de inteligencia artificial, generan rutas personalizadas de aprendizaje, adaptan contenidos, producen analíticas y facilitan la toma de decisiones docentes basadas en datos. La investigación adoptó un enfoque mixto (CUAN > CUAL) y un diseño cuasiexperimental con muestras emparejadas para asegurar la comparabilidad entre grupos. Los resultados mostraron diferencias estadísticamente significativas en el desempeño académico a favor de los grupos experimentales respecto a los grupos control. Además, estudiantes y profesores evaluaron positivamente la EAA en términos de utilidad y experiencia de aprendizaje. Los hallazgos sugieren que una estrategia adaptativa, integral y mediada por tecnología constituye una alternativa eficaz para abordar la nivelación académica en cursos universitarios iniciales, ofreciendo ventajas sustanciales sobre los métodos remediales tradicionales.

Palabras clave: aprendizaje adaptativo; inteligencia artificial; nivelación; educación superior.

INTRODUCTION

In recent decades, Adaptive Learning (AL) has emerged as an innovative educational strategy aimed at personalizing the teaching-learning process, responding to the needs, rhythms, and styles of each student (Ochukut et al., 2023). Its development has been closely linked to advances in educational technology, data analytics, and the development of Artificial Intelligence (AI), which has made it possible to design more dynamic and personalized learning experiences. These innovations offer the possibility of addressing one of the significant challenges of contemporary education: transforming teaching and learning practices to accelerate educational progress (UNESCO, 2019).

In this context, AL is presented as a solid and promising alternative to address one of the most persistent challenges in higher education: academic leveling (Dagunduro et al., 2024; du Plooy et al., 2024; Shi & Liu, 2025). Upon entering university, a considerable number of students have deficiencies in fundamental knowledge, particularly in STEM (Science, Technology, Engineering, and Mathematics) disciplines, where gaps in key concepts, logical reasoning, or quantitative skills often limit their academic performance (Ross et al., 2022). Traditional leveling approaches, based on standardized remedial courses, have demonstrated limitations by not considering the particularities of each student. Therefore, a different educational strategy is required, which focuses on the individual requirements and the real performance of each student. Before addressing the Adaptive Learning Strategy (ALS) proposed in this study, some theoretical foundations that support it are presented.

Academic leveling: a persistent challenge

The lack of prior knowledge in the initial years of professional careers worsened after the COVID-19 pandemic restrictions. Classical theories of learning emphasize the cruciality of previous knowledge to facilitate the understanding of new concepts. Wood et al. (1976) stated that students require structured support to build on their existing knowledge; however, when this prior knowledge is insufficient, the scaffolding process intensifies and becomes more complex.

Learning is most effective when instruction is calibrated just beyond students' current level, provided there is sufficient prior knowledge on which to build (Vygotsky, 1978). In the absence of such prior knowledge, remedial instruction is required to enable learning. Sweller (1988) indicated that the lack of prior knowledge increases the cognitive load, which hinders the learning process. This idea aligns with Vygotsky's theory regarding the need to apply leveling strategies to support students.

The meta-analysis carried out by Hattie (2008) compiled more than 800 studies on the factors that influence student learning, highlighting that prior knowledge is fundamental because it provides the base on which new learning is built. Students with relevant background knowledge tend to have a deeper and more effective understanding of new concepts. In fact, Hattie (2017) found that prior ability has an effect size of 0.94 on academic performance, which is considered highly significant.

Research shows that engineering students with deficiencies in algebra or calculus are at higher risk of dropout (Klingbeil et al., 2007; McKenna et al., 2000). Drawing on cognitive load theory, Sweller et al. (1998) argue that insufficient prior knowledge can overwhelm students when they encounter complex topics (e.g., design

or programming). Therefore, at the university level, bridging strategies—such as remedial courses, reinforcement programs, tutoring, and introductory modules—together with pedagogical approaches that scaffold learning, are needed to ensure that students master the necessary prerequisites before progressing in their programs (Felder & Brent, 2005; Seymour & Hewitt, 1997).

Some universities have adopted multiple strategies to support the academic leveling of new students. For example, California State University has a peer-mentoring STEM program in which students in advanced semesters mentor first-time students, providing them with information about careers and sharing their personal experiences. Having a peer who offers guidance and support during the initial semesters has proven to be an effective strategy to improve retention rates and facilitate a faster, more positive integration into university life (Taha et al., 2015; Rockinson-Szapkiw et al., 2021).

Other universities, such as Massachusetts Institute of Technology (MIT) with its MIT Introduction to Technology, Engineering, and Science (MITES) program or the National Autonomous University of Mexico (UNAM) and its program inspiring women to enter STEM, among many others, have implemented specific programs to strengthen STEM careers. These programs not only offer academic advice to help students level up but also provide emotional support to those who begin their university education, to strengthen their knowledge, and prepare them to face the challenges of these disciplines.

Adding to the structural factors that explain the lack of prior knowledge in particular areas, the COVID-19 pandemic significantly exacerbated this problem (Filatova et al., 2023; López del Puerto et al., 2021). In response, several institutions developed remote programs during and after the pandemic, with a special focus on supporting academic leveling, particularly in STEM courses (Lasater et al., 2021; Schaal et al., 2021).

However, some of the traditional strategies used for leveling, such as preparatory courses before the formal start of the program's subjects, need to be rethought. In the study carried out by González-Beltrán et al. (2018), data mining techniques were applied to analyze whether these leveling courses significantly affected the performance or approval of students in Calculus subjects. However, the results showed no evidence to support such a correlation.

Although leveling courses before university admission have been one of the most common strategies implemented by universities, they are not always the most effective. One of the primary reasons is that these courses tend to address all the contents of the previous educational level in a general way, without considering that students do not all have the same knowledge gaps. This can demotivate students and lead to their negative perception of the usefulness of the content. However, thanks to technological advances, Adaptive Learning (AL) emerges as a promising alternative to achieve academic leveling as it allows the learning process to be personalized, focusing only on the topics that each student needs to reinforce (du Plooy et al., 2024; Taylor et al., 2021).

Beyond traditional teaching: adaptive learning to recover and enhance knowledge.

AL has been implemented for several years, classified under the umbrella of Personalized Learning, according to Jing et al. (2023). The AL approach aims to

personalize learning and materialize the principle of student-centered education. Similarly, Hernandez et al. (2022) define Personalized Learning as an educational approach that adjusts the learning experience of each student according to their individual needs, strengths, abilities, and interests, being flexible regarding what, when, how, and where to learn throughout their curriculum and various formative experiences. In this way, multiple strategies and technologies provide different degrees of autonomy and options, allowing students to play an active, leading role in their learning within these environments.

Digital technologies enable multiple approaches to personalize learning. The broader and more thoughtful their use, the more precise and responsive personalization can become. Adaptive Learning is an educational strategy that utilizes data-driven technology to tailor learning itineraries for personalized learning, providing effective and efficient content pre-designed by the teacher based on the student's performance level, profile, and learning needs. This allows teachers to identify gaps in content comprehension to improve actions and adjust their educational practice, optimizing student performance (Rincon-Flores et al., 2024). Thus, the teacher's role shifts from content provider to facilitator of learning. However, to achieve these efficiencies, teachers must use technological systems appropriately and efficiently (Cavanagh et al., 2020).

Based on a computer science and information technology survey in higher education, Simon and Zeng (2024) agreed that most university leaders and faculty display a positive attitude towards Adaptive Learning (AL) and consider it a strategy with high potential to improve student success. In addition, Rincon-Flores et al. (2023) demonstrated that technology-assisted AL effectively optimizes learning. Similarly, Sarıyalçınkaya et al. (2021) argue that AL allows the teaching trajectory of each student to be defined according to their learning style, based on the evaluation of their previous knowledge.

In the field of technological systems, one must understand the relevance of data and its processing. Long and Siemens (2011) highlighted that the strong drive to increase efficiency and reform higher education has made it natural to orient learning towards new digital formats. This has led to an explosion of compiled educational data, including activity flows, student interactions with peers and mentors, administrative records, and other valuable sources of information. The possibility of collecting this massive volume of data, analyzing it in real time, and applying machine learning techniques to identify patterns is primarily due to the support of artificial intelligence (Miao et al., 2021). Today, these elements are integrated into adaptive learning platforms that incorporate artificial intelligence (AI) as the core of their operations.

Fernández-Morante et al. (2022) reported that learning analytics has enormous potential to optimize educational processes; however, for this potential to produce meaningful results, the focus must shift from technology to learning-centered instructional design. Similarly, Kara and Sevim (2013) added that adaptive platforms use algorithms, assessments, feedback, adjustments, instructor intervention, and different means to offer new learning materials to students who have reached the expected level or provide reinforcement activities to those who have not yet achieved it.

One of the main advantages of adaptive platforms lies in their ability to generate learning analytics and monitor student progress at both the individual and group levels. These tools collect real-time data on performance, study habits, and areas of opportunity, enabling more timely and effective pedagogical decision-making, both

during and after implementation. The strategic use of adaptive platforms with AI integration not only optimizes teaching intervention but also strengthens personalization and promotes equity in STEM courses, characterized by high academic heterogeneity (Koester et al., 2016; Li et al., 2023). These technologies facilitate differentiated learning paths adjusted to the specific needs of each student, thus increasing their chances of academic success.

In this sense, Moskal et al. (2017) highlighted that adaptive platforms employ a data-based and, in some cases, non-linear instructional approach, which promotes academic leveling and continuous content adaptation. This perspective sees adaptive platforms as a high-value pedagogical resource for addressing the challenges of teaching in complex and diverse educational environments.

Regarding instructional design, other methodologies, such as Self-Regulated Learning and Microlearning, can complement AL (Rincon-Flores et al., 2024). The former requires the student to use self-regulation strategies, respond to feedback, and maintain interdependent motivations (Zimmerman, 1990), while the latter offers brief, focused content, facilitating the assimilation of key concepts (Allela et al., 2020). An effective ALS integrates structured content design, relevant teaching strategies, a coherent implementation model, and technologies that enhance personalization (Rincon-Flores et al., 2024; Simon & Zeng, 2024; Shi & Liu, 2025).

A successful example of an ALS application in STEM courses has been at the University of Central Florida (UCF), which redesigned its physics courses by incorporating appropriate technology to serve 4,500 students, achieving improvements in student success, increased learning outcomes, and greater engagement with content (Dubey et al., 2023). Similarly, India's state of Haryana, aware of AL's positive impact, became the first to implement it on a large scale, selecting an educational technology partner to provide relevant software and content for 500,000 tablets distributed to public school students across the country (Press Trust of India, 2022).

On the other hand, the results of AL are not always successful or conclusive. An example is the study by Wang et al. (2023), where, although students in the experimental group (adaptive learning) improved significantly between pre-test and post-test, their results were not statistically significant from the control group (teacher-led learning). Similar findings were reported by Eau et al. (2022), who found benefits only in some students, particularly those with high performance. Similarly, Olmos-López et al. (2023) evaluated an approach in which teachers personalized instruction using a predictive model of student performance; however, the predictions were primarily useful for estimating group performance. Thus, instruction was not personalized at the individual level but was adapted to group needs.

Intelligence leveling model: adaptive learning with AI in STEM programs

The ALS presented in this study integrates adaptive leveling modules within first-year courses in STEM programs, addressing basic knowledge that students should have mastered by the time they graduate from high school. Each module begins with a mandatory diagnostic exam that defines a personalized learning path, allowing only the necessary content to be taken. This occurs in the first few weeks but remains available for consultation throughout the course. Participation is not optional and represents 5% of the final grade, encouraging its use and ensuring that all students reach the required level.

The ALS presents four key components to ensure its pedagogical and technological effectiveness: content design, the didactic model, the role of the teacher, and the technological platform.

Content design

Content design is based on micro-learning, breaking content down into small lessons with specific goals. Each lesson includes contextualization, recognition material, examples, exercises, and a final quiz. The knowledge-check resources include one base and two optional reinforcements. The contents are presented in various formats, including videos, infographics, readings, and animations. The quiz has items of varying difficulty depending on the student's level.

Didactic model

Based on principles of self-regulated learning and self-study, the didactic model allows students to control their learning and adapt it to their needs. The pedagogical structure is designed for autonomous learning, managing time, pace, and needs. The student begins with an initial diagnostic examination. Based on the results, each reflects and follows a personalized learning path that indicates lessons to be reinforced. Each lesson includes context, videos, explanations, examples, exercises, and a quick quiz. A final diagnostic exam is presented at the end of all lessons, and the grade is recorded.

Role of the teacher

Although the self-study model promotes student autonomy, the teacher is key to the student's success. The teacher motivates the class, provides guidance regarding the adaptive module, and ensures that diagnostic tests are performed in class. The teacher reviews AI-generated outcome analytics on the technology platform to identify and track specific needs, thus acting as a strategic companion in the leveling process.

Technology platform

To implement the ALS, it was necessary to select a platform that would allow teachers to create content, integrate various materials, and use AI for analysis, monitoring, and generation of personalized paths. **RealizeIT** was selected after the assessment of different options for its integration with the LMS, variety of content formats, detailed analytics, user-friendly interface, and its *Adaptive Intelligence Engine*. This engine uses mathematical and statistical models and algorithms to record and analyse the progress, strengths, weaknesses, and preferences of each student (Howlin & Lynch, 2014). It detects learning patterns and offers personalized and adaptable paths based on each student's interaction with the content, as well as recommendations for resources and practices to facilitate the mastery of concepts.

METHODOLOGY

The study adopted a mixed approach, **CUAN > CUAL** (Creswell, 2015), where qualitative data complemented quantitative data. A quasi-experimental design was

applied, with control and experimental groups in four first-year courses: Computational Thinking and Fundamental Mathematical Modeling in the School of Engineering and Sciences (SES), Mathematical Reasoning in the School of Business (SB), and Mathematical Thinking in the School of Social Sciences and Government (SSG). The research was carried out between August and December 2024 at a private university with multiple campuses throughout Mexico among students of medium and high socioeconomic status.

Research objective

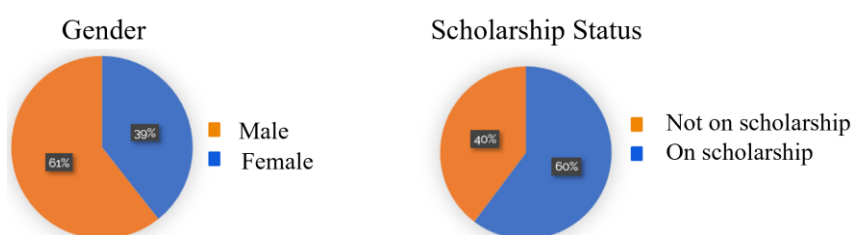
To evaluate the impact of the Adaptive Learning Educational Strategy (ALS), implemented through the technological platform, on the **leveling** of students' prior knowledge in the Mathematical Thinking (MT), Mathematical Reasoning (MR), Mathematical Modeling (MM), and Computational Thinking (CT) training units, and to analyze the **learning experiences** of teachers and students.

Sample

The non-probabilistic sampling process resulted in a sample comprising 1,309 students from the Schools of Engineering and Sciences, School of Business, and School of Social Sciences and Government. Samples per course were matched; for this, the participants of the experimental and control groups with equivalent characteristics were selected using SPSS software (version 29). The variables considered for the pairing were: final high school grade, gender, type of school of origin (public or private), and grade obtained in the pre-test. As a result, the matched sample comprised 734 students: 120 from the *Mathematical Thinking* course, 50 from *Mathematical Reasoning*, 410 from *Mathematical Modeling*, and 154 from *Computational Thinking*. Figure 1 shows some sociodemographic characteristics of the total sample.

Figure 1

Sociodemographic statistics concerning gender and scholarship status



In addition, 88% of students came from private schools and 12% from public schools. Regarding the teaching staff, 40 teachers participated in both treatments; however, only 11 responded to the final survey. Their average age was 40, and the average teaching experience was 10 years.

Instruments

A pre- and post-test of previous knowledge was applied in each course. These instruments were created by the teams of teachers in charge of developing the contents of the leveling modules. Subsequently, a pilot study was conducted to verify that item

difficulty and discrimination indices were appropriate and that the scales showed adequate internal consistency (Cronbach's alpha), see Table 1.

Table 1

Psychometric results of the instruments

Course	Number of students	Difficulty Index	Discrimination Index	Cronbach's alpha
Computational Thinking	28	0.71	0.42	0.8
Mathematical Reasoning	36	0.61	0.48	0.86
Mathematical Thinking	41	0.71	0.29	0.78
Mathematical Modeling	34	0.57	0.33	0.74

Each instrument was piloted with twice as many items as planned. Following the analysis, the 16 items with the best psychometric results were selected, resulting in each instrument consisting of 16 items.

At the end of each course, a survey was applied to students and teachers in the control and experimental groups. The objective of this questionnaire was to collect information on the learning experience in relation to the ALS for leveling. The dimensions of this instrument were: *Learning Path, Learning Process, Teaching Process, Content, Engagement, Usability, User Experience, and Satisfaction*. The instrument utilized a continuous-interval Likert scale with a range of 0 to 100 points to capture nuances in the intensity of each response.

Statistical process

ANOVA tests were performed to analyze the results of the pre- and post-test, after verifying the assumptions of normality and homogeneity of variances. Likewise, a Difference-in-Differences (DiD) analysis was applied to estimate the real effect of the treatments. As for the survey, since it used a continuous Likert-type scale, chi-square tests identified differences between the teachers in the experimental groups. Also, the *use of data* provided by the adaptive platform was analyzed. Finally, the open questions in the questionnaires were analyzed through thematic categorization to complement and enrich the quantitative results.

Ethical aspects

The study was carried out in accordance with the ethical principles set out in the Declaration of Helsinki. All students, professors, and academic authorities were informed about the objectives and procedures of the study. Participants were explicitly notified that all information collected would be held to strict confidentiality standards, ensuring absolute anonymity. Moreover, all participants gave their informed consent before starting their participation.

RESULTS

Table 2 presents the results of the comparative analysis between the control and experimental (adaptive) groups in the post-test of prior knowledge for each course. The assumptions of normality (Shapiro–Wilk test) and equality of variances (Levene test)

were verified. Since no statistically significant differences were identified between treatments in the pre-test, it was decided to perform a one-factor ANOVA without covariates, dispensing with the use of ANCOVA.

Table 2

Mean difference test with paired samples

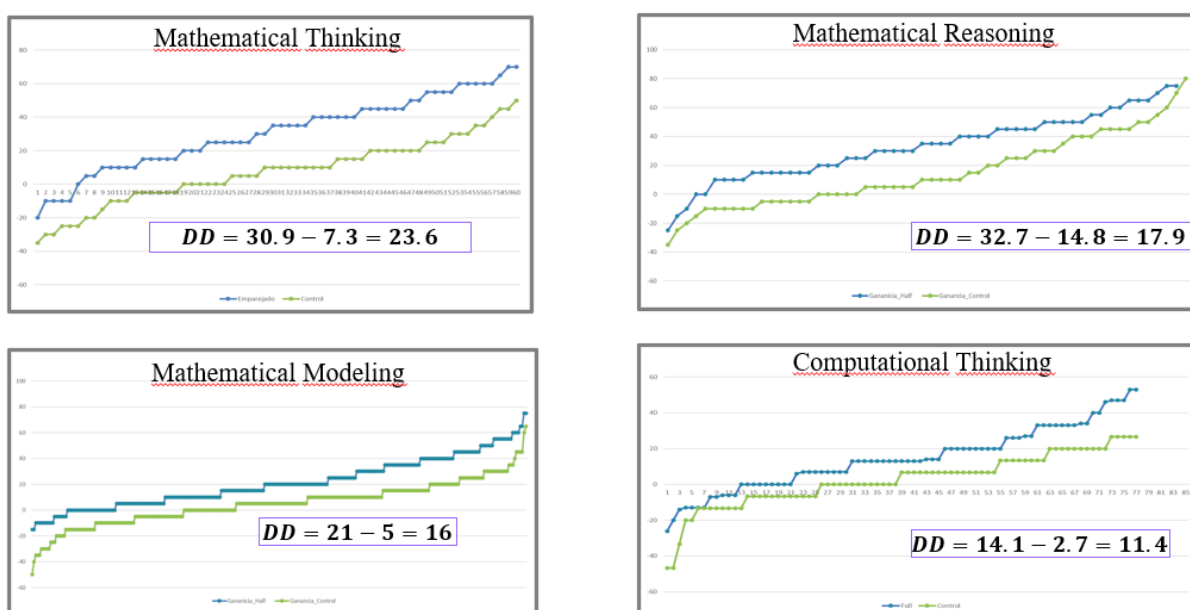
Course	Treatment	n	Mean (SD)	df	F	p	η^2	Interval
Computational Thinking	Adaptive	74	84.7 (13.3)	146	32	.001	0.18	(.08, .29)
	Control	74	70.8 (16.2)					
Mathematical Reasoning	Adaptive	25	72 (19.3)	48	7.9	.007	0.14	(.01, .32)
	Control	25	56.4 (19.4)					
Mathematical Thinking	Adaptive	44	68.3(24.1)	86	26	.001	0.23	(.09, .37)
	Control	44	45.3(17.6)					
Mathematical Modeling	Adaptive	196	73.8 (19.4)	394	45	.001	0.1	(.05, .16)
	Control	196	59.9 (21.7)					

Note: SD = standard deviation.

The results showed statistically significant differences between the control and experimental groups, with a higher performance in the experimental groups ($\beta = 17.2$, CI 95% [11.4, 23.6], $p < .001$). Figure 2 shows the learning gains estimated by the Difference in Differences (DiD) model, where the blue line represents the performance of the experimental groups and the green line corresponds to the control groups. The adaptive intervention yielded net increases (difference-in-differences, DiD) of 11.4, 17.9, 23.6, and 16.0 points in Computational Thinking, Mathematical Reasoning, Mathematical Thinking, and Mathematical Modeling, respectively. The weighted average additional gain was approximately 16.2 points relative to the control group. This pattern is consistent with an effect attributable to the intervention.

Figure 2

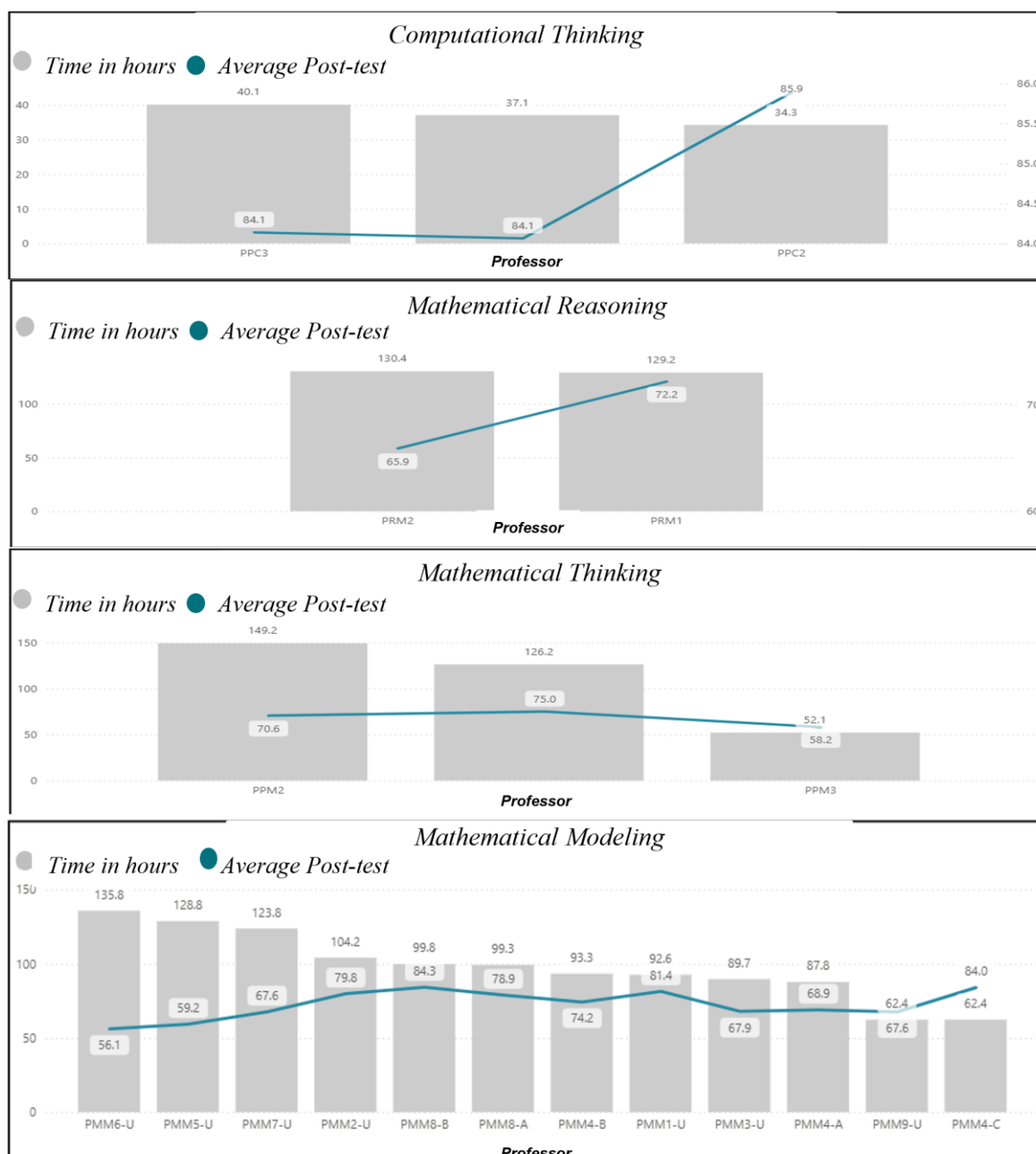
DiD Test Results (Blue line: experimental group. Green line: control group)



The descriptive analysis of data use could identify no clear correlation between the total time of use per group (in hours) and the average score obtained in the post-test. This suggests that the platform adapts to the individual needs of each student: while some required more time to achieve the expected learning, others achieved it in less time. For example, students in *Mathematical Reasoning* and *Computational Thinking* recorded fewer hours of use and obtained higher grades, while those in *Mathematical Modeling* and *Mathematical Thinking* spent more time and achieved better performance (see Figure 3).

Figure 3

Time spent using data and average Post-test vs. Professor on the adaptive platform



A survey was sent to students via email to assess the ALS. Participation was voluntary and, although the teachers made efforts to motivate them, the response rate was lower than expected. The results are presented in Table 3. In the survey, seven dimensions related to ALS were established, that is, to the entire process involved, from the content to the user experience. The scale employed continuous Likert responses. The interpretation of the scores fell into five categories: 0–19: strongly disagree; 20–39: disagree; 40–59: neutral; 60–79: agree; and 80–100: totally agree. This coding allowed for a finer analysis of the variability in participants' perceptions.

Table 3
Students' evaluation of the adaptive learning strategy

Dimensions	Computational Thinking n = 7 Mean (SD)	Mathematical Reasoning n = 22 Mean (SD)	Mathematical Thinking N = 41 Mean (SD)	Mathematical Modeling N = 193 Mean (SD)
Learning path	75 (22)	78.9 (16.9)	76.6 (18.7)	75.9 (22.9)
Learning process	73 (24)	78.2 (21.6)	75.2 (25.2)	73.5 (27)
Teaching Process	71 (26.6)	81.7 (19.7)	85.6 (19)	75.8 (27.4)
Content	75 (22)	81.2 (17.1)	76.8 (25)	78.7 (24.2)
Engagement	76 (23)	79.9 (19.3)	78.6 (23.9)	78 (24.5)
Usability	81 (20)	85.8 (16.3)	83 (19.4)	82.9 (22.4)
User Experience	77 (23)	83.3 (16.7)	78.2 (25.5)	79 (26)
Overall satisfaction with the platform	77.6 (24.1)	80 (18.2)	77.9 (23.9)	78 (25.3)

Note: SD = standard deviation.

According to the established scale, the results fell above the "Agree" level. No statistically significant differences were found between the courses, indicating that the evaluations were consistent among students in the different subjects. An additional finding, derived from a survey question on the use of resources external to the adaptive platform, was that 57% of the students reported not having used additional materials to recover previous knowledge.

The most frequent comments were categorized into the following strengths and areas of opportunity:

- Students valued the ability to access the platform at any time and progress at their own pace.
- They emphasized that the exercises were well organized, starting from the basics to the advanced.
- The videos were highlighted as an effective and easy-to-understand tool.
- The platform was seen as practical and understandable, with good material to review and reinforce knowledge.
- The platform facilitates remembering and reviewing forgotten topics efficiently.

- The use of videos was valued, but they suggested that they be more interactive and have step-by-step explanations.
- They proposed a better explanation of errors in the exercises to understand how to correct them.

On the other hand, with respect to teachers, only 11 responded to the survey since it was not mandatory (see Table 4).

Table 4

Evaluation of the Adaptive Learning Strategy by the professors

Dimensions	Evaluation
Usability	90.9
Use of Analytics	93.8
Teaching	95.3
User Experience	95.8
Pre-class study	91.6
Leveling	90.6
Overall satisfaction	95.9

As can be seen in Table 4, the teachers positively evaluated the Adaptive Learning Strategy, highlighting the dimensions of Teaching and User Experience. The most frequent comments were the following:

- It improves the teaching-learning process innovatively.
- It promotes greater student engagement and participation.
- It facilitates pre-class study and the leveling of basic knowledge, which implies a time saving in classes.
- It allows students to learn from home.
- It offers data analysis to make decisions on how to support lower-performing students and recognize higher-performing students.
- It increases students' security and confidence by corroborating answers and gaining solid knowledge.

DISCUSSION

The results of this study show significant differences in academic performance in favor of experimental groups that used ALS, compared to control groups with a traditional approach. Table 2 shows that, in the four training units, all groups with adaptive modules outperformed the control groups. Likewise, these experimental groups registered superior net learning gains, between 11.4 and 23.6 points, as illustrated in Figure 2. These findings suggest that ALS can effectively facilitate the recovery of previous knowledge when implemented comprehensively, incorporating a solid content design, a coherent didactic model, an active teaching role, and an appropriate technological platform. This coincides with the studies of Rincon-Flores et al. (2024), Shi and Liu (2025), and Simon and Zeng (2024), who highlighted that the impact of AL is enhanced when approached with appropriate teaching strategies and platforms that facilitate these strategies.

The AI technology platform is a vital component of the ALS, as it enables dynamic personalization of learning. By accurately diagnosing each student's prior knowledge, AI adapts learning paths to their specific needs in real-time. As Moskal et al. (2017) pointed out, adaptive platforms employ a data-driven—and, in some cases, non-linear—approach to instruction and leveling, dynamically adjusting to student interactions and achievement levels. This results in content delivery in the most appropriate sequence and time to support individual progress. Figure 3 illustrates how the platform adjusts to the particularities of each student: while some require more time to achieve the expected learning, others need less.

Regarding the assessment of the ALS, Table 3's analysis shows the respondents' positive perception in all the dimensions evaluated, which included pedagogical (content, teaching, and learning process) and technological (engagement, user experience, and usability) aspects, as well as general satisfaction. All exceeded the agreed-upon values, with no significant differences between teachers, indicating a homogeneous perception by the students. In addition, 57% of the students did not require external resources to recover previous knowledge, evidencing the sufficiency of the material. The qualitative comments highlighted the organization of the content, the usefulness of the exercises and videos, and the effectiveness of the platform to review and consolidate learning, in line with the findings of the University of Central Florida, which indicated greater student success, better learning outcomes, and more engagement (Dubay & Chen, 2023). However, suggestions to increase the interactivity of audiovisual resources and offer more detailed feedback coincided with du Plooy et al. (2024), who pointed out the potential of AL to improve both academic performance and student engagement.

Similarly, the teachers positively evaluated the ALS. (See Table 4, which presents the outstanding scores above 90/100 in all dimensions, especially in *Teaching* and *User Experience*.) The qualitative comments highlighted that the strategy innovatively improves the teaching-learning process and encourages more student engagement and participation. These findings, both in perceptions and in academic results, support the effectiveness of the ALS for leveling, using short modules integrated into the training units. They also confirmed that its success lies in a comprehensive design that combines well-structured content, relevant teaching strategies, an adequate teaching model, and appropriate technology, as pointed out by Rincon-Flores et al. (2024) and suggested by Shi and Liu (2025).

However, it is also essential to consider studies that show divergent results. For example, the research of Wang et al. (2023) and Eau et al. (2022) indicated that although the adaptive system favored academic improvement, no significant differences were detected between the groups in absolute terms. Similarly, the study by Rincon-Flores et al. (2024) found substantial improvements in learning gain, academic results, and other indicators in favor of adaptive groups compared to control groups; however, in the area related to the achievement of disciplinary competencies, no significant differences were found between the two groups, regardless of the course or stage evaluated.

Although the results of this research are promising, its limitations in external validity should be noted. The study was conducted at a private Mexican university, implying that the specific conditions of this environment could bias the findings. Therefore, generalizations to other educational contexts, especially those where resources and conditions differ significantly, should be ventured cautiously. Likewise, the interest of both teachers and students in using technological tools could be greater

in this type of institution, which usually offers access to advanced resources and specialized support. Thus, future research should delve into the factors that enhance or limit the impact of ALS in different educational contexts.

On the other hand, the transferability of this ALS with adaptive modules to virtual environments and distance education is fully viable thanks to its digital format, which allows 24/7 access to the content from anywhere with an internet connection. The technological platform employed also facilitates its integration with different learning management systems (LMS), which extends its applicability to various educational contexts, provided that the necessary technological infrastructure is in place. Note that although adaptive modules have also been applied in distance courses, they were not included in the present study due to the small sample size, which included only face-to-face groups.

CONCLUSIONS

The ALS proposed in this study is a promising alternative to close gaps in students' prior knowledge acquired in upper secondary (high school) education. Its implementation through short modules integrated into regular courses resulted in significant recovery without resorting to parallel or additional remedial courses. Under this approach, each student addresses only the content they need, when they require it, avoiding the unnecessary repetition of topics already mastered.

The results suggest that the articulation of the key components of the strategy—content structured in microlearning, a didactic model focused on self-study and self-regulation, an active teaching role in motivating and follow up, and a technological platform with AI to personalize learning paths and monitor progress—positively impacted academic performance, the recovery of previous knowledge, and the favorable perception of the participating students and teachers. In addition, the formal incorporation of the modules as part of the evaluation encouraged participation and compliance with the leveling process.

While the results are encouraging, areas for improvement include increasing the number of videos, providing more detailed feedback in exercises, and encouraging teacher use of the platform's analytics to identify learning patterns and personalize feedback. In some cases, the resistance of certain professors limited the full integration of the module with the dynamics of the course and the application of measurement instruments.

In conclusion, an integrally designed Adaptive Learning Strategy (ALS)—delivered as short modules embedded within regular courses and mediated by an AI platform—can effectively level first-year STEM students' prior knowledge, offering a viable alternative to traditional remediation models. Future research should examine ALS implementations across diverse contexts (public institutions, rural settings, and fully online modalities) and conduct longitudinal studies to assess sustained effects on academic performance, retention, and learner autonomy.

Finally, it is pertinent to promote educational policies that integrate adaptive leveling strategies into the curriculum and encourage continuous teacher training in the use of learning analytics to strengthen accompaniment and feedback.

Acknowledgments

The authors acknowledge the financial support of the Writing Lab, Institute for the Future of Education, Tecnológico de Monterrey, Mexico, in the production of this work.

REFERENCES

- Allela, M. A., Oganje, B. O., Junaid, M. I., & Charles, P. B. (2020). Effectiveness of multimodal microlearning for in-service teacher training. *Journal of Learning for Development*, 7(3), 384-398. <https://doi.org/10.56059/jl4d.v7i3.387>
- Cavanagh, T., Chen, B., Ait Maalem Lahcen, R., & Paradiso, J. R. (2020). Constructing a design framework and pedagogical approach for adaptive learning in higher education: A practitioner's perspective. *The International Review of Research in Open and Distributed Learning*, 21(1), 173-195. <https://www.irrodl.org/index.php/irrodl/article/view/4557>
- Creswell, J. W. (2015). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (5th ed.). Pearson.
- Dagunduro, A. O., Chikwe, C. F., Ajuwon, O. A., & Ediae, A. A. (2024). Adaptive learning models for diverse classrooms: Enhancing educational equity. *International Journal of Applied Research in Social Sciences*, 6(9), 2228-2240. <https://doi.org/10.51594/ijarss.v6i9.1588>
- du Plooy, E., Casteleijn, D., & Franzsen, D. (2024). Personalized adaptive learning in higher education: A scoping review of key characteristics and impact on academic performance and engagement. *Heliyon*, 10(21), e39630. <https://doi.org/10.1016/j.heliyon.2024.e39630>
- Dubey, A., Chen, B., & Lloyd, J. (2023, 4 May). Personalized adaptive learning in algebra-based introductory physics courses at the University of Central Florida. *WCET Frontiers*. <https://wcet.wiche.edu/frontiers/2023/05/04/personalized-adaptive-learning-in-algebra-based-intro-physics-courses-at-ucf/>
- Eau, G., Hoodin, D., & Musaddiq, T. (2022). Testing the effects of adaptive learning courseware on student performance: An experimental approach. *Southern Economic Journal*, 88(3), 1086-1118. <https://doi.org/10.1002/soej.12547>
- Felder, R. M., & Brent, R. (2005). Understanding student differences. *Journal of Engineering Education*, 94(1), 57-72. <https://doi.org/10.1002/j.2168-9830.2005.tb00829.x>
- Fernández-Morante, C., Cebreiro-López, B., Rodríguez-Malmierca, M.-J., & Casal-Otero, L. (2022). Adaptive learning supported by learning analytics for student teachers' personalized training during in-school practices. *Sustainability*, 14(1), 124. <https://doi.org/10.3390/su14010124>
- Filatova, E., Chen, Y.-W., & Li, H. (2023, June). Analysis of the COVID-19 impact on students' enrollment, performance, and retention. In *Proceedings of the 2023 ASEE Annual Conference & Exposition* (Baltimore, MD, United States). <https://doi.org/10.18260/1-2--42660>
- González-Beltrán, B. A., González-Brambila, S. B., Sánchez-Guerrero, L., Ardón-Pulido, I., & Figueroa-González, J. (2018). Estudio del impacto de un curso de nivelación en el desempeño de alumnos de ingeniería utilizando minería de datos educacional. *Research in Computing Science*, 147(5), 159-172. <https://doi.org/10.13053/rscs-147-5-12>
- Hattie, J. (2008). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge. <https://doi.org/10.4324/9780203887332>
- Hattie, J. (2017). *Visible Learning Plus: 252 influences and effect sizes related to*

- student achievement [PDF]. Visible Learning. <https://visible-learning.org/wp-content/uploads/2018/03/VLPLUS-252-Influences-Hattie-ranking-DEC-2017.pdf>
- Hernandez Cardenas, L. S., Castano, L., Cruz Guzman, C., & Nigenda Alvarez, J. P. (2022). Personalised learning model for academic leveling and improvement in higher education. *Australasian Journal of Educational Technology*, 38(2), 70-82. <https://doi.org/10.14742/ajet.7084>
- Howlin, C., & Lynch, D. (2014, October). Learning and academic analytics in the Realizeit system. In *Proceedings of the World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education (E-Learn 2014)* (New Orleans, LA, United States). Association for the Advancement of Computing in Education (AACE). <https://www.learntechlib.org/primary/p/148952/>
- Jing, Y., Zhao, L., Zhu, K., Wang, H., Wang, C., & Xia, Q. (2023). Research landscape of adaptive learning in education: A bibliometric study on research publications from 2000 to 2022. *Sustainability*, 15(4), 3115. <https://doi.org/10.3390/su15043115>
- Kara, N., & Sevim, N. (2013). Adaptive learning systems: Beyond teaching machines. *Contemporary Educational Technology*, 4(2), 108-120. <https://doi.org/10.30935/cedtech/6095>
- Klingbeil, N. W., Rattan, K. S., Raymer, M. L., Reynolds, D. B., Mercer, R., Kukreti, A., & Randolph, B. (2007, June). A national model for engineering mathematics education. In *Proceedings of the 2007 ASEE Annual Conference & Exposition* (Honolulu, HI, United States). <https://doi.org/10.18260/1-2--2104>
- Koester, B. P., Grom, G., & McKay, T. A. (2016). *Patterns of gendered performance difference in introductory STEM courses* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.1608.07565>
- Lasater, K., Smith, C., Pijanowski, J., & Brady, K. P. (2021). Redefining mentorship in an era of crisis: Responding to COVID-19 through compassionate relationships. *International Journal of Mentoring and Coaching in Education*, 10(2), 158-172. <https://doi.org/10.1108/IJMCE-11-2020-0078>
- Li, T., Kirk, C., & Oseguera, L. (2023). Heterogeneous academic achievement profiles of initially STEM-intending students over the college years. *Journal of College Student Development*, 64(6), 728-735. <https://doi.org/10.1353/csd.2023.a917027>
- Long, P., & Siemens, G. (2011). Penetrating the fog: Analytics in learning and education. *EDUCAUSE Review*, 46(5), 31-40. <https://er.educause.edu/articles/2011/9/penetrating-the-fog-analytics-in-learning-and-education>
- López del Puerto, C., Bellido, C. M., Suárez, O. M., Alfaro, M., & Jiménez, M. A. (2021, July). Championing Hispanic student success following natural disasters in Puerto Rico. In *Proceedings of the 2021 ASEE Annual Conference & Exposition* (evento virtual). <https://doi.org/10.18260/1-2--36790>
- McKenna, A., Agogino, A. M., & McMartin, F. (2000, October). What students say about learning physics, math, and engineering. In *Proceedings of the 30th Annual Frontiers in Education Conference* (Kansas City, MO, United States). IEEE. <https://doi.org/10.1109/FIE.2000.897580>
- Miao, F., Holmes, W., Huang, R., & Zhang, H. (2021). *Inteligencia artificial y educación: Guía para las personas a cargo de formular políticas*. UNESCO. <https://unesdoc.unesco.org/ark:/48223/pf00000379376>
- Moskal, P., Carter, D., & Johnson, D. (2017, 4 January). *7 things you should know about... adaptive learning*. EDUCAUSE Learning Initiative. <https://library.educause.edu/resources/2017/1/7-things-you-should-know-about-adaptive-learning>

- Ochukut, S. A., Oboko, R. O., Miriti, E., & Maina, E. (2023). Research trends in adaptive online learning: Systematic literature review (2011–2020). *Technology, Knowledge and Learning*, 28, 431-448. <https://doi.org/10.1007/s10758-022-09615-9>
- Olmos-López, O., Rincon-Flores, E. G., Mena, J., Román, O., & Camacho-López, E. (2023). Artificial intelligence as a way to improve educational practices. In M. Cebal-Loureda, E. G. Rincon-Flores, & G. Sanchez-Ante (Eds.), *What AI can do: Strengths and limitations of artificial intelligence* (pp. 135-151). Taylor & Francis. <https://doi.org/10.1201/b23345-10>
- Press Trust of India. (2022, 3 May). Haryana govt to distribute over 5 lakh free tablets to Class 10, 12 students. *The Indian Express*. <https://indianexpress.com/article/education/haryana-govt-to-distribute-over-5-lakh-free-tablets-to-class-10-12-students-from-may-5-7898959/>
- Rincon-Flores, E. G., Castano Sánchez, L., Guerrero Solís, S. L., Olmos López, O., Rodríguez Hernández, C. F., Castillo Lara, L. A., & Aldape Valdes, L. P. (2024). Improving the learning-teaching process through adaptive learning strategy. *Smart Learning Environments*, 11, 27. <https://doi.org/10.1186/s40561-024-00314-9>
- Rincon-Flores, E. G., Rivera Vázquez, N., Castano Sánchez, L., Guerrero Solís, S. L., Aldape Valdes, L. P., & Castillo Lara, L. A. (2023, April). Improving face-to-face learning using an adaptive technology. In *Proceedings of the 2023 IEEE Global Engineering Education Conference (EDUCON)* (Kuwait City, Kuwait). IEEE. <https://doi.org/10.1109/EDUCON54358.2023.10125276>
- Rockinson-Szapkiw, A., Herring Watson, J., Gishbaugher, J., & Wendt, J. L. (2021). A case for a virtual STEM peer-mentoring experience for racial and ethnic minority women mentees. *International Journal of Mentoring and Coaching in Education*, 10(3), 267-283. <https://doi.org/10.1108/IJMCE-08-2020-0053>
- Ross, P. M., Scanes, E., Poronnik, P., Coates, H., & Locke, W. (2022). Understanding STEM academics' responses and resilience to educational reform of academic roles in higher education. *International Journal of STEM Education*, 9(1), 11. <https://doi.org/10.1186/s40594-022-00327-1>
- Sarıyalçınkaya, A. D., Karal, H., Altınay, F., & Altınay, Z. (2021). Reflections on adaptive learning analytics: Adaptive learning analytics. In A. Azevedo, J. M. Azevedo, J. O. Uhomoibhi, & E. Ossianilsson (Eds.), *Advancing the power of learning analytics and big data in education* (pp. 61-84). IGI Global. <https://doi.org/10.4018/978-1-7998-7103-3.ch003>
- Schaal, N., Chan, S. E., & Saint Clair, J. K. (2021, July). Insights gleaned from the GAIN peer-mentoring program pilot. In *Proceedings of the 2021 ASEE Annual Conference & Exposition* (evento virtual). <https://peer.asee.org/insights-gleaned-from-the-gain-peer-mentoring-program-pilot>
- Seymour, E., & Hewitt, N. M. (1997). *Talking about leaving: why undergraduates leave the sciences*. Westview Press.
- Shi, P., & Liu, W. (2025). Adaptive learning oriented higher educational classroom teaching strategies. *Scientific Reports*, 15, 15661. <https://doi.org/10.1038/s41598-025-00536-y>
- Simon, P. D., & Zeng, L. M. (2024). Behind the scenes of adaptive learning: A scoping review of teachers' perspectives on the use of adaptive learning technologies. *Education Sciences*, 14(12), 1413. <https://doi.org/10.3390/educsci14121413>
- Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2), 257-285. https://doi.org/10.1207/s15516709cog1202_4
- Sweller, J., van Merriënboer, J. J. G., & Paas, F. G. W. C. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, 10(3), 251-296.

- <https://doi.org/10.1023/A:1022193728205>
- Taha, H., Hanim, F., Johar, M., & Shah, M. (2015). Peer mentoring module: The effect of an intervention of academic mentoring program towards motivation and self-esteem among foundation students in Malaysia. *Universal Journal of Psychology*, 3(3), 80-83. <https://doi.org/10.13189/ujp.2015.030304>
- Taylor, D. L., Yeung, M., & Bashet, A. Z. (2021). Personalized and adaptive learning. In J. Ryoo & K. Winkelmann (Eds.), *Innovative learning environments in STEM higher education: Opportunities, challenges, and looking forward* (pp. 17-34). Springer. https://doi.org/10.1007/978-3-030-58948-6_2
- UNESCO. (2019). *Consenso de Beijing sobre la inteligencia artificial y la educación* [Documento de programa]. <https://unesdoc.unesco.org/ark:/48223/pf00000368303>
- Vygotsky, L. S. (1978). *Mind in society: the development of higher psychological processes*. Harvard University Press.
- Wang, S., Christensen, C., Cui, W., Tong, R., Yarnall, L., Shear, L., & Feng, M. (2023). When adaptive learning is effective learning: Comparison of an adaptive learning system to teacher-led instruction. *Interactive Learning Environments*, 31(2), 793-803. <https://doi.org/10.1080/10494820.2020.1808794>
- Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry*, 17(2), 89-100. <https://doi.org/10.1111/j.1469-7610.1976.tb00381.x>
- Zimmerman, B. J. (1990). Self-regulated learning and academic achievement: An overview. *Educational Psychologist*, 25(1), 3-17. https://doi.org/10.1207/s15326985ep2501_2

Date of reception: 1 June 2025

Date of acceptance: 21 August 2025

Date of approval for layout: 7 October 2025

Date of publication in OnlineFirst: 27 October 2025

Date of publication: 1 January 2026

A language model-based recommender assessment system for short-answer questions in the intellectual property domain

Recomendador de evaluación para preguntas cortas utilizando modelos de lenguaje en propiedad intelectual

-----  -----

 David Bañeres Besora - *Universitat Oberta de Catalunya, UOC (Spain)*
 Ana-Elena Guerrero Roldán - *Universitat Autònoma de Barcelona, UAB (Spain)*
 M. Elena Rodríguez González - *Universitat Oberta de Catalunya, UOC (Spain)*

ABSTRACT

The use of Artificial Intelligence (AI) in education is growing rapidly, transforming the teaching-learning process as well as the assessment process. This work introduces SLASys, a tool to recommend the assessment of short-answer questions using semantic AI techniques. Unlike other works based on generative AI, SLASys uses the lightweight BERT language model, which better understands specific domain language concepts, improves computational efficiency, and reduces ethical and privacy concerns. SLASys implements semantic comparison and predictive classification models based on BERT. A mixed research methodology was followed, combining action research with a design and creation approach, to develop and refine SLASys over four editions of a master's-level course on patent examination within the intellectual property domain. SLASys has been integrated into Moodle, enabling its use by teachers without technical expertise, and has been tested by 120 students. The results demonstrate its effectiveness even with small datasets and limited participants within the described experience and according to existing literature. Additionally, it has been positively evaluated by both teachers and students. This work shows the feasibility of using AI in higher education, in both hybrid and online environments, offering a practical solution to improve assessment and feedback for short-answer questions in real learning contexts.

Keywords: assessment; feedback; Moodle; test; short-answer; artificial intelligence.

RESUMEN

El uso de la Inteligencia Artificial (IA) en educación está creciendo rápidamente, transformando el proceso de enseñanza-aprendizaje y también el proceso de evaluación. Este trabajo presenta SLASys, una herramienta para recomendar al profesorado la evaluación de preguntas cortas mediante técnicas de IA semántica, difiriendo de otros trabajos basados en IA generativa por el uso del modelo de lenguaje BERT que es más ligero, comprende mejor los conceptos en un contexto específico, mejora la eficiencia computacional y reduce los problemas éticos y de privacidad. SLASys implementa comparación semántica y modelos predictivos de clasificación de respuestas basados en BERT. Se ha seguido una metodología de investigación mixta, combinando investigación de acción con un enfoque de diseño y creación, para desarrollar y perfeccionar SLASys a lo largo de cuatro ediciones de un curso de nivel de máster sobre examen de patentes en el contexto de la propiedad intelectual. SLASys se ha integrado en Moodle, permitiendo su uso por parte de profesorado sin conocimientos técnicos, y ha sido probada por 120 estudiantes. Los resultados muestran su efectividad, tanto en el marco de la experiencia descrita como según la literatura existente, incluso con conjuntos de datos reducidos y un número limitado de participantes, y ha sido valorada positivamente por el profesorado y el estudiantado. Este trabajo contribuye a mostrar la viabilidad del uso de la IA en la educación superior, tanto en entornos híbridos como en línea, ofreciendo una solución para mejorar la evaluación y el *feedback* en preguntas cortas en contextos reales de aprendizaje.

Palabras clave: evaluación; *feedback*; Moodle; test; preguntas cortas; inteligencia artificial.

INTRODUCTION

Nowadays, educational institutions (including traditional and online) generally adopt information and communication technologies to support their teachers and students. Automating the grading task is one trend followed in many Learning Management Systems (LMS) for efficiently assessing activities and providing timely feedback for formative purposes (György & Vajda, 2007). However, many LMS limit the type of questions to be automatically graded to choice-based or simple text due to the complexity of assessing short-answer questions and essay-type questions.

Although Automatic Short-Answer Grading (ASAG) has been extensively explored in the literature (Burrows et al., 2015), contributions were limited to exploring their accuracy on public datasets without considering their effectiveness in real educational settings. Additionally, the resurgence of Artificial Intelligence (AI) due to the accessibility to computational resources and specific tools for using AI techniques has evolved classic LMS. AI algorithms can power LMS to support the learning process. Recommendations, immediate feedback, or ASAG, are examples of tasks that AI can enhance.

This research aims to provide a starting point for a smart learning system capable of recommending assessment and feedback for short-answer questions using AI techniques. This paper shows how the system is used in a hybrid educational setting and presents the obtained findings. Despite Large Language Models (LLM) and derived generative tools, such as ChatGPT (OpenAI, 2024) currently showing their broad potential in any academic task, they still have limitations for specific tasks or domains like Intellectual Property (IP). Therefore, this work proposes an alternative solution, using a lightweight BERT Language Model (LM) that understands the answers' meaning. Instead of delegating the assessment decision for a short-answer question to a Generative AI tool (GenAI) by defining the assessment criteria as a complex prompt, this work uses BERT, which understands the meaning of any correct and incorrect answer, and can decide whether a student's answer is correct. This latter model is currently more accurate for assessment purposes. It is smaller, contributing to efficient computational and sustainable use; and it can be handled in private instances (i.e., no enterprise solutions are needed), reducing ethical and privacy issues. Thus, this work integrated this AI solution into a Moodle LMS to make it easy and usable for teachers to provide a meaningful assessment to students.

The action research methodology combined with a design and creation approach is used to specify the Smart Learning Automated System (SLASys), explicitly suited for an IP training course, to define a learning system to help students succeed in their learning process using AI techniques for IP training. The study has been carried out in the European Patent Office (EPO) educational area.

This paper is structured as follows. First, it presents the theoretical background of language models and their educational uses, and focusing on ASAG. Then, it details the research design, including methodology, research questions, the SLASys tool, and data analysis. The results section highlights key findings, and the paper concludes with a discussion, limitations, conclusions, and future research.

LITERATURE REVIEW

Large Language Models, BERT, and their application in education

LLMs have emerged as a transformative technology in recent years, demonstrating remarkable capabilities across Natural Language Processing (NLP) tasks. Their rapid progress has led to significant advancements in machine translation, text summarization, and even code generation (Husein et al., 2025; Pang et al., 2025; Zhang et al., 2025). Despite their impressive performance, LLMs face limitations and challenges in real educational settings, such as misinterpretations in specific domain languages (Arefeen et al., 2024) or their interpretability and explainability (Zhao et al., 2024). Recent architectural innovations have improved their performance, reaching the state-of-the-art GenAI tools, such as GPT-4o (OpenAI, 2024). Such tools demonstrate impressive learning capabilities, applicability to limited reasoning tasks, and the ability to generalize across diverse task domains.

BERT (Bidirectional Encoder Representations from Transformers) is a state-of-the-art NLP model developed by Google in 2018 (Devlin et al., 2019). BERT is a transformer-based language model pre-trained on a large collection of textual documents. BERT was considered an LLM but is currently considered an LM compared to its GenAI successors. Its architecture relies on text representation. A sentence is encoded in a special way (i.e., *embeddings*) that stores its semantic meaning, enabling the straightforward application in NLP tasks. Its continued development and application in multiple knowledge domains demonstrate versatility and potential for further advancements in AI and language understanding. BERT has been used for document classification (Adhikari et al., 2019), grammatical category recognition (Souza et al., 2019), logical entailment between pairs of sentences (Wang et al., 2018), or text summarization (Zhang, Cai et al., 2019).

Both BERT and GenAI are being used in education with similar challenges. Scalability and integration with existing infrastructures require significant investment and technical expertise (Xu & Zhu, 2023). Additionally, principles concerning ethics, security, and alignment with core educational objectives must be considered (García-Peñalvo et al., 2024) and developing AI literacy among teachers is essential for fostering critical use (Petridou & Lao, 2024). Moreover, although they seem applicable in any task, GenAI tools are mostly applicable to text generation and analysis, such as concept definition, exercise generation, or improving writing style; while BERT is valid for text classification, sentiment analysis, understanding students' doubts, or providing detailed explanations (Qiu & Jin, 2024).

Assessment is also a relevant task for both models. On the one hand, GenAI tools have started to be explored in assessment as “black boxes” without knowing how they internally work. They have been used to produce feedback, but also for automatic assessment. Automated assessment benefits some domains, such as language learning (Escalante et al., 2023), writing skills (Banihashem et al., 2024), or multimedia designs (Almasre, 2024). On the other hand, BERT can be used on specific assessment tasks. Teachers can generate questions from existing texts, which can be used for tests and exams (Nguyen et al., 2022). BERT models can assess student essays' coherence, relevance, and grammar, helping students improve their writing skills (Wang et al., 2022). BERT-based applications were also designed to understand idiomatic expressions and cultural nuances (Bahdanau et al., 2015), and specific recommenders

can help students practice speaking and comprehending new languages (Zhang, Zhang et al., 2019).

Although automated assessment seems to be one of the future features of GenAI in education, currently, it is not recommended by institutional or even governmental policies (European Commission, 2024; Dai et al., 2025; González Fernández et al., 2025), because GenAI tools are error-prone due to hallucinations (Jia et al., 2024) and their utilization might raise ethical issues. Additionally, GenAI tools may fail on domains with specific language. For instance, some questions related to specific concepts in patent examination in the IP domain, such as Novelty and Clarity, can sometimes result in a GenAI tool interpreting the question as whether the answer is new and clear in natural language. However, such IP-specialized concepts refer to how novel and clear an invention is described in a patent application (or *claim*). GenAI tools could analyze the answer by improving previous context (i.e., *contextualization*), but they sometimes mix patent guidelines from different sources with outdated information. Thus, this requires adding specific documents to the model (i.e., *fine-tuning*) to improve the quality of a proprietary GenAI model, which implies some privacy concerns and discussion about whether it is a cost-effective solution. BERT-based application addresses the assessment task by understanding the answers provided to a question. BERT could be used by performing *semantic comparison* to assess how close the meaning of two sentences are (i.e., the student's answer compared to the teacher's correct answer) or train an *answer classification* AI model to predict whether a sentence meaning is close to a correct or incorrect set of sentences (i.e., decide whether the student's answer is correct or incorrect depending on a set of correct and incorrect answers). Such approaches avoid the variability drawback that GenAI tools currently exhibit.

Automated short-answer questions grading and feedback provision

To assess the acquisition of knowledge, skills, and competences, feedback becomes one of the essential components in the teaching-learning process (Evans, 2013; Hattie & Timperley, 2007). Feedback is a key element to engage students and enhance their learning (Winstone et al., 2017). It aims to provide specific information related to a learning activity to fill the gap between the desired and the actual understanding (Sadler, 1989). Additionally, feedback is crucial in online settings to guarantee a meaningful and timely learning process (Nicol & Macfarlane-Dick, 2006). Elaborating and delivering valuable feedback requires a considerable teacher's effort, which limits its personalization when there are time constraints or a big number of enrolled students (Dhananjaya et al., 2024). However, personalized feedback motivates students to a higher level, better regulates their behavior, contributes to better knowledge acquisition, and enhances engagement with learning materials (Kim, 2023; Wang & Lehman, 2021).

Automating feedback generation has been sought as a tentative solution to support teachers. Although different terminology has been used for denoting such tools in the literature, e.g., intelligent tutoring system, teaching platform, automated feedback system, online feedback tools; all of them focus on enhancing teachers' efficiency while helping students during their learning process (Xie & Li, 2018). Previous works have presented applicable techniques, such as feedback generation for essays (Akçapinar, 2015), for programming exercises (Messer et al., 2024), or collaborative tasks (Zheng et al., 2023).

Automated test grading has also been extensively researched due to its benefits for providing automatic but also formative assessment (Burrows et al., 2015), and the positive impact on learning due to feedback provision (Gaona et al., 2018; Rezaei, 2015). Tests have been mostly simplified using choice-based questions, and short- and essay-based questions were usually evaded since they required higher teacher effort and intervention during the assessment phase, or alternative methods were used, such as peer review (Huisman et al., 2017). However, there is previous work related to ASAG. Answer comparison against one sentence (Siddiqi & Harrison, 2008), against a set of sentences with similar vocabulary (Klein et al., 2011), or against primary keywords (Saha et al., 2018) have been used to detect correct answers. Although these techniques have evidenced good results, they have some deficiencies since semantics is not considered (i.e., one pattern, similar vocabulary, or a set of keywords may not gather all possible correct answers). With GenAI tools, authors investigated their applicability to the ASAG (Aggarwal et al., 2025; Grévisse, 2024). Although its applicability for essay grading seems promising (Senthilnathan et al., 2025), the ASAG results are inconclusive for supporting students' assessment. Hallucinations, decontextualization, outdated information, and limited critical thinking still affect their utilization (De La Cruz Martínez et al., 2024).

Since NLP and semantic LM models are still state-of-the-art techniques for ASAG, this work used BERT because of its capabilities, potential accuracy, and better fit for real educational settings. Two types of models are explored in the literature. Some models seek to predict whether the student's answer is correct (Camus & Filighera, 2020; Liu et al., 2019; Lun et al., 2020; Padó et al., 2024; Schneider et al., 2023; Sung et al., 2019; Wang et al., 2019), while others go further by predicting the grade, demonstrating their potential on public datasets (Baral et al., 2021; del Gobbo et al., 2023; Gaddipati et al., 2020; Metzler et al., 2024; Soulimani et al., 2024). Still, as far as the authors of this research know, no work has tested such tools in real educational settings due to acceptability issues or human intervention needs for their utilization (Hustad & Arntzen, 2013; Xavier et al., 2025). Therefore, this work proposes SLASys as a recommender tool for ASAG, which has been tested in a hybrid educational environment during four editions of a fundamental patent examination course over a year, with 120 students, to gain insights about their utilization. Thus, this research aims to answer the following research questions:

- RQ1. How accurate is answer classification compared to semantic comparison for ASAG recommendation?
- RQ2. How accurate is SLASys in a real educational setting?
- RQ3. Can SLASys be used without previous students' answers?
- RQ4. What is the opinion of the students and teachers?

METHODOLOGY

Research design

The SLASys system employs a mixed research methodology that combines action research (Oates, 2006) with a design and creation approach (Kuechler & Vaishnavi, 2012). Action research is appropriate because the system is developed and tested within real educational settings, aiming to address practical challenges in teaching and learning. This methodology emphasizes solving real-world problems through a collaborative, iterative cycle of planning, acting, and reflecting, while collecting data to evaluate outcomes. The design and creation approach complements this by focusing on the development of innovative technological artifacts. It follows a five-step, iterative problem-solving process (Kuechler & Vaishnavi, 2012) to guide the creation and refinement of the system.

This paper focuses on the first two iterations of developing the recommender for ASAG. The first one involved the creation of the first artifact and evaluating the technical solution, while the second evaluated its integration into the teaching-learning process. This last iteration involved testing the artifact in a fundamental course in patent examination for the IP domain. This course, with a duration of 6 weeks, is part of the courses oriented towards the students' training for patent examiners, which comprises six courses with a 2-year duration. There are specific requirements to be eligible for this training. Students should have a master's degree in physics, chemistry, engineering, or natural sciences since knowledge of specific technical fields is required to examine patent application proposals.

The course objective is to prepare the students for the patent examination by introducing the fundamental concepts, the European Patent Convention (EPC) guidelines, systems available for the task, and exercises related to patent examination. The learning methodology combines a flipped classroom strategy (Bergmann & Sams, 2012) with just-in-time teaching (Novak, 2012). Self-learning activities and reading learning resources are performed individually offline. There are also synchronous online sessions with all the students and the teacher, where offline activities are discussed, and practical exercises, questions, and group activities are performed.

Teachers reported issues applying just-in-time teaching since they did not have enough time to assess the students' short-answer activities and provide meaningful feedback before the online sessions. Thus, SLASys answers these needs by providing a solution to efficiently assess and give feedback to each student.

Figure 1
Research design

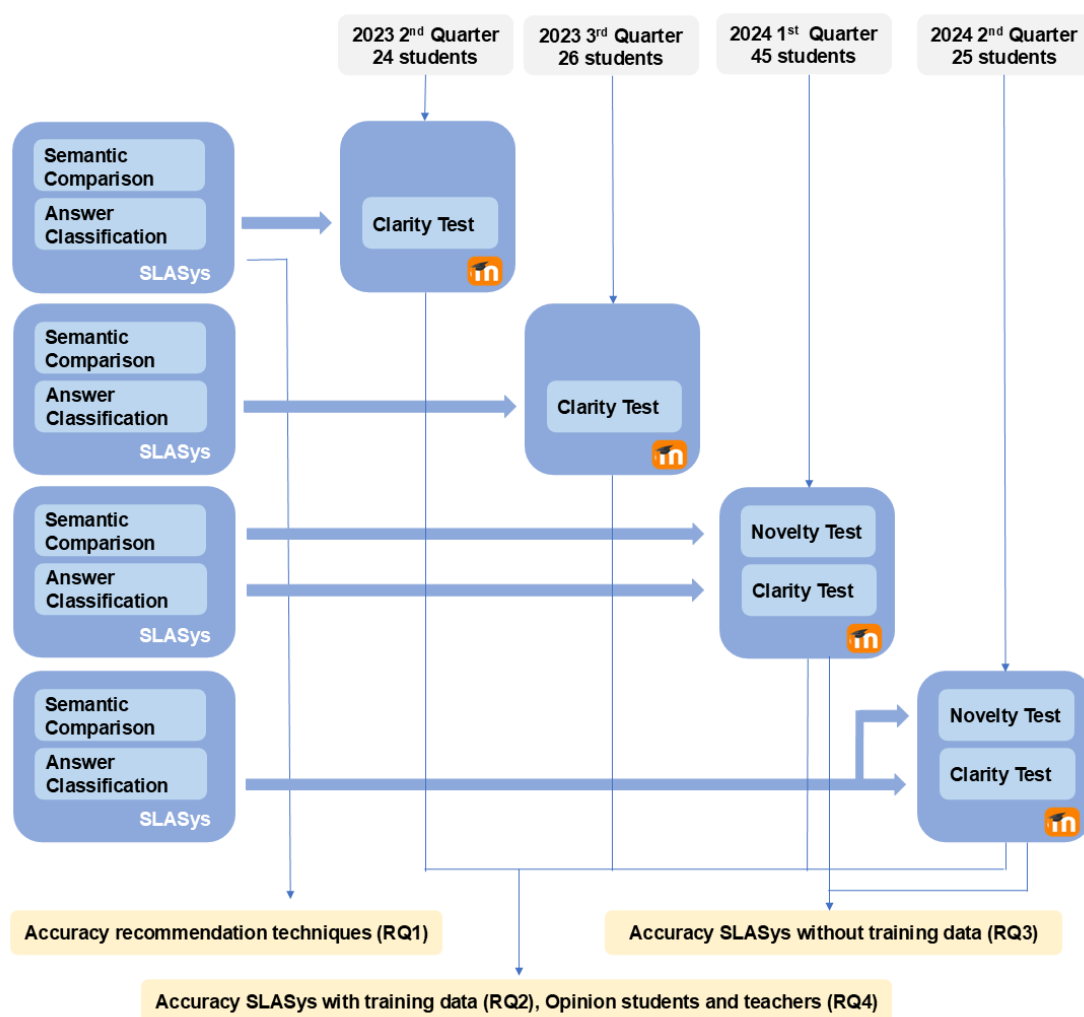


Figure 1 shows the research design. First, the designed approaches are compared to see their accuracy for answering RQ1. Second, during four-course editions, the answer classification approach was tested in a test related to the patent Clarity concept for answering RQ2. In total, 120 students interacted with the tests. Third, a new test concerning the Novelty concept was designed in the last two editions without any training data (i.e., previous students' answers). In this case, the two designed approaches were used to assess the performance of the recommender to answer RQ3. In this case, 70 students answered this second test. Additionally, the students' opinions and the teachers' experiences were collected across the editions to answer RQ4.

The testing with real students was performed online synchronously with a custom Moodle LMS where SLASys was used to recommend the assessment results. The Clarity and Novelty tests comprised 8 and 11 questions, respectively, and students had one hour to answer them. Note that they aim to improve students' knowledge but not students' grades. Students are assessed twice during the complete formation with onsite exams.

SLASys: A Recommender for ASAG

SLASys was designed to find whether a student's short answer is correct, depending on the available information related to the question. Figure 2 shows the two techniques used to detect whether an answer is correct. The question is about the concept of Novelty in the IP domain, asking why the claim is not novel. The comparison is performed by using BERT *embeddings*. Although an embedding is a complex mathematical matrix, it could be seen as an encoding of the sentence's meaning by a set of symbols that define their meaning.

Figure 2

Question example about Novelty

Why is this claim not novel?

A (mobile) phone with dialling means, a microphone and a loudspeaker, characterised in that the dialling means are made from aluminium ...

Golden answer

The claim is unclear because of the bracketed expression (mobile). It is unclear whether the feature in brackets should be limiting or not. See Guidelines F-IV, 4.19.

(★, ☎, ..., ☀)

	Embedding	Cosine similarity	Probability of being classified correct
Answer 1 (correct) Feature in brackets which is not a reference sign it is not clear whether this is optional or not (F-IV, 4.19)	(★, ☎, ..., ☀)	32.2%	99.9%
Answer 2 (incorrect) Expression in brackets are not allowed (I'm not sure if it's falling under the exemption in this technical field).	($\sqrt{x}$, ☎, ..., ⚙)	23.3%	0%

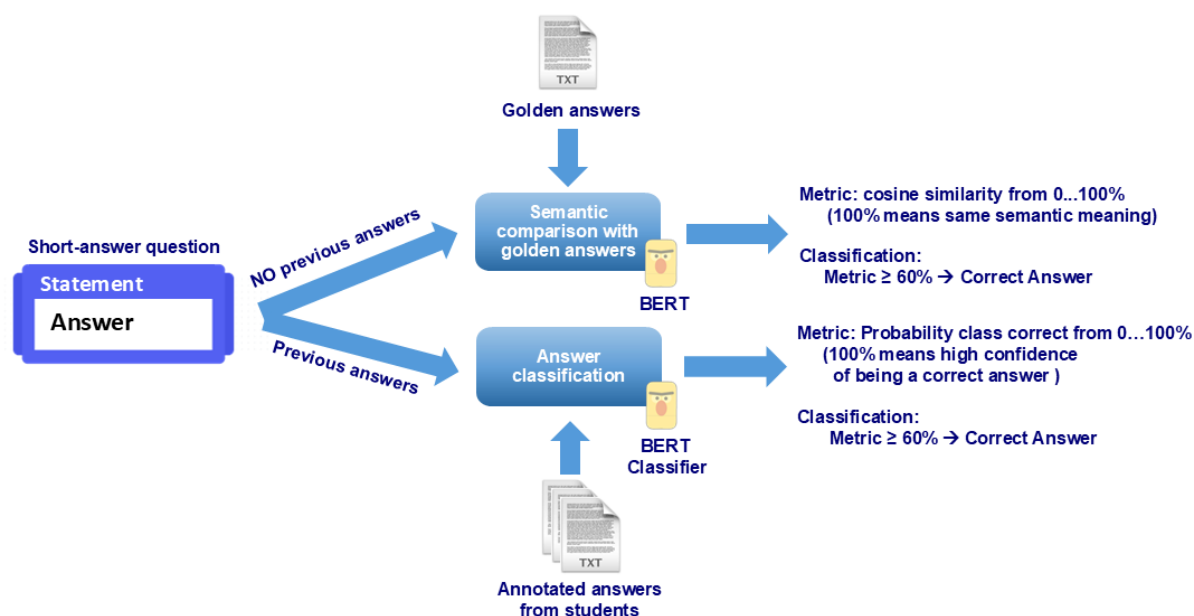
The first technique uses the correct answer embedding provided by teachers, denoted as the *golden answer*, to compare with the answers provided by students. The *cosine similarity* metric is used to evaluate how similar the student's sentence is to the golden answer (i.e., how many symbols are equal in both embeddings). The comparison output (i.e., *metric score*) is a similarity probability, where 100% means the same meaning as the golden answer. Figure 2 shows that this method fails for the correct answer, since there are different correct answers, and the golden answer does not cover all of them.

However, this process could be enhanced by comparing students' answers with a set of correct and incorrect answers (i.e., extracted from real students of previous course editions after being assessed by teachers), denoted as *annotated answers*. Each answer can be annotated, whether it is correct or incorrect. Although the comparison can be done by cosine similarity for each annotated answer, there is a more efficient technique that involves training an AI classifier and fine-tuning based on the embeddings of the annotated information. Fine-tuning specializes the classifier to focus on the specific annotated answers gathered for a question. The model's output is

the probability of the embedding being classified as a correct answer, where 100% means the model is highly confident that the student's answer is correct.

Since a short-answer question could be created from scratch or reused from other learning contexts, the system considers these two possible scenarios (Figure 3). Semantic comparison is used when the question is created and no previous student data are available. The comparison uses cosine similarity with a set of correct answers provided by teachers. The answer classification trains an AI classifier when previous answers are available, providing a prediction of being correct. Since both metrics are percentages, a threshold is defined to decide whether a student's submission is a correct answer. An optimization problem solves the quality threshold identification during performance analysis since the objective is to maximize the threshold while correct classification does not worsen significantly. After some experiments and teachers' validation as domain experts, it was concluded that a threshold greater than 60% mostly identified correct answers.

Figure 3
SLASys technical design

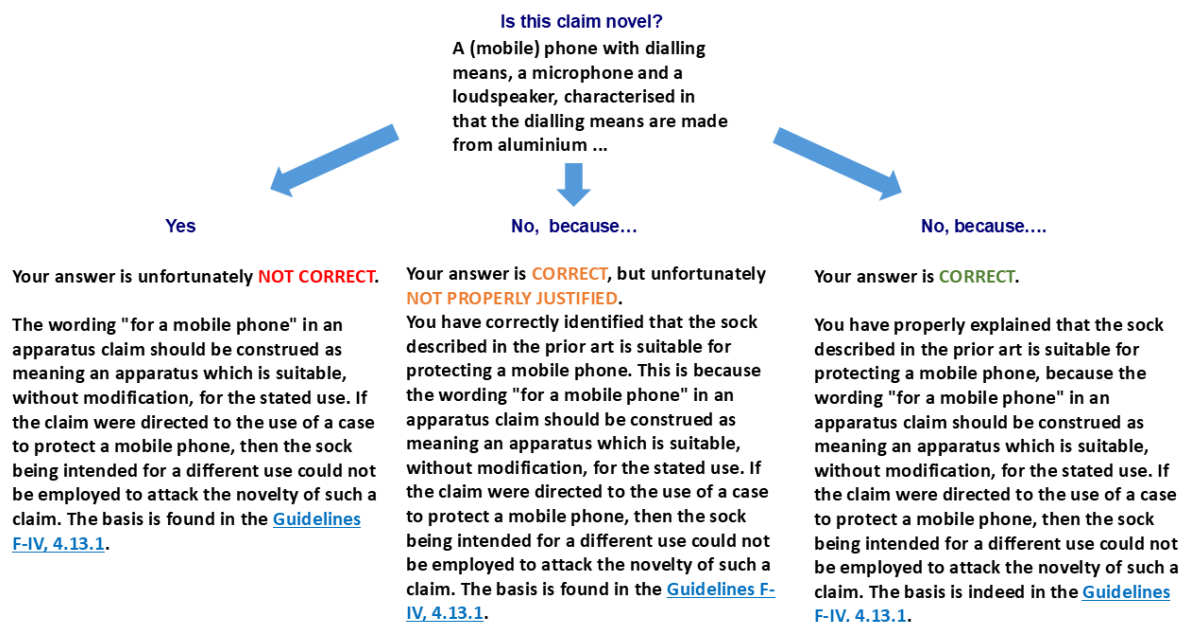


SLASys Moodle Integration for assessment and feedback provision

Since SLASys offers an automatic process to recommend assessment, the assessment process could be enhanced by providing automated feedback. Figure 4 shows an example of the feedback provided on the question associated with the claim of Figure 2. Two types of questions were defined: a simple short answer or a choice question with a justification. This second type is relevant for teachers in patent examination since questions usually state some claim related to a patent definition, and students should examine whether the claim satisfies EPC guidelines, such as novelty or clarity. Figure 4 shows the second type with the feedback provided and the link to the EPC guidelines depending on the assessment result. The feedback is predefined for each option and configured on the question design, and it provides

meaningful information for students to understand the assessment result and improve their learning.

Figure 4
Feedback example about Novelty question



The SLASys system incorporates an API with different operations that can be integrated into any LMS. This work used Moodle integration by developing a custom plugin to access SLASys capabilities. Complexity issues were hidden, providing an easy-to-use interface for teachers' management. A new question type has been integrated into Moodle by adapting the question configuration and the assessment interface for teachers, as well as the interface to submit the answers and review the assessment results for students. The design minimizes teachers' workload with an extremely small learning curve.

A new question can be added to a test by providing mandatory information: type of question, the statement, and the set of golden answers. Optional annotated answers from previous students can be added or retrieved from previous course editions. When this information is not provided, SLASys uses the semantic comparison approach. Otherwise, the corresponding answer AI classifier is automatically trained without technical intervention (Figure 3).

The most powerful part comes from the assessment and feedback configuration. The question has embedded a rubric (i.e., not available for short-answer questions in Moodle), unifying the assessment criteria for teachers and delivering the corresponding feedback depending on assessment results.

Figure 5.a) illustrates the student's interface for submitting an answer to a question with the choice option enabled. The figure also shows each level's rubric and the associated grade. The submission is sent to SLASys, and the recommendation is returned to Moodle when finished. The recommendation is shown in the assessment interface (Figure 6). The teacher sees the golden answers, the SLASys recommendation, the explainable information from the answer classification model,

and the metric score. Such explainable information shows in colors which words impact positive or negative classification decisions. Although this information sometimes includes irrelevant words, teachers can still identify incorrect words that affect how students' answers are classified. The interface also proposes the rubric grading that is automatically selected depending on the recommendation. The grading and the feedback are assigned depending on the rubric selection. The teacher must review this information and select the correct rubric level.

After being assessed, the test results can be reviewed in the students' interface (Figure 5.b). Each student sees the grading and the feedback based on the selected rubric option, validated by the teacher.

Figure 5
Moodle question interface for students

a) Attempt interface

In a telephone apparatus with an automatic dialler, a dial tone detector and a feature controller, the dial tone detector comprising ...

Statement

Select answer

☒ Compliant with the requirements of Art. 84 EPC

☐ Not compliant with the requirements of Art. 84 EPC, because:

Choice answer

Justification answer

Student's answer

Rubric (review, no manual comment)

Response	The option answer is incorrect	The option answer is correct, but the justification answer is incorrect	The option and justification answers are correct
	0 points	0 points	10 points

b) Feedback interface

In a telephone apparatus with an automatic dialler, a dial tone detector and a feature controller, the dial tone detector comprising ...

Select answer

☐ Compliant with the requirements of Art. 84 EPC

☒ Not compliant with the requirements of Art. 84 EPC, because: ✓

Assessment choice answer

Justification answer

The claim does not clearly defines the subject-matter of the invention, as the term "in" makes it unclear whether the entire telephone apparatus is included in the scope of protection.

Comment: Well done. The provided answer is correct.

Feedback

The correct solution is The expression "in" makes it unclear whether protection is sought for the dial tone detector per se or a telephone apparatus (see Guidelines F-IV 4.15).

Rubric (review)

Response	The option answer is incorrect	The option answer is correct, but the justification answer is incorrect	The option and justification answers are correct
	0 points	0 points	10 points

Rubric assessment

Grading

Figure 6
Moodle assessment interface for teachers

Model answer
The expression "in" makes it unclear whether protection is sought for the dial tone detector per se or a telephone apparatus (see Guidelines F-IV 4.15).

Golden answer

Assessment Recommendation
Well done. The provided answer is correct.

Legend: ■ Negative □ Neutral ■ Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
1	1 (1.00)	The claim does not clearly defines the subject-matter of the invention, as the term "in" makes it unclear whether the entire telephone apparatus is included in the scope of protection.	-0.25	[CLS] the claim does not clearly defines the subject-matter of the invention, as the term " in " makes it unclear whether the entire telephone apparatus is included in the scope of protection . [SEP]

Rubric (grading)

Response	Points
☐ The option answer is incorrect	0 points
☐ The option answer is correct, but the justification answer is incorrect	0 points
☒ The option and justification answers are correct	10 points

Comment

Well done. The provided answer is correct.

The correct solution is The expression "in" makes it unclear whether protection is sought for the dial tone detector per se or a telephone apparatus (see Guidelines F-IV 4.15).

Mark **Grade** **score** **Metric**
10.00 out of 10.00 1.000

Data analysis

Data from two sources were used to answer the RQ. First, information from SLASys has been used to analyze the number of correct assessments on answers used for training (i.e., *performance*) (RQ1). Next, using the Moodle plugin, the recommendations provided by SLASys and the assessment performed by the teachers were retrieved to analyze the correct recommendations during the test with students (RQ2 and RQ3). The performance during the training phase and test with students can be computed with the following metrics:

$$TNR = \frac{TN}{TN + FP} \quad ACC = \frac{TP + TN}{TP + FP + TN + FN}$$

$$TPR = \frac{TP}{TP + FN} \quad F_{1.5} = \frac{(1 + 1.5^2)TP}{(1 + 1.5^2)TP + 1.5^2FN + FP}$$

Where TP denotes the number of correct answers correctly identified, TN the number of incorrect answers correctly identified, FP the number of incorrect answers not correctly identified, and FN the number of correct answers not correctly identified. These four metrics allow to compute the accuracy when detecting correct and incorrect answers (ACC), the accuracy when detecting correct answers (True Positive Rate—TPR), the accuracy when distinguishing incorrect answers (True Negative Rate—TNR), and the percentage of correct answers penalizing the incorrect identifications (F score - F_{1.5}). These metrics are used to evaluate the performance between the semantic comparison and answer classification approaches (RQ1) and whether SLASys recommends the expected teacher's assessment in the real educational setting (RQ2 and RQ3).

Finally, the average access time to the feedback interface and students' ratings of the results were gathered. The Moodle plugin added a Likert scale to collect students' appraisals within the feedback interface. The scale ranged from 1 to 5 (5 means the feedback was highly useful). Regarding the teachers' experience, only their feedback on the system usage was collected because only three teachers were involved (RQ4).

RESULTS

RQ1: How accurate is answer classification compared to semantic comparison for ASAG recommendation?

Knowing SLASys' accuracy before the first pilot with students is crucial before analyzing the system's applicability in a real educational setting. The results are detailed in Table 1 where the eight questions from the Clarity test were analyzed.

Table 1
Performance metrics on Clarity training test

Exercise	Correct	Incorrect	Semantic comparison				Answer classification			
	Train / Test	Train / Test	ACC	TPR	TNR	F _{1.5}	ACC	TPR	TNR	F _{1.5}
Question 1	36 / 8	107 / 22	37 %	100 %	14 %	58 %	87 %	88 %	86 %	81 %
Question 2	47 / 10	7 / 2	67 %	70 %	50 %	75 %	100 %	100 %	100 %	100 %
Question 3	3 / 2	46 / 4	42 %	100 %	30 %	48 %	83 %	0 %	100 %	0 %
Question 4	34 / 7	17 / 20	27 %	0 %	75 %	0 %	64 %	86 %	25 %	79 %
Question 5	67 / 14	100 / 20	53 %	100 %	20 %	74 %	82 %	71 %	90 %	75 %
Question 6	82 / 17	36 / 8	84 %	94 %	63 %	91 %	88 %	100 %	63 %	95 %
Question 7	33 / 7	33 / 7	57 %	86 %	29 %	73 %	79 %	71 %	86 %	75 %
Question 8	68 / 14	19 / 4	83 %	93 %	50 %	91 %	83 %	100 %	25 %	94 %
Average			56 %	80 %	41 %	64 %	83 %	77 %	72 %	75 %

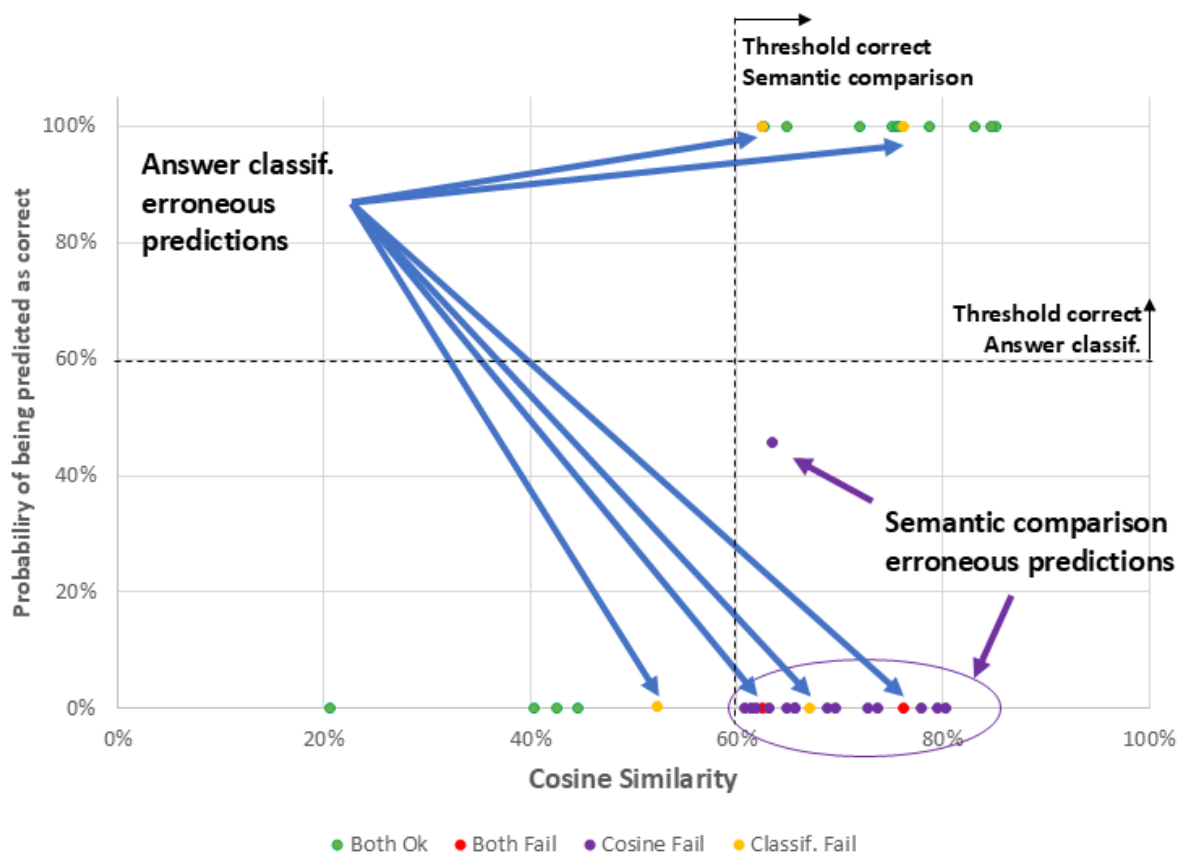
The table summarizes the number of annotated answers used for fine-tuning the AI classifiers (i.e., *training dataset*), testing their accuracy (i.e., *testing dataset*) and the metrics results. AI classifiers were fine-tuned by taking 80% of the annotated answers (i.e., there is a classifier for each question). The remaining 20% was used for testing and obtaining the results of the table. The split process considers correct and

incorrect answers in both datasets. Note that there is no fine-tuning phase on the semantic comparison approach. In such cases, three random correct answers were compared with the testing answers.

Answer classification outperforms semantic comparison because annotated answers provide examples of correct and incorrect answers. Semantic comparison commonly fails when there are multiple ways to answer a question, and they are semantically far from the golden answers. Focusing on metrics to detect correct and incorrect answers (i.e., TPR and TNR), quality results highly depend on the number of answers provided by the teachers for training and the balance between correct and incorrect. Classifiers with few answers have difficulties recommending. It can be observed that semantic comparison can detect more correct answers on average (i.e., 80% compared to 77%). However, this increment is caused by many incorrect answers being incorrectly identified as correct (i.e., 41% compared to 72%).

Individual predictions were analyzed to better understand the techniques' behavior. Figure 7 depicts the analysis on the testing dataset for Question 5. The respective metrics, cosine similarity and probability of being predicted as correct, are plotted for semantic comparison and answer classification, respectively. Different colors were used to identify which predictions are correct using both metrics (green), which ones are incorrectly predicted by semantic comparison (purple), which ones are incorrectly predicted by answer classification (yellow), and which ones cannot be predicted by both methods (red).

Figure 7
Accuracy comparison for Question 5



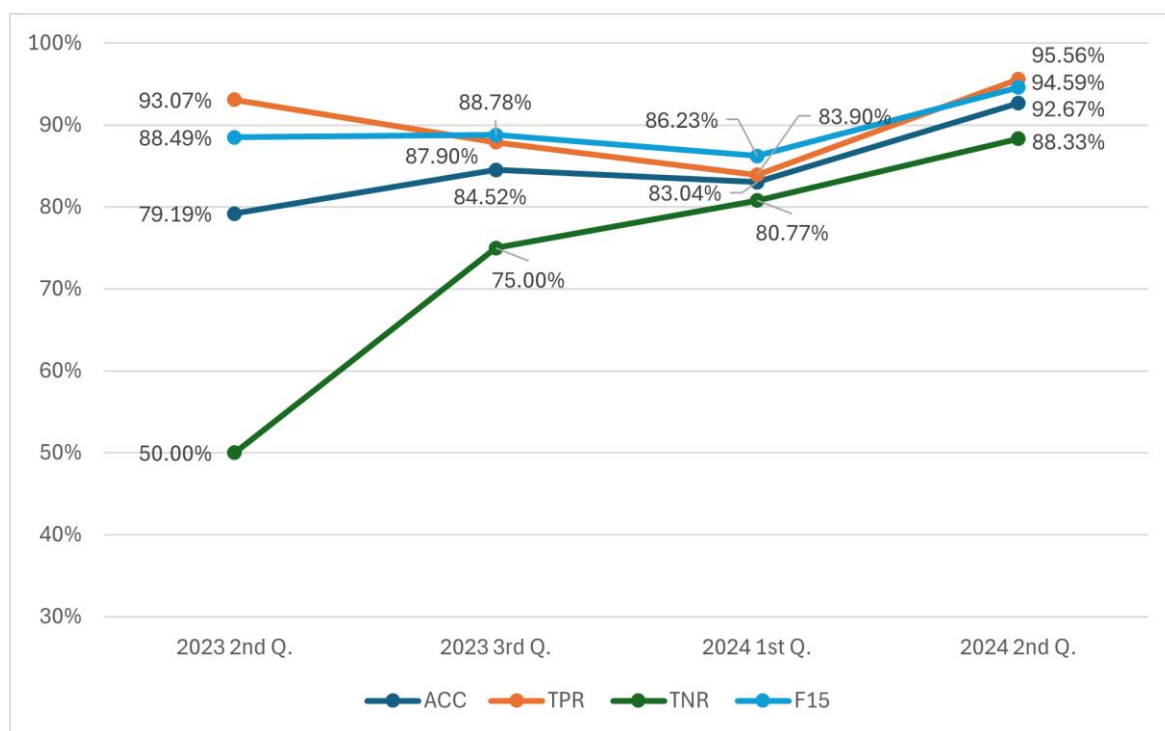
Most errors appear in semantic comparison. Many students' answers are considered correct (i.e., semantically similar to the correct answer provided by the teachers), ranging from 60% to 85%, but are incorrect. Thus, slight changes in the sentence's meaning can generate incorrect predictions. The answer classification also has some erroneous predictions in detecting correct answers, but in fewer cases. Additionally, cases where both methods fail can be observed. Teachers realized that some of these answers were incorrectly predicted because they were not covered in the answers used for the fine-tuning and golden answers.

RQ2: How accurate is SLASys in a real educational setting?

Figure 7 shows some erroneous cases. Adding such answers to the AI classifiers could enhance the identification in new course editions. This is the fundamental idea for enhancing classifiers: classifiers are retrained, including the assessment results performed in the previous editions. Thus, they are refined with new annotated answers that improve identification. RQ2 analyzes whether SLASys answer classification recommends the teachers' assessment during the four-course editions by considering this refinement on each edition. Aggregated results for the different metrics have been summarized in the plot depicted in Figure 8 for the Clarity test. The results for all questions have been aggregated to simplify the evaluation.

Figure 8

Performance metrics on Clarity test



A significant improvement can be observed in the second edition due to new annotated answers from the first edition, which improved the accuracy on questions with an unbalanced number of correct and incorrect answers (i.e., questions 2, 4, and

8 of Table 1). Although there is a slight reduction in the correct detection of correct answers (i.e., TPR), this is due to the refining process.

The following course editions contributed to enhancing the overall accuracy, reaching values larger than 88% in the last edition. Thus, including new data can effectively improve the quality of the recommender, which implies that it can be effectively used for assessment purposes. Better accuracy means a recommender with fewer errors that teachers can trust during assessment.

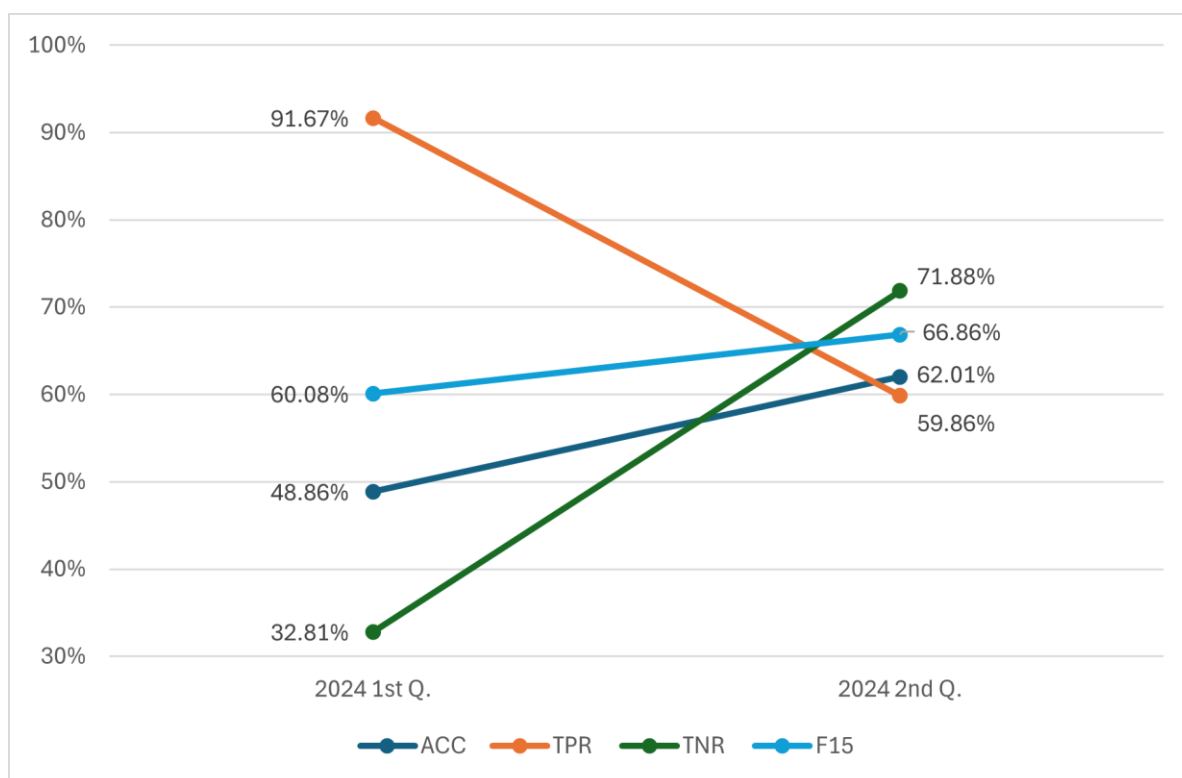
RQ3: Can SLASys be used without previous students' answers?

Good results in the previous section were obtained because teachers performed an initial task to annotate previous students' answers, which contributed to getting initial classifiers.

However, it was relevant to know how SLASys performs on new questions without previous answers. In this case, only semantic comparison can be used with golden answers provided by the teachers in the first edition (i.e., three golden answers were included per question). Nevertheless, in the second edition, answer classification can be applied using the assessment conducted in the first one. The experiment with real students was conducted in the last two editions, i.e., 2024 1st and 2nd quarter, where a new test was designed for the Novelty concept. The experiment depicted in Figure 9 shows how much improvement was reached by the SLASys system when changing from semantic comparison to answer classification in a real educational setting.

Figure 9

Performance metrics on Novelty test.



A significant difference between the two approaches can be observed. Global accuracy (i.e., ACC) improves from 48.86% to 62.01%. The meaningful increment in detecting wrong answers caused this improvement (i.e., TNR increased from 32.81% to 71.88%). However, the detection of correct answers decreases. Semantic comparison tends to recommend that most students' answers are correct since it cannot disaggregate slight semantic differences. This produces the unwanted countereffect that semantic comparison cannot correctly detect most incorrect answers. Thus, changing from semantic comparison to answer classification positively impacts the recommender.

The relevant insight in this experiment is that the answer classification models can produce high-quality recommendations in a few editions, even with a cold start with the semantic comparison approach.

RQ4: What is the opinion of the students and teachers?

Finally, it is of interest whether the new tests and feedback were helpful. Table 2 summarizes the information gathered from students, including the number who accessed the feedback results, the average time spent on the page, and the final rating given to the feedback. Not all students accessed the feedback results, probably because the tests did not impact the course assessment. Additionally, the average time is relatively low. Checking individual access time revealed that the time highly correlates with the number of wrong answers. Thus, students with wrong answers spent more time reading the feedback. Thus, it can be deduced that they are checking what they must learn for improvement. Finally, the average appraisal score is significantly high (i.e., larger than 70%) in all editions and tests.

Table 2
Students' feedback revision time and appraisal

Edition	Clarity				Novelty	
	2023 2 nd Q	2023 3 rd Q	2024 1 st Q	2024 2 nd Q	2024 1 st Q	2024 2 nd Q
Students who Accessed (%)	83.33	84.62	88.89	80.00	77.78	76.00
Avg. Access Time (minutes)	2.33	1.73	3.51	2.91	4.74	4.68
Appraisal (1-5)	3.94	3.84	3.63	3.85	3.55	3.77

Table 3
Comparison with related work

Reference	Technique	Prediction	Feedback	ACC
(Saha et al., 2018)	NLP	Correctness	No	66 %
(Soulimani et al., 2024)	BERT-Classifier	Grade (0-4)*	No	71 %
(Padó et al., 2024)	BERT-Classifier	Correctness	No	72 %
(Wang et al., 2019)	BERT-Classifier	Correctness	No	80 %
(Camus y Filighera, 2020)	BERT-Classifier	Correctness	No	80 %
(Lun et al., 2020)	BERT-Classifier	Correctness	No	82 %
(Sung et al., 2019)	BERT-Classifier	Correctness	No	84 %

Bañeres Besora, D., Guerrero Roldán, A.-E., & Rodríguez González, M. E. (2026). A language model-based recommender assessment system for short-answer questions in the intellectual property domain. [Recomendador de evaluación para preguntas cortas utilizando modelos de lenguaje en propiedad intelectual]. *RIED-Revista Iberoamericana de Educación a Distancia*, 29(1), 321-352. <https://doi.org/10.5944/ried.29.1.45541>

Reference	Technique	Prediction	Feedback	ACC
(Schneider et al., 2023)	BERT-Classifier	Correctness	No	86 %
(Liu et al., 2019)	BERT-Classifier	Correctness	No	89 %
(Grévisse, 2024)	GenAI (GPT4)	Grade (0-10)*	Generated	64 %
(Aggarwal et al., 2025)	GenAI (Mistral)	Correctness	Generated	75 %
SLASys	BERT-Classifier	Correctness	Predefined	83 %

(*) Accuracy is computed by assuming correct when the grade is larger than the middle of the maximum points

Furthermore, teachers collected some opinions on the online sessions. Students mostly agreed that the assessment grade, feedback about correct answers, and EPC guidelines to review were highly valuable. Also, obtaining them nearly immediately was positive. However, students complained that the feedback should be more personalized, describing why their answer was wrong.

Teachers also shared their experiences with the recommender. Initially, they were reluctant to use it, arguing that relying on automatic grading could result in wrong assessments. However, the tool was highly accepted as a recommender, where teachers' final revision was mandatory. After assessing the different editions, teachers agreed they gained efficiency since they also acquired experience over time. Related to the experiment with the Novelty test without annotated answers, they complained in the first edition (i.e. 2024 1st Q) since they expressed no gain in workload reduction, since the recommender produced many wrong recommendations. However, they realized the positive benefit in the last edition when the answer classification started to operate.

DISCUSSION

The collected insights allow the research questions to be answered. Related to RQ1, as other work states (Burrows et al., 2015), BERT AI classifiers (i.e., answer classification) show greater precision compared to semantic comparison, and they obtain similar results to other ASAG system evaluations. Table 3 compares with state-of-the-art systems in public datasets, summarizing the technique used, the prediction objective, type of provided feedback, and the obtained accuracy. SLASys' answer classification meets the accuracy performance obtained by related works (i.e., accuracy ranging from 71% to 89%), focusing on predicting the correctness of the answer. However, it cannot be compared with works that focus on predicting the grade since they use other metrics to evaluate how close the predicted grade is to the correct one (i.e., root square mean error). Such techniques have moderate to high error rates ranging from 0.57 (Baral et al., 2021) to values larger than 1 (Gaddipati et al., 2020; Metzler et al., 2024), making their use in a real setting difficult. Note that predicting the grade is a more difficult task since there is more variability in the result (del Gobbo et al., 2023). Therefore, the method proposed in this work simplifies the process by giving the assessment result to the teacher, who decides the grade depending on a rubric. Additionally, the experimental results show insights into the reduced dataset size needed to get a fine-tuned classifier for short-answer questions (Mehrafarin et al., 2022). Typically, it is recommended to have a minimum of 500 answers (i.e., *instances*) to have a proper classifier. However, Table 1 shows that a good classifier can be trained even with fewer answers. Moreover, a comparison was conducted between

both metric computations for a specific question (Figure 7). Teachers can access this explainable information within Moodle to better understand which answers are not well-classified. One should recall that explainable AI is one of the recommendations for AI tools (Zhao et al., 2024). Table 3 also compares with GenAI approaches, which currently perform worse than BERT ones. Although GenAI tools generate great enthusiasm, they still have some drawbacks for assessment. They may have hallucinations (Jia et al., 2024) or conceptual limitations in specific domains (la Cruz Martínez et al., 2024) that can decrease their efficacy for assessment. Some of these GenAI approaches can be used locally. However, enterprise solutions are widely used, which operate remotely, offer a bigger processing capacity, but at higher costs with sustainability issues (van Wynaert, 2021). Additionally, they require the analysis of ethical issues and the rethinking of educational and data protection policies. Higher education institutions are encouraged to review García-Peñalvo et al. (2024)'s manifesto for a safe, ethical, and secure system. In this case, SLASys operates locally, reducing processing costs, maintaining its functionality, and not providing data to third parties. Additionally, it is stateless without collecting students' sensitive data, keeping them safe within the LMS.

Concerning RQ2, accurate progression over four editions of the course was shown. As the previous research question claims, the small sets of annotated answers produced highly accurate classifiers, making it available for use in a real educational setting. By generating accurate classifiers, SLASys is an example of an AI tool that can benefit both actors. On the one hand, teachers can benefit from a recommender for ASAG, increasing assessment efficiency and unifying assessment criteria (Xavier et al., 2025). Besides, benefits for teachers are clear as their workload is reduced, allowing them to devote time to other qualitative tasks related to checking feedback, designing new questions, or better-designed learning activities. Due to the integration with Moodle, SLASys provides formative feedback to students compared to other techniques described in Table 3. Related works only provide a recommendation without feedback, which, in the case of deployment, would need additional methods to deliver it. According to Gaddipati et al. (2021), one option could be using GenAI tools that show potential for feedback generation. On the other hand, students gain quality feedback when asked questions that require reasoning (Calimeris & Kosack, 2020). With this feedback, they can learn better with comprehensive and detailed information about their assessment task. Ultimately, such feedback becomes a personalized component that enhances student knowledge (Abu Khurma et al., 2024).

Related to RQ3, results demonstrate that the system can be used even without training data. In the first edition of a new question, the teachers will require more attention due to the low accuracy recommendations. However, the AI classifier will be ready for use in the next edition. It is worth noting the technical simplicity of the system. SLASys follows the good assessment practices described in Petridou & Lao (2024).

Finally, concerning RQ4, opinions and results indicate that recommender mispredictions do not impact students since they obtain the teachers' reviewed assessment. Note that human supervision is mandatory because mistakes in assessment may have negative effects on students (Li et al., 2023). As in Sangapu (2018), opinions are mostly positively related to introducing IA tools in education. Teachers are not overwhelmed with technical issues since the system is embedded in the well-known Moodle LMS. The system will autonomously change from one

technique to another without teachers' intervention. Note that SLASys has been designed as an assessment recommender following the recommendations of the European AI Act (European Commission, 2024), avoiding their utilization as an automatic AI grader.

This study presents limitations related to sample size and non-probability sampling that impact external validity. Due to its application in a real educational setting, a small dataset from a specific domain was used, which restricts the generalizability of the findings and may lead to sample bias. This can be mostly observed on RQ3, where the sample size for the last edition is only 25 students. However, the findings on RQ2 demonstrate that the system improves accuracy and remains applicable and effective over time (i.e., four editions) even when trained on limited data. Regarding non-probability sampling, all students participated in the study, making the knowledge acquisition evaluation with the new tests unfeasible. Note that the first exam is after the second course during the complete training process. A longitudinal analysis until this exam would enrich the understanding of the effect of the new tests on the acquired knowledge. Sample size also affects RQ4. Students' positive opinions could be biased due to the sample size. Additionally, teachers' opinions were collected with simple semi-structured interviews. A qualitative analysis with focus groups would give deeper insights.

However, the obtained experimental results could not lead to generalization, although particular emphasis is placed on the straightforward transferability to any domain with specific language, such as law, history, philosophy, or social science. Transferability would depend on an initial workload to design the questions, add the corresponding golden answers, or gather and review the annotated answers. SLASys allows the creation of questions where students should apply the acquired knowledge instead of answering based on theoretical concepts, leading to an authentic assessment (Villarroel et al., 2018).

CONCLUSIONS

The contribution of this research is threefold. Firstly, a short-answer assessment recommender system guided by three core principles was developed: free BERT techniques suitable for private deployment, running on servers with limited resources, and easy-to-use without requiring technical AI expertise. Secondly, a notable feature is the seamless integration with Moodle, enabling students' tracking, grading, and visualization of learning outcomes through SLASys. Results have shown that, although these results cannot be generalized, answer classification can lead to better recommendations. Nevertheless, SLASys allows the recommender to be refined to produce better results for new cohorts of students. Thirdly, the work underscores the role of AI as a supportive tool in education, empowering teachers to use a tool designed for recommending assessment and feedback that can work in several domains. Teachers can use integrative strategies by highlighting the convergence of technology and pedagogy to enhance the students' learning experience.

Future work will focus on increasing the recommender utilization in other tests in the same fundamental course on patent examination in the IP domain. Such an increment will help to evaluate the system's capability for other concepts. Additionally, how feedback can be personalized based on students' wrong answers and the available learning resources will be explored. BERT can also be used to create semantic search tools based on the course learning materials. Such tools can automatically identify the

resources needed for a specific question, which could leverage the teacher's effort on feedback generation.

Acknowledgments

This work is part of the project PID2023-147592OB-I00 funded by MCIU/AEI/10.13039/501100011033/ FEDER, UE, and the Academic Research Programme of the European Patent Office Grant Agreement No. 2021/8404.

REFERENCES

- Abu Khurma, O., Albahti, F., Ali, N., & Bustanji, A. (2024). AI ChatGPT and student engagement: Unraveling dimensions through PRISMA analysis for enhanced learning experiences. *Contemporary Educational Technology*, 16(2).
<https://doi.org/10.30935/cedtech/14334>
- Adhikari, A., Ram, A., Tang, R., & Lin, J. (2019). *DocBERT: BERT for document classification*. arXiv.
<https://doi.org/10.48550/arXiv.1904.08398>
- Aggarwal, D., Sil, P., Raman, B., & Bhattacharyya, P. (2025). "I understand why I got this grade": Automatic short answer grading with feedback. In *Lecture Notes in Computer Science*.
https://doi.org/10.1007/978-3-031-98420-4_22
- Akçapınar, G. (2015). How automated feedback through text mining changes plagiaristic behavior in online assignments. *Computers & Education*, 87, 123-130.
<https://doi.org/10.1016/j.compedu.2015.04.007>
- Almasre, M. (2024). Development and evaluation of a custom GPT for the assessment of students' designs in a typography course. *Education Sciences*, 14(2), Article 148.
<https://doi.org/10.3390/educsci14020148>
- Arefeen, M. A., Debnath, B., & Chakradhar, S. (2024). LeanContext: Cost-efficient domain-specific question answering using LLMs. *Natural Language Processing Journal*, 7, 100065.
<https://doi.org/10.1016/j.nlp.2024.100065>
- Bahdanau, D., Cho, K., & Bengio, Y. (2016). Neural machine translation by jointly learning to align and translate. *arXiv*.
<https://doi.org/10.48550/arXiv.1409.0473>
- Banihashem, S. K., Kerman, N. T., Noroozi, O., Moon, J., & Drachsler, H. (2024). Feedback sources in essay writing: Peer-generated or AI-generated feedback? *International Journal of Educational Technology in Higher Education*, 21(1), 23.
<https://doi.org/10.1186/s41239-024-00455-4>
- Baral, S., Botelho, A. F., Erickson, J. A., Benachamardi, P., & Heffernan, N. T. (2021). Improving automated scoring of student open responses in mathematics. In *Proceedings of the 14th International Conference on Educational Data Mining (EDM 2021)*.
- Bergmann, J., & Sams, A. (2012). *Flip your classroom: Reach every student in every class every day*. International Society for Technology in Education.
- Burrows, S., Gurevych, I., & Stein, B. (2015). The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education*, 25(1), 60-117.
<https://doi.org/10.1007/s40593-014-0026-8>
- Calimeris, L., & Kosack, E. (2020). Immediate feedback assessment technique (IF-AT) quizzes and student performance in microeconomic principles courses. *Journal of Economic Education*, 51(3-4), 304-319.

- <https://doi.org/10.1080/00220485.2020.1804501>
- Camus, L., & Filighera, A. (2020). Investigating transformers for automatic short answer grading. In I. I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, & E. Millán (Eds.), *Artificial intelligence in education* (Lecture Notes in Computer Science, Vol. 12164, pp. 43-48). Springer. https://doi.org/10.1007/978-3-030-52240-7_8
- Dai, Y., Lai, S., Lim, C. P., & Liu, A. (2025). University policies on generative AI in Asia: Promising practices, gaps, and future directions. *Journal of Asian Public Policy*, 18(2), 260-281. <https://doi.org/10.1080/17516234.2024.2379070>
- del Gobbo, E., Guarino, A., Cafarelli, B., & Grilli, L. (2023). GradeAid: A framework for automatic short answers grading in educational contexts-Design, implementation and evaluation. *Knowledge and Information Systems*, 65(10), 4479-4507. <https://doi.org/10.1007/s10115-023-01892-9>
- De La Cruz Martínez, G., Eslava-Cervantes, A.-L., & Ramírez, S. (2024, July 1). Analysis of solutions of ChatGPT to logic problems based on critical thinking. In *Proceedings of the 16th International Conference on Education and New Learning Technologies (EDULEARN24)* (pp. 10324-10331). IATED. <https://doi.org/10.21125/edulearn.2024.2525>
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4171-4186). Association for Computational Linguistics. <https://doi.org/10.18653/v1/N19-1423>
- Dhananjaya, G. M., Goudar, R. H., Kulkarni, A. A., Rathod, V. N., & Hukkeri, G. S. (2024). A digital recommendation system for personalized learning to enhance online education: A review. *IEEE Access*, 12, 33591-33615. <https://doi.org/10.1109/ACCESS.2024.3369901>
- Escalante, J., Pack, A., & Barrett, A. (2023). AI-generated feedback on writing: Insights into efficacy and ENL student preference. *International Journal of Educational Technology in Higher Education*, 20(1), 40. <https://doi.org/10.1186/s41239-023-00425-2>
- European Commission. (2024). *Regulation (EU) 2024/1689 of the European Parliament and of the Council*. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32024R1689>
- Evans, C. (2013). Making sense of assessment feedback in higher education. *Review of Educational Research*, 83(1), 70-120. <https://doi.org/10.3102/0034654312474350>
- Gaddipati, S. K., Nair, D., & Plöger, P. G. (2020). Comparative evaluation of pretrained transfer learning models on automatic short answer grading. *arXiv*. <https://arxiv.org/abs/2009.01303>
- Gaddipati, S. K., Plöger, P., Hochgeschwender, N., & Metzler, M. (2021, April 5). *Automatic formative assessment for students' short text answers through feature extraction* [Doctoral dissertation, Hochschule Bonn-Rhein-Sieg].
- Gaona, J., Reguant, M., Valdivia, I., Vásquez, M., & Sancho-Vinuesa, T. (2018). Feedback by automatic assessment systems used in mathematics homework in the engineering field. *Computer Applications in Engineering Education*, 26(4), 921-934. <https://doi.org/10.1002/cae.21950>
- García-Peñalvo, F. J., Alier, M., Pereira, J., & Casany, M. J. (2024). Safe, transparent, and ethical artificial intelligence: Keys to quality sustainable education (SDG4). *International Journal of Educational Research and Innovation*, 22, 1-21. <https://doi.org/10.46661/ijeri.11036>
- González Fernández, M. O., Romero-López, M. A., Sgreccia, N. F., & Latorre Medina, M. J. (2025). Marcos normativos

- para una IA ética y confiable en la educación superior: Estado de la cuestión. *RIED-Revista Iberoamericana de Educación a Distancia*, 28(2), 181-208. <https://doi.org/10.5944/ried.28.2.43511>
- Grévisse, C. (2024). LLM-based automatic short answer grading in undergraduate medical education. *BMC Medical Education*, 24(1), 1060. <https://doi.org/10.1186/s12909-024-06026-5>
- György, A., & Vajda, I. (2007). Intelligent mathematics assessment in eMax. In *IEEE AFRICON Conference* (pp. 1-7). <https://doi.org/10.1109/AFRICON.2007.4401512>
- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112. <https://doi.org/10.3102/003465430298487>
- Huisman, B., Saab, N., van Driel, J., & van den Broek, P. (2017). Peer feedback on college students' writing: Exploring the relation between students' ability match, feedback quality and essay performance. *Higher Education Research & Development*, 36(7), 1433-1446. <https://doi.org/10.1080/07294360.2017.1325854>
- Husein, R. A., Aburajouh, H., & Catal, C. (2025). Large language models for code completion: A systematic literature review. *Computer Standards & Interfaces*, 92, 103917. <https://doi.org/10.1016/j.csi.2024.103917>
- Hustad, E., & Arntzen, A. A. B. (2013). Facilitating teaching and learning capabilities in social learning management systems: Challenges, issues, and implications for design. *Journal of Integrated Design and Process Science*, 17(1), 33-46. <https://doi.org/10.3233/JID-2013-0003>
- Jia, Q., Cui, J., Xi, R., Liu, C., Rashid, P., Li, R., & Gehringer, E. (2024). On assessing the faithfulness of LLM-generated feedback on student assignments. In B. Paaßen & C. D. Epp (Eds.), *Proceedings of the 17th International Conference on Educational Data Mining* (pp. 491-499). International Educational Data Mining Society. <https://doi.org/10.5281/zenodo.12729868>
- Kim, T. W. (2023). Application of artificial intelligence chatbots, including ChatGPT, in education, scholarly work, programming, and content generation and its prospects: A narrative review. *Journal of Educational Evaluation for Health Professions*, 20, 38. <https://doi.org/10.3352/jeehp.2023.20.38>
- Klein, R., Kyrilov, A., & Tokman, M. (2011). Automated assessment of short free-text responses in computer science using latent semantic analysis. In *Proceedings of ITiCSE '11: The 16th Annual Conference on Innovation and Technology in Computer Science Education* (pp. 158-162). <https://doi.org/10.1145/1999747.1999793>
- Kuechler, W., & Vaishnavi, V. (2012). A framework for theory development in design science research: Multiple perspectives. *Journal of the Association for Information Systems*, 13(6), 395-423. <https://doi.org/10.17705/1JAIS.00300>
- Li, T. W., Hsu, S., Fowler, M., Zhang, Z., Zilles, C., & Karahalios, K. (2023). Am I wrong, or is the autograder wrong? Effects of AI grading mistakes on learning. In *Proceedings of the 2023 ACM Conference on International Computing Education Research (ICER '23)* (pp. 85-97). <https://doi.org/10.1145/3568813.3600124>
- Liu, T., Ding, W., Wang, Z., Tang, J., Huang, G. Y., & Liu, Z. (2019). Automatic short answer grading via multiway attention networks. In *Lecture Notes in Computer Science* (Vol. 11626, pp. 376–388). https://doi.org/10.1007/978-3-030-23207-8_32
- Lun, J., Zhu, J., Tang, Y., & Yang, M. (2020). Multiple data augmentation strategies for improving performance on automatic short answer scoring. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34, pp. 13381-13388). <https://doi.org/10.1609/aaai.v34i09.7062>

- Mehrafarin, H., Rajaei, S., & Pilehvar, M. T. (2022). On the importance of data size in probing fine-tuned models. In *Findings of the Association for Computational Linguistics* (pp. 239–248). <https://doi.org/10.18653/v1/2022.findings-acl.20>
- Messer, M., Brown, N. C. C., Kölling, M., & Shi, M. (2024). Automated grading and feedback tools for programming education: A systematic review. *ACM Transactions on Computing Education*, 24(1), Article 1. <https://doi.org/10.1145/3636515>
- Metzler, T., Plöger, P. G., & Hees, J. (2024). Computer-assisted short answer grading using large language models and rubrics. In *INFORMATIK 2024: AI@WORK* (pp. 1383-1393). Gesellschaft für Informatik e.V. https://doi.org/10.18420/inf2024_121
- Nguyen, H., Bhat, S., Moore, S., Bier, N., & Stamper, J. (2022). Towards generalized methods for automatic question generation in educational domains. In S. Isotani, E. Millán, A. Ogan, P. Hastings, B. McLaren, & R. Luckin (Eds.), *Intelligent Tutoring Systems. ITS 2022. Lecture Notes in Computer Science* (Vol. 13450, pp. 272-284). Springer. https://doi.org/10.1007/978-3-031-16290-9_20
- Nicol, D., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, 31(2), 199-218. <https://doi.org/10.1080/03075070600572090>
- Novak, G. M. (2012). Just-in-time teaching. *New Directions for Teaching and Learning*, 2012(128), 63-73. <https://doi.org/10.1002/tl.469>
- Oates, B. J. (2006). *Researching information systems and computing*. SAGE Publications Ltd.
- OpenAI. (2024). *GPT-4o system card*. <https://openai.com/index/gpt-4o-system-card/>
- Padó, U., Eryilmaz, Y., & Kirschner, L. (2024). Short-answer grading for German: Addressing the challenges. *International Journal of Artificial Intelligence in Education*, 34(4), 1488-1510. <https://doi.org/10.1007/s40593-023-00383-w>
- Pang, J., Ye, F., Wong, D. F., Yu, D., Shi, S., Tu, Z., & Wang, L. (2025). Salute the classic: Revisiting challenges of machine translation in the age of large language models. *Transactions of the Association for Computational Linguistics*, 13, 73-95. https://doi.org/10.1162/tacl_a_00730
- Petridou, E., & Lao, L. (2024). Identifying challenges and best practices for implementing AI additional qualifications in vocational and continuing education: A mixed methods analysis. *International Journal of Lifelong Education*, 43(4), 385-400. <https://doi.org/10.1080/02601370.2024.2351076>
- Qiu, Y., & Jin, Y. (2024). ChatGPT and finetuned BERT: A comparative study for developing intelligent design support systems. *Intelligent Systems with Applications*, 21, 200308. <https://doi.org/10.1016/j.iswa.2023.200308>
- Rezaei, A. R. (2015). Frequent collaborative quiz taking and conceptual learning. *Active Learning in Higher Education*, 16(3), 189-204. <https://doi.org/10.1177/1469787415589627>
- Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, 18(2), 119-144. <https://doi.org/10.1007/BF00117714>
- Saha, S., Dhamecha, T. I., Marvaniya, S., Sindhgatta, R., & Sengupta, B. (2018). Sentence-level or token-level features for automatic short answer grading? Use both. In *Lecture Notes in Computer Science* (Vol. 10947, pp. 475-486). https://doi.org/10.1007/978-3-319-93843-1_37
- Sangapu, I. (2018). Artificial intelligence in education: From a teacher and a student perspective. *SSRN*. <https://doi.org/10.2139/ssrn.3372914>
- Schneider, J., Richner, R., & Riser, M. (2023). Towards trustworthy autograding of short, multi-lingual, multi-type answers. *International Journal of Artificial Intelligence in Education*, 33(1), 1-29. <https://doi.org/10.1007/s40593-022-00289-z>

- Senthilnathan, V., Sakthi Vaibhav, M., & Alexander, R. (2025). *Semantic refined prompting based automated essay scoring system*. In *Proceedings of the 2025 International Conference on Electronics and Renewable Systems (ICEARS)* (pp. 1344-1348). IEEE. <https://doi.org/10.1109/ICEARS64219.2025.10940227>
- Siddiqi, R., & Harrison, C. (2008). A systematic approach to the automated marking of short-answer questions. In *Proceedings of IEEE INMIC 2008: 12th International Multitopic Conference* (pp. 281-286). <https://doi.org/10.1109/INMIC.2008.477758>
- Soulamani, Y. A., El Achaak, L., & Bouhorma, M. (2024). Deep learning-based Arabic short answer grading in serious games. *International Journal of Electrical and Computer Engineering*, 14(1), 841-853. <https://doi.org/10.11591/ijece.v14i1.pp841-853>
- Souza, F., Nogueira, R., & Lotufo, R. A. (2019). *Portuguese named entity recognition using BERT-CRF* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.1909.10649>
- Sung, C., Ma, T., Dhamecha, T. I., Reddy, V., Saha, S., & Arora, R. (2019). Pre-training BERT on domain resources for short answer grading. In *Proceedings of EMNLP-IJCNLP 2019* (pp. 6076-6086). <https://doi.org/10.18653/v1/D19-1628>
- van Wynsberghe, A. (2021). Sustainable AI: AI for sustainability and the sustainability of AI. *AI and Ethics*, 1(3), 213-218. <https://doi.org/10.1007/s43681-021-00043-6>
- Villarroel, V., Bloxham, S., Bruna, D., Bruna, C., & Herrera-Seda, C. (2018). Authentic assessment: Creating a blueprint for course design. *Assessment & Evaluation in Higher Education*, 43(5), 840-854. <https://doi.org/10.1080/02602938.2017.1412396>
- Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (pp. 353-355). <https://doi.org/10.18653/v1/W18-5446>
- Wang, H., & Lehman, J. D. (2021). Using achievement goal-based personalized motivational feedback to enhance online learning. *Educational Technology Research and Development*, 69(2), 807-836. <https://doi.org/10.1007/s11423-021-09940-3>
- Wang, Y., Wang, C., Li, R., & Lin, H. (2022). On the use of BERT for automated essay scoring: Joint learning of multi-scale essay representation. In *Proceedings of NAACL 2022* (pp. 3432-3444). <https://doi.org/10.18653/v1/2022.naacl-main.249>
- Wang, Z., Lan, A. S., Waters, A. E., Grimaldi, P., & Baraniuk, R. G. (2019). A meta-learning augmented bidirectional transformer model for automatic short answer grading. In *Proceedings of the 12th International Conference on Educational Data Mining (EDM 2019)*.
- Winstone, N. E., Nash, R. A., Parker, M., & Rowntree, J. (2017). Supporting learners' agentic engagement with feedback: A systematic review and a taxonomy of recipience processes. *Educational Psychologist*, 52(1), 17-37. <https://doi.org/10.1080/00461520.2016.1207538>
- Xavier, C., Rodrigues, L., Costa, N., Neto, R., Alves, G., Falcão, T. P., Gašević, D., & Mello, R. F. (2025). Empowering instructors with AI: Evaluating the impact of an AI-driven feedback tool in learning analytics. *IEEE Transactions on Learning Technologies*, 18, 498-512. <https://doi.org/10.1109/TLT.2025.3562379>
- Xie, X., & Li, X. (2018). Research on personalized exercises and teaching feedback based on big data. In *Proceedings of the ACM International Conference* (pp. 217-221). <https://doi.org/10.1145/3232116.3232143>
- Xu, Z., & Zhu, P. (2023). Using BERT-based textual analysis to design a smarter classroom mode for computer teaching in higher education institutions.

- International Journal of Emerging Technologies in Learning*, 18(19), 120-133.
<https://doi.org/10.3991/ijet.v18i19.42483>
- Zhang, H., Cai, J., Xu, J., & Wang, J. (2019). Pretraining-based natural language generation for text summarization. In *Proceedings of CoNLL 2019* (pp. 789-798). <https://doi.org/10.18653/v1/K19-1074>
- Zhang, H., Yu, P. S., & Zhang, J. (2025). A systematic survey of text summarization: From statistical methods to large language models. *ACM Computing Surveys*.
<https://doi.org/10.1145/3731445>
- Zhang, Z., Zhang, Z., Chen, H., & Zhang, Z. (2019). A joint learning framework with BERT for spoken language understanding. *IEEE Access*, 7, 168849–168858.
<https://doi.org/10.1109/ACCESS.2019.2954766>
- Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., Wang, S., Yin, D., & Du, M. (2024). Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2), Article 26.
<https://doi.org/10.1145/3639372>
- Zheng, L., Long, M., Chen, B., & Fan, Y. (2023). Promoting knowledge elaboration, socially shared regulation, and group performance in collaborative learning: An automated assessment and feedback approach based on knowledge graphs. *International Journal of Educational Technology in Higher Education*, 20(1), 12.
<https://doi.org/10.1186/s41239-023-00415-4>

Date of reception: 1 June 2025

Date of acceptance: 6 August 2025

Date of approval for layout: 16 September 2025

Date of publication in OnlineFirst: 14 October 2025

Date of publication: 1 January 2026

Measuring writing skills in Spanish as a foreign language with generative artificial intelligence

Medición de la habilidad escrita en español como lengua extranjera con inteligencia artificial generativa



María-Victoria Cantero Romero - *Universidad de Jaén, UJA (Spain)*

María-Teresa Martín-Valdivia - *Universidad de Jaén, UJA (Spain)*

Ana María Ortiz-Colón - *Universidad de Jaén, UJA (Spain)*

Salud María Jiménez-Zafra - *Universidad de Jaén, UJA (Spain)*

ABSTRACT

The emergence of Generative Artificial Intelligence (GAI)—particularly Large Language Models (LLMs) such as ChatGPT—is transforming the educational landscape, especially in the field of foreign language instruction. This article explores the potential of these technologies to automate the assessment of writing proficiency in Spanish as a Foreign Language (SFL), a task that is especially time-consuming at the beginning of university-level courses for Erasmus students. The study is based on three experiments conducted using the Spanish Learner Corpus compiled by the Instituto Cervantes. The first experiment applied a zero-shot learning approach by prompting the model with level descriptors from the Instituto Cervantes's Curriculum Plan. In the second and third experiments, the model was adjusted through fine-tuning using 90% and 80% of the corpus, respectively, with the remaining data reserved for testing and validation. The results indicate that the fine-tuned models significantly outperform the zero-shot configuration in identifying the correct proficiency levels of learner texts. These findings demonstrate that LLMs can be effectively employed to streamline the initial placement process in SFL courses, thus reducing the workload of instructors and improving efficiency. The study concludes that GAI can serve as a valuable complementary tool in multilingual and multicultural educational settings, provided its use is guided by sound pedagogical principles.

Keywords: language instruction; writing; artificial intelligence; teaching method innovations.

RESUMEN

La irrupción de la Inteligencia Artificial Generativa (IAG), y en particular de los Modelos de Lenguaje de gran tamaño (LLMs) como ChatGPT, está transformando el ámbito educativo, especialmente en la enseñanza de lenguas extranjeras. Este artículo analiza el potencial de estas tecnologías para automatizar la evaluación de la competencia escrita en español como lengua extranjera (ELE), una tarea especialmente laboriosa al inicio de los cursos universitarios dirigidos a estudiantes Erasmus. La metodología se basa en tres experimentos con el Corpus de Aprendices de Español del Instituto Cervantes. En el primero, se utilizó la técnica de *zero-shot learning*, proporcionando al modelo un *prompt* con los descriptores del Plan Curricular del Instituto Cervantes. En el segundo y tercer experimentos, se ajustó el modelo mediante *fine-tuning* con el 90 % y el 80 % del corpus, respectivamente, reservando el resto para validación y prueba. Los resultados muestran que los modelos ajustados son capaces de identificar el nivel de competencia escrita con una precisión significativamente superior al enfoque sin entrenamiento previo. Estos hallazgos evidencian que los LLMs pueden emplearse para agilizar procesos de evaluación inicial en cursos de ELE, reduciendo la carga docente y mejorando la eficiencia. Se concluye que la IAG representa una herramienta complementaria valiosa en contextos educativos multiculturales y multilingües, siempre que su uso esté guiado por criterios pedagógicos sólidos.

Palabras clave: enseñanza de lenguas; expresión escrita; inteligencia artificial; innovación pedagógica relevante.

INTRODUCTION

The evolution of Artificial Intelligence has marked a turning point in all areas, especially in education (Aparicio Gómez, 2023; Hernández-León & Rodríguez-Conde, 2024; Zambrano Campozano, 2025). Advances in this discipline are opening up a field of new research and work with which to implement Artificial Intelligence, specifically generative AI, in the classroom. As Bolaño-García and Duarte-Acosta (2024) point out, generative Artificial Intelligence has gained attention in education because it improves the personalisation of learning and real-time feedback. Zapata Ros (2024) supports this idea of personalisation in addition to the availability of information. Likewise, Fajardo et al. (2023) highlight the use of these tools for personalising learning in university education, adapting it to the preferences and needs of each student through guided and virtual tutorials. García-Peñalvo et al. (2024) point out the importance of preparing both teachers and learners in the use of generative artificial intelligence, as it will be present in all aspects of life. Furthermore, as Moreno (2019) notes, it is important to highlight the potential of generative Artificial Intelligence to transform education by creating adaptive learning environments tailored to student performance, such as for students with special educational needs. In their study on generative Artificial Intelligence tools, Area-Moreira et al. (2024), in addition to the functions already mentioned, indicate that these tools can be used to automate tasks to support teachers in their work and even as anti-plagiarism systems. In line with the above, Chan and Tsi (2023) point to the use of generative Artificial Intelligence as a supplementary tool for teachers and not as a replacement for them. Moreno (2019) also points to three approaches to work with in education: generative Artificial Intelligence, educational robotics, and self-learning platforms. We will focus on the first of them in this study. Both Barroso-Osuna and Cabero-Almenara (2025) and Owan et al. (2023) identify a number of benefits of using Artificial Intelligence, specifically generative AI, in education, including the optimisation of teaching time and automated and accurate assessment. In this regard, Crespo Mendoza et al. (2024) point out that it can improve the accuracy and reliability of assessments.

As mentioned above, the main focus of this work is generative Artificial Intelligence, specifically Language Models (LLMs), which have provided a new tool for conducting studies aimed at improving teaching tasks. LLMs are natural language generation tools trained with a large amount of text (Wang, 2024). García-Peñalvo et al. (2024) identify various functions that LLMs can perform, such as supporting research, creating educational content, generating chatbots to interact with students by offering self-directed feedback, complementing search engines, paraphrasing text, teaching languages, and generating exams and questionnaires. It is in connection with these last two functions that the present study is framed: placement in language teaching, and specifically in Spanish as a foreign language.

Generative Artificial Intelligence offers the opportunity to carry out adaptive assessments for each student with immediate and specific feedback, suggesting possible solutions (Barroso-Osuna & Cabero-Almenara, 2025). Likewise, Language Models can be used to automate (García-Peñalvo, 2024) the correction of both multiple-choice tests and open-ended responses (Moreno, 2019). The advantages of using LLMs for assessment include efficiency (Area-Moreira et al., 2024) and plagiarism detection (García-Peñalvo et al., 2024). García-Peñalvo et al. (2023) also point out that LLMs improve teacher productivity.

With regard to the use of Language Models and language teaching, their usefulness is backed by various studies (Baskara & Mukarto, 2023; Salguero Romero, 2023; Wang, 2024). Baskara and Mukarto (2023) point out how ChatGPT is capable of generating realistic texts that bring students closer to the reality of the language. On the other hand, Hong (2023) points out the advantage of being able to use these tools to speed up exam correction, freeing teachers from workload and giving them the opportunity to focus on lesson preparation.

One of the challenges of language teaching, and in this case, teaching Spanish as a foreign language, is the appropriate level assessment of students when they begin Spanish courses. It is essential for a student to be at the right level of Spanish learning in order to progress appropriately, since if they are placed at a higher level than what is appropriate for them, they may become frustrated, and conversely, if they are placed at a lower level, they may lack motivation. That is why level assessment is key when starting language courses.

Until now, this level assessment has been carried out through multiple-choice tests or interviews (Biedma Torrecillas et al., 2012). These tests mainly focus on determining the student's level of oral and written expression and comprehension. Existing Spanish level tests, such as the Instituto Cervantes level test, consist of answering a series of written multiple-choice questions (true-false; matching or ordering) of increasing difficulty (Centro Virtual Cervantes, n.d.). This same typology is currently observed in international contexts, such as in the Spanish level tests at Columbia University and the University of Wisconsin–Madison, which are also based on multiple-choice exercises without including free written production (Columbia University, n.d.; University of Wisconsin–Madison, n.d.). However, Spanish language proficiency tests have not focused directly on written expression tests involving the writing of complete texts, but rather indirectly, through item responses, due to the limitations this entails, such as the lack of immediacy or the complexity of level assessment processes when dealing with large groups.

To solve this problem, this study will focus on the level assessment of the written expression test. As mentioned above, Language Models are capable of automating the grading of written work (García-Peñalvo et al., 2024), and they do so quickly, saving teachers time (Area-Moreira et al., 2024). This is the main reason why written expression tests have not been included in current level tests.

In this context, it is necessary to establish a conceptual basis for understanding the strengths and limitations of automated writing assessment and, in particular, the linguistic foundations that underpin the classification of levels A1–C1. The following section develops this theoretical framework, which will serve as a basis for the methodological proposal of this study.

This article is structured as follows. After the introduction and theoretical framework, the technological basis and choice of model, the objectives of the study and the methodology are presented. This is followed by the results, followed by a section on the pedagogical relevance of the model. Finally, the article concludes with a discussion and conclusions, to which a subsection on ethical aspects and licences for use has been added.

THEORETICAL FRAMEWORK

This section begins with a review of the state of the art in automated writing assessment and the linguistic foundations underpinning the level classification of the Common European Framework of Reference (CEFR) and the Instituto Cervantes Curriculum Plan (PCIC). This conceptual basis will serve to contextualise the study proposal and, finally, to present the objectives guiding the research.

Automated handwriting assessment: current developments and approaches

The automatic classification of learners' texts using language models is part of a broader tradition of automated writing assessment (AWE), the development of which has given rise to multiple tools and systems that should be taken into account in order to contextualise this work. AWE has evolved considerably in recent decades, becoming an increasingly common tool in educational contexts. Pioneering systems such as e-rater (Burstein et al., 2003), developed by ETS, have been widely used in standardised tests, using linguistic, grammatical and discursive metrics to estimate textual quality. Coh-Metrix (McNamara et al., 2014), meanwhile, allows for detailed analysis of cohesion, syntactic complexity and readability, providing a multifactorial approach to written discourse. Recent studies, such as that by Zhang (2021), offer systematic reviews of these systems, highlighting their transition from rule-based approaches to models driven by machine learning and natural language processing. Along the same lines, Wang et al. (2022) analyse current approaches to the evaluation of argumentative texts, focusing on discursive components such as reasoning structure, evidence, and organisation. Other proposals, such as Writing Mentor (Burstein et al., 2018), integrate automatic assessment with formative feedback, promoting self-regulation processes in academic writing. Likewise, tools such as Write & Improve, developed by the University of Cambridge, exemplify how it is possible to provide immediate automatic feedback on texts produced by foreign language learners, facilitating autonomous and guided learning (Cambridge English, n.d.).

Linguistic foundations of the A1–C1 level classification

The classification of texts produced by learners of Spanish as a foreign language at levels A1–C1 is based on the descriptors established by the Common European Framework of Reference for Languages (Council of Europe, 2002) and the Instituto Cervantes Curriculum Plan (Instituto Cervantes, 2006). These documents define in detail the linguistic, pragmatic and sociolinguistic competences associated with each level, providing a solid basis for assessment.

At level A1, the lexical repertoire is very limited and restricted to basic transactions and everyday expressions. Texts are very short and simple, with simple sentences and a low average number of words per utterance. The use of regular forms of the present indicative predominates and an elementary grammatical repertoire is employed (Council of Europe, 2002; Instituto Cervantes, 2006).

At level A2, learners can produce short texts that convey simple information on familiar topics. A larger lexical repertoire and slightly more complex structures are observed, incorporating past indicative tenses (preterite, imperfect and indefinite) and

some irregular forms of the present tense. The use of the affirmative imperative also appears (Council of Europe, 2002; Instituto Cervantes, 2006).

At level B1, the lexical repertoire is broader and allows for the creation of texts that fulfil a specific communicative task, maintaining a coherent structure. Grammatically, tenses such as the simple future, the simple conditional and the past perfect are handled with a certain fluency, in addition to introducing the present subjunctive and the negative imperative. Discourse shows greater cohesion and a wider variety of connecting devices (Council of Europe, 2002; Instituto Cervantes, 2006).

At level B2, the user has a broad and precise linguistic repertoire, capable of sustaining complex arguments and detailed descriptions. Compound and subordinate clauses are used fluently, as well as confident use of indicative tenses (present, past, future and conditional) and subjunctive tenses (present, imperfect, perfect and pluperfect). Textual cohesion is consistent and lexical nuances appropriate to different registers are used (Council of Europe, 2002; Instituto Cervantes, 2006).

At level C1, the linguistic and non-linguistic repertoire is sufficiently broad and flexible to handle any type of communicative transaction or interaction, even in demanding academic or professional contexts. The learner is able to produce long, complex texts with a clear, well-organised structure, using all tenses accurately and a wide range of syntactic and lexical resources (Council of Europe, 2002; Instituto Cervantes, 2006).

TECHNOLOGICAL BASIS AND MODEL SELECTION

This study used version 3.5 of the ChatGPT language model developed by OpenAI, released in November 2022 (OpenAI, 2022). Although this version is not open access, it allows fine-tuning through OpenAI's API. The choice of version 3.5 over the latest version is justified by the possibility of performing this process, which is not possible in the latest version of ChatGPT.

For this research, GPT-3.5 was chosen over more recent or open-source models due to a combination of technical, economic, and methodological factors. As this is a novel task—the automatic level assessment of written proficiency in Spanish as a foreign language—it was considered appropriate to evaluate the performance of a widely tested generalist model, such as GPT-3.5, both in its base configuration and through fine-tuning. This model, accessible through the OpenAI API, does not require advanced computational infrastructure and offers a robust architecture with good coverage of Spanish (Li, 2023; Pourpanah et al., 2023). In addition, it offers an adequate balance between performance, cost and response speed, key factors for validating the viability of the approach in this exploratory phase (Roumeliotis et al., 2024). In contrast, at the time the experiments were carried out, models such as GPT-4 involved a considerably higher economic cost and longer inference times, which reinforced the decision to use GPT-3.5 as the initial reference.

However, it should be noted that GPT-3.5 has limitations compared to more recent models, such as GPT-4, which offer more accurate contextual understanding, training with larger and more heterogeneous datasets, and greater reasoning ability. These characteristics make them particularly suitable candidates for future research, both in the automated assessment of written competence aligned with CEFR levels and in the generation of high-quality synthetic corpora for training specialised systems.

STUDY OBJECTIVE

To carry out this study, the descriptors established by the Common European Framework of Reference for Languages (Council of Europe, 2002) and the Instituto Cervantes Curriculum Plan have been adapted in order to establish clear and operational criteria that allow a language model, in this case ChatGPT, to automatically classify the written productions of Spanish learners according to their level of competence.

Despite the aforementioned advances in automatic writing assessment, few studies have focused on the level assessment of written texts produced by foreign language learners and, specifically, on the specific case of teaching Spanish as a foreign language. Existing research tends to address automatic feedback or grading, but not the classification of written work according to CEFR or Instituto Cervantes Curriculum levels, especially in initial assessment tasks. This gap is particularly relevant in contexts such as universities, where the heterogeneity of international students, as in Erasmus programmes, requires efficient tools for level assessment. This study proposes an innovative solution based on generative language models (ChatGPT), which allows for the automatic classification of written texts by learners of Spanish as a foreign language (ELE) according to their level of proficiency, reducing the teaching load and improving the management of workload at the start of the academic year. Thus, this work not only complements previous research focused on textual improvement, but also broadens the scope of automated assessment to include initial diagnostic tasks in second language teaching contexts.

That is why, as the general objective of the study, we have set out to evaluate the effectiveness of the ChatGPT language model as an innovative tool for assessing the written expression level of Spanish learners. We also hope to achieve a series of specific objectives: i) to find out whether ChatGPT, with its prior training, is capable of carrying out level assessment adequately; ii) to verify whether ChatGPT is capable of level assessment if it is adjusted with a corpus levelled with the Instituto Cervantes Curriculum Plan and the Common European Framework of Reference for Languages; and iii) to determine the impact of ChatGPT on the efficiency of the level assessment of Spanish as a foreign language.

METHODOLOGY

To carry out this research, we used the Corpus of Spanish Learners (hereinafter CAES) (Palacios Martínez et al., 2019), developed by the Instituto Cervantes in collaboration with the University of Santiago de Compostela. We also used version 3.5 of the ChatGPT language model, developed by OpenAI (OpenAI, 2022).

Three types of experimental tests were carried out in this study: one using the zero-shot learning technique and the other two using model fine-tuning.

The fine-tuning procedure was carried out with a single training epoch, keeping the rest of the parameters at their default values according to the OpenAI API. To ensure reproducibility, a fixed random seed (value 42) was used, i.e., a reference value that allows experiments to be reproduced with the same results every time. No class balancing techniques were applied, as it was decided to preserve the actual distribution of the corpus, thus reflecting authentic levelling conditions in the classroom. In this way, the system's results are better suited to the challenges of real educational scenarios, without introducing artificial modifications to the representation of levels.

However, it is recognised that in future research, compensation strategies could be applied to compare their effect on the fairness and robustness of the model. Cross-validation was also not used, in line with the initial and experimental nature of the study.

Regarding the technical parameters of the training, the default values of OpenAI's API were maintained in aspects such as batch size (i.e., the number of examples processed at a time), learning rate (which indicates the speed at which the model adjusts its parameters during training), and loss functions (metrics that measure the difference between the model's prediction and the expected result). No additional regularisation techniques or early stopping strategies (early termination of training to avoid overfitting) were applied, as the main objective was to validate the feasibility of the approach rather than to optimise the model for maximum performance. This decision is in line with the exploratory nature of the research, which aims to test the applicability of the model to the task of automatic text levelling.

CAES Corpus

The CAES corpus was compiled by the University of Santiago de Compostela and funded by the Instituto Cervantes. Computerised data was collected from October 2011 to December 2020 from centres, mostly universities, in different countries such as Spain, the United States, Brazil, Egypt, Ireland and Portugal. The students who took part in the project had eleven different native languages (English, Mandarin Chinese, Portuguese, Arabic, Russian, German, French, Greek, Italian, Japanese and Polish).

This study used version 2.1 of the CAES corpus from March 2022, which updated the first data collection, which had a smaller amount of data, with a total of 1,423 participants, compared to the 2,544 participants in the 2022 update.

The corpus contains examples of different levels of Spanish according to the Common European Framework of Reference for Languages, from A1 to C1. At levels A1, A2 and B1, texts belonging to three different types of tasks that students had to write, ranging from 30 to 200 words in length, were collected. Levels B2 and C1 have a sample of two tasks per level, ranging from 275 to 500 words in length.

The topics identified at each level are as follows: At A1, the first task consists of an email about changing jobs, with a total of 728 samples; the second task consists of an email about their family, with 703 samples; the last task consists of a note about being late, with a sample of 705. At level A2, the first task is a biography, with 673 samples; the second task is booking a hotel room, with a sample of 603 texts; and the last task is writing a postcard about your holidays, with a sample of 701 texts. As it can be seen, around 2,000 examples were collected per level at the initial levels, providing a significant sample of these levels. Likewise, the texts' themes are matched to the functions and corresponding textual products of each level according to the Instituto Cervantes Curriculum Plan (Instituto Cervantes, 2006).

At level B1, there are also three different tasks. The first is to write a letter to a friend, with a sample of 528 texts; the second task is to write an email about a complaint to an airline, with a sample of 454; and the last task is to narrate a story, with a sample of 382. It should be noted that these tasks, like those at levels A1 and A2, are matched to the functions described in the Curriculum Plan.

At levels B2 and C1, the tasks are reduced to two. At level B2, the first task is to write an application for admission, with a total of 375 samples; while the second task

at level B2 is to write a text arguing the case for smoking in public places, with a sample size of 356.

At level C1, the first task consists of writing a complaint to a gas company, with a sample of 169; and in the second task, they must write a film review, with a sample of 184 texts. It can be seen how the sample size is significantly reduced at these levels as the number of tasks is reduced. It is also notable how the samples are smaller at level C1, with fewer than 400 samples at this level.

As in the previous levels, it can be seen that the tasks in levels B2 and C1 also correspond to the functions described in the Curriculum Plan.

The corpus was annotated by specialists in teaching Spanish as a foreign language at the University of Santiago de Compostela. Each text was classified into one of the CEFR levels according to the criteria established in the Instituto Cervantes's Curriculum and the guidelines defined in the project itself (Palacios Martínez et al., 2019; University of Santiago de Compostela, n.d.). To ensure the reliability of the classification, the texts were evaluated independently by several annotators and then reviewed jointly until a consensus was reached. This procedure ensures that level assessment is coherent and consistent, making the corpus a solid resource for research and automatic assessment of written production.

Table 1 presents a summary of the levels with respect to the tasks addressed and the total number of samples collected.

Cantero (2024) also conducted a study of this corpus, from which the following results can be drawn. At level A1, the average number of words per sentence is between 10.9 and 11.7, and the most frequent words are simple connectors and a limited lexicon. At level A2, the average number of words per sentence is higher, ranging from 12.5 to 14.6, with a simple, frequent lexicon. At the next level, B1, the average number of words per sentence increases to 12.0-16.5, and the vocabulary is broader and more complex than at previous levels. At level B2, the average number of words per sentence is even more complex, ranging from 17.7 to 21.5, and in terms of vocabulary, more complex connectors and more specialised vocabulary are used. Finally, at level C1, the average number of words per sentence is between 20.7 and 23.3, with complex and varied vocabulary.

Table 1
Sample summary

Level	Task	Exhibit
A1	Email regarding change of job	728
	Family email	703
	Note: arriving late	705
A2	Biography of a person you admire	673
	Book a hotel room	603
	Holiday postcard	701
B1	Letter to a friend	528
	Funny story	382
	Airline claim	454
B2	Application for admission	375
	Smoking in public places	356
C1	Film review	184
	Gas company complaint	169

Prompt used

Once the corpus had been analysed, a specific prompt was developed to carry out the tests with the model. As Morales-Chan (2023) points out, a good prompt can guarantee the success of the task. Therefore, it is important to define the objective and provide sufficient context.

The following prompt was used in the three tests carried out in this study —zero-shot learning (Test 1) and fine-tuning (Tests 2 and 3)— in order to maintain methodological consistency in the evaluation criteria. The prompt¹ design was based on the linguistic descriptors of the Cervantes Institute Curriculum Plan and the Common European Framework of Reference for Languages.

Tú eres un experto lingüista especializado en enseñanza de español como lengua extranjera. Tu tarea es indicar el nivel de español como lengua extranjera de los textos siguiendo el Plan Curricular del Instituto Cervantes.

Aquí tienes una descripción de los distintos niveles.

Niveles A1 y A2 Transacciones básicas relacionadas con su entorno.

A1: Repertorio limitado de léxico, textos muy breves y sencillos, un promedio de 10 palabras por oración. Formas regulares del presente de indicativo.

A2: Textos breves con información sencilla, un promedio de 12 palabras por oración. Tiempos verbales del pasado de indicativo: Pretérito perfecto, imperfecto e indefinido. Formas irregulares de presente de indicativo. Imperativo afirmativo.

Niveles B1 y B2 desenvolverse con textos sobre temas de su interés gustos y preferencias.

B1: Vocabulario amplio pero sencillo, realizar textos con una tarea concreta. Presente de indicativo, pretérito perfecto, imperfecto e indefinido de indicativo, futuro simple, condicional simple, pretérito pluscuamperfecto de indicativo, presente de subjuntivo. Imperativo negativo.

B2: Repertorio lingüístico amplio, oraciones subordinadas. Tiempos verbales de indicativo: presente, pretérito perfecto, imperfecto, indefinido, futuro simple y compuesto, condicional simple y compuesto, pretérito pluscuamperfecto. Tiempos verbales de subjuntivo: presente, pretérito imperfecto, pretérito perfecto y pluscuamperfecto.

C1 transacciones de todo tipo. Disponen de un repertorio de recursos lingüísticos y no lingüísticos lo suficientemente amplio y rico. Pueden enfrentarse a una amplia serie de textos extensos y complejos. Todos los tiempos verbales de indicativo y de subjuntivo el presente, pretérito perfecto, imperfecto y pluscuamperfecto.

Ahora vas a recibir un TEXTO y teniendo en cuenta lo explicado anteriormente y los errores gramaticales indica al final de tu respuesta con la etiqueta 'NIVEL:' el nivel del TEXTO (A1, A2, B1, B2 o C1).

TEXTO: "..."

With this prompt, brief descriptive information has been added for each level, following the analysis of the corpus mentioned above. Likewise, following the Instituto Cervantes Curriculum Plan (Instituto Cervantes, 2006), a description of the verb

¹ English version of the prompt in Appendix 2

tenses used in each of the levels and the types of texts has been included, following the indications of the textual products. In this way, the model is provided with a broader context so that its response is more accurate and tailored to the needs requested.

Testing with the ChatGPT Language Model

A total of three different tests were carried out in the study to assess the model's ability to assess the level of the texts in the corpus.

The first test was zero-shot learning. In this process, the model does not receive specific examples, but rather relies on its prior knowledge. To carry out this test, only the prompt mentioned in the previous section was used.

In tests two and three, fine-tuning was performed. This process consists of specialising a pre-trained model to perform a specific task, adapting it to a specific set of data provided to it. We obtained this dataset from the CAES corpus, given that it provides with a set of clear examples for the model. To perform fine-tuning, the model is equipped with an input-output specifying the input and the type of output we want it to provide. In this case, the prompt is the one mentioned in the previous section. Furthermore, the output requested was the level of Spanish. Likewise, to perform fine-tuning, part of the dataset was reserved to verify the response.

Tests 2 and 3 differ from each other in the division of the corpus. For the second test, the corpus was divided into 90%, 5% and 5%. Ninety per cent of the corpus was used to train the model, 5% to validate it and the remaining 5% to test it. In the third test, an 80%–20% division was made, using 80% of the corpus for training and 20% for testing.

RESULTS

To analyse the results, three evaluation measures widely used in classification tasks have been employed:

- **Precision:** indicates the percentage of examples that the model classified at a certain level and that actually belong to that level.
- **Recall:** indicates the proportion of examples of a specific level that the model correctly identified. For example, the percentage of A1-level texts detected as A1 out of the total number of A1 texts in the corpus.
- **F1-score:** A single value that combines accuracy and coverage through its harmonic mean, providing a balanced measure of performance. This metric is particularly useful when it is important for the model not only to be correct, but also to detect all possible cases in each category.

The results of the three experiments are presented below:

Zero-shot learning experimentation

As mentioned above, in this technique, the model is not provided with any examples; it is carried out using the prompt developed. In this case, the results obtained are as follows:

Table 2
ZSL experimental results

Tags	Precision	Recall	F1-score
A1	0.9375	0.1402	0.2439
A2	0.3333	0.7677	0.4648
B1	0.2400	0.2609	0.2500
B2	0.4286	0.8110	0.1364
C1	0.0000	0.0000	0.0000

Table 2 shows that, at level A1, accuracy is high, i.e. texts classified as level A1 have a high probability of being at this level. However, coverage is quite low, as the model has detected only a small percentage of texts in the corpus as A1. Therefore, the F1 score is low.

At level A2, the opposite phenomenon to that described at level A1 occurs. Accuracy is low, so it is less successful, but coverage is high. Therefore, we can say that at this level, although it detects A2 texts to a greater extent, it has low accuracy when it comes to detecting the correct level.

In relation to level B1, both factors, precision and coverage, are low. At this level, the model has problems both detecting B1-level texts and identifying them at their correct level.

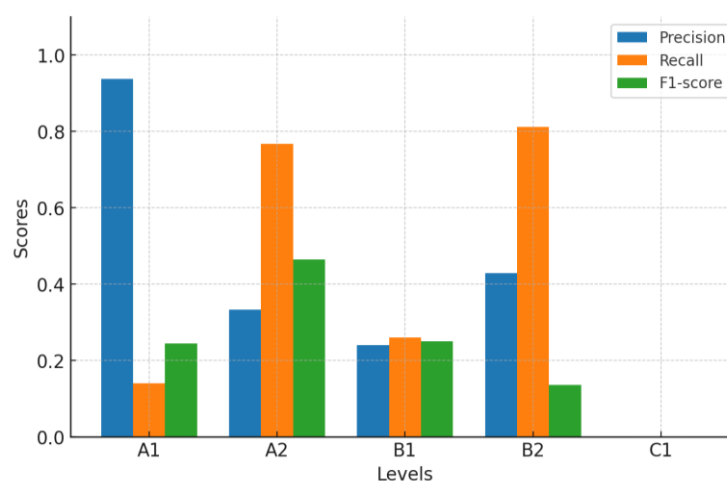
With regard to level B2, coverage is high, detecting most B2-level texts, but precision is moderate, with most classifications being incorrect.

Finally, the case of level C1 is striking, as both accuracy and coverage are 0; it does not detect any texts at this level.

Furthermore, as this was an experiment using a prompt, the model not only gave the level of the text in its responses, but also added comments on each one about the errors found. Complete examples of these responses, including the original texts and the corrections proposed by the model, are presented in Appendix 1.

Figure 1 complements this information by showing a comparison of the accuracy, coverage and F1-score values for each level evaluated in the zero-shot learning configuration.

Figure 1
Results by level in zero-shot learning experimentation



Fine-tuning experimentation 90-5-5

In this second experiment, as mentioned above, the corpus was fine-tuned and divided into 90%, 5% and 5% for training, validation and testing of the model. The results of this experiment are shown below:

Table 3

Fine-tuning experimentation 90-5-5

Tags	Precision	Recall	F1-score
A1	0.9905	0.9720	0.9811
A2	0.9519	1.0000	0.9754
B1	1.0000	0.9565	0.9778
B2	1.0000	1.0000	1.0000
C1	1.0000	1.0000	1.0000

Table 3 shows that the results have higher levels of accuracy and coverage than those obtained with zero-shot learning. At level A1, we can see that the model makes almost no errors and detects almost all texts at level A1.

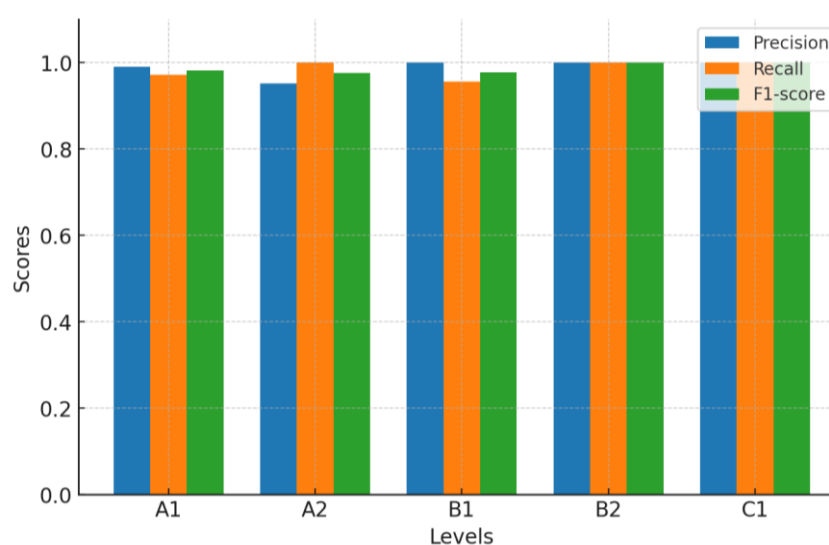
At level A2, as we can see, the model detects all texts at level A2 with very high accuracy.

With regard to level B1, the model correctly predicts all texts and also has high coverage, with a small percentage undetected.

Finally, with levels B2 and C1, the results show that the model detects all texts at these levels and is correct on all occasions. However, the results of this experiment show that the model can correctly predict all levels with values close to or equal to 1. Figure 2 complements this information by showing a comparison of the accuracy, coverage and F1-score values for each level evaluated in the 90-5-5 fine-tuning configuration.

Figure 2

Results by level in the 90-5-5 fine-tuning experiment



Fine-tuning experimentation 80-20

In the third experiment, fine-tuning was performed by dividing the corpus into two percentages, 80% for training and 20% for model validation. The results obtained are shown below.

In this experiment, as shown in Table 4, at level A1, both precision and coverage show a high level of accuracy and text detection.

Levels A2 and B1 show a similar result to A1, although with a slightly lower result in accuracy at level A2 and coverage at level B1.

With regard to levels B2 and C1, we observe that the accuracy of the model is excellent, as it correctly predicts these levels, although the coverage is moderately lower.

Table 4

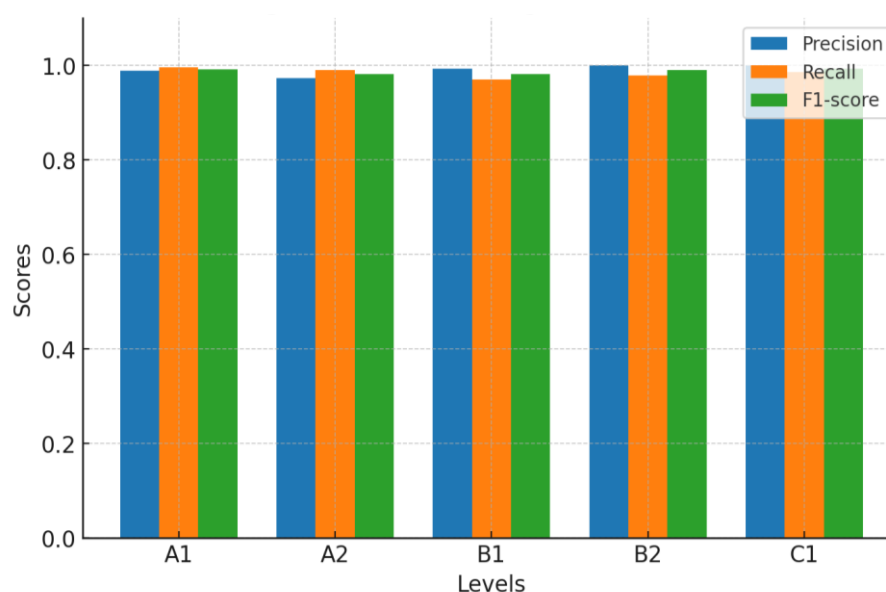
Fine-tuning experimentation 80-20

Tags	Precision	Recall	F1-score
A1	0.9884	0.9953	0.9918
A2	0.9727	0.9899	0.9812
B1	0.9925	0.9707	0.9815
B2	1.0000	0.9795	0.9896
C1	1.0000	0.9859	0.9929

Figure 3 complements this information by showing a comparison of the precision, coverage, and F1-score values for each level evaluated in the 80-20 fine-tuning configuration.

Figure 3

Results by level in the 80-20 fine-tuning experiment



Statistical analysis

To assess the statistical significance of the results, 95% confidence intervals (95% CI) were estimated for the Macro-F1 values using **bootstrapping**, a resampling technique with replacement that allows confidence intervals to be calculated without assuming a specific statistical distribution. One thousand replicates were performed to obtain these intervals, which indicate the range within which the true value of Macro-F1 is expected to fall with a 95% probability. Additionally, hypothesis tests were performed to verify the results. The zero-shot learning experiment obtained a Macro-F1 value of 0.2190 with a confidence interval $CI_{95\%} = [0.1785, 0.2617]$, while the fine-tuning experiments achieved values close to perfection: a Macro-F1 value of 0.9869 with a confidence interval $CI_{95\%} = [0.9748, 0.9959]$ on the 90-5-5 split and a Macro-F1 value of 0.9874 (95% CI = $[0.9812, 0.9933]$) on the 80-20 split. The comparisons were made using non-parametric bootstrap-based hypothesis tests—i.e., statistical contrasts that do not require assuming a specific data distribution and are based on multiple random resamples. These analyses showed that both fine-tuning experiments statistically significantly outperformed the zero-shot experiment ($\Delta^1 \approx 0.77$; $p < 0.001$ in both cases). In contrast, no significant differences were observed between the two fine-tuning experiments ($\Delta = -0.0005$; 95% CI = $[-0.0133, 0.0113]$; $p = 0.964$). These results confirm that fine-tuning substantially improves the model's ability to assess written competence in ELE.

PEDAGOGICAL RELEVANCE OF THE MODEL AND ITS APPLICATION IN BOTH FACE-TO-FACE AND DISTANCE LEARNING CONTEXTS

The results obtained show that a language model adjusted with a specialised corpus can achieve very high performance in the automatic classification of written texts according to CEFR levels. This capability has clear potential to facilitate teaching, especially in the initial stage of student placement, constituting a pedagogical innovation in the use of artificial intelligence for language teaching.

In a real learning environment, the system could be integrated into a Learning Management System (LMS) as a diagnostic assessment module. The flow of use would be simple: the student submits a text, the model automatically classifies it by level and presents the result to the teacher using a rubric aligned with the descriptors of the Instituto Cervantes Curriculum Plan (PCIC). In this way, teachers could assign students to the course corresponding to their actual level, optimising the adjustment of groups and avoiding gaps that could affect learning progress.

In future implementations, and based on more exhaustive analyses, the system could generate more detailed reports based on the PCIC, identifying specific linguistic areas to be reinforced (e.g. use of verb tenses, textual cohesion or lexical repertoire). This information would allow the course to be tailored not only to the student's level, but also to their main weaknesses, guiding the teaching programme towards improving these aspects.

In distance education and online learning contexts, integration into an LMS would have the same diagnostic function, automatically classifying students from their first access. This would allow for efficient group organisation even in non-face-to-face environments, which is especially relevant in massive courses or virtual programmes with continuous enrolment.

Compared to tools such as Write & Improve (Cambridge English, n.d.), which focus mainly on corrective feedback and formative assessment, the proposal presented here offers a complementary approach: automatic classification by CEFR levels. This feature, combined with the possibility of direct integration into educational platforms, reinforces its value as a teaching support tool for optimising initial placement and streamlining course planning.

In all cases, the system is designed as a teaching aid and not as a replacement, with the aim of streamlining initial assessment tasks and freeing up time for higher value-added activities, such as individualised feedback or progress monitoring.

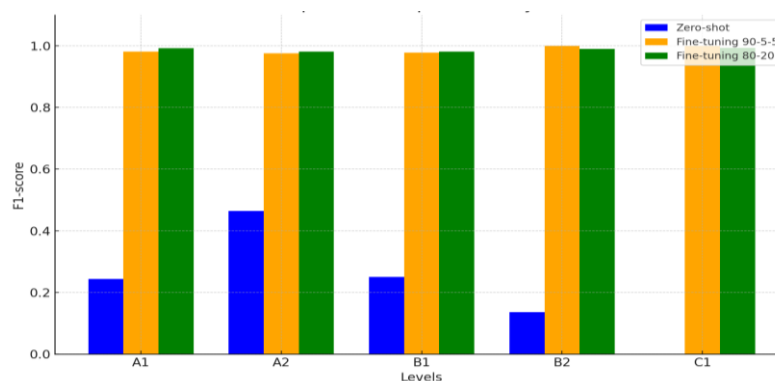
Among the possible future improvements are expanding the corpus to include more genres and topics, adding additional lexical and pragmatic descriptors, experimenting with other language models, and adapting the system to other languages and educational levels. It would also be relevant to evaluate its real impact on student performance and motivation, as well as its smooth integration into different LMSs and educational contexts.

DISCUSSION AND CONCLUSIONS

Once the tests have been carried out and the results obtained, there is a clear contrast between the model's performance in the zero-shot learning configuration and in the fine-tuning experiments. In the former, the model does not achieve adequate accuracy at any of the levels, with only a relatively high value in A1, albeit with very low coverage. At levels A2 and B2, coverage is medium-high (0.76 and 0.81), indicating that the model detects most of the texts at those levels, but with low accuracy. Level C1 is particularly problematic, with zero values for both accuracy and coverage. These data confirm that, with its original pre-training, the model is not capable of reliably assessing the level of ELE texts according to the Instituto Cervantes Curriculum Plan.

In contrast, the two fine-tuning tests (90-5-5 and 80-20) provided accuracy and coverage values very close to 1 at all levels, with particular strength at B2 and C1. The comparison between the two experiments suggests that a larger volume of training data can improve performance, although even with less data the model maintains a high predictive capacity. This behaviour coincides with that observed in previous work on model adaptation for specific natural language processing tasks ((García-Peñalvo, 2024; García-Peñalvo et al., 2024), where customisation of the system significantly increases its effectiveness.

Figure 4
Comparison of experimental results



Compared to tools such as Write & Improve (Cambridge English, n.d.), which focus on formative feedback and error detection, this study offers a complementary approach: automatic classification by CEFR levels, which could be integrated into learning environments to optimise the initial placement of students and streamline teaching. Likewise, studies such as those by Burstein et al. (2003, e-rater) and McNamara et al. (2014, Coh-Metrix) had already shown the usefulness of combining linguistic features and automatic metrics to evaluate writing; our results reinforce this line of thinking, showing that a model adjusted with specific corpora achieves very high performance levels.

On the other hand, in the zero-shot learning experiment, the model not only indicated the texts' level but also generated comments on errors and suggestions for correcting the texts. Although this functionality is not relevant to the main objective of this study, it could be explored in future research as a teaching support resource to help students identify and correct their errors.

In terms of limitations, the study was conducted on a single corpus due to the scarcity of specialised and level-assessed ELE corpora that meet homogeneous criteria and are managed by experts. Furthermore, the classification was based on grammatical descriptors present in the prompt (e.g., verbal repertoire, average number of words per sentence, use of certain tenses and modes), without systematically integrating complex syntactic structures, specific lexical resources, or pragmatic aspects such as appropriateness or discursive coherence, which are particularly relevant at intermediate and advanced levels.

As lines of future work, we propose:

1. Design and validate new corpora for Spanish as a foreign language, level-assessed by experts, with greater thematic diversity and representativeness of levels.
2. Explore the application of the method with other language models to compare their performance.
3. Expand the set of linguistic descriptors, incorporating syntactic, lexical, and pragmatic indicators.
4. Experiment with controlled variations of the prompt to evaluate their impact on classification.
5. Validate the system with authentic student texts in real educational contexts and integrate the tool into learning management systems (LMS).

Overall, the results achieved confirm that generative artificial intelligence, and in particular language models adjusted with specific data, can become an effective resource for streamlining the initial level assessment of ELE students, supporting teaching work, and optimizing teaching-learning processes.

Within this reflection, it is also pertinent to address the ethical and legal aspects involved in the use of learner corpora, which are discussed in the following subsection.

Ethical and licensing issues

This study used the Corpus of Spanish Learners (CAES), which is available online for academic and research purposes (University of Santiago de Compostela, n.d.). Its texts have been anonymised beforehand so that they do not include any identifiable personal data, and they were compiled in regulated educational contexts

and with expert validation, which guarantees the appropriate treatment of the information.

From an ethical point of view, it is necessary to consider that the expansion or combination of the corpus with other resources must address both clarity in usage licences and the prevention of biases associated with mother tongue or sociocultural context. These factors are essential to ensure fair and reproducible classification of texts and, ultimately, to guarantee the responsible use of artificial intelligence in educational environments.

Acknowledgements

This work is funded by the Ministerio para la Transformación Digital y de la Función Pública and Plan de Recuperación, Transformación y Resiliencia - Funded by EU – NextGenerationEU within the framework of the project Desarrollo Modelos ALIA. This work has also been partially supported by Project CONSENSO (PID2021-122263OB-C21), Project MODERATES (TED2021-130145B-I00) and Project SocialTox (PDC2022-133146-C21) funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR, Project ROMANET (CERV-2024-CHAR-LITI-101215052), funded by the European Union under the Citizens, Equality, Rights and Values programme, and Project HEART-NLP-UJA (PID2024-156263OB-C21) funded by MICIU/AEI/10.13039/501100011033 and by ERDF/EU. The research work conducted by Salud María Jiménez-Zafra is part of the grant RYC2023-044481-I, supported by MICIU/AEI/10.13039/501100011033 and by ESF+.

NOTES

1. To quantify the differences between models, the symbol Δ (delta) was used, which is conventionally employed in statistics to indicate the difference between two values. Thus, $\Delta = \text{Macro-F1}(\text{fine-tuning}) - \text{Macro-F1}(\text{zero-shot}) \approx 0.77$ means that the difference in Macro-F1 between the experiments being compared is approximately 0.77.

REFERENCES

- Aparicio Gómez, W. O. (2023). La inteligencia artificial y su incidencia en la educación: transformando el aprendizaje para el siglo XXI. *Revista Internacional de Pedagogía e Innovación Educativa*, 3(2), 217-229. <https://dialnet.unirioja.es/servlet/articulo?codigo=9624350>
- Area-Moreira, M., Del Prete, A., Sanabria-Mesa, A. L., & Sannicolás-Santos, M. B. (2024). No todas las herramientas de IA son iguales: análisis de aplicaciones inteligentes para la enseñanza universitaria. *Digital Education Review*, 45, 141-149. <https://doi.org/10.1344/der.2024.45.141-149>
- Barroso-Osuna, J., & Cabero-Almenara, J. (2025). Potencialidades de la inteligencia artificial en la personalización de la educación. In P. Román-Graván, J. Barroso-Osuna, J. Cabero-Almenara, & C. Llorente-Cejudo (Eds.), *Visiones sobre la integración educativa de la inteligencia artificial* (1st ed.). Dykinson. <https://doi.org/10.14679/4177>
- Baskara, R., & Mukarto, M. (2023). Exploring the implications of ChatGPT for language learning in higher education. *Indonesian Journal of English Language Teaching and Applied Linguistics*, 7(2), 343-358. <https://doi.org/10.21093/ijeltal.v7i2.1387>

- Biedma Torrecillas, A., Chamorro Guerrero, M. D., Lozano, G., & Sánchez Cuadrado, A. (2012). Diseño y validación de las pruebas de nivel del CLM de la Universidad de Granada. In *Actas del VII Congreso ACLES: Multilingüismo en los centros de lengua universitarios: evaluación, acreditación, calidad y política lingüística* (pp. 26-37). ACLES. <https://dialnet.unirioja.es/servlet/libro?codigo=501925>
- Bolaño-García, M., & Duarte-Acosta, N. (2024). Una revisión sistemática del uso de la inteligencia artificial en la educación. *Revista Colombiana de Cirugía*, 39(1), 51-63. <https://doi.org/10.30944/20117582.2365>
- Burstein, J., Elliot, N., Beigman Klebanov, B., Madnani, N., Napolitano, D., Schwartz, M., Houghton, P., & Molloy, H. (2018). Writing mentor: Writing progress using self-regulated writing support. *Journal of Writing Analytics*, 2, 285-313. <https://doi.org/10.37514/JWA-J.2018.2.1.12>
- Burstein, J., Kukich, K., Wolff, S., Lu, C., Chodorow, M., Braden-Harder, L., & Harris, M. D. (2003). E-rater as a diagnostic tool for writing instruction. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Demonstrations* (pp. 79-81). Association for Computational Linguistics.
- Cambridge English. (n. d.). *Write & Improve*. <https://writeandimprove.com>
- Cantero, M. V. (2024). Aproximación a un posible uso de ChatGPT para nivelar la expresión escrita en ELE. In F. M. Sirignano, R. Martínez Roig, & A. López Padrón (Eds.), *Enseñanza y aprendizaje en la era digital desde la investigación y la innovación* (pp. 55-64). Octaedro.
- Centro Virtual Cervantes. (n. d.). *Ítem de respuesta cerrada*. https://cvc.cervantes.es/ensenanza/biblioteca_ele/diccio_ele/diccionario/itemrespuestacerrada.htm
- Chan, C. K. Y., & Tsi, L. H. Y. (2023). *The AI revolution in education: Will AI replace or assist teachers in higher education* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2305.01185>
- Columbia University, Department of Latin American and Iberian Cultures. (n. d.). *Spanish placement exam*. Recuperado el 25 de julio de 2025, de <https://laic.columbia.edu/content/spanish-second-language-placement-exam>
- Council of Europe. (2002). *Marco común europeo de referencia para las lenguas: aprendizaje, enseñanza, evaluación*. Instituto Cervantes; Ministerio de Educación, Cultura y Deporte. https://cvc.cervantes.es/ensenanza/biblioteca_ele/marco/cvc_mer.pdf
- Crespo Mendoza, R., Rodríguez López, W., Montenegro Patrel, M., & Tomalá Tomalá, G. (2024). IA: una herramienta para asistir a los docentes en la evaluación de los estudiantes. *Conocimiento Global*, 9(2), 305-323. <https://doi.org/10.70165/cglobal.v9i2.423>
- Fajardo, G. M., Ayala, D. C., Arroba, E. M., & López, M. (2023). Inteligencia artificial y la educación universitaria: una revisión sistemática. *Magazine de las Ciencias: Revista de Investigación e Innovación*, 8(1), 109-131. <https://doi.org/10.33262/rmc.v8i1.2935>
- García-Peñalvo, F. J. (2024). Cómo afecta la inteligencia artificial generativa a los procesos de evaluación. *Cuadernos de Pedagogía*, (549).
- García-Peñalvo, F. J., Llorens-Largo, F., & Vidal, J. (2024). La nueva realidad de la educación ante los avances de la inteligencia artificial generativa. *RIED-Revista Iberoamericana de Educación a Distancia*, 27(1), 9-39. <https://doi.org/10.5944/ried.27.1.37716>
- Hernández-León, N., & Rodríguez-Conde, M. J. (2024). Inteligencia artificial aplicada a la educación y la evaluación educativa en la universidad: Introducción de sistemas de tutorización inteligentes, sistemas de reconocimiento y otras tendencias futuras. *Revista de Educación a Distancia (RED)*, 24(78), Artículo 6. <https://doi.org/10.6018/red.594651>
- Hong, W. C. H. (2023). *The impact of ChatGPT on foreign language teaching and learning: Opportunities in education*

- and research. *Journal of Educational Technology and Innovation*, 5(1), 38-53. <https://doi.org/10.61414/jeti.v5i1.103>
- Instituto Cervantes. (2006). *Plan curricular del Instituto Cervantes: Niveles de referencia para el español* (3 vols.). Biblioteca Nueva. https://cvc.cervantes.es/ensenanza/biblioteca_ele/plan_curricular/
- Li, Y. (2023). *A practical survey on zero-shot prompt design for in-context learning* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2309.13205>
- McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511894664>
- Morales-Chan, M. A. (2023). *Explorando el potencial de ChatGPT: Una clasificación de prompts efectivos para la enseñanza*. Universidad Galileo. <https://biblioteca.galileo.edu/tesario/handle/123456789/1348>
- Moreno, R. D. (2019). La llegada de la inteligencia artificial a la educación. *Revista de Investigación en Tecnologías de la Información*, 7(14), 260-270. <https://doi.org/10.36825/riti.07.14.022>
- OpenAI. (2022). *ChatGPT (versión 3.5) [Artificial intelligence language model]*. <https://openai.com>
- Owan, V. J., Abang, K. B., Idika, D. O., Etta, E. O., & Bassey, B. A. (2023). Exploring the potential of artificial intelligence tools in educational measurement and assessment. *EURASIA Journal of Mathematics, Science and Technology Education*, 19(8), em2307. <https://doi.org/10.29333/ejmste/13428>
- Palacios Martínez, I., Barcala Rodríguez, F. M., & Rojo, G. (2019). El corpus de aprendices de español (CAES) y sus aplicaciones para la enseñanza y aprendizaje del español como lengua extranjera. In M. Blanco, H. Olbertz, & V. Vázquez Rozas (Eds.), *Corpus y construcciones: Perspectivas hispánicas* (pp. 273-301). Universidade de Santiago de Compostela (Verba, Anexo 79). <https://doi.org/10.15304/9788417595876>
- Pourpanah, F., Abdar, M., Luo, Y., Zhou, X., Wang, R., & Lim, C. P. (2023). A review of generalized zero-shot learning methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4), 4051-4070. <https://doi.org/10.1109/TPAMI.2022.3182926>
- Roumeliotis, K. I., Tselikas, N. D., & Nasiopoulos, D. K. (2024). Next-generation spam filtering: Comparative fine-tuning of LLMs, NLPs, and CNN models for email spam classification. *Electronics*, 13(11), 2034. <https://doi.org/10.3390/electronics13112034>
- Salguero Romero, P. (2023). *La traducció pedagògica i l'ús de ChatGPT-3 a classes d'anglès com a segona llengua per a nens i nenes* [Trabajo de fin de grado, Universitat Autònoma de Barcelona]. Repositorio UAB. <https://ddd.uab.cat/record/279383>
- Universidad de Santiago de Compostela. (n. d.). *Corpus de aprendices de español (CAES)*. <https://galvan.usc.es/caes>
- University of Wisconsin–Madison, Testing and Evaluation Services. (n. d.). *Spanish placement test*. University of Wisconsin–Madison. <https://testing.wisc.edu/centerpages/spanishtest.html>
- Wang, L. (2024). *El uso de ChatGPT y Gemini para la preparación de actividades de comprensión oral de CLE: Un estudio comparativo* [Trabajo de fin de máster, Universitat Autònoma de Barcelona]. Repositorio UAB. <https://ddd.uab.cat/record/300353>
- Wang, X., Lee, Y., & Park, J. (2022). *Automated evaluation for student argumentative writing: A survey* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2205.04083>
- Zambrano Campo, R. F. (2025). Impacto de la inteligencia artificial en la transformación digital de la educación. *Serie Científica de la Universidad de las Ciencias Informáticas*, 18(1), 177-192. <https://scielo.sld.cu/scielo.php?script=s>

[ci_arttext&pid=S2306-24952025000100177](#)

Zhang, S. (2021). Review of automated writing evaluation systems. *Journal of China Computer-Assisted Language Learning*, 1(1), 170-176. <https://doi.org/10.1515/jccall-2021-2007>

Zapata Ros, M. (2024). IA generativa y ChatGPT en educación: Un reto para la evaluación y ¿una nueva pedagogía? *Revista Paraguaya de Educación a Distancia (REPED)*, 5(1), 12-44. <https://doi.org/10.56152/reped2024-vol5num1-art2>

APPENDICES

Appendix 1. Complete examples of model responses.

Corpus text:

"Hola Eva.

Que hace dos meses, hemos estudiado espanol en madrid.

He vuelt a Japon la semana pasada.

Glacias por tu ayuda.

Yo escribo mi familia, mi mujer, hija y hijo.

Mi mujer trabaya en Idioma.

Hija trabaya en una tienda de palo, hijo en un enterprise electrolico.

Vivimos en yokohama cerca de tokyo.

Vamos a ir a madrid que mes viene

Por favor, vernos en madrid.

Hasta plonto." (CAES, 2022)

Model response:

Errores encontrados en el texto:

- "Que hace dos meses, hemos estudiado espanol en madrid." La forma correcta sería "Hace dos meses, estudiamos español en Madrid."

- "He vuelt a Japon la semana pasada." La forma correcta sería "He vuelto a Japón la semana pasada."

- "Glacias por tu ayuda." Debería ser "Gracias por tu ayuda."

- "Yo escribo mi familia, mi mujer, hija y hijo." Se debería decir "Te escribo sobre mi familia: mi mujer, mi hija y mi hijo."

- "Mi mujer trabaya en Idioma." La forma correcta sería "Mi mujer trabaja en Idiomas."

- "Hija trabaya en una tienda de palo, hijo en un enterprise electrolico." Debería ser "Mi hija trabaja en una tienda de ropa, y mi hijo en una empresa electrónica."

- "Vamos a ir a madrid que mes viene." La forma correcta sería "Vamos a ir a Madrid el mes que viene."

- "Hasta plonto." Debería ser "Hasta pronto."

NIVEL: A2 (OpenAI, 2022)

—

Corpus text:

"El 24 de Noviembre

Estimados señores

El 22 de noviembre pasado, viajé con Tonta Airlines de Badaros a Madrid con el vuelo TA 222.

Estoy aquí en Madrid hasta el 27 para algunas citas profesionales.

Cuando embarqué en Badaros, la azafata me dijo que mi maleta estaba demasiado grande y que tenía que facturarla.

La maleta esta de marca Samsonite, de modelo XYZ1234, color negra, con mis nombre y dirección muy visibles.

Cuando llegamos a Madrid, la maleta no fue encontrada.

Hace a dos días que estoy esperando pero nadie de su compañía me ha llamado con información.

Evidentemente, no puedo quedar sin cambiarme o lavarme me.

Entonces compré algunas cosas y me parece que Tonta Airlines tenga que indemnizarme.

Yo incluido la lista de mis compras y les pide que se pongan en contacto para decirme si sean de acuerdo de pagar el 50% des estas compras.

Se pueden contactarme por correo electronico o llamarme en el movil 660 12 34 45 y evidentemente, en cualquier caso que Tonta Airlines encuentre mi maleta, me daría mucha alegría.

Un saludo” (CAES, 2022)

Model response:

El texto presenta varios errores gramaticales y algunos elementos propios de un nivel de español como lengua extranjera B1:

- Uso incorrecto de tiempos verbales: "tenia" en lugar de "tenía".
- Errores en el uso de los artículos: "la maleta esta" en lugar de "la maleta está".
- Uso incorrecto de pronombres: "Hace a dos días" en lugar de "Hace dos días".
- Errores de concordancia de género y número: "me parece que Tonta Airlines tenga que indemnizarme" debería ser "me parece que Tonta Airlines debe indemnizarme".

NIVEL: B1 (OpenAI, 2022)

Appendix 2. English version prompt.

You are a linguist expert specializing in teaching Spanish as a foreign language. Your task is to indicate the level of Spanish as a foreign language in the texts, following the Instituto Cervantes Curriculum Plan.

Here is a description of the different levels.

Levels A1 and A2 Basic transactions related to their environment.

A1: Limited vocabulary, very short and simple texts, an average of 10 words per sentence. Regular forms of the present indicative.

A2: Short texts with simple information, an average of 12 words per sentence. Past indicative tenses: present perfect, imperfect, and indefinite. Irregular forms of the present indicative. Affirmative imperative.

Levels B1 and B2: ability to handle texts on topics of interest, tastes, and preferences.

B1: Extensive but simple vocabulary, writing texts with a specific task. Present indicative, past perfect, imperfect, and indefinite indicative, simple future, simple conditional, past perfect indicative, present subjunctive. Negative imperative.

B2: Extensive linguistic repertoire, subordinate clauses. Indicative verb tenses: present, past perfect, imperfect, indefinite, simple and compound future, simple and compound conditional, past perfect. Subjunctive verb tenses: present, imperfect, past perfect and past perfect.

C1 transactions of all kinds. They have a sufficiently broad and rich repertoire of linguistic and non-linguistic resources. They can deal with a wide range of long and complex texts. All indicative and subjunctive verb tenses: present, past perfect, imperfect, and pluperfect.

Now you are going to receive a TEXT. Taking into account the above and the grammatical errors, indicate at the end of your answer with the label 'LEVEL:' the level of the TEXT (A1, A2, B1, B2, or C1).

TEXT: “...”

Date of reception: 1 June 2025

Date of acceptance: 9 September 2025

Date of approval for layout: 8 October 2025

Date of publication in OnlineFirst: 28 October 2025

Date of publication: 1 January 2026

Participatory methodology for creating microlearning digital resources with generative AI

Metodología participativa para la creación de recursos digitales de microaprendizaje con inteligencia artificial generativa



 Bell Manrique-Losada - *Universidad de Medellín, UdeMedellín (Colombia)*
 Sandra Isabel Arango-Vásquez - *Universidad de Medellín, UdeMedellín (Colombia)*
 Mariutsi Alexandra Osorio-Sanabria - *Instituto Tecnológico Metropolitano, ITM (Colombia)*

ABSTRACT

In recent years, advances in information technology and the Internet have transformed education by creating new opportunities for learning and collaboration in the production of Digital Educational Resources (DERs). Both students and teachers are leveraging the potential of models, methods, and approaches derived from Generative Artificial Intelligence (GenAI). The study analyzes and describes a microlearning-based methodology to guide interaction between actors and support the co-creation of Microlearning Resources (abbreviated microDERs) using GenAI. Adopting a qualitative approach, the study integrates Participatory Action Research (PAR) and the Creative Social Problem Solving (CSPS) methodology. It unfolds in three phases: (1) recognizing the context and conducting exploratory, participatory, thematic delimitation; (2) collaboratively designing microDERs with high school and higher education students and teachers, including the evaluation of GenAI tools; and (3) validating the methodology through co-creation workshops, bootcamps, and qualitative analysis. Data are collected and analyzed through co-creation workshops with students and teachers from elementary, high school, and higher education. Findings on the contributions of the actors in each phase indicate that combining PAR and CSPS improves interaction and knowledge building in the co-creation of DERs using GenAI.

Keywords: education; learning; artificial intelligence; information technology.

RESUMEN

En los últimos años, los avances en las Tecnologías de la Información (TI) e Internet han transformado el ámbito educativo. Estos avances han creado nuevas oportunidades de aprendizaje e interacción colaborativa en la producción de recursos educativos digitales. Tanto estudiantes como docentes reconocen esta oportunidad y están aprovechando el potencial de los modelos, métodos, TI y enfoques derivados de la inteligencia artificial generativa (IAGen). Este estudio analiza y describe el proceso de implementación de una metodología basada en principios de microaprendizaje para guiar la interacción entre actores y lograr la cocreación de recursos de microaprendizaje (microRED) con IAGen. La investigación adopta una perspectiva cualitativa y se estructura metodológicamente desde la articulación de la Investigación Acción Participativa (IAP) y la Metodología de Solución Creativa de Problemas Sociales (SCPS). Este estudio se desarrolla en tres fases: 1) Reconocimiento del contexto y delimitación temática participativa exploratoria; 2) Diseño colaborativo de microRED con estudiantes y docentes de secundaria y educación superior, integrando retos y evaluación de herramientas de IAGen; y, 3) Validación de la metodología propuesta a través de talleres de cocreación, *bootcamp* y análisis cualitativo. La recopilación y el análisis cualitativo de los datos se realiza a través de talleres de cocreación con estudiantes y docentes de básica secundaria y educación superior. Como resultado del análisis de las contribuciones de los actores en cada fase, se concluye que los enfoques de IAP y SCPS mejoran la interacción y la generación de conocimiento asociado al uso de la IAGen para la cocreación de RED.

Palabras clave: educación; aprendizaje; inteligencia artificial; tecnología de la información.

INTRODUCTION

Reviewing educational strategies to improve student learning outcomes is essential for promoting creativity and critical thinking. In this regard, microlearning—which provides brief and accessible content through Digital Educational Resources (DERs)—emerges as a valuable alternative (Denojean-Mairet et al., 2024; Torgerson & Iannone, 2019).

The integration of Generative Artificial Intelligence (GenAI) further enhances its impact by enabling the personalization and creation of materials tailored to specific contexts (Miao & Holmes, 2024; Piantari et al., 2024). At the same time, the co-creation of DERs between students and teachers fosters autonomy and the social appropriation of knowledge, which is the process by which communities produce, use, and apply knowledge. Consequently, this strengthens democratization and supports social transformation (Gonnet, 2019).

This study analyzes the implementation of a participatory methodology in educational contexts to facilitate the co-creation of Microlearning Resources (abbreviated microDERs in Spanish) using GenAI tools. The research was carried out in three phases: (1) recognizing the context and mapping of actors, needs, challenges, and knowledge; (2) collaboratively designing microDERs with high school (10th and 11th grades) and higher education (9th semester) students, integrating challenges and evaluating GenAI tools; and (3) validating the methodology through co-creation workshops, bootcamps, and qualitative analysis.

Adopting the Participatory Action Research (PAR) approach and the Creative Social Problem Solving (CSPS) method, the study involves students and teachers in reflective and participatory cycles to collaboratively build knowledge (Kemmis & McTaggart, 2005). These approaches promote knowledge exchange and community participation (Ritzer, 2008), as well as ensure that microDERs co-creation responds to the needs and perspectives of the participants.

The article is organized into four sections. First, the conceptual framework section focuses on the fundamentals of PAR, CSPS, microlearning, and GenAI. Next, the methodology section outlines the three phases of the microDERs co-creation process. Then, the results section presents the activities, actors, and artifacts from each phase. Lastly, the last section contains discussion, conclusions, and future work.

CONCEPTUAL FRAMEWORK

Participatory Action Research (PAR)

PAR is a methodology that combines theory, action, and participation to understand social phenomena. It adapts to regional contexts to foster the collective building of knowledge (Fals Borda, 2001; Sirvent & Rigal, 2012). PAR views individuals as agents capable of transforming their realities through collaborative research, not as objects of study (Fals Borda, 2001). Furthermore, it strengthens social ownership and supports the transition of social ecosystems by integrating democratic practices, knowledge, and systemic perspectives on problems (Arnanz et al., 2023).

Creative Social Problem Solving (CSPS)

CSPS is a participatory methodology based on *design thinking* that focuses on the collaborative design of contextualized educational and social solutions (Mirón, 2022). Through creativity, critical analysis, and co-design tools, CSPS fosters co-creation, critical thinking, and stakeholder empowerment for identifying problems, designing strategies, and assessing impact. When integrated with PAR, CSPS broadens diagnostic capacity and strengthens solution design in educational, community, and organizational contexts.

Generative Artificial Intelligence (GenAI)

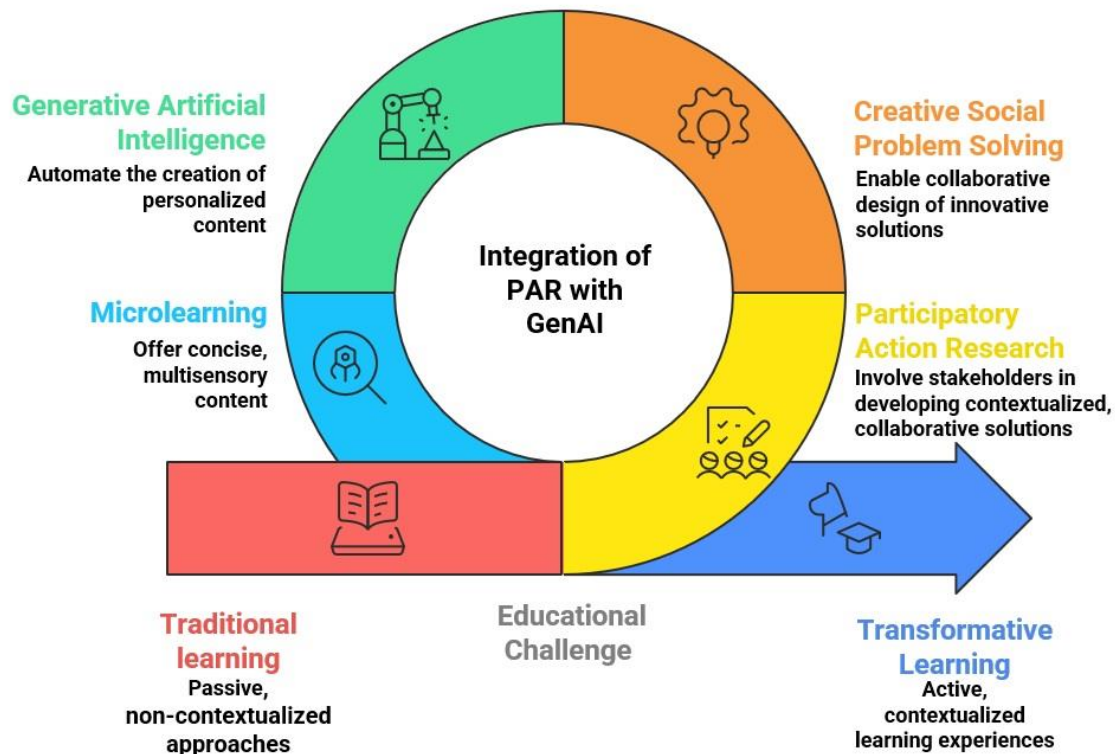
GenAI is a branch of Artificial Intelligence (AI) that uses deep neural networks and natural language processing to automate information processing and present results as symbolic representations of human thought. It produces semi-finished knowledge products in multiple formats (Cain, 2024; Cabeza-Rodríguez, 2025; Samoili et al., 2021). In education, GenAI enhances the creation of personalized teaching materials, increases teacher productivity in designing resources and learning environments, and supports learning experiences tailored to students' needs and interests (Abdous, 2023; Cain, 2024; Dickey & Bejarano, 2023).

Microlearning

Microlearning is an educational strategy that delivers short, focused digital content, incorporating multisensory and multimodal elements to facilitate comprehension and adaptability (Denojean-Mairet et al., 2024; Dolasinski & Reynolds, 2020; Durán & Escudero, 2023; Torgerson & Iannone, 2019). Microlearning structures DERs into seven dimensions—time, content, curriculum, form, process, mediality, and learning type (Hug et al., 2006)—and presents them in formats such as videos, texts, micropodcasts, and short social media messages (Díaz Redondo et al., 2021; Mateus-Nieves & Moreno, 2021).

Figure 1 illustrates the functional interaction of the four conceptual pillars—PAR, CSPS, GenAI, and microlearning—that underpin the proposed methodology. The process begins with defining a *process approach* based on educational challenges in the classroom. It then advances through stakeholder interaction and knowledge exchange within the frameworks of PAR and CSPS. Next, GenAI supports the creation of personalized content, which ultimately materializes as microlearning resources.

Figure 1
Functional interaction of the study's conceptual pillars



Source: Prepared by the authors. Software: Napkin.

Background and related studies

The rapid development of AI has led to more personalized and automated educational models, processes, and strategies, providing opportunities to transform pedagogy, resource development, and assessment. However, challenges remain regarding adaptation to individual characteristics (Yang, 2024), dynamic learning processes (Tu et al., 2025), empathy (Zhang et al., 2024), and literacy goals for students (Adeshola & Adepoju, 2023).

From a microlearning perspective, GenAI strengthens collaboration between teachers and technology. This collaboration supports the development of students' cognitive, psychomotor, and affective skills. It also positions students at the center of the educational process (Memon & Kwan, 2025; Bygstad et al., 2022). In parallel, advances have been made in generating adaptive microcontent using GenAI (Boumalek et al., 2024).

Within this context, two main lines of intervention emerge. The first redefines the teacher's role by shifting the focus from instructor-centered to student-centered learning and recognizing GenAI tools as integral to the process (Bozkurt, 2023). The second line leverages emerging technologies to design and deliver engaging, relevant content that aligns with the needs and interests of contemporary learners (Holmes et al., 2023; Karras et al., 2020; Luckin et al., 2016; Sengar et al., 2025). This study addresses both perspectives by proposing a collaborative model for creating microDERs through the integration and use of GenAI tools tailored to young students.

METHODOLOGY

This study takes a qualitative and participatory approach, combining PAR and CSPS. This convergence enables the addressing of complex problems in a situated, collaborative, and transformative manner across four dimensions: (1) promoting the social appropriation of knowledge by fostering collaboration between those experiencing the problem and end users of microDERs; (2) generating action-oriented outcomes, as knowledge is translated into artifacts and concrete community actions (e.g., building and using microDERs); (3) encouraging creativity and innovation in searching for solutions to identified challenges; and (4) strengthening communication and dialogue by encouraging open interaction, collaboration, and appropriation processes.

The research is guided by the hypothesis that implementing a microDERs co-creation methodology—with the participation of teachers and young students and supported by GenAI—fosters the development of computational thinking and social interaction. This addresses the central research question: *What is the best way to design a co-creation methodology that uses GenAI tools to help teachers and students build microDERs?*

Data collection relied on three instruments: (i) focus group guides, which were applied in co-creation workshops with students and teachers and validated through pilot processes; (ii) semi-structured questionnaires, which were used to gather perceptions about the methodology and the resulting products; and (iii) analysis matrices and research logs, which documented interactions, decisions, and adjustments. Content validity was ensured through participant feedback and cross-reviews by the research team.

The data analysis was conducted through a progressive process of open and axial coding, which shaped the emerging categories and revealed their connections. Comparative matrices were used to organize information, identify patterns, and connect findings with the research question. The process was participatory; interpretations were contrasted with the participants, and triangulation strategies reinforced coherence and transparency. Theoretical saturation was achieved in the final workshops when no new categories emerged, confirming the analysis consistency and solidity.

Triangulation and comparison further strengthened the analysis. The co-created microDERs, systematization matrices, and research logs were examined to verify the coherence of the emerging categories across sources. Additionally, interpretations were aligned with the conceptual frameworks of PAR, CSPS, GenAI, and microlearning. This ensured that the findings were theoretically supported and not solely dependent on group dynamics.

Participants

In line with Manzini's (2015) vision of collaborative design in social innovation, stakeholders from various educational levels were involved, including the research team, high school teachers and students in grades 10 and 11, as well as higher education teachers and ninth-semester students affiliated with a research incubator at the University of Medellín in Colombia. The study relied on a diverse and representative purposive sample. Table 1 shows the educational level, stakeholder, age range, and number of participants involved in designing and validating the methodology.

Table 1
Study participant demographics

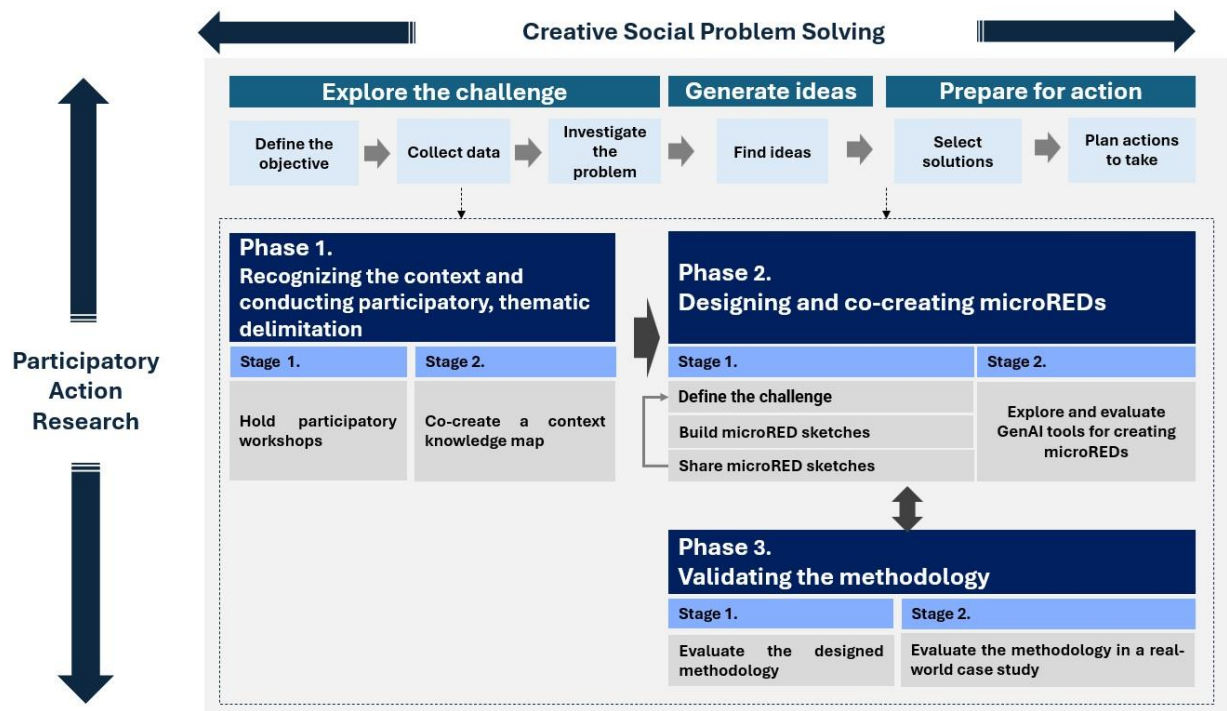
Educational level	Actor	Age range	Participants 2024-2	Participants 2025-1
High school	Students (Grades 10 and 11)	13 – 16	33	32
	Teachers	40 – 50	2	2
Higher education	Ninth-semester students	18 – 22	13	19
	Research incubator students	18 – 22	8	5

Source: Data collected by the authors.

Procedure

The research was conducted in three phases: (1) recognizing the context and delimitating the theme based on PAR; (2) designing and co-creating microDERs, and (3) validating the methodology. Figure 2 illustrates these phases and their methodological articulation.

Figure 2
Methodological process



Source: Prepared by the authors.

Phase 1: Recognizing the context and conducting participatory, thematic delimitation

Developed under the principles of PAR, this phase prioritized understanding of the educational context through the experience and voices of stakeholders. During the second semester of 2024, co-creation workshops were held with students and faculty from participating institutions.

Phase 2: Designing and co-creating microDERs

Conducted over two semesters (2024-2 and 2025-1), this phase aligned with PAR and CSPA principles to guide the work toward collaborative and creative solutions. Using the knowledge map, higher education students formulated and developed microDER sketches, which were shared and enriched by high school students in participatory sessions. This interaction enabled the progressive co-creation of prototypes. In parallel, research incubator students explored GenAI tools based on scope, license type, and generable resources to determine their suitability for microDER building.

Phase 3: Validating the methodology

The goal of this phase was to verify the structural (stages and activities) and functional (tasks) coherence of the microDER co-creation methodology in a real-life context. A pilot case with students from a university research incubator was used for this purpose. Validation took place in two stages. The first stage assessed the organization and sequencing of the stages and activities. The second stage applied the methodology to the pilot case to evaluate its effectiveness and relevance under real conditions.

Ethical procedure

The methodology was applied within an ethical framework designed to protect the integrity of participants, especially the minors. This included informed consent and confidentiality and anonymity protocols. The consent forms were written in accessible language for high school students and required signatures from parents or guardians, as well as assent from the minors. This ensured that they understood the purpose of the research and the conditions of participation.

Measures to ensure confidentiality include using codes or pseudonyms in records and reports, storing data securely with restricted access for the research team, and deleting personal information after the retention period. The procedure complied with Colombian Law 1581 of 2012 (Congreso de la República de Colombia, 2012) and 1377 of 2013 (Presidente de la República de Colombia, 2013) regarding the protection of personal data, ensuring minimal data collection and the ethical use of images, videos, and audio recordings.

RESULTS

Results are presented by methodological phase.

Phase 1: Recognizing the context and conducting a participatory, thematic delimitation

This phase was structured in two stages, which enabled a collaborative diagnostic process and collective knowledge-building.

Stage 1: Participatory workshops

During semesters 2024-2 and 2025-1, the research team conducted two workshops with 10th grade and three workshops with 11th grade students. These sessions were designed as co-creation spaces that explored the students' educational, social, and emotional realities. The participatory dynamics revealed three priority issues: lack of vocational guidance, limited access to educational opportunities (*e.g.*, scholarships and entrepreneurship programs), and the importance of emotional management for fostering a positive school environment.

Stage 2: Context knowledge map

The findings were synthesized into a knowledge map that highlights training needs, student potential, and school environment dynamics. This output guided the subsequent methodological phases, defining challenges and establishing two thematic axes for building microDERs: vocational guidance and emotional management.

Phase 2: Designing and co-creating microDERs

In this phase, the needs identified in the knowledge map were translated into products that integrate microDERs. These products address the challenges outlined in Phase 1. The development process involved collaborative work between higher education and high school students, as well as the pedagogical use of GenAI tools. The results were organized into two stages.

Stage 1: MicroDERs co-creation

During semesters 2024-2 and 2025-1 semesters, higher education students designed sketches in response to the challenges identified in Phase 1. The sketches were shared with high school students, who provided feedback based on guiding questions related to their experiences. This interaction resulted in four products aimed at raising awareness and supporting training processes adapted to the school context. Table 2 details the characterization of the work by academic semester, and Figures 3 and 4 present examples of the co-created products.

Table 2
Stage 1: MicroDER co-creation

Semester	Participants	Actions	Guiding questions	Products
2024-2	13 undergraduate students and 33 high school students (10 th grade)	▪ Review challenges ▪ Design sketches ▪ Share sketches ▪ Review and propose adjustments	▪ How can vocational uncertainty be reduced to prevent dropouts and strengthen professional development? ▪ How can emotion management strategies be implemented to improve assertive communication in high school students?	▪ Vía Pro: A career guidance website with thematic capsules, career choice simulators, and student testimonial videos. ▪ Moodies: A mobile app for emotion management with self-regulation exercises, interactive reflection content, and guides for assertive communication.
2025-1	19 higher education students and 32 high school students (11 th grade)	▪ Review challenges and sketches from the previous semester ▪ Propose a new challenge ▪ Identify improvements to previous challenges	▪ How can interpersonal relationships be improved among students through communication strategies that foster coexistence? ▪ How can the Vía Pro tool be improved to provide relevant information for academic continuity? ▪ How can the Moodies app be enhanced to encourage student participation?	▪ Life is not a game: A board game prototype promoting positive school environment, with analogous microDERs, such as thematic cards and printed guides for collective reflection. ▪ EnRuta: A web platform that enhances the usability of Vía Pro with interactive vocational capsules. ▪ Moodies+: A gamified web platform with microDERs challenges, self-assessments, and multimedia content to encourage student participation.

Source: Data collected by the authors.

Figure 3

Co-created prototypes in 2024-2



a) *Via Pro prototype*



b) *Moodies mobile app*

Source: Data collected by the authors.

Figure 4

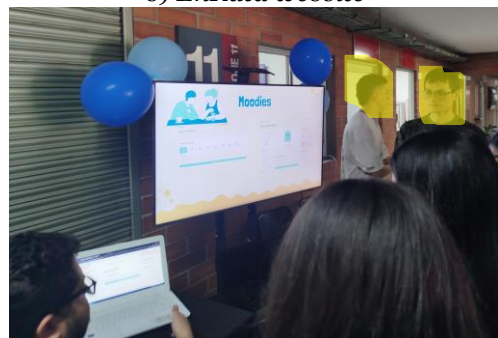
Co-created prototypes in 2025-1



a) *Life Is Not a Game analog game*



b) *EnRuta website*



b) *Moodies+ platform dissemination*

Source: Data collected by the authors.

Stage 2. Exploration and evaluation of GenAI tools

Students in the research incubator explored and evaluated 24 GenAI tools. The tools were classified by resource type (text, image, audiovisual, and code), scope, and license (free or proprietary).

Six tools were analyzed for textual resources: ChatGPT, Claude AI, Google Gemini, DeepL, Hugging Face, and Notion AI. Textual resources were the most explored category, and Hugging Face was identified as the only open-source option. While the others are useful for pedagogical writing, multilingual translation, and training guide development, they require proprietary licenses. Additional tools, such

as Grammarly, Acrobat AI, and Humata, were also tested for academic text review and synthesis.

Five text-to-image generators were reviewed for visual resources: Adobe Firefly, DALL·E, Magic Design, Midjourney, and Stable Diffusion. Only Stable Diffusion offers a free license. These tools support creating digital images in various formats, including 2D graphics, visual effects, automated designs, and high-resolution illustrations. However, exploration was limited due to restricted access, requiring further advanced evaluation.

Four tools were explored for creating and editing videos with avatars, text-to-video conversion, and visual effects: Lumen5, Pictory, RunwayML, and Synthesia. These tools are proprietary and licensed, which may limit their use by institutions. Additionally, two complementary tools were analyzed for customizing the creation of educational resources: Whisper, a free audio-to-text transcription tool; and ElevenLabs, a realistic voice synthesis platform.

Finally, GitHub, Copilot, Tabnine, Testim, and Applitools were evaluated for their ability to generate, assist with, or validate code snippets in the category of software code generation and automated testing tools. While they cannot be used directly in school education, they have the potential to be used in technical training activities and processes.

In line with the research hypothesis, this phase of the study demonstrated that:

- Teachers and young students can actively participate in collaborative microDER creation.
- Students can take on a leadership role in their learning when supported by a methodology that incorporates their participation, teacher mediation, and GenAI tools.
- A co-creation methodology supported by GenAI tools enables effective design and implementation of microDERs.
- The design and co-creation process fosters and strengthens computational thinking and social interaction among students.

Phase 3: Validating the methodology

The validation phase documented the microDER co-creation process through illustrations, analysis excerpts, photographs of student activities, and the generated artifacts. Table 3 summarizes the actors involved, detailing the time, session, modality, participants, and activities.

Table 3

Stages and actors involved in the validation phase

Stage	Session	Modality	Participants/Actors	Activity
1. Evaluation of the methodology structure	S1	In-person	▪ Research team ▪ Research incubator coordinator ▪ Higher education students (research incubator members)	Creative exchange of ideas
	S2	In-person	▪ Research team ▪ Higher education students (research incubator members)	Restructuring and adjustments

	S3	In-person	▪ Research team	Visual organization
2. Validation of the methodology implementation	S1	In-person	▪ Research team ▪ Research incubator coordinator ▪ Higher education students (research incubator members)	Methodology implementation and evaluation

Source: Prepared by the authors.

Stage 1: Designed methodology evaluation

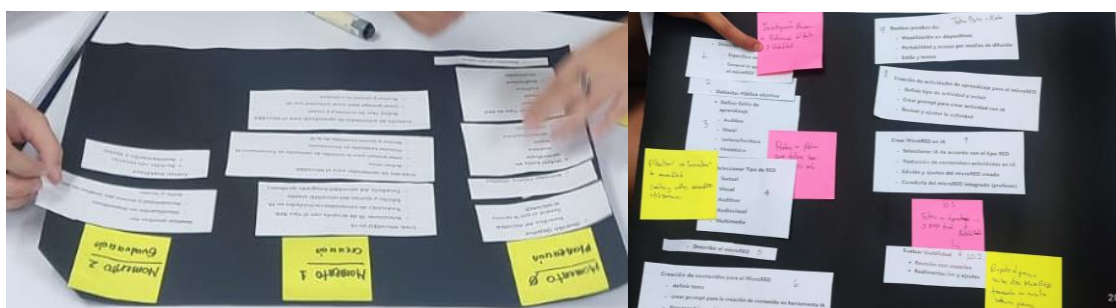
This stage included three sessions that engaged participants in developing the proposed methodology according to their needs, context, and perspectives.

In Session 1, the research team, students, and research incubator coordinator collaborated to exchange ideas. The students were divided into three groups, and the professor provided guidance while the research team facilitated the process. The activities began with individual reflection, followed by sharing, analysis, and collective building. Key tasks included ideation, process analysis, element prioritization, and reflection on previous experiences in creating resources.

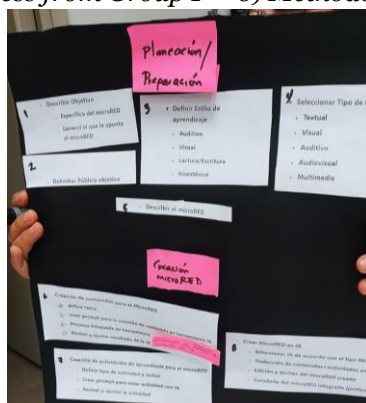
Throughout the session, students defined work paths and action flows for different stakeholders by focusing on co-creating microDERs. These contributions were incorporated into the methodology, fostering a reflective approach to the process. The outcome was the first version of the methodology, which integrated both the researchers' reference elements and the students' contributions.

Figure 5

First version of the MicroDER co-creation methodology from Session 1



a) Methodology progress from Group 1 b) Methodology progress from Group 2



c) Methodology progress from Group 3

Source: Data collected by the authors.

In Session 2, students from the research incubator and the research team came together to review the contributions from Session 1, compare proposals, and collaboratively develop a unified version of the microDER co-creation methodology. The session consisted of three activities: sharing proposals, conducting a comparative analysis, and developing a unified version (see Table 4). Table 5 documents the evolution of these proposals and the adjustments made, tracing the contributions and changing management throughout the process.

Table 4
Session 2 activities

No	Activity	Description
1	Sharing proposals	Each group presented its preliminary version of the methodology, emphasizing steps, GenAI tools, roles, and general observations. Presentations were supported by diagrams or maps to facilitate collective understanding.
2	Comparative analysis	Using an analysis matrix, participants identified similarities, differences, strengths, and irrelevant aspects among the proposals. The research team moderated discussion, fostering argumentation and reflection.
3	Collaborative building process	Based on the comparative analysis, participants developed a unified version of the methodology (V1.0). This version was created in a shared document and integrated contributions from all groups. It also refined phases, activities, resources, roles, and observations to ensure overall coherence in the methodology.

Source: Data collected by the authors.

Table 5
Traceability of methodology changes

Elements	Preliminary (Vo.1)	V1.0
Date	Sept 2024 – Mar 2025	Apr – May 2025
Phases	Three stages defined: P1: Recognizing the context P2: Building a MicroDER P3: Validating the methodology	- Three stages are kept - Names are specified: P1: Context definition (Planning/Preparation) P3: Validation/Evaluation
Subphases	Defined for P2 and P3: P2: Building a microDER P2.SP1: Content creation P2.SP2: Creation of learning activities P2.SP3: Creation of a microDER using AI P3: Validating the methodology P3.SP1: Testing P3.SP2: Usability	- Subphases are kept for P2 and P3 - Names are adjusted for P2.SP3, P3.SP1, and P3.SP2 P2: Building a microDER P2.SP3: Creation of a microDER using AI P3: Validation/Evaluation P3.SP1: Test development P3.SP2: Evaluation of microDER usability
Activities	Activities are defined for P1, P2, and P3:	- One activity is added to P1 - Names of 15 activities are adjusted

Elements	Preliminary (Vo.1)	V1.0
		P1: Context definition P1.A1: Identify a learning problem P1.A2: Describe the general learning objective of the microDER P1.A3: Describe the specific objective of the microDER P1.A4: Define/Delimit the target audience P1.A5: Define the learning style (auditory, visual, reading/writing, or kinesthetic) P1.A6: Select the type of microDER (textual, visual, auditory, audiovisual, or multimedia)
P1: Context P1.A1: Target audience P1.A2: Description P1.A3: Objective (specific and general) P1.A4: Learning style (auditory, visual, reading/writing, or kinesthetic) P1.A5: Type of DER (textual, visual, auditory, audiovisual, or multimedia)	P2: Building a microDER P2.SP1: Content creation P2.SP1.A1: Define the topic P2.SP1.A2: Create an AI-generated content prompt P2.SP1.A3: Process searches in an AI tool and adjust and/or refine prompts P2.SP1.A4: Review and adjust AI-generated results P2.SP2: Creation of learning activities P2.SP2.A1: Define the type of activity to include P2.SP2.A2: Create a prompt for activity generation with AI P2.SP2.A3: Review and adjust the activity P2.SP3: Creation of a microDER using AI P2.SP3.A1: Test AI tools according to the type of microDER and select the most appropriate one P2.SP3.A2: Translate and integrate content and activities in AI P2.SP3.A3: Edit and adjust the AI-generated microDER P2.SP3.A4: Curate the integrated microDER (teacher)	P2: Building a microDER P2.SP1: Content creation P2.SP1.A3: Use an AI tool to process searches (adjust and refine prompts) P2.SP3 MicroDERs creation using AI P2.SP3.A1: Test AI tools by DER type to select the most appropriate one P2.SP3.A2: Edit and adjust the AI-generated microDER P2.SP3.A4: Curate the integrated microDER (teacher)
P3: Validation P3.SP1: Testing P3.A1: Testing P3.SP2: Usability P3.A2: Validation		P3: Validation/Evaluation P3.SP1: Test development P3.SP1.A1: Perform display tests on devices P3.SP1.A2: Perform portability and media access tests P3.SP1.A3: Perform style and text tests P3.SP2: Evaluation of microDER usability P3.SP2.A1: Assess usability with users P3.SP2.A2: Provide feedback and adjust microDER

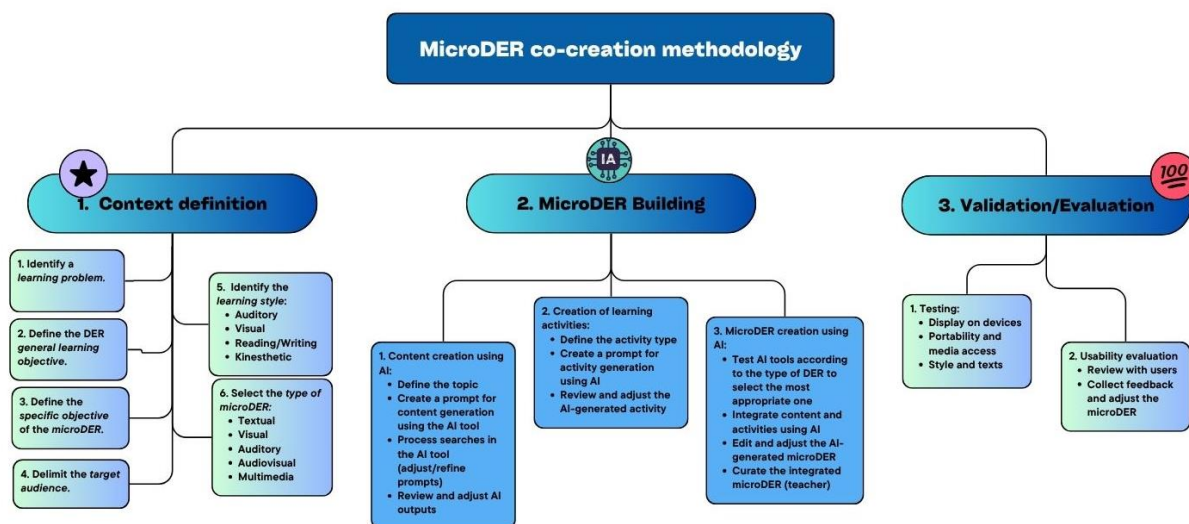
Source: Data collected by the authors.

In Session 3, the research team organized and visually represented the unified version of the microDER co-creation methodology. The goal was to translate the agreements from Sessions 1 and 2 into a clear, functional, and accessible visual representation to support its implementation and adoption by diverse audiences.

Based on the integrated version from Session 2—which included phases, activities, roles, and implementation conditions—the team explored graphic visualization models for the final representation. The team then conducted an internal validation to assess the model's clarity, coherence, and design fidelity. This process resulted in a graphical version of the methodology, ready for validation in application settings (see Figure 6).

Figure 6

Graphic representation of the proposed methodology

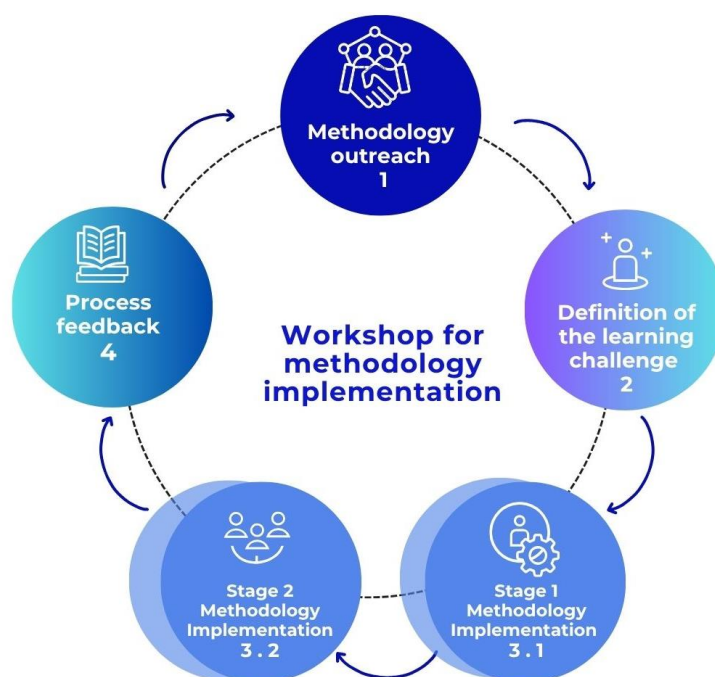


Source: Prepared by the authors.

Stage 2. Methodology implementation

This stage involved applying the designed methodology to a pilot group consisting of the research team, a professor, and students from the participating research incubator. The purpose was to evaluate the methodology's relevance, applicability, and coherence in creating microDERs using GenAI. A bootcamp-style workshop was conducted (see Figure 7), during which participants developed products that were evaluated to adjust the process and define improvement criteria. The session included four activities: dissemination of methodology, definition of the learning context, implementation of the methodology, and feedback on the process (see Table 6). Figure 8 shows evidence of the microDER pilot planning process, and Figures 9 and 10 present examples of the built microDERs.

Figure 7
Workshop structure for methodology implementation



Source: Prepared by the authors.

Table 6
Session 1 activities

No	Activity	Description
1.	Methodology dissemination	The research team introduced the project and explained the activities corresponding to the three defined phases, which support the proposed methodology.
2.	Definition of the learning context	The challenge was framed around video game design. Participants used their prior knowledge from the research incubator to define the topic and review related content and activities.
3.	Methodology implementation	Stage 1: Students were organized into working groups and provided with a printed guide and a graphic representation of the methodology (Figure 8.a) to serve as a work plan. Each group carried out the activities autonomously. See Figure 8.b for evidence. Stage 2: Groups developed their microDERs using GenAI tools (see Figures 9 and 10).
4.	Process feedback	Groups presented the initial versions of their microDERs and described the creation process and the GenAI tools they used. The audience provided feedback on both the process and products, generating reflections and best practices for future methodology validations.

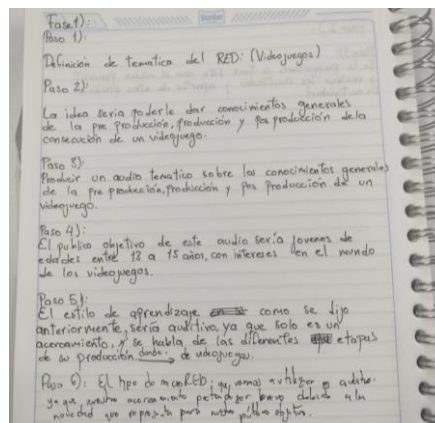
Source: Data collected by the authors.

Figure 8
Evidence of microDER planning in Stage 1

PLANEACIÓN

1. Problema de aprendizaje: diseño de videojuegos.
2. Objetivo general: Identificar qué género de videojuego es apto para hacer realidad una idea
3. Objetivo específico:
 - Diseñar un test que ayude a las personas que planeen el desarrollo de un videojuego a identificar cuáles son el género y tipo más acertados para su proyecto
 - Facilitar a las personas en la toma de decisiones sobre el género y tipo de videojuego que se quiere desarrollar
 - Dar a conocer los tipos de videojuegos que existen en la actualidad y cuáles son sus diferencias
4. Público objetivo: desarrolladores y diseñadores de videojuegos principiantes de 16 a 22 años de edad
5. Estilo de aprendizaje: Escritura/lectura
6. Tipo de microRED: Multimedia (texto e imagen)

a) Planning phase from Group 1



b) Planning phase from Group 2

Source: Data collected by the authors.

Figure 9
Evidence of microDER building in Stage 2

Fase 2: Construcción de MicroRED

1. Creación de Contenidos

*Definir el tema: Conocimientos generales de la preproducción, producción y posproducción de los videojuegos.

El prompt: "*Para la generación de una actividad MicroRED quiero explicar de manera breve e introductoria conocimientos básicos de la preproducción, producción y post producción de los videojuegos, para adolescentes de entre 13 y 15 años. Podrías generar un texto sobre esto*".

*Procesar Búsqueda en Herramientas de IA: Para el ejercicio pusimos el mismo prompt en tres IA diferentes para comparar sus resultados.

*Generar resultados:

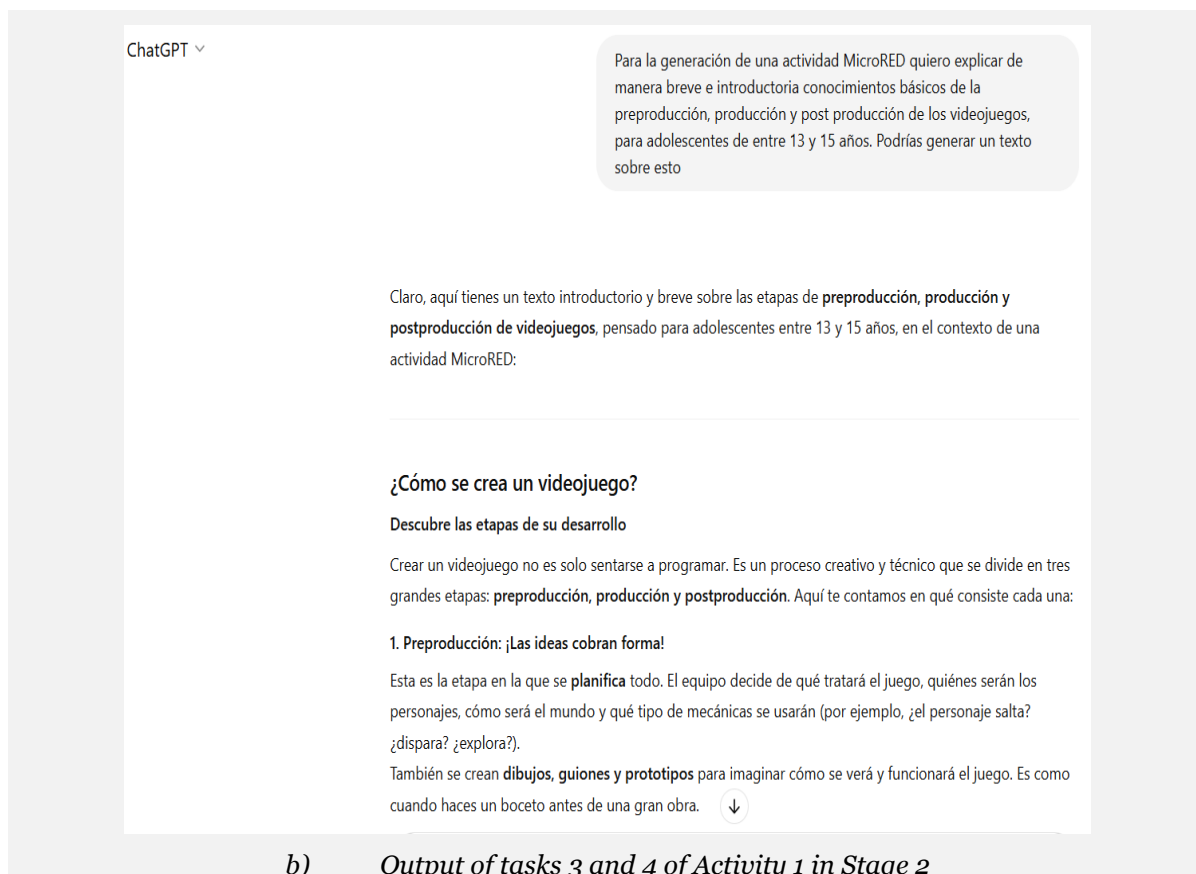
Chat GPT

¿Cómo se crea un videojuego?

Conoce las etapas básicas: Preproducción, Producción y Postproducción

Crear un videojuego es como hacer una película o montar una obra de teatro: necesita planificación, trabajo en equipo y mucha creatividad. Aquí te contamos las tres etapas principales que hacen posible que un juego llegue a nuestras manos:

a) Tasks 1, 2, and 3 of Activity 1 in Stage 2



Source: Data collected by the authors.

Figure 10
Example of a microDER built in Stage 2



Source: Data collected by the authors.

To conclude this stage, the stakeholders' roles and action paths were analyzed, as well as the curation formats. In addition, the level of human intervention required to review the outputs of each GenAI was examined across the proposed phases of the methodology implementation process. This discussion enabled adjustments to be made and a consensus to be reached on the operational effectiveness of the process. The resulting contributions were systematized into an updated version of the methodology, which is planned for evaluation with larger pilot groups.

DISCUSSION AND CONCLUSIONS

This study presents a methodology for microDER co-creation supported by GenAI tools. Grounded in PAR and CSPS methodologies, the process relies on collaborative strategies, such as co-creation sessions, bootcamps, and knowledge-sharing with community stakeholders. The problem-oriented, learning outcome-driven process fosters thematic exploration, peer interaction, and innovative practices that differ from traditional curricular approaches (Ortiz Jaramillo, 2022). In these spaces, students experiment, pursue their interests, and develop curiosity for new topics and skills.

The PAR approach is at the core of methodology, providing a foundation for social appropriation in participatory learning contexts. It promotes the pedagogical use of GenAI to redesign teaching and learning environments and support complex DER production processes. In line with Gimenez and Siqueira (2024), the framework facilitates the fragmentation, organization, sequencing, and adaptation of educational content to enhance students' learning experiences.

The methodology is structured into three interconnected phases. In Phase 1, students collaborated to characterize the school context and gather relevant information from educational stakeholders. This process guided the definition of the research theme and strengthened interaction among participants, consistent with the work of Fals Borda (2001) and Arnanz et al. (2023).

Phase 2 built on these foundations by combining the technical and reflective skills of higher education and high school students, who then explored tools and resources in the classroom. These outputs served as application spaces where concrete, microDER-based solutions were designed and evaluated.

Phase 3 focused on the initial validation of the methodological proposal by incorporating criteria of relevance, applicability, and coherence. This validation is a key input for the next stage: designing and publishing microDERs in open-access virtual environments to assess their impact on computational thinking development in young students.

The findings confirm that the methodology is feasible in real classroom settings and promotes meaningful learning in high school and higher education settings. Consistent with Leal-Urueña's (2020) approach, it incorporates activities, actors, evaluation procedures, and feedback mechanisms that refine the final product, the microDERs.

Its formal and integrative structure, particularly in the initial and final phases, distinguishes it from other models. Examples include Zhang and West's (2020) organization of instructional events around a key topic and Kohler et al. (2021) curation of microlearning elements based on connected concepts.

Unlike Monib et al. (2025) instructional design framework, which focuses on the individual, situational, and explicit characteristics of resource beneficiaries, this methodology is strengthened by a central phase that uses GenAI to create microDERs. Systematizing contributions generated in each phase reinforces the proposal's viability and guides its expansion through pilot group implementation and the publication of microDERs on open-access platforms. References include platforms such as 7taps.com and SafetyCulture.com, which enable the creation of microlearning courses optimized for smartphone use and distribution via social media.

Within this framework, the methodology demonstrates both transferability and scalability with a strategy for social appropriation mediated by GenAI. It also enables students to play an active role in their learning and develop computational thinking and social interaction skills.

Through the promotion of co-creation workshops with the educational community, the involvement of the research incubator, and the development of classroom projects, the methodology has demonstrated its operational feasibility and educational relevance in diverse learning environments.

A growing body of research on microlearning and the use of GenAI in education underscores their transformative potential for pedagogical processes and practices (Gimenez & Siqueira, 2024). However, adapting or redesigning these practices requires addressing challenges related to social interaction, stakeholder collaboration, and specific aspects of DER design, such as parameterization, fragmentation, and reliability.

Additionally, integrating innovative components into pedagogical practices demands incorporating new tools, methodologies, and approaches that support learning outcomes by solving educational problems (Villarreal et al., 2022). As Stanford et al. (2016) note, innovative practices must contribute to achieving objectives, differentiate from traditional approaches, and foster stakeholder interaction and the effective use of resources.

The findings of this study have theoretical and practical implications, as they offer insights into the value of microlearning and the role of GenAI in education. These results also shed light on the evolving roles of teachers and students in new learning environments. The results confirm the effectiveness of microlearning as an instructional approach that delivers brief, targeted content aligned with cognitive load theory. This theory emphasizes that minimizing cognitive load enhances learning (Sweller, 2011).

The proposed methodology functions as both a framework for organizing the microDERs co-creation process and a tool for rethinking learning environments through collaboration and the responsible use of GenAI.

The microDERs co-creation methodology addresses key challenges and guide educators, students, instructional designers, curriculum developers, and content creators in producing AI-driven resources. Its implementation in virtual environments, such as learning management systems and massive open online courses, is strengthened through collaborative tools that encourage interaction, observation, and systematization of contributions from all participants at every stage of the process.

The findings suggest that integrating GenAI complements, rather than replaces, pedagogical processes by enhancing creativity, thematic exploration, and collaborative knowledge building—all of which are essential for strengthening computational thinking. Implementing GenAI fosters dynamic, adaptive learning settings where

students develop cross-disciplinary skills, such as problem-solving, critical analysis, and innovation. Thus, GenAI serves as a strategic resource that promotes inclusion, participation, and more meaningful and sustainable learning experiences.

Combining microlearning with participatory approaches, such as PAR and CSPS, makes this methodology innovative and transformative. This integration enables social appropriation and the development of educational resources adaptable to different school settings. It also fosters co-creation and interdisciplinary teaching. Thus, microlearning broadens the possibilities of technological mediation and contributes to more inclusive, flexible, and competency-oriented educational practices that can address contemporary challenges.

Future work should validate the methodology in more high school and higher education institutions and modalities (in-person and virtual environments) to enhance its replicability. As this study was limited to two cases and focused only on design and creation, further research should assess its pedagogical effectiveness and examine the ethical implications of GenAI use in vulnerable settings or limited-access contexts.

Conflict of interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Acknowledgments

This article is part of the preliminary results of research project 1375-2024 "Metodología de cocreación de recursos digitales de microaprendizaje activo para la enseñanza de pensamiento computacional y la interacción social por jóvenes estudiantes de informática" funded by Universidad de Medellín and Instituto Tecnológico Metropolitano (ITM) Medellín-Colombia. We would like to thank ITM Translation Agency (traducciones@itm.edu.co) for translating the manuscript in English.

REFERENCES

- Abdous, M. (2023, 22 March). How AI is shaping the future of higher ed. *Inside Higher Ed*. <https://www.insidehighered.com/views/2023/03/22/how-ai-shaping-future-higher-ed-opinion>
- Adeshola, I., & Adepoju, A. P. (2023). The opportunities and challenges of ChatGPT in education. *Interactive Learning Environments*, 32(10), 6159-6172. <https://doi.org/10.1080/10494820.2023.2253858>
- Arnanz Monreal, L., García Montes, N., & Villasante Prieto, T. R. (2023). La investigación y acción participativa y transformadora basada en los grupos motores. *Prisma Social*, (43), 34-56. <https://revistaprismasocial.es/article/view/5176>
- Boumalek, K., El Mezouary, A., Hmedna, B., & Bakki, A. (2024). Transforming microlearning with generative AI: Current advances and future challenges. In M. Lahby, Y. Maleh, A. Bucchiarone, & S. E. Schaeffer (eds.), *General aspects of applying generative AI in higher education*. Springer Nature. https://doi.org/10.1007/978-3-031-65691-0_13
- Bozkurt, A. (2023). Generative artificial intelligence (AI) powered conversational educational agents: The inevitable paradigm shift. *Asian Journal of Distance Education*, 18(1). <https://www.asianjde.com/ojs/index.php/AsianJDE/article/view/718>
- Bygstad, B., Øvrelid, E., Ludvigsen, S., & Dæhlen, M. (2022). From dual

- digitalization to digital learning space: Exploring the digital transformation of higher education. *Computers & Education*, 182, 104463. <https://doi.org/10.1016/j.compedu.2022.104463>
- Cabeza-Rodríguez, M.-Á. (2025). Asistentes ChatGPT en educación superior en línea y satisfacción del alumnado: Un caso de estudio. *RIED-Revista Iberoamericana de Educación a Distancia*, 28(2). <https://doi.org/10.5944/RIED.28.2.43552>
- Cain, W. (2024). Prompting change: Exploring prompt engineering in large language model AI and its potential to transform education. *TechTrends*, 68(1), 47-57. <https://doi.org/10.1007/s11528-023-00896-0>
- Congreso de la República de Colombia. (2012, 17 de octubre). *Ley 1581 de 2012, por la cual se dictan disposiciones generales para la protección de datos personales* [Ley estatutaria; Diario Oficial 48.587, 18 de octubre de 2012]. <https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=49981>
- Denojean-Mairet, M., López-Pernas, S., Agbo, F. J., & Tedre, M. (2024). A literature review on the integration of microlearning and social media. *Smart Learning Environments*, 11, 46. <https://doi.org/10.1186/s40561-024-00334-5>
- Díaz Redondo, R. P., Caeiro Rodríguez, M., López Escobar, J. J., & Fernández Vilas, A. (2021). Integrating micro-learning content in traditional e-learning platforms. *Multimedia Tools and Applications*, 80(2), 3121-3151. <https://doi.org/10.1007/s11042-020-09523-z>
- Dickey, E., & Bejarano, A. (2023). *A model for integrating generative AI into course content development* [Preprint]. arXiv. <https://arxiv.org/abs/2308.12276>
- Dolasinski, M., & Reynolds, J. (2020). Microlearning: A new learning model. *Journal of Hospitality and Tourism Research*, 44(3), 551-561. <https://doi.org/10.1177/10963480209015>
- Durán Alcalá, M., & Escudero Nahón, A. (2023). Microlearning en el entorno educativo. *Revista de Investigación Educativa de la REDIECH*, 14, e1763. https://doi.org/10.33010/ie_rie_rediech.v14i0.1763
- Fals Borda, O. (2001). Transformaciones del conocimiento social aplicado: Lo que va de Cartagena a Ballarat. *Análisis Político*, 42, 93-101. <https://revistas.unal.edu.co/index.php/anpol/article/view/75309>
- Gimenez, P., & Siqueira, S. (2024). *Microlearning e IA generativa na educação online: Fragmentação, sequenciamento e adaptação de conteúdos para aprendizagem centrada no aluno*. In *Anais Estendidos do XIII Congresso Brasileiro de Informática na Educação* (pp. 314-319). SBC – Sociedade Brasileira de Computação. https://sol.sbc.org.br/index.php/cbie_estendido/article/view/31804
- Gonnet, J. P. (2019). ¿Por qué la interacción? Una reconstrucción de los escritos tempranos de Erving Goffman. *Revista Reflexiones*, 99(1), 0-15. <https://doi.org/10.15517/rr.v99i1.35308>
- Holmes, W., Bialik, M., & Fadel, C. (2023). *Artificial intelligence in education*. Globethics Publications. <https://doi.org/10.58863/20.500.12424/4276068>
- Hug, T., Lindner, M., & Bruck, P. A. (eds.). (2006). *Micromedia & e-learning 2.0: Gaining the big picture* [Microlearning Conference Proceedings, 2006]. Innsbruck University Press. <http://library.oapen.org/handle/20.500.12657/39646>
- Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). Analyzing and improving the image quality of StyleGAN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 8107-8116). Computer Vision Foundation/IEEE. <https://doi.org/10.1109/CVPR42600.2020.00813>
- Kemmis, S., & McTaggart, R. (2005). Participatory action research: Communicative action and the public sphere. In N. K. Denzin & Y. S. Lincoln

- (Eds.), *The SAGE handbook of qualitative research* (3^a ed., pp. 559–603). SAGE.
- Kohler, M., Gamrat, C., Raish, V., & Gross, E. (2021). Microlearning and micro-credentials in higher education. In J. R. Corbeil, B. H. Khan, & M. E. Corbeil (eds.), *Microlearning in the digital age: The design and delivery of learning in snippets* (pp. 109-128). Routledge. <https://doi.org/10.4324/9780367821623-10>
- Leal-Urueña, L. A. (2020). *Definición de las especificaciones de diseño*. CIDET Moodle (Universidad Pedagógica Nacional). <https://cidetmoodle.pedagogica.edu.co/mod/page/view.php?id=94698>
- Luckin, R., Holmes, W., Griffiths, M., & Forcier, L. B. (2016). *Intelligence unleashed: An argument for AI in education*. Pearson. <https://edu.google.com/pdfs/Intelligence-Unleashed-Publication.pdf>
- Manzini, E. (2015). *Design, when everybody designs: An introduction to design for social innovation*. MIT Press. <https://doi.org/10.7551/mitpress/9873.001.0001>
- Mateus-Nieves, E., & Moreno Moreno, E. (2021). Microlearning como estrategia para una educación asincrónica. *Épsilon: Revista de la Sociedad Andaluza de Educación Matemática "THALES"*, (109), 41-58. https://thales.cica.es/epsilon/sites/default/files/2023-02/epsilon109_3.pdf
- Memon, T. D., & Kwan, P. (2025). A collaborative model for integrating teacher and GenAI into future education. *TechTrends*. <https://doi.org/10.1007/s11528-025-01105-w>
- Miao, F., & Holmes, W. (2024, April). *Guía para el uso de IA generativa en educación e investigación*. UNESCO. <https://unesdoc.unesco.org/ark:/48223/pf0000389227>
- Mirón, L. (2022). La Escuela Creativa de Innovación Social. In *Jornadas Hacia una Nueva Cultura Científica* (pp. 25-34). Editorial Universitat Politècnica de València. <https://doi.org/10.4995/NCC2022.2022.15773>
- Monib, W. K., Qazi, A., & Apong, R. A. (2025). Microlearning beyond boundaries: A systematic review and a novel framework for improving learning outcomes. *Heliyon*, 11(2), e41413. <https://doi.org/10.1016/j.heliyon.2024.e41413>
- Ortiz Jaramillo, I. (2022). *Diseño de bootcamp de programación y pensamiento computacional con enfoque transdisciplinar* [Master's Thesis, Universidad de los Andes]. Repositorio Universidad de los Andes. <https://hdl.handle.net/1992/64050>
- Piantari, E., Al Husaeni, D. N., Anisyah, A., & Erlangga, E. (2024). Analysis of publication trends on artificial intelligence in education: A systematic literature review. In *Proceedings of the 9th Mathematics, Science, and Computer Science Education International Seminar (MSCEIS 2023)* (pp. 430-443). Atlantis Press. <https://www.atlantispress.com/proceedings/msceis-23/126003203>
- Presidente de la República de Colombia. (2013, 27 June). *Decreto 1377 de 2013, por el cual se reglamenta parcialmente la Ley 1581 de 2012* [Diario Oficial 48.834, 27 de junio de 2013]. <https://www.alcaldiabogota.gov.co/sisjur/normas/Norma1.jsp?i=53646>
- Ritzer, G. (2008). *Teoría sociológica moderna* (4th edition). McGraw-Hill.
- Samoili, S., López Cobo, M., Delipetrev, B., Martínez-Plumed, F., Gómez Gutiérrez, E., & De Prato, G. (2021). *AI Watch. Defining artificial intelligence 2.0*. Publications Office of the European Union. <https://doi.org/10.2760/019901>
- Sengar, S. S., Hasan, A. B., Kumar, S., & Carroll, F. (2025). Generative artificial intelligence: A systematic review and applications. *Multimedia Tools and Applications*, 84, 23661-23700. <https://doi.org/10.1007/s11042-024-20016-1>
- Sirvent, M. T., & Rigal, L. A. (2012). *Investigación acción participativa: Un desafío de nuestros tiempos para la construcción de una sociedad democrática*. Proyecto Páramo Andino. <http://hdl.handle.net/11336/187628>

- Stanford, C., Cole, R., Froyd, J., Friedrichsen, D., Khatri, R., & Henderson, C. (2016). Supporting sustained adoption of education innovations: The Designing for Sustained Adoption Assessment Instrument. *International Journal of STEM Education*, 3, Article 1. <https://doi.org/10.1186/s40594-016-0034-3>
- Sweller, J. (2011). Cognitive load theory. In J. P. Mestre & B. H. Ross (eds.), *Psychology of Learning and Motivation* (Vol. 55, pp. 37-76). Academic Press. <https://doi.org/10.1016/B978-0-12-387691-1.00002-8>
- Torgerson, C., & Iannone, S. (2019). *Designing microlearning* (What Works in Talent Development). ATD Press (Association for Talent Development). <https://www.td.org/product/book--designing-microlearning/111919>
- Tu, Y., Chen, J., & Huang, C. (2025). Empowering personalized learning with generative artificial intelligence: Mechanisms, challenges and pathways. *Frontiers of Digital Education*, 2, 19. <https://doi.org/10.1007/s44366-025-0056-9>
- Villarreal Contreras, R., Salas Álvarez, D., & Alemán, A. (2022). *Prácticas educativas innovadoras desde un enfoque STEAM+A*. Fondo Editorial Universidad de Córdoba. <https://repositorio.unicordoba.edu.co/entities/publication/d29d8a9e-1990-4682-b97d-adb53844525b>
- Yang, J. (2024). Deep integration of generative AI with higher education: Scenarios, risks and recommendations. *China Higher Education*, 5, 52-56.
- Zhang, J., & West, R. E. (2020). Designing microlearning instruction for professional development through a competency-based approach. *TechTrends*, 64(2), 310-318. <https://doi.org/10.1007/s11528-019-00449-4>
- Zhang, Y., Kong, F., Wang, P., Sun, S., Wang, L., Feng, S., Wang, D., Zhang, Y., & Song, K. (2024). *STICKERCONV: Generating multimodal empathetic responses from scratch* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2402.01679>

Date of reception: 1 June 2025

Date of acceptance: 2 September 2025

Date of approval for layout: 3 October 2025

Date of publication in OnlineFirst: 24 October 2025

Date of publication: 1 January 2026

Active learning in higher engineering education: a decalogue for the design of post-lecture online quizzes

Aprendizaje activo en la educación superior de ingeniería: un decálogo para diseñar quizzes online post-clase



Dubravka Novkovic - *La Salle-Universidad Ramon Llull, URL (Spain)*

Ricardo Torres Kompen - *La Salle-Universidad Ramon Llull, URL (Spain)*

Josep Petchamé - *La Salle-Universidad Ramon Llull, URL (Spain)*

ABSTRACT

Quizzes are a valuable addition to the ever-growing toolkit of active learning strategies in higher education. In distance learning contexts, quizzes are frequently used to assess students' level of knowledge, to encourage their participation and to support knowledge retention or academic performance. Despite a wealth of studies about quizzes in general, and pre-lecture quizzes in particular, there is relatively little empirical work examining post-lecture online quizzes. This paper discusses Information and Communication Technology (ICT) Engineering students' experiences with weekly post-lecture online quizzes in a management subject. Participants took a single attempt, automatically graded, quiz on Moodle Learning Management System (LMS) after each subject unit. Once the learning experience was completed, qualitative research was carried out through a Bipolar Laddering (BLA) tool, designed to extract positive and negative perceptions of the students on the implemented online quiz model. Overall, our analysis shows that students view post-lecture online quizzes as a valuable active learning strategy, especially useful as a practice for high-stakes exams. Regular low-stakes quizzing was perceived to foster ongoing engagement with the subject material, to support continuous review and self-directed learning. Building on the student-reported insights, the researchers formulated ten design considerations that may inform the development of effective post-lecture online quizzes, which can be easily implemented in both remote and face-to-face learning contexts.

Keywords: active learning; learning strategy; online quiz; post-lecture test; engineering; Bipolar Laddering (BLA).

RESUMEN

Los *quizzes* son una valiosa herramienta entre las diversas estrategias de aprendizaje activo en la educación superior. En contextos de aprendizaje a distancia se utilizan con frecuencia para evaluar el nivel de conocimientos de los estudiantes, fomentar su participación y apoyar la retención de conocimientos o el rendimiento académico. Existen numerosos estudios sobre los *quizzes* en general, y los *quizzes pre-clase*, existiendo escasa investigación empírica sobre los *quizzes post-clase*. Este artículo analiza las experiencias de estudiantes de Ingeniería en Tecnologías de Información y Comunicación (TIC) con *quizzes online post-clase* semanales en una asignatura de gestión. Los participantes realizaron un *quiz* de un solo intento, calificado automáticamente, en el Sistema de Gestión del Aprendizaje (en inglés, LMS) Moodle de la institución después de cada unidad temática. Completada la experiencia de usuario del alumnado, se llevó a cabo una investigación cualitativa mediante la herramienta *Bipolar Laddering (BLA)*, diseñada para extraer las percepciones positivas y negativas del alumnado sobre el modelo de *quizzes* implementado. En general, nuestro análisis muestra que los estudiantes consideran los *quizzes online post-clase* una valiosa estrategia de aprendizaje activo, especialmente útil para la preparación de exámenes. Los encuestados percibieron que los *quizzes* semanales fomentan tanto el *engagement* como el aprendizaje autónomo. Basándose en las observaciones de los estudiantes, los investigadores formularon diez recomendaciones de diseño, que podrían orientar el desarrollo de los *quizzes online post-clase* eficaces que se pueden implementar fácilmente tanto en contextos de aprendizaje remoto como presencial.

Palabras clave: aprendizaje activo; estrategia de aprendizaje; *quiz* en línea; examen posterior a la clase; ingeniería; *Bipolar Laddering (BLA)*.

INTRODUCTION

Present-day education has been undergoing a fundamental transformation through the integration of Information and Communication Technologies (ICT), essentially reshaping how knowledge is accessed, delivered and constructed. Within this context, online learning environments, which can be implemented across online, hybrid and face-to-face education models, have expanded and strengthened key elements of flexible learning —such as unrestricted access in terms of time, location and pace (Veletsianos et al., 2021). Grounded in active learning theory —a theoretical framework that emphasizes learners’ active participation in the construction of knowledge— the present study seeks to understand ICT Engineering student perceptions of one of the active learning tools —i.e., online quizzes— implemented in a management course and delivered through a Learning Management System (LMS). As a first step, a research background on active learning and quizzes is presented. Thereafter, a research work aimed at recognizing the student perceptions of the quizzes was developed to understand the actual learner experience. The researchers used Bipolar Laddering (BLA) —a qualitative method in the format of an open-ended questionnaire— to identify positive and negative elements that facilitate and hinder learning. Based on the analysis of student responses, this study offers practical guidance for the effective implementation of post-lecture online quizzes.

The term ‘active learning’ formally emerged in higher education literature over three decades ago as a transformative pedagogical approach (Bonwell & Eison, 1991), which shifted the traditional content-centred approach from passive knowledge transmission to student-centred instruction (Hartikainen et al., 2019). In terms of an active learning definition, a recent peer-reviewed article related to biology education literature and teaching community (Driessen et al., 2020) found that the great bulk of the active learning literature they reviewed (148 articles) failed to clarify what the term actually means. Active learning for the purpose of this study is defined as “a student-centred approach to the construction of knowledge focused on activities and strategies that foster higher-order thinking” (Doolittle et al., 2023, p.1).

As far as the benefits of active learning are concerned, a landmark meta-analysis of 225 studies across Science, Technology, Engineering and Mathematics (STEM) disciplines (Freeman et al., 2014) found that replacing traditional lectures with active learning led to significantly higher participant examination scores and lower failure rates. While increased student engagement and academic performance have been solidly established benefits of active learning (Prince, 2004), it was later confirmed that it also promotes communication and interactivity, fosters stronger sense of community and connectedness, enhances student satisfaction with the learning experience and provides greater flexibility in the education process (Allsop et al., 2020).

On the other hand, there are studies in engineering which reported mixed response to active learning as per Tharayil et al. (2018). Based on observation of engineering classrooms and a qualitative analysis of interviews with 17 undergraduate engineering professors in the United States, the authors found that students were more likely to perceive positively those new ways of teaching only when instructors explained benefits of active learning and provided skilled facilitation of classroom activities, which in turn led to higher participation and course evaluation. Deployed as an inquiry-based lab experience in the context of essential science practices (Owens et al., 2020), active learning attracted student resistance which stemmed from more effort

required from all participants. The authors suggested that instructors may lessen the problem by scaffolding practices and consistently giving clear feedback, which helps curb the frustration students may feel when struggling with challenging or unfamiliar scientific tasks. Interestingly, in another STEM study in Texas, Andrews et al. (2020) compared instructors' and students' attitudes and behaviours regarding active learning. The instructor survey assessed their attitudes toward active learning, the strategies they employed to reduce student resistance, and instructors' perceptions of student behaviour. A corresponding student survey invited students to evaluate their instructors' teaching methods and reflect on their own attitudes and behaviours during class. Both data sets were complemented by classroom observations. The key finding was instructors' tendency to overestimate student resistance to active learning in their class. Instructors occasionally feared negative reaction from students, particularly those instructors whose tenure, promotion and merit ratings depended considerably on their end-of-semester course evaluations. However, the collected student data painted a different reality: students appreciated active learning activities, engaged with them willingly, stayed focused, and intended to rate both the course and the instructor favourably.

Many instructional strategies fall under the umbrella of active learning. For example, previously cited research of Doolittle et al. (2023) lists fifteen different approaches that appeared at least ten times across 547 reviewed articles, most frequently mentioned being problem-based learning, the flipped classroom, and small group discussions. Other literature reviews also highlight a wide range of instructional activities categorized per different aspects of interest (Arthurs & Kreager, 2017; Kozanitis & Nenciovici, 2023; Mitchell et al., 2017). In engineering education, Villalonga Pons et al. (2023) cite strategies such as problem-based learning, project-based learning, cooperative learning, and learning by doing, considered key indicators of quality in creating effective classroom experiences, especially in online learning environments. Similarly, Gosavi and Arora (2022) observed that active learning strategies often work in tandem, with the implementation of one frequently triggering the use of others.

Specifically on quizzes, Mitchell et al. (2017) discussed the instructional strategy in Information System courses. A recent systematic review of online assessments in higher education (Heil & Ifenthaler, 2023) identified quizzes as a type of automated assessment. The word 'quiz' first emerged around 1780, and over time, it became established alongside other evaluation terms such as 'test' or 'examination' (Dengri et al., 2021). These authors reviewed eleven studies that used quizzes as teaching and learning tools and reported an almost universal increase in post-test scores, together with positive comments from students. Earlier studies had explored the use of pre-lecture quizzes as a strategy for incentivizing and sometimes rewarding students' class preparation (Brink, 2013; Cook & Babon, 2017; Howell, 2021). A more recent study comparing active lectures with traditional lectures across six faculties at a public university in the United States, highlighted students' appreciation of both pre-lecture and post-lecture quizzes as valuable tools for improving focus during classes and supporting self-assessment of their learning (Klein et al., 2023). A study conducted by Ross et al. (2018) revealed that 88.6% students enrolled in an online tertiary program at an Australian university agreed or strongly agreed that their quiz results highlighted the areas where their knowledge was lacking. When enhanced by technology and delivered digitally, Raes et al. (2020) showed that quizzes have positively influenced student intrinsic motivation across different learning environments. Indeed, quizzes

were perceived as useful and engaging by students regardless of whether they participated remotely or attended in person. Staying in the digital context, in an experimental course for graduate and undergraduate students in Brazil, Rossi et al. (2021) described quizzes as an effective tool for reinforcing knowledge, in particular because answers were corrected immediately. This approach helped solidify key concepts introduced during lectures. Although the study did not measure the isolated effect of quizzes, the researchers underscored the educational benefit of tools that enable immediate feedback on student performance. In another recent study, Gowele (2024) recommended integrating quizzes into mathematics curricula to boost student engagement and performance, identifying online quiz platforms as superior tools for accessibility and student progress monitoring. Indeed, recent research on game-based quiz platforms like Kahoot! reported that consistent use of such tools led to notable gains in learning (Nuci et al., 2021). Moreover, Kirzner et al. (2021) administered Kahoot! quizzes at the end of each class, covering the content of the day. Both quantitative and qualitative results of that study confirmed that students acknowledged the link between gamification and improved learning and retention.

While the existing evidence largely supports the educational benefits of quizzes, by contrast, some studies offer critical perspectives or factors that may impact the effectiveness of testing. Critics point, for example, to test anxiety and stress quizzes cause for some students (Stoyanova & Giannouli, 2022). This suggests that effective implementation may require attention to both quiz frequency and maintaining their low-stakes nature. Nevertheless, the authors found that students' test anxiety diminished, and they were more motivated when they were allowed multiple quiz attempts with the corrective feedback. These and similar concerns were investigated in a meta-analysis of the 'testing effect' conducted by a team of Asian scholars (Yang et al., 2021). Their findings reveal that the benefits of testing spanned 18 academic disciplines, supporting not only the retention of information but also a deeper understanding and practical application of knowledge.

To frame the present study, it is worthwhile mentioning that La Salle-Universidad Ramon Llull (URL) has implemented an educational model called the New Learning Context (NLC) with the objective of meeting the needs of contemporary learners, emphasizing flexibility, active participation and technology integration (Petchamé et al., 2021). Introduced during the 2018-2019 academic year, the NLC model is grounded in five pedagogical principles (Interiority; Mind, Body & Movement; Thought Construction; Self-Regulated Behaviour; Social Dimension of Learning) which are operationalized through five learning environments (Welcoming, Seminar, Workshop, Project and Closure), each designed to support different stages of the learning process. As such, the NLC model is fully aligned with the use of active learning methodologies.

The NLC is supported by a series of technological infrastructure and application enablers of active learning that have been implemented at La Salle-URL Campus to provide students with flexible attendance options, allowing them to attend classes either remotely (through a videoconferencing system) or in person on Campus (Petchamé et al., 2023). The cornerstone of this Smart Classroom system is a Moodle LMS recommended by López-Tocón (2021) as an effective method for learning scientific content.

Unlike previous studies largely related to the use of pre-lecture quizzes (Evans et al., 2021; Gyllen et al., 2021), the present work analyses post-lecture quizzes as a tool for continuous and formative assessment of engineering students. This article

examines the active participation of undergraduate students from an ICT Engineering program enrolled in a course on 'Financial Economics' which was distributed into seven units and taught over one semester. To promote consistent engagement and reinforce lecture content, students are assigned a post-lecture quiz after each subject unit, approximately once a week. The quizzes consist of 15-20 multiple-response questions (MRQs), administered and automatically graded via a Moodle LMS. Students are expected to complete each quiz individually (single attempt per quiz) within five days of the lecture, allowing the instructor sufficient time to review aggregate results prior to the next class session. While individual quiz grades are reviewed for feedback purposes, they are not factored directly into the course grade. Instead, the quiz participation —defined as timely quiz submission— is what counts toward the 10% of the final grade. This approach emphasizes formative learning over performance pressure, encouraging regular study habits and comprehension checks. Finally, to support exam preparation, all quizzes are reopened four weeks before the final examination. Moodle automatically reshuffles MRQs at each new quiz attempt. At that stage, students are granted unrestricted access and unlimited quiz attempts, supporting them as a self-study resource. The final exam may feature up to two MRQs adapted from the unit quizzes within the 10-question exam structure. Summative course assessment is carried out based on the weighted formula computed as follows: three paper-based tests during the semester (50%); a final written exam (40%); and timely quiz submission (10%).

The aim of this research work is twofold: (1) to deepen the understanding of student perceptions of post-lecture online quizzes as an active learning instrument implemented in a business-related subject of an ICT Engineering undergraduate degree, and (2) to offer practical guidance to instructors seeking to integrate post-lecture online quizzes into their teaching in any type of learning environment.

METHODOLOGY

The present study investigated the ICT Engineering undergraduates experience with post-lecture quizzes using the BLA methodology developed by Pifarré and Tomico (2007) and featured in several higher education research studies (Horna-Saldaña & Canaleta, 2024; Petchamé et al., 2021, 2023, 2024). As defined in Pifarré and Tomico (2007), it is a qualitative research approach that combines open-ended questioning with traditional laddering interview techniques. Following a Socratic-tabula-rasa principle, with no prior prompting, students are first asked to freely identify both positive and negative aspects of their quiz experience. These bipolar elements, representing strengths and weaknesses, are then rated by students on a 10-point scale of importance (1=minimum; 10=maximum) and justified through follow-up 'Why?' questions. In the final phase, students are encouraged to propose improvements, suggesting how to enhance positive or address negative aspects.

This method enables the analysis of user experience while minimizing potential bias, as responses arise from students' direct experiences rather than from leading questions. As such, this approach would be suitable for diverse environments, including digital contexts examined in this research. In bipolar laddering, large samples are not a requirement since the goal is to capture the range and hierarchy of user perceptions, with saturation typically reached after a limited number of responses (Hennink & Kaiser, 2022; Petchamé et al., 2021). The sample purpose is to reveal distinctive cognitive dimensions for their posterior clustering and analysis, instead of

building generalizable models. Therefore, rather than statistical representativeness, the focus is on achieving qualitative saturation —the point where additional responses no longer introduce new elements or themes to the emerging constructs. On the other hand, since the active learning tools we analyse in this research (online quizzes) are technology-based and not tied to any specific subject or discipline, the recommendations presented can be applied broadly across different higher education contexts.

For the present study, data points were labelled as described in the article by Petchamé et al. (2024). Positive elements of students' feedback were labelled as Positive (P) Elements (E) that are either Common (C) —if repeated spontaneously by more than one student— or Unique (U) —if mentioned by a single student. The corresponding labelling scheme has been applied to the negative statements, in this case however, with the letter N (Negative) used instead of the letter P (Positive). Both positive and negative elements were then ladderred according to the number of Mentions, which represent the number of times each element was cited by students as their perception of the quiz experience. We then calculated the average for each element, representing the simple mean of the individual importance ratings assigned by the respondents, where ratings ranged from 1 (minimum importance) to 10 (maximum importance). Finally, the variance was calculated to indicate the level of heterogeneity among individual assessments of the same element.

Given the brief nature of student responses, no specialized qualitative analysis tool was required. Hence, three researchers individually analysed all the collected short answers and subsequently met to reach a consensus about each one of the elements (Petchamé et al., 2021). This approach included a triangulation process aimed at improving the result validity (Vaismoradi & Snelgrove, 2019).

RESULTS

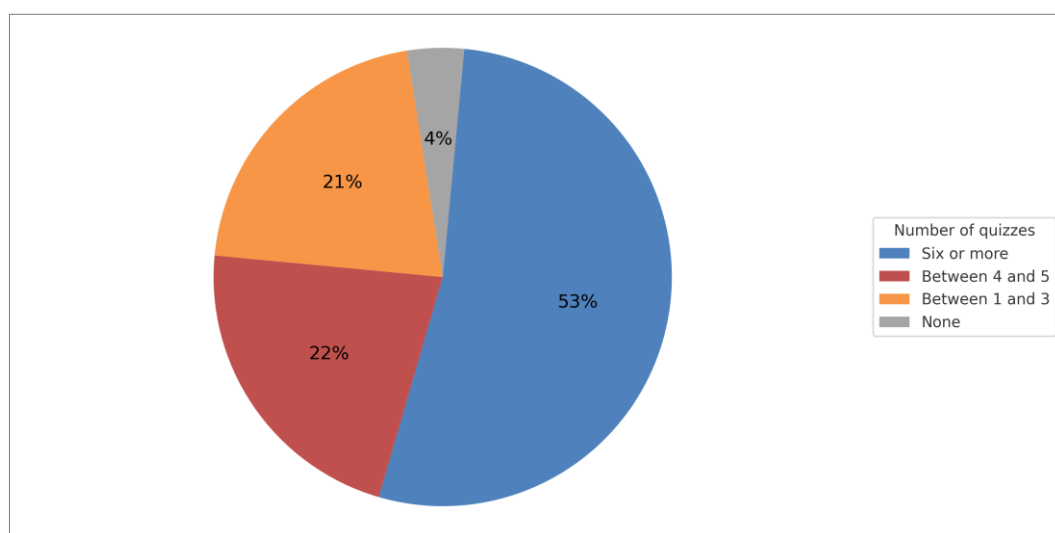
Our BLA survey was carried out as anonymous at the end of the 2024-2025 first academic term among four groups of the second-year ICT Engineering students at La Salle-URL. Prior to the survey, all participants completed their learning experience in a business-related subject 'Financial Economics'. The total number of students enrolled in the course was 209 and 51 of them responded to the BLA survey. Only two male respondents were in the age group 23-25 years old. In addition to age, the surveyed student cohort shared other similarities such as the level of digital literacy, educational background, and socioeconomic status. Respondent demographic data is shown in Table 1 below.

Note that the BLA survey should be administered during the final class session to ensure the students completed full learning experience before providing feedback. Conducting the survey at that time allows for a more comprehensive assessment of the entire learning process and aligns with the usual class attendance patterns observed in most subjects at that stage. Although the number of responses is not a limitation in BLA methodology, the survey could be distributed before the end of the last class session or the time limit for survey completion could be extended to obtain more responses.

Table 1*BLA survey respondents by gender and age group*

	Number	%	Age M	Age DE
Male	38	75	20.2	0.9
Female	11	21	20	0
Prefers not to declare	2	4	20	0
Total	51	100	20.2	0.8

As shown in Figure 1 below, over half of the respondents (n=27; 53%) had considerable experience in post lecture quizzes.

Figure 1*Number of quizzes taken by the respondents*

Through the survey, the researchers identified eight Positive Common Elements (PCE-1 to PCE-8) and five Positive Unique Elements (PUE-1 to PUE-5), as per Table 2 below. It displays the positive elements of the quiz experience respondents reported in the form of statements. The 'Mentions' column shows the number of times each element was cited, and the associated percentages based on the total number of mentions (n=105). Because each participant could report up to three positive elements, 'Mentions' percentages do not sum to 100% of respondents. These results indicate that, for students, the most impactful aspects of the post-lecture online quiz experience were the ability to deepen their understanding of the course material and to help them in exam preparation (e.g., "because it seems to me to be a simple and quick way to prepare for the exam"). Participants also highlighted the immediate feedback online quizzes provide, allowing them to rapidly recognize their knowledge gaps. PCE-4 and PCE-5 received fewer mentions suggesting that those elements may play a more supportive role: regular quizzing to stay current with course topics ("It is positive because it is a mandatory way to continuously be up to date with the subject as well as it contributing to the grade"), as well as online quiz accessibility, availability and variety of question formats. A few students highlighted the dynamic and engaging nature of quizzes which makes learning more enjoyable ("it is fun and motivating"). Mentions of PUE elements

were rare, each accounting for just under 1% of the total. Respondents were encouraged to rate on a scale 1 to 10 (1=minimum; 10=maximum) the importance of each reported statement, which is shown below in the 'Average' column, while the 'Variance' reflects the degree of variation among those individual ratings.

Table 2

Student perceptions of elements that enhance their learning

Statement	Element	Mentions n	Mentions %	Average	Var
Quizzes help me to better understand course material, review and reinforce core subject concepts.	PCE-1	28	27	8.6	1.17
Quizzes are a valuable practice tool for exam preparation.	PCE-2	25	24	8.2	1.92
Quizzes help me identify gaps in my understanding by means of immediate feedback.	PCE-3	14	13	8.0	2.43
Weekly quizzes help me stay up to date with the course.	PCE-4	9	9	8.4	1.36
Quizzes are short and flexible, with varying question formats and enough time to complete them.	PCE-5	9	9	8.2	1.95
Quizzes are organized by unit, easy to access and always open.	PCE-6	7	7	9.0	0.86
Quizzes have a positive impact on my final grade.	PCE-7	5	4	8.2	1.36
Quizzes are dynamic and engaging, which makes my learning more enjoyable and motivating.	PCE-8	3	2	8.0	0.67
Quizzes include questions beyond the lecture slides which helps me consolidate newly introduced topics.	PUE-1	1	1	10.0	
Quizzes encouraged me to review the lecture slides, focus on key points, and remember the content afterward.	PUE-2	1	1	10.0	
Quizzes foster active learning by applying theory to real-life scenarios.	PUE-3	1	1	10.0	
Quizzes create a competitive element that motivates me to excel.	PUE-4	1	1	7.0	
Randomizing the order of quiz questions helps prevent memorizing patterns and promotes real understanding.	PUE-5	1	1	6.0	

Note: Elements (E) are identified as Positive (P), Common (C) or Unique (U)

On the other hand, 51 students shared a total of 63 elements they found challenging in their learning experience. Again, because each participant could report up to three negative elements, 'Mentions' percentages do not sum up to the total number of respondents but rather to the total number of mentions (n=63). All negative perceptions are grouped into nine commonly reported themes (NCE-1 to NCE-9) and five unique themes (NUE-1 to NUE-5), as shown in Table 3 below. Students expressed several concerns about post-lecture quizzes, offering suggestions that could potentially increase the educational value of this activity. The most frequently raised point (n=17; 27%) involved time constraints for quiz completion, particularly when questions felt long, repetitive or unclear (n=10; 16%). Other statements were accompanied by ideas to improve learning experience, such as: allow multiple quiz attempts with continued access upon completion; align quiz content more closely with the exam material; include more practical examples and reasoning-based questions; provide more detailed feedback, especially explanations for incorrect answers (as one student noted: "I don't like that it only tells you if your answers are correct or wrong without giving you any arguments to support the answer"). While these observations reflect the experiences of certain students, they may offer useful starting points for implementation of this active learning task. Other points were mentioned less frequently.

Table 3

Student perceptions of elements that hinder their learning

Statement	Element	Mentions n	Mentions %	Average	Var
Some quizzes require too much time to complete within the available time window.	NCE-1	17	27	3.5	1.31
Some quiz questions are too long, unclear or repeated.	NCE-2	10	16	4.0	0.60
Quizzes don't allow multiple attempts nor access after quiz completion.	NCE-3	6	9	4.3	0.56
Some quiz questions do not align well with exams.	NCE-4	6	9	4.2	2.14
Quizzes lack practical examples and reasoning-based questions.	NCE-5	5	8	3.4	2.64
The number of quiz questions is insufficient.	NCE-6	4	6	5.0	0.00
Missing or underperforming on a mandatory quiz can negatively affect the final grade.	NCE-7	4	6	4.3	0.19
Quizzes lack feedback such as explanations for error.	NCE-8	4	6	3.3	1.69
Some graphs in the quizzes are not visible.	NCE-9	2	3	1.5	0.25

Statement	Element	Mentions n	Mentions %	Average	Var
Quizzes do not include randomized question sets.	NUE-1	1	2	5.0	
Lack of deadline reminders by email leads to missed quiz attempts.	NUE-2	1	2	5.0	
Quizzes have little weight in the final grade.	NUE-3	1	2	4.0	
Some quiz questions cover material that was not addressed during class sessions.	NUE-4	1	2	2.0	
Quizzes were only somewhat helpful for long-term retention of theoretical concepts.	NUE-5	1	2	1.0	

Note: Elements (E) are identified as Negative (N), Common (C) or Unique (U)

Additionally, students proposed improvements for each perceived element of their quiz experience, offering suggestions to either strengthen benefits or mitigate limitations. For example, increasing the number of quizzes (e.g., “There is only one (quiz) per unit, and perhaps it would be beneficial to add one or two more for additional practice” or “Add more questions and make them change from test to test”); notifying students when quizzes are launched (e.g., “Students should be notified via email when a new quiz is published”); making quizzes suitable for practice (e.g., “Unlimited access to the quiz review could be granted immediately following the submission deadline”, or “Add exercises to the quizzes”).

DISCUSSION

The present study researched engineering student perceptions of post-lecture online quizzes as an active learning tool, highlighting perceived benefits as well as potential areas for refinement. While the findings could inform instructors seeking to implement and optimize this pedagogical approach, it is important to interpret student feedback thoughtfully and in context.

Most students recognized post-lecture quizzes as a valuable study tool, particularly for exam preparation. This aligns with previous research demonstrating that regular testing can promote knowledge retention through the ‘testing effect’ (Roediger & Karpicke, 2006). In that regard, quizzes may serve as a scaffolding tool, providing support that is closely aligned with the intended learning outcomes. The overall coherence of the tool could be enhanced by ensuring better alignment between the objectives of formative assessments and the evaluative criteria of summative assessments. Several students noted the benefit of the immediate feedback online quizzes provide, which they felt allowed them to quickly identify and address knowledge gaps—a view consistent with the Hattie and Timperley (2007) model of effective feedback in learning environments. While students tend to emphasize scores, the literature suggests that the feedback itself—not the score—is the true accelerator of learning. Responses to our BLA survey further suggested that weekly quizzes could encourage students to interact more regularly with the course content, potentially

reducing the tendency to relay on last-minute exam preparation. Although these perceptions align with the evidence that the use of distributed practice through regular low-stakes quizzes can be more effective for long term knowledge retention than massed practice (Dunlosky et al., 2013), additional evidence would be needed to confirm whether this benefit was consistently achieved in practice. Student comments also highlighted potential motivational benefits of regular assessment, while a few participants noted the quizzes were entertaining, suggesting that the gamification elements of quizzes might positively contribute to the student experience (Dicheva et al., 2015; Majuri et al., 2018).

The mean scores for the importance of Positive Common Elements ranged from 8.0 to 9.0, indicating generally appreciative feedback across main items. The variance values ranged from 0.67 for the item 'Quizzes are dynamic and engaging' to 2.43 for 'Quizzes help me identify gaps in my understanding by means of immediate feedback', suggesting that some items elicited more divergent opinions than others. The item 'Quizzes are organized by unit, easy to access, and always open,' received the highest mean score of 9.0 (SD=0.86), indicating strong consensus, albeit among only seven students who mentioned it.

Participant feedback also revealed several opportunities for improvement. The most frequently mentioned concerns pertained to time constraints and question design. Students found lengthy, repetitive or ambiguously phrased questions as obstacles to effective learning. This echoes the research by Gillmor et al. (2015) stating that poorly constructed assessments can impose unnecessary cognitive load that interferes with intended learning outcomes. Respondents also expressed preference for unlimited quiz attempts and immediate access to results, which may reflect their desire for quizzes to function primarily as learning tools rather than evaluative instruments. This perspective aligns with the formative assessment paradigm described by Black and Wiliam (1998), which emphasized assessment as a way of guiding learning rather than as a performance measure. Another enhancement that students commented on was the opportunity to link theoretical concepts with their practical application. Respondents asked for more reasoning-based questions, which points to higher-order learning objectives in Bloom's taxonomy of learning (Tofade et al., 2013). The statement calls for reducing the focus on lower-order knowledge-recall questions, and instead prioritizing application, analysis and evaluation-level questions to facilitate deeper understanding and long-term knowledge retention. Last but not least, students commented on the feedback mechanism, particularly noting the absence of explanations for incorrect answers. The importance of explanatory feedback on errors was studied by Shute (2008), who demonstrated that elaborate feedback was more effective than verification feedback alone. Implementing appropriate explanations for both correct and incorrect responses could be a meaningful way to enhance learning, requiring minimal effort.

Looking at mean scores and variances of the Negative Common Elements, the most frequently mentioned statement on the time to complete quizzes (n=17; 27%) received a moderate 3.5 average of importance, suggesting that, while often acknowledged, opinions about it were somewhat mixed. On the other hand, the statement 'The number of quiz questions is insufficient' received a maximum 5.0 average (SD=0.00), indicating complete agreement albeit among only four students who touched on it. Two other items attracted our attention because of the range of views they elicited: the statement on the lack of reasoning-based questions (SD=2.64)

and the low level of alignment with exams ($SD=2.14$) —some respondents considered them to be very important, while others gave low ratings.

In terms of the limitations of this study, its very nature, which focuses on identifying positive and negative elements (elicited through the BLA tool) associated with post-lecture online quizzes, led to deliberate exclusion of data on respondents' academic performance. Future research could incorporate objective measures of academic results, expand the sample size and employ methodological triangulation to strengthen the validity and applicability of the findings. It could also investigate the correlation between quiz engagement patterns and examination performance to provide more robust evidence for optimizing quiz design. Likewise, experimental studies comparing different feedback mechanisms could offer more precise guidance for maximizing the pedagogical effectiveness of post-lecture online quizzes. Additionally, this study was conducted with the second-year cohort of ICT Engineering students. Future research could include participants from different academic levels, for example first-year, third-year or final year students to assess whether the findings hold true across varying levels of experience and knowledge. Researchers could replicate this study to other disciplines that use the online quizzes, to evaluate whether results are specific to engineering or generalizable across fields. To increase sample diversity, future studies could also include international students or participants from other faculties inside or outside the institution.

CONCLUSIONS AND RECOMMENDATIONS

This study explored the use of weekly post-lecture online quizzes as a formative assessment tool for ICT Engineering undergraduates on a management course. While pre-lecture quizzes have been well-studied, post-lecture quizzes received comparatively less attention. In this intervention, students completed automatically graded MRQ quizzes on Moodle LMS after each subject unit, with timely participation accounting for 10% of their final grade. This research integrates an active learning theoretical framework with a qualitative methodology grounded in Constructivist Psychology, aiming to examine user experiences, classify participant perceptions into strengths and weaknesses of the implemented model along with corresponding importance ratings, and collect ideas for enhancing the model benefits and mitigating its limitations.

Findings from the anonymous Bipolar Laddering (BLA) survey suggest that, from the student perspective, post-lecture online quizzes may be a useful active learning strategy, particularly in supporting exam preparation and ongoing engagement with course material. Students reported that frequent low-stakes quizzes can facilitate content recall and knowledge retention, encourage self-directed learning, and promote regular content review. Furthermore, students indicated that active learning may be enhanced when quizzes extend beyond mere information recall and include questions that encourage deeper thinking. Also, student perceptions of quiz effectiveness were closely linked to quiz usability (e.g., intuitive access, manageable length, diverse formats) and system-level support (e.g., timely notifications, visual clarity, etc.). Respondents noted that the effectiveness of post-lecture online quizzes may partly depend on balancing structure (e.g., by-unit organization, consistent availability and feedback) and flexibility (e.g., multiple attempts, explanatory feedback, post-completion access). Finally, our findings suggest that quiz design features can foster student motivation, particularly when the objectives of formative assessments

are perceived to be aligned with the summative assessment criteria and post-lecture quiz questions are well-aligned with the course examinations.

Building on these student-reported insights, this study offers the following ten recommendations for designing post-lecture quizzes that may support both online and in-person education formats:

- Quiz design: Develop concise, clearly worded questions that align with intended learning objectives, while avoiding unnecessary complexity or length.
- Grade weighting: Consider an appropriate weight of quiz scores within the overall assessment framework to incentivize regular engagement with the course content, without increasing performance anxiety.
- Exam alignment: Where possible, align quiz content with exam blueprints and in-class coverage to strengthen perceived value for exam preparation.
- Question quality: Include practical and scenario-based questions that may facilitate transfer of learning to real-world contexts. Where feasible, incorporate reasoning-based items that encourage deeper thinking.
- Question format: Use a variety of question formats (e.g., multiple-choice, true/false, short answer, blank filling, etc.) to address different learning approaches and dimensions of understanding.
- Answer feedback: Provide brief explanations or annotated solutions where possible and consider offering at least one quiz retry opportunity to transform errors into learning opportunities.
- Quiz access: Reopen completed quizzes in a 'study mode', i.e., with unlimited attempts, to emphasize their formative role and reduce their perception as high-stakes assessments.
- Time considerations: Balance quiz comprehensiveness with reasonable time requirements to minimize fatigue and maintain a manageable cognitive load. Where appropriate, consider creating multiple quizzes per course unit.
- Randomization: Develop question banks with randomization features to discourage rote memorization strategies and promote broader conceptual understanding across attempts.
- Reflection: Include short post-quiz reflective activities to encourage students to process results metacognitively and identify areas for further study.

These recommendations should be interpreted in the context of the study limitations, as they are primarily based on student perceptions. Future research could test their impact on learning outcomes using objective performance measures and larger, more diverse, samples.

REFERENCES

- Allsop, J., Young, S. J., Nelson, E. J., Piatt, J., & Knapp, D. (2020). Examining the benefits associated with implementing an active learning classroom among undergraduate students. *International Journal of Teaching and Learning in Higher Education*, 32(3), 418-426.
- Andrews, M. E., Graham, M., Prince, M., Borrego, M., Finelli, C. J., & Husman, J. (2020). Student resistance to active learning: Do instructors (mostly) get it wrong? *Australasian Journal of Engineering Education*, 25(2), 142-154. <https://doi.org/10.1080/22054952.2020.1861771>

- Arthurs, L. A., & Kreager, B. Z. (2017). An integrative review of in-class activities that enable active learning in college science classroom settings. *International Journal of Science Education*, 39(15), 2073-2091.
<https://doi.org/10.1080/09500693.2017.1363925>
- 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>
- Bonwell, C. C., & Eison, J. A. (1991). *Active learning: Creating excitement in the classroom* (1991 ASHE-ERIC Higher Education Reports). ERIC Clearinghouse on Higher Education, The George Washington University.
<https://eric.ed.gov/?id=ED336049>
- Brink, A. G. (2013). The impact of pre- and post-lecture quizzes on performance in intermediate accounting II. *Issues in Accounting Education*, 28(3), 461-485.
<https://doi.org/10.2308/iace-50445>
- Cook, B. R., & Babon, A. (2017). Active learning through online quizzes: Better learning and less (busy) work. *Journal of Geography in Higher Education*, 41(1), 1--14.
<https://doi.org/10.1080/03098265.2016.1185772>
- Dengri, C., Gill, A., Chopra, J., Dengri, C., Koritala, T., Khedr, A., Korsapati, A. R., Adhikari, R., Jain, S., Zec, S., Chand, M., Kashyap, R., Pattan, V., Khan, S. A., & Jain, N. K. (2021). A review of the quiz, as a new dimension in medical education. *Cureus*, 13(10), e18854.
<https://doi.org/10.7759/cureus.18854>
- Dicheva, D., Dichev, C., Agre, G., & Angelova, G. (2015). Gamification in education: A systematic mapping study. *Journal of Educational Technology & Society*, 18(3), 75-88.
- Doolittle, P., Wojdak, K., & Walters, A. (2023). Defining active learning: A restricted systematic review. *Teaching and Learning Inquiry*, 11, Article 25.
<https://doi.org/10.20343/teachlearningq.11.25>
- Driessen, E. P., Knight, J. K., Smith, M. K., & Ballen, C. J. (2020). Demystifying the meaning of active learning in postsecondary biology education. *CBE-Life Sciences Education*, 19(4), ar52.
<https://doi.org/10.1187/cbe.20-04-0068>
- Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, 14(1), 4-58.
<https://doi.org/10.1177/1529100612453266>
- Evans, T., Kensington-Miller, B., & Novak, J. (2021). Effectiveness, efficiency, engagement: Mapping the impact of pre-lecture quizzes on educational exchange. *Australasian Journal of Educational Technology*, 37(1), Article 1.
<https://doi.org/10.14742/ajet.6258>
- Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences*, 111(23), 8410-8415.
<https://doi.org/10.1073/pnas.1319030111>
- Gillmor, S., Poggio, J., & Embretson, S. (2015). Effects of reducing the cognitive load of mathematics test items on student performance. *Numeracy*, 8(1), Article 4.
<https://doi.org/10.5038/1936-4660.8.1.4>
- Gosavi, C. S., & Arora, S. (2022). Active learning strategies for engaging students in higher education. *Journal of Engineering Education Transformations*, 36(S1), 1-7.
<https://doi.org/10.16920/jeet/2022/v36is1/22167>
- Gowe, J. E. (2024). The impact of quizzes on improving mathematics performance among higher education students. *The Journal of Informatics*, 4(1), Article 1.
<https://doi.org/10.59645/tji.v4i1.397>
- Gyllen, J. G., Stahovich, T. F., Mayer, R. E., Entezari, N., & Darvishzadeh, A. (2021). Priming productive study strategies with preparatory quizzes in an engineering course. *Applied Cognitive Psychology*, 35(1), 169-180.
<https://doi.org/10.1002/acp.3750>

- Hartikainen, S., Rintala, H., Pylväs, L., & Nokelainen, P. (2019). The concept of active learning and the measurement of learning outcomes: A review of research in engineering higher education. *Education Sciences*, 9(4), Article 276. <https://doi.org/10.3390/educsci9040276>
- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81-112. <https://doi.org/10.3102/003465430298487>
- Heil, J., & Ifenthaler, D. (2023). Online assessment in higher education: A systematic review. *Online Learning*, 27(1), 187-218. <https://doi.org/10.24059/olj.v27i1.3398>
- Hennink, M., & Kaiser, B. N. (2022). Sample sizes for saturation in qualitative research: A systematic review of empirical tests. *Social Science & Medicine*, 292, 114523. <https://doi.org/10.1016/j.socscimed.2021.114523>
- Horna-Saldaña, C., & Canaleta, X. (2024). Application of universal design for learning and digital fabrication in the creation of a tool for inclusive teaching of the ordering of chemical elements. *Journal of Chemical Education*, 101(12), 5261-5271. <https://doi.org/10.1021/acs.jchemed.4c00679>
- Howell, R. A. (2021). Engaging students in education for sustainable development: The benefits of active learning, reflective practices and flipped classroom pedagogies. *Journal of Cleaner Production*, 325, 129318. <https://doi.org/10.1016/j.jclepro.2021.129318>
- Kirzner, R. S., Alter, T., & Hughes, C. A. (2021). Online quiz as exit ticket: Using technology to reinforce learning in face-to-face classes. *Journal of Teaching in Social Work*, 41(2), 151-171. <https://doi.org/10.1080/08841233.2021.1898521>
- Klein, K., Calabrese, J., Aguiar, A., Mathew, S., Ajani, K., Almajid, R., & Aarons, J. (2023). Evaluating active lecture and traditional lecture in higher education. *Journal on Empowering Teaching Excellence*, 7(2), Article 2. <https://doi.org/10.26077/ba42-a5cc>
- Kozanitis, A., & Nenciovici, L. (2023). Effect of active learning versus traditional lecturing on the learning achievement of college students in humanities and social sciences: A meta-analysis. *Higher Education*, 86(6), 1377-1394. <https://doi.org/10.1007/s10734-022-00977-8>
- López-Tocón, I. (2021). Moodle quizzes as a continuous assessment in higher education: An exploratory approach in physical chemistry. *Education Sciences*, 11(9), Article 500. <https://doi.org/10.3390/educsci11090500>
- Majuri, J., Koivisto, J., & Hamari, J. (2018). Gamification of education and learning: A review of empirical literature. In *Proceedings of the 2nd International GamiFIN Conference* (pp. 11-19). CEUR Workshop Proceedings, 2186.
- Mitchell, A., Petter, S., & Harris, A. (2017). Learning by doing: Twenty successful active learning exercises for information systems courses. *Journal of Information Technology Education: Innovations in Practice*, 16, 21-46. <https://doi.org/10.28945/3643>
- Nuci, K. P., Tahir, R., Wang, A. I., & Imran, A. S. (2021). Game-based digital quiz as a tool for improving students' engagement and learning in online lectures. *IEEE Access*, 9, 91220-91234. <https://doi.org/10.1109/ACCESS.2021.3088583>
- Owens, D. C., Sadler, T. D., Barlow, A. T., & Smith-Walters, C. (2020). Student motivation from and resistance to active learning rooted in essential science practices. *Research in Science Education*, 50(1), 253-277. <https://doi.org/10.1007/s11165-017-9688-1>
- Petchamé, J., Iriondo, I., Korres, O., & Paños-Castro, J. (2023). Digital transformation in higher education: A qualitative evaluative study of a hybrid virtual format using a smart classroom system. *Heliyon*, 9(6), e16675. <https://doi.org/10.1016/j.heliyon.2023.e16675>

- Petchamé, J., Iriondo, I., Torné, O., & Solanellas, F. (2024). Student perceptions of project-based learning when studying information systems in an ICT engineering program: Taking advantage of active learning. *Education Sciences*, 14(11), Article 1148. <https://doi.org/10.3390/educsci1411148>
- Petchamé, J., Iriondo, I., Villegas, E., Fonseca, D., Romero Yesa, S., & Aláez, M. (2021). A qualitative approach to help adjust the design of management subjects in ICT engineering undergraduate programs through user experience in a smart classroom context. *Sensors*, 21(14), 4762. <https://doi.org/10.3390/s21144762>
- Pifarré, M., & Tomico, O. (2007). Bipolar laddering (BLA): A participatory subjective exploration method on user experience. In *Proceedings of the 2007 Conference on Designing for User eXperiences* (pp. 2-13). ACM. <https://doi.org/10.1145/1389908.1389911>
- Prince, M. (2004). Does active learning work? A review of the research. *Journal of Engineering Education*, 93(3), 223-231. <https://doi.org/10.1002/j.2168-9830.2004.tb00809.x>
- Raes, A., Vanneste, P., Pieters, M., Windey, I., Van Den Noortgate, W., & Depaepe, F. (2020). Learning and instruction in the hybrid virtual classroom: An investigation of students' engagement and the effect of quizzes. *Computers & Education*, 143, 103682. <https://doi.org/10.1016/j.compedu.2019.103682>
- Roediger, H. L., III, & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, 17(3), 249-255. <https://doi.org/10.1111/j.1467-9280.2006.01693.x>
- Ross, B., Chase, A.-M., Robbie, D., Oates, G., & Absalom, Y. (2018). Adaptive quizzes to increase motivation, engagement and learning outcomes in a first year accounting unit. *International Journal of Educational Technology in Higher Education*, 15(1), Article 30. <https://doi.org/10.1186/s41239-018-0113-2>
- Rossi, I. V., de Lima, J. D., Sabatke, B., Nunes, M. A. F., Ramirez, G. E., & Ramirez, M. I. (2021). Active learning tools improve the learning outcomes, scientific attitude, and critical thinking in higher education: Experiences in an online course during the COVID-19 pandemic. *Biochemistry and Molecular Biology Education*, 49(6), 888-903. <https://doi.org/10.1002/bmb.21574>
- Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, 78(1), 153-189. <https://doi.org/10.3102/0034654307313795>
- Stoyanova, S., & Giannouli, V. (2022). Online testing as a means of enhancing students' academic motivation during the coronavirus pandemic. *Education Sciences*, 13(1), Article 25. <https://doi.org/10.3390/educsci13010025>
- Tharayil, S., Borrego, M., Prince, M., Nguyen, K. A., Shekhar, P., Finelli, C. J., & Waters, C. (2018). Strategies to mitigate student resistance to active learning. *International Journal of STEM Education*, 5(1), Article 7. <https://doi.org/10.1186/s40594-018-0102-y>
- Tofade, T., Elsner, J., & Haines, S. T. (2013). Best practice strategies for effective use of questions as a teaching tool. *American Journal of Pharmaceutical Education*, 77(7), Article 155. <https://doi.org/10.5688/ajpe777155>
- Vaismoradi, M., & Snelgrove, S. (2019). Theme in qualitative content analysis and thematic analysis. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, 20(3), Article 23. <https://doi.org/10.17169/fqs-20.3.3376>
- Veletsianos, G., Kimmons, R., Larsen, R., & Rogers, J. (2021). Temporal flexibility, gender, and online learning completion. *Distance Education*, 42(1), 22-36. <https://doi.org/10.1080/01587919.2020.1869523>
- Villalonga Pons, J., Besalú, M., Samà Camí, A., & Sancho-Vinuesa, T. (2023). Estrategias de aprendizaje de estudiantes de ingeniería en línea. *RIED-Revista Iberoamericana de Educación a*

Distancia, 26(2), 237-256.
<https://doi.org/10.5944/ried.26.2.36257>
Yang, C., Luo, L., Vadillo, M. A., Yu, R., &
Shanks, D. R. (2021). Testing (quizzing)

boosts classroom learning: A systematic
and meta-analytic review. *Psychological
Bulletin*, 147(4), 399-435.
<https://doi.org/10.1037/bul0000309>

Date of reception: 1 June 2025

Date of acceptance: 6 August 2025

Date of approval for layout: 16 September 2025

Date of publication in OnlineFirst: 13 October 2025

Date of publication: 1 January 2026