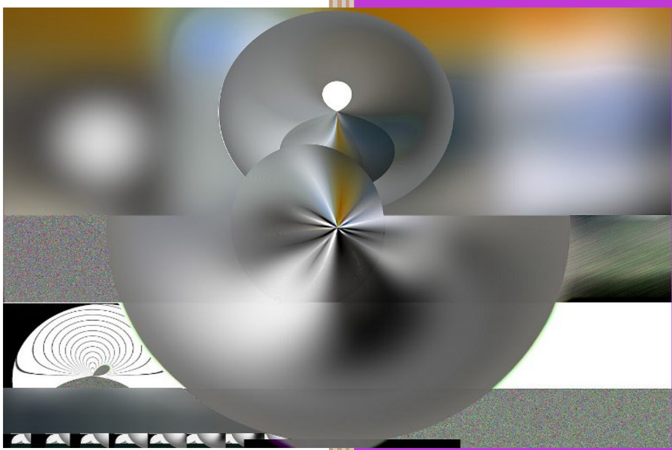
2026

EJEL Volume 24, Issue 3



Editors

Heinrich Söbke and Alessandro Pagano

Published by Academic Publishing International Limited
Curtis Farm, Kidmore End, Nr Reading, RG4 9AY, United Kingdom
karen.harris@academic-publishing.org

eISSN: 1479-4403

Cover artwork: Anastasis Kratsios, Public domain, via Wikimedia Commons

EJEL Volume 24, Issue 3

Designing a Gamified Retirement Planning Application: Evidence from a Workplace Study of Millennial Employees

Chrizaan Grobbelaar[1] and Liezel Alsemgeest[2]

[1]Department of Accounting and Auditing, Central University of Technology, Free State, Bloemfontein, South Africa

[2]School of Financial Planning Law, University of the Free State, Bloemfontein, South Africa

chrizaan@cut.ac.za (Corresponding Author)
alsemgeestl@ufs.ac.za

https://doi.org/10.34190/ejel.24.3.4375

Abstract: The phenomenon of a global retirement crisis has gained increasing attention in recent years. A growing number of individuals struggle to secure sufficient financial resources to sustain themselves during retirement. Financial knowledge not only increases the tendency to save for retirement, but it also affects an individual's everyday money practices, including borrowing, saving, and investing decisions. Millennials are currently the largest cohort in the workforce and should be constantly made aware of the urgent need to save from the earliest age possible. The development and implementation of a unique gamified retirement planning application aimed at improving retirement preparedness was critically examined for its design rationale, user experiences, and observed successes and shortcomings. A randomised control experiment was conducted to observe any changes the gamified tool might have on the players' retirement preparedness. Participants, millennial employees at a higher education institution, aged between 24 and 43 were randomly divided into one of three groups (gamification group, education group, and control group), ranging between 29 and 34 participants. The gamified application created a platform where the gamification group could experience the long-term effects of their decisions in a low-risk setting by modelling real-life financial decisions, which helped them develop a more accurate grasp of retirement planning. The education group was only exposed to infographics on retirement planning, and the control group was not exposed to any intervention. Data were analysed and tested using a nonparametric ANCOVA. The gamified application tends to be more effective at improving basic financial literacy, promoting interest, fostering awareness, and building confidence. However, a decrease was observed in respondents' perceptions of how well they are prepared for retirement and whether they will achieve the income they need in retirement, which may reflect a "reality check" effect. The infographics appeared to be more effective at improving financial retirement literacy and resulted in a more positive perception of how well they are prepared for retirement, which may reflect overconfidence in their knowledge, leading them to believe they are better prepared than they actually are. Despite some content and technical limitations of the gamified application, it exhibits promise as a financial education tool, motivating individuals to take proactive steps toward retirement planning. Collectively, the evidence suggests that gamification enhances financial literacy and interest in retirement planning, prompts self-reflection, and serves as a "reality check" on respondents' perceived confidence in retirement preparedness. Consequently, e-learning environments for financial education interventions should not only consider gamification for motivation, but also balance engagement, knowledge development, and behavioural awareness, ensuring a meaningful and realistic development of financial proficiency and behaviour. The limitations of the study suggest that the sample size limits statistical power and restricts the generalisability of the findings, the self-reported measures may be subject to social desirability bias or overconfidence, particularly among participants exposed to educational material, and, lastly, due to the time limit of the study, long-term knowledge retention and sustained behavioural change could not be assessed.

Keywords: Financial education, Retirement planning, Gamification, Millennials

1. Introduction

The phenomenon of a global retirement crisis has attracted increasing attention in recent years as demographic shifts, economic factors, and social policy inadequacies converge. This crisis is characterised by the inability of an increasing number of individuals to secure sufficient financial resources to sustain themselves during retirement (Aegon, 2017). Population ageing is a global reality regarded as one of the main economic and social changes shaping the 21st century. People are living longer, and fewer children are born. A smaller number of people will be working to support an increasingly ageing population dependent on their government for financial support in retirement (Bodnár & Nerlich, 2022). Low levels of financial knowledge, combined with retirement unpreparedness, are perceived as a significant problem among millennials (Ndou, 2023). The traditional classroom setting may have worked for previous generations, but it only discourages millennials (McHaney, 2023). The conventional approach of seminars, workshops, and other educational programmes on financial

Cite this article: Grobbelaar, C. and Alsemgeest, L. 2026. "Designing a Gamified Retirement Planning Application: Evidence from a Workplace Study of Millennial Employees", *Electronic Journal of e-Learning*, 24(3), pp 1-15, https://doi.org/10.34190/ejel.24.3.4375

education has become insufficient to motivate millennial employees to save and plan for retirement (Sutton, 2019).

It has been argued that the focus of financial education should be on millennial employees (those younger than 43), given that they currently form the largest segment of the workforce and still have the potential to impact their retirement situation (Hill, 2017). In addition, experimentation and active participation are a great part of millennials' learning style. To gain their interest and attention, they must be constantly stimulated and challenged since they approach learning as a "plug-and-play" (effortless connection of devices to a computer system to operate) experience that allows them frequent, fast interaction with content. Although gamification has increasingly gained traction as a promising approach to financial education, the evidence remains fragmented (Yulianto et al., 2024).

Prior research suggests that gamification may increase engagement, motivation, and short-term interest in financial content, especially among younger learners. However, evidence of sustained gains in financial literacy and behavioural outcomes remains inconsistent. Such outcomes appear highly contingent on the quality of game design, the contextual relevance, and individual user characteristics. Much of the available literature is focused solely on the acquisition of basic financial knowledge (Ahmad et al., 2025), with comparably limited attention being paid to retirement planning specifically, especially among full-time working individuals. The literature indicates that current financial education initiatives do not necessarily enhance millennials' engagement in finances, and the retirement situation is becoming increasingly dire due to inadequate planning and saving for retirement (Yakoboski, Lusardi, & Sticha, 2024). This growing challenge puts the financial well-being of millions of young working adults at risk and calls for innovative educational approaches to address this problem. Accordingly, further context-specific research is needed to understand whether gamification supports retirement education and whether the development of the gamified application influences engagement and perceived preparedness. This study addresses this gap by examining a gamified retirement planning application for millennial employees. This study seeks to answer two questions:

Will a gamified retirement planning application enhance financial literacy, increase interest in retirement planning, and improve perceived retirement preparedness?

What are the potential strengths and weaknesses of such an application?

The article aims to critically examine the development and implementation of a gamified retirement planning application to enhance the retirement preparedness of full-time millennial employees (ages 24 to 43) based at a tertiary education institution in South Africa, with a particular focus on its design rationale, user experiences, and observed successes and shortcomings. This article will highlight the impact of the global retirement crisis, the future implications for millennials specifically, the need for financial literacy and retirement preparedness, and explore the use of gamification and digital financial education. It will also elaborate in detail on the design and implementation of the gamified application, focusing specifically on the theories, player types, design, content, and game elements applied. The successes, shortcomings, and lessons learned will be examined, after which concluding remarks will be made.

2. Literature Review

The literature review will provide an overview of the global retirement crisis, financial and retirement literacy, and the importance of targeting millennials specifically.

2.1 The Global Retirement Crisis

The global reality is that people are increasingly being forced to retire earlier than planned, are steadily growing older each year, and reach retirement with little to no savings, which places immense fiscal pressure on governments worldwide. Employed individuals will pay higher taxes, old-age support will be reduced to match current tax and saving rates, and people will have to work longer and retire later to accumulate more income (Bodnár & Nerlich, 2022; Natixis, 2022). Low-wage workers (ages 50 to 60) are likely to be the first to lose their jobs in times of economic downturn (for example, the COVID-19 pandemic), forcing them to retire sooner than planned with little to no savings. These households enter retirement with more debt than they had a decade ago. Pensions are eroded relentlessly and other avenues of retirement income become more unreliable due to fluctuating financial markets, creating a great inequality between economic classes as they age. Currently, half of American households have no retirement savings. The retirement crisis dooms one to observe the rapid increase in the number of elderly people becoming poorer, fruitlessly searching for jobs, and adding to the ranks of the unemployed (Ghilarducci, 2024).

2.2 Financial Literacy

The complex financial landscape today calls for financial sophistication to build sufficient wealth and increase the individuals' tendency to plan for retirement (OECD, 2023). Financially sophisticated individuals are more likely to be aware of and occupied with saving for retirement. Evidently, planning and preparing for retirement leads to more wealth in retirement. Financial knowledge also impacts an individual's everyday money practices, including borrowing, saving, and investing decisions (Bai, 2023). Young people are frequently advised to start saving as early as possible, but this can pose a daunting and challenging task, as the younger workforce may lack sufficient discretionary income to put away in savings accounts, and they might lack the necessary knowledge and skills, or perceive savings as unimportant while still young (Rappaport, 2019). Research has shown that a lack of financial literacy leads to the absence of retirement planning. A financially illiterate individual with low income is most vulnerable to poor financial decisions (Sundarasen, Rajagopalan, & Ibrahim, 2024). The most recent survey conducted by the Global Financial Literacy Excellence Centre (GFLEC) assessing the financial literacy levels of US citizens, shows no improvement in overall financial literacy. The results of the survey (percentage of questions answered correctly) for the last eight years (2017-2024) show a decrease of 1% (49% in 2017 and 48% in 2024) (Yakoboski et al., 2024). The most recent financial literacy survey conducted by the OECD, which included 40 countries (excluding the US), showed an overall financial literacy score of 60% (percentage of questions answered correctly) (OECD, 2023). Comparing it to the OECD financial literacy survey in 2015, the overall score for 30 countries was 63%, indicating a 3% decrease (2015: 63% to 2023: 60%) (OECD, 2023; 2016). The OECD/INFE has been actively working towards strengthening the financial literacy of adults globally through financial education initiatives, but the effect thereof on the financial literacy levels of individuals is not evident yet, at least not in the OECD and GFLEC scores (OECD, 2023; 2019; Yakoboski et al., 2024).

2.3 Millennials

Millennials are currently the largest cohort of employees in the workforce (born between the 1980s and the beginning of the 2000s). They should be constantly made aware of the urgent need to save from the earliest possible age. A millennial's most expensive mistake is delaying retirement savings (Hill, 2017; Sanlam, 2017). Compared to previous generations, millennials grew up in a more protective environment with unrealistic expectations for the future. Due to ongoing demographic, political, and macroeconomic factors, millennials seem oblivious to the less comfortable retirement awaiting them compared to what previous generations experienced. Their limited understanding of financial matters might stem from a lack of information about the stark reality they may encounter upon retirement, as well as from a general disinterest in financial matters (Bank of New York Mellon, 2015). Despite millennials being the most educated generation to date, they are also less involved in retirement planning and tend to focus on their immediate personal accomplishments, while saving and investing for retirement are put on the back burner. However, they firmly believe that they will not be able to retire comfortably and will be worse off than previous generations (De Lima, Nery, & Miranda, 2024). The millennials' delay in saving for retirement was also evident in the results from the National Institute for Retirement Survey in the United States, where 78% of millennials indicated that they would have to continue working past retirement age due to their financial insecurity (Doonan & Kenneally, 2021). Working past retirement age may not be an option given the looming retirement crisis and the high percentage of retirees entering retirement involuntarily (Ghilarducci, 2024). Millennials will face a deeply troubling retirement outlook as they are not saving nearly enough or not saving at all for retirement due to reasons such as depressed wages, high student debt, and the lack of access to employer retirement plans (Doonan & Kenneally, 2021). The literature indicates that young people entering the workforce need financial education and guidance from their employers to equip them with sufficient knowledge to manage their income and choose the best investment options for retirement (Michaud, 2021). Motivating millennials to take part in the learning process means engaging them with visuals and involving them in the process. Millennials are more likely to choose online (distance) learning formats due to their immersion in technology (McHaney, 2023). Millennials are in favour of free, easily accessible, and easy-to-understand online financial education resources, but this alone is not enough to gain their attention (Lusardi & Hasler, 2019). In addition, millennials prefer personalised ways to learn how to manage their money (Huang, Lassu, & Chan, 2018).

2.4 Gamification and Digital Financial Education

Gamification is defined as the use of game elements in non-game contexts to promote learning by adding a layer of interest to the task at hand, increasing engagement and motivating action (Kapp, 2012). Online gamification programmes are a convenient way that has proven to be successful to teach content, making learning engaging, relevant, personalised, and motivational for millennial employees (Kapp, 2012; Mazzo, 2015). Using gamification as an educational tool can build confidence among millennials to take action to enhance their financial knowledge and start saving towards retirement (Grobbelaar & Alsemgeest, 2024). Furthermore, gamification can create a fictional setting where individuals can experiment without real-life consequences. This safe space allows individuals to strategise and learn from their mistakes through immediate feedback, which can foster sound behaviour in real life (Kapp, 2012). Difficult concepts can be explained in smaller sections of content to avoid the feeling of being overwhelmed. An essential benefit of gamification is the alleviation of psychological factors, such as fear and anxiety one experiences when managing finances. These psychological factors can be managed through gamification by creating smaller tasks and goals to achieve, rather than focusing on the end goal that may seem unachievable (Van Raaij, 2016). Gamification has gained significant attention across diverse fields – including education, health services, marketing, and personnel management (Rodrigues et al., 2018) – due to its ability to engage fundamental drivers of psychological motivation in addition to intrinsic and extrinsic incentives (Uppaluri, 2025). Uppaluri (2025) conducted a comparative analysis on users of gamified systems and non-gamified systems to investigate changes in users' engagement and saving behaviours. The results showed that after financial applications assimilated gamification, the daily active users increased by 191.67%, financial goals set by users increased by 180%, and the number of users meeting their financial goals increased by 85.71%. Thus, by incorporating game elements into financial applications, user activity increases, users are more motivated to set financial goals and objectives, and users are supported in meeting their set goals. The need to combine education and gamification into financial management platforms is essential to enhancing user performance and fostering financial stability. However, gamification must be applied responsibly to avoid the ethical dilemma of exploiting users instead of empowering them, while maintaining the integrity of financial guidance within these platforms (Uppaluri, 2025).

3. Methodology

The article aims to explore the design rationale, user experiences, and observed successes and shortcomings of a gamified retirement planning application developed and implemented to enhance financial literacy, increase interest in retirement planning, and improve perceived retirement preparedness of full-time millennial employees who fall between the ages of 24 and 43 at a tertiary education institution in South Africa. The best approach to determine changes in individuals' financial literacy levels, interest in the subject and perceptions, is for the researcher to be objective and independent of the research, which is also called a positivist philosophy. The primary aim of positivism is to draw conclusions through empirical research, namely observation and experimentation. Causal relationships between variables are established through causal laws and are connected to logical reasoning and linked with quantitative methods, where quantitative numerical data are statistically analysed (Collis & Hussey, 2014). Thus, an observation for successes and shortcomings was conducted through a randomised control experiment that focused on the cause-and-effect relationship between the independent variable (cause – financial education in retirement planning) and the dependent variable (effect – retirement preparedness).

The dependent variable was affected by offering financial education on retirement planning through a gamified application. The experiment involved three groups, ranging between 29 and 34 respondents per group (total respondents = 96), one group exposed to the gamified tool, one group exposed to educational infographics (also available in the gamified tool), and one control group. The recommended sample size for experimental research is 15 to 30 participants per group (Daniel, 2012; Gay, Mills, & Airasian, 2012). It is important to note that despite meeting the recommended minimum sample sizes for experimental designs, the relatively small group sizes limit the statistical power of the analyses and the generalisability of the findings beyond the specific context of the study.

The experiment used a pre-test, intervention, and post-test and lasted over a two-month period. The pre- and post-test data were collected through an online questionnaire that all three groups had to complete. The responses in the pre-test and post-test for the financial literacy and retirement preparedness questions were compared across the three experimental groups, and any differences among the groups' scores were linked to the intervention involved. Kapp (2012), however, cautioned that using a gamified application will not necessarily turn individuals into subject experts. Thus, an increase in the financial literacy score for a group exposed to an intervention exceeding that of the control group suggests a noteworthy impact, regardless of whether the difference is statistically significant. The difference can be traced back to the specific educational initiative involved.

All output data from the completed pre- and post-tests were extracted in the form of a Microsoft Excel report. The data collected were then imported, coded, and analysed using the statistical programme SPSS (Version:

29.0.0.0). The ANCOVA (analysis of covariance) test was applied to determine whether the pre- and post-test mean scores differed significantly across groups. In order to do a parametric ANCOVA test, two of the assumptions must be met, namely: the assumption of homogeneity of variance and the assumption of normality. For each assumption, there is a test to be conducted and if either one of the assumptions is not met, a non-parametric ANCOVA test must be performed. For either of the tests to be met, the desired p-value should be more than 0.05 (Field, 2018). The results of both the homogeneity test and the normality test indicated that some variables' p-values were greater than 0.05 and others were less. Thus, none of the assumptions were met, and a non-parametric ANCOVA test was performed. The non-parametric ANCOVA test used to analyse the data will be discussed later on. Internal and external validity were retained throughout the experiment and protected against history threats (influence by outside events), maturation threats (development of participants), and repeated testing and instrumentation threats, through the use of a randomised pre-test-post-test control group and experimental group design and being conducted in a real-life setting (Babbie, 2016). The development and implementation of the gamified application was part of a larger PhD study.

4. Design and Implementation

The design of the gamified application was based on the determinants of retirement planning, including financial literacy, retirement goals, and future time perspective (Anuar, Ismail, & Samad, 2023). Financial literacy is the ability to allocate financial resources more effectively by analysing financial data and making informed decisions when saving and building wealth over a lifetime (Ifeanyi, Rena, & Prinsloo, 2019). Yong, Yew, and Wee (2018) described it as the ability to apply financial skills such as personal financial management, budgeting, and saving that enable individuals to make sound financial decisions. A retirement goal is a long-term objective that provides a clear direction of the preferred living standard and quality in retirement to be achieved and serves as a driving force that inspires one to start planning (Jiun, Satar, & Ishak, 2021). Future time perspective refers to the extent to which an individual considers the future, anticipates potential outcomes, and plans before taking any actions (Kooij et al., 2018).

The purpose of the gamified retirement planning application was to provide players with a gameful experience of real-world financial matters related to retirement planning, what being prepared for retirement entails, and to raise awareness of the need to plan and save for retirement. Players were tasked with making decisions on everyday financial and retirement matters and viewed the impact of their decisions and overall performance through immediate, constant feedback. In reality, the impact of a financial decision can only be experienced after a certain time has elapsed (Jain & Dutta, 2019). The expectation was that after playing the game, players would show an increase in their financial literacy levels, and that their perception of how well they are prepared for retirement would be more accurate in terms of attitudes and behaviours towards retirement planning.

Taking into account the determinants of retirement planning and the intended goal to be achieved through a gamified retirement planning application, several theories relevant to gamification, including both learning and motivational theories, were explored to act as the foundation for the development of the gamified application. These theories inform how individuals acquire knowledge and skills, how to influence attitudes, and change behaviours. It also serves as guidance for educators in designing effective gamified learning environments by incorporating appropriate game elements (Kapp, 2012). Identifying the appropriate type of player within the gamified retirement planning application environment can also contribute to the development of the application. The type of player can also indicate which game elements to incorporate into the gamified application (Marczewski, 2018). The following section explores the relevant theories and player types that support the intended outcomes of the gamified application.

4.1 Theories

The design and implementation of the gamified application were built on four gamification theories. The first is the theory of gamified learning (Landers et al., 2019), which specifies that game elements can be used in psychological processes to affect learning. During the development of the gamified application, specific game elements were incorporated to influence players' behaviours and attitudes towards retirement planning and retirement preparedness. Goals were incorporated to enhance players' understanding of how retirement planning can help them achieve their retirement goals. Narrative elements created an environment that players could relate to and change their attitude towards planning and preparing for retirement. The social learning theory allows players to observe a model's desired behaviour, process it internally, and afterwards exhibit the learned behaviour (Kapp, 2012). Through participation in the gamified application, players observe and practice the process required to prepare for retirement and, ultimately, reach the retirement goal. The play environment was safe and secure, allowing the participants to make financial decisions and learn from their mistakes without experiencing the consequences in real life. Distributed practice allows for the educational content within the gamified application to be spread across manageable portions at the relevant stage in the application, to prevent information overload, and develop long-term information retention abilities (Kapp, 2012). The self-determination theory focuses on the fundamental purpose of human motivation to perform a task by fulfilling one's psychological needs (autonomy, competence, and relatedness). Players had control (autonomy) over their financial decisions through immediate feedback, effort was required to achieve the retirement goal (competence), and a narrative of everyday financial matters and unexpected life events allowed them to connect with the gamified environment (relatedness) (Richter, Raban, & Rafaeli, 2015).

4.2 Player Types

Marczewski (2018) developed six types of players for gamified applications: socialisers, free spirits, achievers, philanthropists, disrupters, and players. The type of player depends on intrinsic motivational factors and behaviours that players are expected to develop when engaging with the gamified application. Planning for retirement is solely the individual's responsibility (Iwry et al., 2021). Keeping this in mind, the gamified application was developed for a single player to engage with the gamified content without competition or cooperation with other players. Thus, the types of players suited for such an application were the achiever and the free spirit. The definition of each player guided the specific game element to incorporate into the application.

Achievers are driven by mastery, by improving themselves, and overcoming challenges is their strength. Achieving their desired retirement goal will motivate them to engage with the gamified content until they succeed. Free spirits want to be in control and choose their own path. Their strength lies in their ability to create wealth within the gamified application and have total control over their finances. Game elements that are identified by these two types of players are challenges, rewards, new skills, levels, choosing their own path, unlocking content, avatars, and resources (virtual currency) (Marczewski, 2018).

4.3 Design

Kapp (2012) suggested a development framework used to document the gamified application. A detailed discussion of the framework employed is provided as follows:

4.3.1 Overview of the gamified retirement planning application

The gamified application was developed to teach individuals the essence of being financially prepared for retirement. It depicted a real-life scenario of an employee over a 35-year period, in which players had to make general financial and savings decisions to reach their chosen retirement goal. At the start of the application, the players chose the retirement goal they wanted to achieve. Construct play allowed players to use their income earned within the gamified application to gather assets, pay expenses, and save towards retirement, ultimately reaching the chosen retirement goal. Infographics were provided to help them make informed decisions when faced with financial events (e.g., deciding whether to buy or rent a house).

4.3.2 Instructional objectives

Instructions were clear and guided the user to reach their retirement goal. The intended outcome of participating in the gamified application was for players to improve their financial literacy and change their financial attitudes and behaviours.

4.3.3 Character description

Players could choose an avatar to play the role of a lecturer at a higher education institution, which closely resonated with their real-world employment environment at the time. Employment started from the age of 26 until 60 years. At the time the gamified application was introduced, the institution's mandatory retirement age was 60. The player earned a salary and contributed to the employer's sponsored retirement fund. Disposable income was the resource available to the player for making financial choices and dealing with unexpected life events.

4.3.4 Gamified environment and gameplay

The gamified application was a web-based, fixed-screen application that displayed the avatar and tabs for the monthly budget, financial ledger, and assets and liabilities. On the same screen, a window displayed the financial choice and the event the player faced. Once the player made a financial decision by clicking the preferred option from the available options, the effect of the decision was immediately shown in the monthly budget and other financial feedback statements.

4.3.5 Reward structure

Rewards are received in the form of retirement savings within the gamified application. Sound financial decisions will increase retirement savings, and adverse financial decisions will deplete or restrain retirement savings growth. Whatever financial decisions were made eventually affected their retirement savings at the end of the application, which then determined whether they reached their retirement goal.

4.3.6 Look and feel of the gamified application

The gamified application had basic aesthetics without any special features or sound. The application was in the form of a click-and-play, single-player setting.

4.3.7 Technical aspects

The gamified application was entirely web-based, and no software needed to be downloaded. The application was compatible with laptops and computers with internet access.

4.4 Content

Players faced day-to-day and unexpected financial events that required them to make financial decisions. The content of these financial events incorporated into the gamified application was gained from literature on the phases of life during employment, financial literacy, retirement planning, and retirement preparedness. The financial topics included the following:

- Salaries and key deductions.
- Retirement funds – rules, regulations, calculations, taxation, life and living annuities, and income replacement rates.
- Medical health products.
- Accommodation – buying or renting: pros and cons, and the costs involved.
- Financial products – various insurance products, investment properties, equity shares, portfolio diversification, and investment risk.
- Debt – personal loans, mortgage loans, and vehicle finance.
- Time value of money.
- Budgets.

4.5 Game Elements Applied

A gamified application does not necessarily use all available game elements; it uses only those that are needed based on the theories and the type of players that form part of the development process. The elements of gamification are divided into three categories: dynamic, mechanics, and components. Game dynamics outline the goal to be achieved with the gamified application – the big picture. Game mechanics are the basic processes that drive movement in the action and player engagement. Game components are the visible indicators of game dynamics and mechanics. The game elements are interlinked and do not necessarily have a specific approach for use in the development of the gamified application. The main dynamics applied within the application included the narrative, emotions, constraints (rules and boundaries), and progression (Werbach & Hunter, 2020).

4.5.1 Narrative

The narrative of the gamified application was based on life as a newly employed worker at a higher education institution. The employee began their virtual life at age 26 and continued until age 60. The players faced life events that were directly aligned with their employment, aiming to create an authentic experience. The flow and sequence of the financial life events acted as curves of interest to hold the players' attention (Kapp, 2012). Due to resource constraints, players could only choose between a male or a female character. The aim was not to link the image (avatar) of the player within the gamified application to a specific ethnic group; they could, however, choose between a green or purple avatar. As the avatar progressed from one level to the next, it grew older to illustrate the approach of retirement and that time is running out to save towards their retirement goal within the application. The ageing avatar was specifically designed based on the finding that participants who observed their avatar growing older seemed more inclined towards changing their saving behaviour (see Figure 1) (Ersner-Hershfield et al., 2009).

Figure 1: The avatar within the gamified application that grew older each level

4.5.2 Emotions

Game aesthetics served as the emotional element in the gamified application. Due to cost constraints, only basic aesthetics were applied (see Figure 2). Visual elements were kept to a minimum, with the focus on the financial decisions the players had to make. The essential elements incorporated to contribute towards the emotions and the look and feel of the gamified application included the ageing avatar, feedback panes (budget, financial ledger, and assets and liabilities), and the life event pane. All of the aesthetics portrayed the importance of preparing for retirement, which was the ultimate purpose of the gamified application.

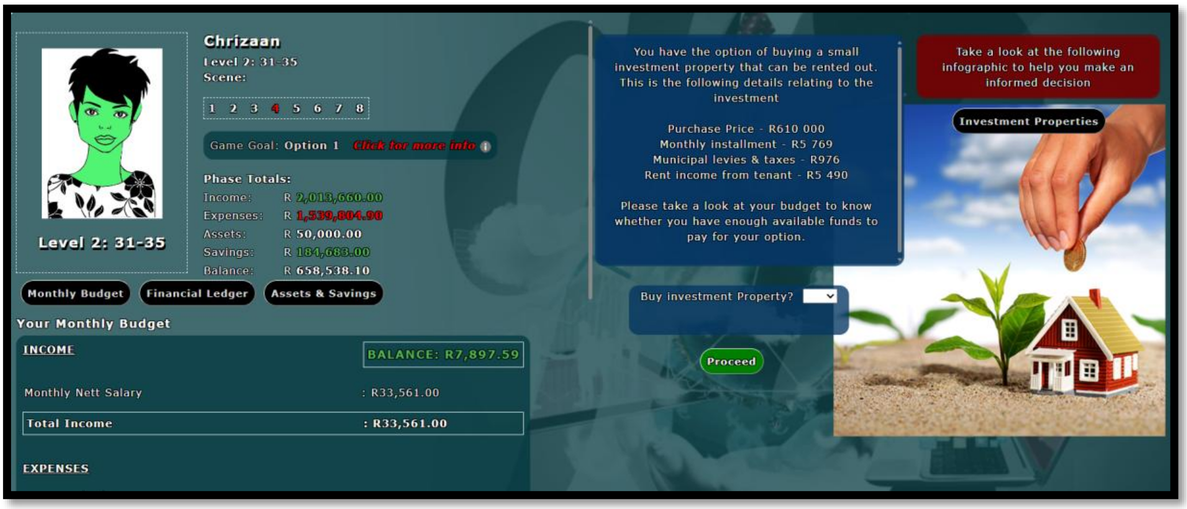


Figure 2: The interface of the Gamified Retirement Planning Application (https://game.go4software.co.za/login/)

4.5.3 Constraints (rules and boundaries)

Constraint elements incorporated in the gamified application included levels, time constraints, infinite gameplay, and transactions. The application's fixed-screen format prevented participants from exploring the gamified environment. The players faced a series of choices in sequential order, representing different phases of life and general life events. Due to time constraints, levels were created, each representing a lifespan of five years, totaling seven levels (starting at age 26 and ending at age 60). The time constraint served as a warning to players that the time they have in real life should be used efficiently to save for retirement. The first level presented players with general financial decisions, including medical aid, accommodation, transport, income protection, and savings. Players were allowed to make changes to their general financial choices at the beginning of each subsequent level. From Level 2 onward, investment properties were introduced as another avenue for saving for retirement. If a player encountered budgeting problems due to costly financial decisions made in previous levels, they could reset the application to the level where they want to change their previous decision.

However, players were not allowed to move to the next level or skip levels if they had not completed the series of financial events in the current level. The levels differed in the types of life events they faced, but not in terms of difficulty. If adverse decisions were made at earlier levels, it would become increasingly difficult in later levels to achieve their retirement goal set at the start of the gamified application. The players were allowed infinite gameplay by resetting the gamified application as many times as they wished and choosing different retirement goals. With each financial decision, a transaction occured that automatically updated their budget, financial ledger, and assets and savings. Rules for calculations in the gamified application were programmed to account for inflation, increases in asset values, compound interest, fixed interest rates, depreciation of vehicles, buying and selling of assets, the type of financing allowed for specific transactions, and different debt pay-off periods. The final retirement savings at the end of all seven levels consisted of retirement fund savings, personal savings, bank account balance, rent income from investment properties, and a deduction for any mortgage loan balance. The total accumulated retirement savings was compared to the retirement goal chosen at the start of the application to determine whether the goal was achieved.

4.5.4 Progression

Goal-focused challenges, resources, cascading information, feedback, and rewards embodied the progression elements employed in the gamified application. Only one goal-focused challenge was incorporated into the application – to achieve the retirement goal chosen at the start of the gamified application. To achieve their chosen goal, players had to complete all seven levels and make appropriate financial decisions. Players could choose from six different retirement goals. The easiest of the retirement goals (goal number one) reflected a retirement with insufficient savings that does not even cover the most basic needs. The better the retirement outlook (with more benefits and luxury), the higher the required retirement savings needed to achieve the goal. Thus, goal number six was the most difficult and nearly impossible to achieve within the gamified application. Players collected resources in the form of a salary. The disposable income left after compulsory deductions (retirement fund contributions and taxes) was available to the player to use when making financial decisions with each life event. More income could be collected if the player decided to buy investment property. Players were also allowed to put aside additional savings that earned interest. In the event of unexpected life events with a financial impact, the money in their bank account or savings account could be utilised. If no funds were available or the player did not want to use their savings or bank account, they could obtain a personal loan, which would have cost them more, as interest was charged at a very high rate. Through cascading, players were exposed to information about the financial event they faced at that specific stage. The information was provided to help players make better-informed decisions. To avoid information overload, the information was given in chunks rather than all at once. The element of unlocking content as the player progressed through the gamified application also meant they did not know what lay ahead in the coming levels. If adverse financial decisions were made, such as declining to obtain income protection, this decision could cost them later due to unexpected life events (e.g. loss of income due to a medical condition). The most crucial game element used in the gamified application was feedback. The impact of every financial decision made by a player is immediately reflected on the budget, financial ledger, and asset and savings statements. The asset statements reported the values of the properties and the vehicle, total employer pension fund savings, the bank account balance, and the mortgage loans balances. Feedback provided the player with information about the impact of the decision they made. In reality, the impact of financial decisions is only discovered in the long term, by which time it may be too late to rectify their financial situation. Thus, players could learn from their mistakes within the gamified application and transfer this knowledge to their real-life (Aegon, 2021; Bell, 2018; Kapp, 2012). During the gamified application, players were rewarded when making sound financial decisions by avoiding the costs of unexpected life events. They could avoid medical costs if they had the appropriate medical aid plan, avoid accident costs if they had car insurance, or avoid income loss if they had income protection. When players reached the end of the gamified application, a message indicated whether they achieved their goal. Reaching the retirement goal set for oneself was the reward in the gamified application. If a player did not reach their retirement goal, the application indicated the type of retirement they could afford based on the amount of retirement savings they accumulated. The result could give the players an idea of how realistic their retirement goal was.

5. Results

The dependent variable, retirement preparedness, was measured in terms of basic financial literacy, retirement financial literacy, interest in retirement planning, and their perception of how well they think they are prepared for retirement (attitudes and behaviours towards their retirement). Basic financial literacy assessed understanding of time value of money, interest rates, compound interest, a safe investment, investment risk, investment diversity, inflation, and prime rates. Retirement financial literacy measured knowledge of pension fund rules (retirement age, withdrawals, and transfers), tax-related matters, differences between funds, percentage of salary allocated to retirement, and scenarios impacting retirement savings.

This article focused on overall improvements in participants' financial literacy and retirement preparedness, emphasising the general trends observed from the gamified application rather than definitive causal conclusions. The gamification group also provided feedback on Likert-scale statements about the application (1 = strongly disagree; 5 = strongly agree) and answered open-ended questions about what they liked and disliked, as well as suggestions for improvement. The data gathered from the pre- and post-test were analysed in SPSS (Version: 29.0.0.0) using a Quade non-parametric ANCOVA to assess differences between groups attributable to the interventions. A p-value below 0.05 indicates a statistically significant change in post-test results. Although the intervention groups showed greater increases in financial literacy scores than the control group, these findings should be interpreted as preliminary trends reflecting the exploratory nature of the study, rather than conclusive evidence. Table 1 illustrates the financial literacy scores before and after the interventions between the groups.

Table 1: Basic financial literacy (BFL) and financial retirement literacy (FRL) scores

Variable tested:	Gamification	Education	Control	Total	P-value
Overall basic financial literacy score					0.383
Pre-test: Average BFL score	73.9%	71.7%	76.5%	74.1%	
Post-test: Average BFL score	81.2%	78.1%	79.7%	79.6%	
Overall increase in BFL score	7.3%	6.4%	3.3%	5.6%	
Overall financial retirement literacy score					*0.001
Pre-test: Average FRL score	63.2%	59.3%	68.1%	63.6%	
Pre-test: Average FRL score	66.1%	72.2%	68.1%	68.9%	
Increase in FRL score	2.9%	12.9%	0.0%	5.3%	
*p-value <0.05 **p-value <0.001					

Both interventions were associated with improvements when compared to the control group. The gamified application appeared more effective at improving basic financial literacy, whereas a focused study of the infographics produced greater gains in retirement financial literacy. This difference may be attributed to the education group's exclusive focus on core retirement concepts. In contrast, the gamification group divided attention between gameplay and supplementary infographic content, with limited engagement with the embedded links.

Table 2 illustrates the participants' interest in retirement planning before and after the interventions (Likert scale of 1 to 5, where 1 indicates low interest and 5 indicates high interest).

Table 2: Interest in retirement planning

Variable tested:	Gamification	Education	Control	Total	P-value
Interest in retirement planning					0.278
Pre-test - mean	4.48	4.55	4.62	4.55	
Post-test - mean	4.62	4.42	4.44	4.49	
Change in mean score	0.14	-0.13	-0.18	-0.06	
% increase/-decrease	3%	-3%	-4%	-1%	
*p-value <0.05 **p-value <0.001					

Participation in the gamified application was associated with a slight increase in interest, whereas the education and control groups experienced decreases. These trends suggest that interactive, engaging experiences may promote short-term motivation to engage with retirement planning, consistent with Self-Determination Theory, which posits that autonomy, mastery, and feedback increase intrinsic motivation. Table 3 provides the percentage changes of each group for the variables tested, indicating how well the participants think they are prepared for retirement before and after the interventions.

Table 3: Perception of how well a participant thinks they are prepared for retirement

Variable tested:	Gamification	Education	Control	Total	P-value
1. Personal responsibility					0.564
% increase/-decrease	-1%	2%	-1%	0%	
2. Level of awareness					0.204
% increase/-decrease	7%	12%	1%	7%	
3. Understanding of financial matters					0.299
% increase/-decrease	5%	18%	13%	12%	
4. Personal retirement plan					0.277
% increase/-decrease	-2%	11%	3%	4%	
5. Saving enough					*0.012
% increase/-decrease	1%	20%	-2%	6%	
6. Confidence towards retirement					0.104
% increase/-decrease	8%	13%	4%	8%	
7. Gross income replacement					1.000
% increase/-decrease	-1%	4%	8%	3%	
8. Achieve income required					0.633
% increase/-decrease	-4%	15%	0%	3%	

*p-value <0.05 **p-value <0.001

The gamified application was associated with increased awareness and confidence, but a decrease in their perception of how well they think their personal retirement plan was developed. This was also the case for the question on whether the gamification group thinks they would achieve the income needed in retirement. Their response decreased after participating in the game. This may reflect a "reality check" effect, where experiential learning exposes participants to the real-life challenges of retirement, prompting more realistic self-evaluations. In contrast, the education group, after being exposed to the infographics, was, overall, more positive in their perception of how well they are prepared for retirement. This result is consistent with the cognitive bias literature, suggesting that individuals who study limited but targeted information may overestimate preparedness. In addition, over-confident individuals are no more likely to be better prepared than those who are less confident (Angrisani & Casanova, 2019).

Table 4 presents the feedback from the gamification group on the Likert scale statements, rated on a scale from 1 (strongly disagree) to 5 (strongly agree).

Table 4: Feedback from the gamification group on the gamified application

Statements	Agree/Strongly Agree	Mean
I know more about retirement planning after participating in the gamified application.	68.9%	3.8
All my concerns regarding retirement planning were covered in the gamified application.	44.8%	3.2
The gamified application was engaging, fun, and highly interactive.	86.2%	4.2
The gamified application motivated me to be more involved in my retirement planning and be prepared for retirement.	82.7%	4.4
I would recommend the gamified application to other employees.	79.3%	4.1
Average mean		4.0

The first and second responses in Table 4 might have appeared more positive if the players actually consulted the infographics in the gamified application, even though it was not a requirement. They were mainly tasked with a retirement goal, playing all the levels in the application (making financial decisions throughout), and trying to achieve their initial retirement goal. The average mean of 4 indicated that the gamification group found the gamified application useful.

Responses to the open-ended questions were predominantly positive, although some negative experiences were also reported. Respondents described the gamified application as practical, realistic, and informative due to real-life scenarios and actual figures incorporated into it. Many indicated that the application made financial planning more engaging and helped place financial decisions and their consequences for retirement savings into a clearer perspective. Many thought the unexpected life events were a meaningful and thought-provoking experience. The feedback statements were particularly well received, as participants could immediately observe the consequences of their financial decisions. This real-time feedback encouraged reflection on personal retirement planning and promoted a more realistic assessment of retirement preparedness. Several participants also reported increased motivation to save for retirement. As one participant commented: "I am more aware of the need to plan sufficiently for my retirement", which was an objective of the gamified application.

Despite the generally positive feedback, participants identified several limitations. Some respondents indicated that the instructions provided at the start of the application could have been clearer. Others noted the absence of additional saving and payment options, as well as the limitation that certain financial choices, such as selecting car insurance or deciding to have children, could only be made once. The budgeting structure allowed participants to spend only available funds, with no option to simulate overdraft or debt, which some felt restricted their ability to explore the consequences of unwise financial decisions. In addition, a small number of participants reported feelings of anxiety when engaging with retirement-related content, while technical issues experienced during gameplay contributed to frustration.

6. Limitations of the study

Several limitations should be acknowledged when interpreting the findings of this study.

First, although the sample size of 29–34 participants per group meets recommended thresholds for exploratory experimental research, it may have constrained statistical power. It restricts the generalisability of the findings beyond the study context. Second, the study relied primarily on self-reported measures of financial literacy, interest, and perceived retirement preparedness. Such measures may be subject to social desirability bias or overconfidence, particularly among participants who have been exposed to educational material. Third, the intervention was evaluated over a two-month period, providing insight into short-term changes only. Long-term knowledge retention and sustained behavioural change could not be assessed. Finally, while open-ended responses were thematically categorised, a more in-depth qualitative approach (such as interviews or focus groups) could provide richer insight into participants' decision-making processes and behavioural responses.

7. Discussion

The findings suggest that gamification can positively influence basic financial literacy and stimulate interest in retirement planning, supporting prior research indicating that interactive and experiential learning environments enhance engagement with complex financial topics (Kapp, 2012; Uppalari, 2025). The increase in interest observed in the gamification group aligns with studies showing that game elements such as feedback, progression, and simulated decision-making can enhance intrinsic motivation among adult learners. The results also indicated a more nuanced outcome. Participation in the gamified application was associated with a decline in participants' perceived retirement preparedness. Rather than suggesting a negative effect, the result may reflect a constructive "reality check", whereby exposure to realistic financial scenarios could prompt participants to reassess their previous, more optimistic perception of their retirement situation. This finding extends the gamification literature by demonstrating that increased engagement does not necessarily lead to greater confidence and may, in some cases, reduce perceived preparedness as participants become more aware of the complexity of retirement planning.

This outcome contrasts with the education group, where exposure to infographics resulted in increased financial retirement literacy, as well as confidence in preparedness. While increased financial literacy is essential, behavioural finance literacy cautions that greater knowledge can also contribute to overconfidence bias, in which individuals overestimate their readiness despite limited practical application (Angrisani & Casanova, 2019). The contrasting results across the two intervention groups suggest a gap between perceived competence and actual preparedness, underscoring the importance of integrating knowledge-based education with experiential learning approaches, such as gamification.

The way forward may be to expand the gamified application with a more comprehensive approach, including a wider range of financial decisions (investment and payment options) and retirement concerns. Players should be allowed to enter their real salaries and expenses to manipulate the starting conditions in a gamified application. The automatic update of calculations prevented participants from being educated on how to perform basic financial calculations which would have contributed to their understanding of financial matters. Providing educational material through a clickable link to assist players with their financial decisions was not successful. A different approach must be followed to address this issue. Scaffolding, as one of many ways, can be applied to compel players to first complete knowledge quizzes or tasks before they can unlock the next level.

Overall, the findings contribute context-specific insight into how gamification influences not only financial literacy and interest but also self-perception and confidence in retirement preparedness. Gamification may extend beyond motivation by acting as a reflective process that challenges assumptions and supports more realistic self-perceptions. This nuanced outcome contributes to existing research by emphasising the need for financial education interventions that balance engagement, knowledge development, and behavioural awareness.

8. Conclusion

The gamified retirement planning application successfully illustrated how interactive learning can promote financial literacy and retirement preparedness. Players experienced the long-term effects of their decisions in a low-risk setting by modelling real-life financial decisions, which helped them develop a more accurate understanding of retirement planning. Participants' self-assessment of their retirement readiness became more realistic as a result of this interactive tool's ability to challenge their prejudices and expand their knowledge. The limitations of the gamified application were a need for clearer instructions, more financial possibilities, and the ability to customise income and expenses to better suit unique situations. Participants also indicated a desire to see the consequences of poor financial decisions, which could be included to enhance the educational impact. These observations imply that, despite its effectiveness, the application might need some improvements to be more relatable and adaptable, possibly including scaffolding strategies to encourage slower learning. In the end, this gamified method exhibits promise as a financial education tool, motivating people to take proactive steps toward retirement planning.

AI Statement: No AI tool has been used in the development of this paper.

Ethics Statement: All subjects gave their informed consent for inclusion before they participated in the study. The study was conducted in accordance with the rules of the University of the Free State, South Africa. The protocol was approved by the General Human Research Ethics Committee (GHREC) with the Ethical Clearance Number of UFS-HSD2019/0023/21/22.

References

Aegon (2017). *Successful Retirement - Healthy Aging and Financial Security: The Aegon Retirement Readiness Survey 2017.* The Hague: Aegon. Available at: https://globalcoalitiononaging.com/wp-content/uploads/2018/05/tcrs2017_sr_successful_retirement_healthy_aging_financial_security.pdf (Accessed: 11 February 2026).

Aegon (2021). *The New Social Contract: Future-Proofing Retirement.* The Hague: Aegon. Available at: https://www.aegon.com/sites/default/files/file/2023-04/the-new-social-contract-future-proofing-retirement.pdf (Accessed: 11 February 2026).

Ahmad, A., Fernandez Galeote, D., Xi, N., & Hamari, J. (2025). Gamification of personal finance: A systematic literature review. In PACIS 2025 Proceedings (Paper 20). Association for Information Systems. https://aisel.aisnet.org/pacis2025/general_topic/general_topic/20/

Angrisani, M. and Casanova, M. (2019). 'What You Think You Know Can Hurt You: Under/Over Confidence in Financial Knowledge and Preparedness for Retirement.' *Journal of Pension Economics and Finance*, 20(4):516–531. https://doi.org/10.1017/S1474747219000131

Anuar, M.E.A., Ismail, S., and Samad, K.A. (2023). 'Financial Planning for Retirement: A Proposed Conceptual Framework.' *International Journal of Academic Research in Accounting Finance and Management Sciences*, 13(1), 177–187. http://dx.doi.org/10.6007/IJARAFMS/v13-i1/16157

Babbie, E. (2016). *The Basics of Social Research*. 7th ed. Hampshire: Cengage Learning.

Bai, R. (2023). 'Impact of financial literacy, mental budgeting and self-control on financial wellbeing: Mediating impact of investment decision making.' *PLoS ONE*, 18(11): e0294466. https://doi.org/10.1371/journal.pone.0294466

Bank of New York Mellon (2015). *Generation lost: Engaging millennials with retirement saving.* Available at: https://www.tisa.uk.com/wp-content/uploads/2020/02/generation-lost.pdf (Accessed: 11 February 2026).

Bell, K. (2018). *Game On!: Gamification, Gameful Design, and the Rise of the Gamer Educator.* Maryland: Johns Hopkins University Press.

Bodnár, K. and Nerlich, C. (2022). *The Macroeconomic and Fiscal Impact of Population Ageing.* Technical Report (No. 296). Frankfurt: Central European Bank. Online: https://www.ecb.europa.eu/pub/pdf/scpops/ecb.op296~aaf209ffe5.en.pdf

Collis, J. and Hussey, R. (2014). *Business Research: A Practical Guide for Undergraduate & Postgraduate Students.* 4th ed. Hampshire: Palgrave Macmillan.

Daniel, J. (2012). *Sampling Essentials: Practical Guidelines for Making Sampling Choices.* London: Sage Publications.

De Lima, D.V., Nery, D.N., and Miranda, C. de S. (2024). 'An analysis of how Millennials view social security.' *Revista Ambiente Contábil*, Universidad, Federal do Rio Grande do Norte, ISSN 2176-9036, 16(1), 442-468. https://doi.org/10.21680/2176-9036.2024v16n1ID34962

Doonan, D. and Kenneally, K. (2021). *Generational Views of Retirement in the United States.* Washington: National Institute on Retirement Security. Available at: https://www.nirsonline.org/wp-content/uploads/2021/07/Generations-Issue-Brief-F4.pdf (Accessed: 11 February 2026).

Ersner-Hershfield, H., Garton, M.T., Ballard, K., Samanez-Larkin, G.R., and Knutson, B. (2009). 'Don't stop thinking about tomorrow: Individual differences in future self-continuity account for saving.' *Judgment and Decision Making*, 4(4), 280-286. https://doi.org/10.1017/S1930297500003855

Field, A. (2018). *Discovering Statistics Using IBM SPSS Statistics.* 5th ed. London: Sage Publishers.

Gay, L.R., Mills, G.E. and Airasian, P.W. (2012). *Educational Research: Competencies for Analysis and Applications.* 10th ed. New York: Pearson.

Ghilarducci, T. (2024). *Work, Retire, Repeat: The Uncertainty of Retirement in the New Economy.* Chicago: University of Chicago Press.

Grobbelaar, C. and Alsemgeest, L. (2024). 'Gamification as an educational tool in retirement planning.' *Adult Learning*, 37(1), 13-23. https://doi.org/10.1177/10451595241286913

Hill, E.R. (2017). *E-learning across generational boundaries: A study of learner satisfaction.* Ph.D. thesis. Minneapolis: Capella University.

Huang, E., Lassu, R. and Chan, K. (2018). 'User-Source Fit and Financial Information Source Selection of Millennials.' *Journal of Financial Counseling and Planning*, 29, 272-289. https://doi.org/10.1891/1052-3073.29.2.272

Ifeanyi, M., Rena, R. and Prinsloo, J.J.H. (2019). 'The National Strategy on Financial Literacy: A Conceptual Review of South African Perspectives.' *Journal of Reviews on Global Economics*, 8, 1121–1134. https://doi.org/10.6000/1929-7092.2019.08.97

Iwry, J.M., John, D.C., Pulliam, C., and Gale, W.G. (2021). *Collective Defined Contributions Plans.* Technical Report. Washington, D.C.: Tax Policy Center, Urban Institute and Brookings Institution. Available at: https://www.urban.org/sites/default/files/publication/104809/collective-defined-contributions-plans.pdf (Accessed: 11 February 2026).

Jain, A., and Dutta, D. (2019). 'Millennials and gamification: Guerilla tactics for making learning fun.' *South Asian Journal of Human Resources Management*, 6(1), 29–44. https://doi.org/10.1177/2322093718796303

Jiun, T.Y., Satar, N.H.M., and Ishak, N.A. (2021). 'Malaysian Undergraduate Students' Expectation of Life Satisfaction After Retirement: The Role of Parental Influence and Financial Knowledge.' *Journal of Wealth Management & Financial Planning*, 8, 3-24. https://jwmfp.mfpc.org.my/wp-content/uploads/2023/01/03-24_Original-Research_Msian-Undergraduate-Student-Expectation_Aug-4.pdf

Kapp, K.M. (2012). *The Gamification of Learning and Instruction: Game-Based Methods and Strategies for Training and Education.* San Francisco: Pfeiffer.

Kooij, D.T.A.M., Kanfer, R., Betts, M., and Rudolph, C.W. (2018). 'Future Time Perspective: A systematic review and meta-analysis.' *Journal of Applied Psychology*, 103(8), 867–893. https://doi.org/10.1037/apl0000306

Landers, R.N., Auer, E.M., Helms, A.B., Marin, S., and Armstrong, M.B. (2019). 'Gamification of adult learning: Gamifying employee training and development', in: Landers, R.N. (Ed.). *The Cambridge Handbook of Technology and Employee Behavior*. Cambridge: Cambridge University Press, pp. 271–295. https://doi.org/10.1017/9781108649636.012

Lusardi, A. and Hasler, A. (2019). *Millennials' Engagement with Online Financial Education Resources and Tools: New Survey Insights and Recommendations*. GFLEC Insights Report. Available at: https://gflec.org/wp-content/uploads/2019/04/Millennial-Engagement-with-Online-Financial-Education-Resources_FINAL.pdf (Accessed: 11 February 2026).

Marczewski, A. (2018). *Even Ninja Monkeys Like to Play: Unicorn Edition*. Gamified UK. ISBN 978-1724017109.

Mazzo, J. (2015). *Gamifying Education for Millennials: It's More Than Just a Video Game.* Master's thesis. School of Journalism and Mass Communications.

McHaney, R. (2023). *The new digital shoreline: How Web 2.0 and millennials are revolutionising higher education.* New York: Routledge. https://doi.org/10.4324/9781003447979

Michaud, R. (2021). *Millennials' Retirement Struggle: The Interaction Between Student Loans, Financial Decisions, and Financial Literacy.* Honors Thesis. Bryant University. Available at: https://digitalcommons.bryant.edu/cgi/viewcontent.cgi?article=1050&context=honors_finance (Accessed: 11 February 2026).

Natixis (2022). *Global Retirement Index - Danger Zone: Global Retirement Security Challenges Come Home to Roost in 2022*. Technical Report. Paris: Natixis Investment Managers. Available at: https://www.im.natixis.com/content/dam/natixis/website/insights/investor-sentiment/2022/2022-global-retirement-index/RC31-0822-2022-Natixis-GRI-Full-Report-F.pdf (Accessed: 11 February 2026).

Ndou, A. (2023). 'The relationship between demographic factors and financial literacy.' *International Journal of Research in Business and Social Science*, 12(1), 155–164. https://doi.org/10.20525/ijrbs.v12i1.2298

OECD (2016). *OECD/INFE International Survey of Adult Financial Literacy Competencies.* Paris: OECD. Available at: https://web-archive.oecd.org/2018-12-10/417183-OECD-INFE-International-Survey-of-Adult-Financial-Literacy-Competencies.pdf (Accessed: 11 February 2026).

OECD (2019). *Smarter Financial Education - Key Lessons from Behavioural Insights for Financial Literacy Initiatives.* OECD, Paris. Available at: https://web-archive.oecd.org/2021-04-20/530340-smarter-financial-education-behavioural-insights.pdf (Accessed: 11 February 2026).

OECD (2023). *OECD/INFE 2023 International Survey of Adult Financial Literacy*. Paris: OECD. Available at: https://www.oecd.org/en/publications/oecd-infe-2023-international-survey-of-adult-financial-literacy_56003a32-en.html (Accessed: 11 February 2026).

Rappaport, A.M. (2019). 'Financial and retirement planning implications of financial Perspectives across generations.' *Benefits Quarterly*, 35(3), 25–33. https://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=137835525&site=ehost-live&scope=site

Richter, G., Raban, D.R., and Rafaeli, S. (2015). 'Studying gamification: The effect of rewards and incentives on motivation', in: T. Reiners, L. C. Wood (Eds.), *Gamification in Education and Business.* Cham: Springer. https://doi.org/10.1007/978-3-319-10208-5_2

Rodrigues, L. F., Oliveira, A., Costa, C. J., and Rodrigues, H. (2018). Gamification to teach and assess financial education: A case study of self-directed bank investors. Proceedings of the 17th Annual Hawaii International Conference on Education, 1851–1882. Available at: https://repositorio.iscte-iul.pt/bitstream/10071/16908/1/Paper%20A_Game_Quiz_in_e-banking_conference_hawaii_2018_.pdf (Accessed: 11 February 2026).

Sanlam (2017). *Benchmark Survey 2017: Research Summary*. Technical Report. Belville: Sanlam. Available at: https://www.sanlamonline.co.za/corporate/retirement/benchmark-symposium/archive#past-benchmark-reports (Accessed: 11 February 2026).

Sundarasen, S., Rajagopalan, U., and Ibrahim, I. (2024). 'Financial Sustainability Through Literacy and Retirement Preparedness.' *Sustainability*, 16, 10692. https://doi.org/10.3390/su16231069

Sutton, H. (2019). 'Integrate different generational needs for a strong adult learner experience.' *Recruiting & Retaining Adult Learners*, 21(6), 4–5. https://doi.org/10.1002/nsr.30447

Uppaluri, R. (2025). Gamification: Revolutionizing financial planning systems. World Journal of Advanced Engineering Technology and Sciences, 14(3), 399–409. https://doi.org/10.30574/wjaets.2025.14.3.0158

Van Raaij, W. F. (2016). Responsible financial behaviour. In Understanding consumer financial behaviour: Money management in an age of financial illiteracy (pp. 127–141). New York, NY: Palgrave Macmillan US.

Werbach, K. and Hunter, D. (2020). *For the Win: The Power of Gamification and Game Thinking in Business, Education, Government, and Social Impact,* 2nd. ed. Philadelphia, P.A.: Wharton School Press.

Yakoboski, P.J., Lusardi, A., and Sticha, A. (2024). *Retirement Literacy and Retirement Fluency: New Insights for Improving Financial Well-Being.* The 2024 TIAA Institute GFLEC Personal Finance Index. TIAA Institute and Global Financial Literacy Excellence Center. Available at: https://gflec.org/wp-content/uploads/2024/04/TIAA_GFLEC_Report_PFin_April2024_07.pdf (Accessed: 11 February 2026).

Yong, C.-C., Yew, S.-Y., and Wee, C.-K. (2018). 'Financial Knowledge, Attitude and Behaviour of Young Working Adults in Malaysia.' *Institutions and Economies*, 10(4). https://ijie.um.edu.my/article/view/13444

Yulianto, A., Pramono, S. E., Wijaya, A. P., & Nasrun. (2024). Gamification in enhancing student financial knowledge, engagement, and enjoyment in financial education. Jurnal Kependidikan, 10(3), 965–975. https://doi.org/10.33394/jk.v10i3.12653

EsyGrade: An AI-Based Essay Assessment System for Economics Education

Ramadzan Defitri Pratama[1], Khresna Bayu Sangka[2] and Cicilia Dyah Sulistyaningrum Indrawati[3]

[1]Master of Economics Education, Faculty of Teachers Training and Education, Sebelas Maret University, Indonesia

[2]Economics Education, Faculty of Teachers Training and Education, Sebelas Maret University, Indonesia

[3]Office Administration Education, Faculty of Teachers Training and Education, Sebelas Maret University, Indonesia

ramadp_24@student.uns.ac.id
b.sangka@staff.uns.ac.id (corresponding author)
ciciliadyah@staff.uns.ac.id

https://doi.org/10.34190/ejel.24.3.4475

Abstract: This study aims to design, develop, and evaluate EsyGrade, a web-based essay grading system integrated with ChatGPT to support the assessment of conceptual understanding in economics education. This study is motivated by the prevalence of multiple-choice questions in Indonesian high schools due to efficiency considerations, while essay assessments better suited to measuring conceptual reasoning are still rarely used because of high assessment load. The study employed a simplified Research and Development (R&D) approach consisting of seven stages. Data were collected through expert validation using Aiken's V, a user response questionnaire, and pretest–post-test, and were later analysed using N-Gain and effect size (Cohen's d). The research findings indicate that EsyGrade has a high level of validity based on the expert evaluation and an excellent acceptance rate in the initial pilot test. In the main field trial, the results indicated a moderate improvement in conceptual understanding based on N-Gain, with a "large" effect size. These findings suggest that the overall change in learning outcomes indicates a fairly strong impact, although the degree of improvement varied among students. In particular, the system's key value lies in its ability to generate structured, reflective feedback, so that it serves not only as a summative assessment tool but also as a means for formative one. This study contributes to the development of e-learning by demonstrating that AI-based essay grading systems can be designed to improve efficiency while supporting deeper learning through the integration of pedagogical principles and technology.

Keywords: Automated essay scoring, ChatGPT, Conceptual understanding, Economics education, Formative assessment, Educational technology

1. Introduction

Economics education at vocational or public high schools emphasizes the importance of in-depth mastery of concepts so that students are able to analyse economic phenomena, understand the interrelationships between variables, and evaluate policies in a real-world context. In Indonesian context, the implementation of the Freedom to Learn curriculum (*Kurikulum Merdeka*) developed in response to the disruption of learning caused by the COVID-19 pandemic marks a significant turning point in reevaluating student learning outcomes. As the pandemic has led to learning loss and gaps in understanding, this curriculum was designated as a solution to emphasize learning recovery through the strengthening of essential competencies, particularly critical thinking skills and conceptual understanding. Thus, it serves as not only a learning framework but also a new benchmark for assessing learning outcomes that are more oriented toward understanding rather than mere mastery of facts.

In economics education, this approach is particularly relevant because economic concepts are abstract and interconnected, requiring advanced analytical and reasoning skills. This approach aligns with the concept of deep learning, which demands that students not only memorize information but also understand, analyse, and apply concepts in context (Marton and Hounsell, 1984; Biggs and Tang, 2011). Therefore, the success of this curriculum's implementation depends heavily on the assessment system's ability to measure deep conceptual understanding, rather than merely superficial learning outcomes.

However, the results of the 2022 Program for International Student Assessment (PISA) indicate that Indonesian students' performance remains relatively low, with scores ranging from 359 to 371 far below the OECD average

Cite this article: Pratama, R.D., Sangka, K.B. and Indrawati, C.D.S. 2026. "EsyGrade: An AI-Based Essay Assessment System for Economics Education", *Electronic Journal of e-Learning*, 24(3), pp 16-32, https://doi.org/10.34190/ejel.24.3.4475

of 472 to 476 (OECD, 2023). Although PISA does not specifically measure economics, the indicators used such as reasoning skills, problem solving, and the application of concepts in real-world situations reflect key aspects of conceptual understanding that are also central to the learning objectives of economics. These low scores indicate that Indonesian students still struggle to connect concepts with real-world practice and to develop higher order reasoning.

Although essay questions are considered more representative for measuring conceptual understanding because they allow students to express their arguments and reasoning (Elliott and Balasubramanyam, 2016; Beldar, 2025), assessment practices in the classroom are still dominated by multiple-choice questions. A study by Kamalia (2023) found that the majority of high school economics teachers in Indonesia prefer objective questions for their efficiency, despite these instruments' limitations in measuring higher-order thinking skills (Petersen, Craig, and Denny, 2016). Preliminary research conducted by researchers in Surakarta reveals that 73% of teachers tend to use multiple-choice questions more frequently, while less than a third of them employ essays, and the main obstacle is that correcting one essay may take an average of six to eight minutes to assess (Pasaribu, Budiman, and Indrarini, 2024), making it difficult for teachers to provide quick and reflective feedback. This situation slightly creates a research gap between curriculum policies that emphasize authentic assessment and the actual practices of classroom assessment.

On the other hand, advancements in digital technology particularly in the context of e-learning and artificial intelligence have opened up new opportunities in educational assessment practices. AI-based systems enable automated assessment, real-time feedback, and personalized learning, all of which are key elements in the modern digital learning ecosystem (Mizumoto and Eguchi, 2023). Previous research from Indonesia has developed an essay assessment system based on Latent Semantic Analysis (LSA), but this system exhibits limitations in comprehending context, argument logic, and language flexibility (Kinanti and Qoiriah, 2020). Meanwhile, international research has begun to utilize Large Language Models (LLM) for automatic assessment, including GPT-3.5 and GPT-4, which have proven to be superior in assessing students' answers semantically and contextually (Pack, Barrett, and Escalante, 2024; Smerdon, 2024). Mizumoto and Eguchi (2023) even show that the integration of ChatGPT with learning analytics can offer reflective feedback and personalize learning experiences. However, research in Indonesia that specifically focuses on developing ChatGPT-based automatic assessment systems for high school and vocational school economics subjects is still quite limited.

This study introduces two novel aspects in the field of economics education. Firstly, it shifts the focus from merely assessing learning outcomes based on facts to evaluating students' conceptual understanding in line with the Merdeka Curriculum's requirements. Secondly, the automated assessment system developed using ChatGPT not only generates scores but also provides reflective feedback that supports formative assessment and differentiated learning. This innovation aims to bridge the gap between curriculum expectations and the limitations of traditional assessment practices.

Based on the background and research gaps, this study aims to (1) design and develop a ChatGPT-based automatic assessment system that is in line with learning outcome indicators, (2) test the feasibility and practicality of the developed system according to experts, teachers, and students, and (3) analyse the development of students' conceptual understanding after using the system through a pretest–post-test design with Normalized Gain (N-Gain) analysis.

2. Literature Review

2.1 Conceptual Understanding in Economics Education

Conceptual understanding is a foundation for meaningful learning, especially in economics where students must integrate abstract concepts and apply them in real contexts. Rittle-Johnson, Schneider, and Star (2015) define conceptual understanding as the ability to connect ideas and use them to solve problems, while Anderson and Krathwohl, (2001) position it as a key stage in the revised Bloom's taxonomy. Within Bloom's framework, conceptual understanding spans from remembering (C1) and understanding (C2), to applying (C3), analysing (C4), evaluating (C5), and creating (C6) (Bloom, 1956; Anderson, and Krathwohl, 2001). These indicators are central to assessing deep learning outcomes in economics education, as students are expected not only to recall definitions but to analyse market dynamics, evaluate policy impacts, and design alternative solutions. However, studies highlight persistent difficulties among Indonesian students in achieving higher-order thinking skills, with many remaining at the level of factual recall (Arini, Dewi, and Wibawa, 2024). Despite its importance, the development of conceptual understanding is often constrained by existing assessment practices. Thus, examining how assessment is implemented in economics education becomes crucial to determine whether it aligns with the goal of fostering higher-order thinking skills.

2.2 Assessment Practices in Economics Education

Despite the centrality of conceptual understanding, classroom assessment practices remain dominated by multiple choice tests. Teachers favour these instruments because they are efficient, quick to grade, and objective (Kamalia, 2023). Yet, such tools are limited in capturing higher-order thinking skills (Petersen et al., 2016). Essay assessments, by contrast, are more representative for evaluating students' reasoning and conceptual mastery (Elliott and Balasubramanyam, 2016; Beldar, 2025). This is consistent with the findings of Walstad (2006) and Al-Obaydi, Pikhart, and Tawafak (2023), which show that essay questions are more effective in measuring the depth of economic understanding than multiple-choice questions. They allow students to articulate arguments, demonstrate critical thinking, and connect theories with socio-economic realities (Van Wyk, 2015). However, the manual grading process is time-consuming, with each essay requiring six to eight minutes on average (Pasaribu, Budiman, and Indrarini, 2024), and in a class of 30–32 students, teachers can spend three to five hours marking a single task (Rafi, Ramadhani, and Sanjaya, 2025). This workload seems to create delays in providing feedback and reduces the potential for formative assessment.

2.3 Teacher Workload and Assessment Challenges

Teacher workload is a recurring issue in the implementation of authentic assessment. According to Runtuwene and Tangkawarow, (2020), teacher responsibilities span lesson planning, instruction, assessment, and administrative duties. Although Minister of Elementary and Secondary Education Regulation No. 11 of 2025 stipulates a standard weekly workload of 37.5 hours (2025), empirical studies show that assessment tasks, especially essay grading, often exceed this limit (Morris *et al.*, 2024, 2026). This heavy workload not only affects teachers' ability to provide timely and reflective feedback but also impacts instructional quality and teacher well-being (Sabon, 2020; Johnson and Coleman, 2025). These challenges underscore the imperative of embracing technology to mitigate administrative burdens while maintaining the integrity of assessment processes.

2.4 Automated Essay Scoring and Artificial Intelligence in Education

To address such challenges, Automated Essay Scoring (AES) systems have been developed, relying on Natural Language Processing (NLP). Early models such as Latent Semantic Analysis (LSA) have been applied in Indonesian contexts (Kinanti and Qoiriah, 2020), yet remain limited in handling complex argumentation and contextual reasoning (Alief, Irawati, and Mude, 2023). The emergence of Large Language Models (LLMs), including GPT-3.5 and GPT-4, introduces greater capability in semantic and logical analysis. International studies have shown that ChatGPT can effectively evaluate essays, provide formative feedback, and support personalised learning (Mizumoto and Eguchi, 2023; Foung, Lin, and Chen, 2024; Latif and Zhai, 2024; Smerdon, 2024). This suggests that the feedback generated by language models can help explain students' conceptual errors and support a deeper learning process. Thus, this development aligns with the goals of the Merdeka Curriculum, which promotes personalized, reflective, and meaningful learning experiences.

2.5 Research Gap and Contribution

Despite the development of various automated grading systems, including GradedPro, CoGrader, and Automark, most of these systems primarily prioritize efficiency through the automation of scoring. However, these systems generally lack explicit integration of pedagogical aspects, such as reflective feedback to rectify conceptual errors, alignment with Bloom's taxonomy, the utilization of structured rubrics, and prompt flexibility tailored to individual learning requirements.

In accordance with this, while international research increasingly emphasizes the potential of artificial intelligence (AI) in essay grading, its application in economics education in Indonesia remains limited. Existing studies have not adequately addressed how AI-based systems can assess conceptual understanding aligned with Bloom's taxonomy while simultaneously reducing teachers' workload. Therefore, this study addresses this gap by developing a ChatGPT-based essay grading system (EsyGrade) specifically designed for high school economics education in Indonesia, and by evaluating its impact on students' conceptual understanding through a pre-test–post-test design with N-Gain analysis, paired Cohen's d, and distribution analysis of N-Gain categories.

3. Methodology

This study employed a Research and Development (R&D) design adapted from the model of Borg and Gall (Bennett, Borg, and Gall, 1984). Educational R&D aims to develop and validate products in iterative cycles of design, expert validation, revision, and field testing (Plomp and Nieveen, 2013). While Borg and Gall originally proposed ten steps, this study streamlined the process into seven stages due to contextual and time constraints, focusing on the development of an automated essay assessment system using ChatGPT for evaluating students' conceptual understanding. The research phases include: (1) needs analysis, (2) design planning, (3) prototype development, (4) expert validation, (5) product revision, (6) field testing, and (7) evaluation and finalization. This simplification aims to maintain a balance between the depth of development and the feasibility of implementation in an educational context.

Besides, students' conceptual understanding in this study was assessed using essay-based tests administered in pre-test and post-test formats. The items were designed based on the revised Bloom's taxonomy (C1–C6) to capture different levels of cognitive processes. Student responses were evaluated using a rubric consisting of four criteria: conceptual mastery, argumentation, relevance of examples, and structure, ensuring alignment between automated scoring and pedagogical standards.

A validation instrument, meticulously crafted by experts, comprising 15 items, was employed to assess the appropriateness of the indicators, the clarity of the rubric, and the feasibility of the system from both an assessment and a technological standpoint. This evaluation was conducted by educational assessment experts and IT experts. The instrument underwent expert evaluation using a Likert scale (ranging from 1 to 4) and was subsequently analysed employing Aiken's V.

The pilot study was then conducted in two phases, a preliminary pilot study (with 10 participants) and a main pilot study (72 respondents). These sample sizes were chosen to represent the exploratory and confirmatory phases of the development research, in which the preliminary pilot study was used to identify initial improvements, while the main pilot study was used to evaluate the system's performance on a broader scale. During the initial pilot phase, questionnaires were distributed to teachers and students to assess usability and practicality, while observations were conducted during implementation to document technical and pedagogical challenges.

Data were collected using a pre-test and post-test design with equivalent question characteristics to measure changes in conceptual understanding after exposure to AI-generated feedback. In addition, questionnaires and observations were employed to capture user responses and implementation challenges.

Data analysis combined quantitative and qualitative approaches. Quantitative analysis included Normalized Gain (N-Gain) to measure relative improvement (Hake, 1998), paired Cohen's d to assess the magnitude of change (Cohen, 1988), and distribution analysis of N-Gain categories (low, medium, high) to examine individual variation. Qualitative data from observations and open-ended responses were analysed descriptively to provide contextual insights into system usability and classroom integration.

4. Results

4.1 Needs Analysis

A previous study worked on a systematic literature review (SLR) on automated essay scoring (AES) between 2020 and 2025, which revealed a notable surge in research in this field, particularly with the adoption of GPT-based large language models (LLMs) such as GPT-3.5, GPT-4, and ChatGPT (Pratama and Sangka, 2025). These models demonstrate high accuracy and reliability, especially when combined with prompt designs based on rubrics (Tang *et al.*, 2024; Yavuz, Çelik and Yavaş Çelik, 2025). However, the literature also highlights challenges related to validity, consistency, and domain-specific appropriateness (Steiss *et al.*, 2024; Gandolfi, 2025). Importantly, existing studies are still concentrated on language education, with limited application in economics education, indicating a gap in domain-specific development (Lee *et al.*, 2024).

Complementing these findings, the needs analysis involving local schools in Surakarta revealed that essay assessments remain underutilised in economics classrooms. Teachers reported a strong preference for multiple-choice questions due to efficiency concerns, with essay assessments accounting for less than one-third of practice. On average, grading a single essay required six to eight minutes, which created a heavy workload in classes of 30 or more students. This burden limited the ability of teachers to provide timely and reflective feedback. These results confirm the broader international literature, which recognises manual essay scoring as labour-intensive, subjective, and inconsistent (Mendonça, Quintal, and Mendonça, 2025).

In the context of e-learning, these limitations become increasingly relevant, as the effectiveness of digital learning depends heavily on the availability of rapid and meaningful feedback. Thus, from both theoretical and practical perspectives, there is a compelling need for AI-based assessment systems that not only enhance the efficiency of assessment but are also capable of facilitating the development of conceptual understanding through reflective and contextual feedback.

4.2 Design Planning and Prototype Development

Drawing inspiration from both international literature and field requirements, the system was meticulously designed to encompass essay input, automated scoring facilitated by ChatGPT, and structured feedback output. A rubric aligned with Bloom's revised taxonomy (C1–C6) was constructed to ensure the evaluation captured different levels of conceptual understanding, from recall to higher-order reasoning (Anderson and Krathwohl, 2001). This choice directly addresses the gap identified in prior research, where AES systems often emphasise surface-level correctness but lack explicit alignment with pedagogical standards (Poole and Coss, 2023).

The prototype system, named EsyGrade, was developed as a web-based platform integrated with the ChatGPT API to support automated essay assessment in economics education. EsyGrade was designed with dual user roles: teachers and students. Those taking the role as teachers can create question packages, design rubrics, embed GPT-based scoring prompts, monitor submissions, and export results. On the other side, students can log in, answer essay questions, and directly access scores and rubric-based feedback.

The system architecture was implemented through four main modules: (1) Essay Input Module, where teachers add essay tasks and students submit answers; (2) Rubric and Prompt Module, which encodes assessment criteria into GPT-compatible instructions; (3) Scoring and Feedback Engine, powered by the GPT-3.5 Turbo API to produce scores and reflective feedback; and (4) Result Display Module, which presents individual results, summary analytics, and export options for teachers and students.

The early prototype of EsyGrade successfully demonstrated real-time essay scoring and transparent rubric-based reporting. Figure 1 presents the homepage, while Figure 2 shows the teacher dashboard with features for question package management, monitoring student answers, and tracking token usage.

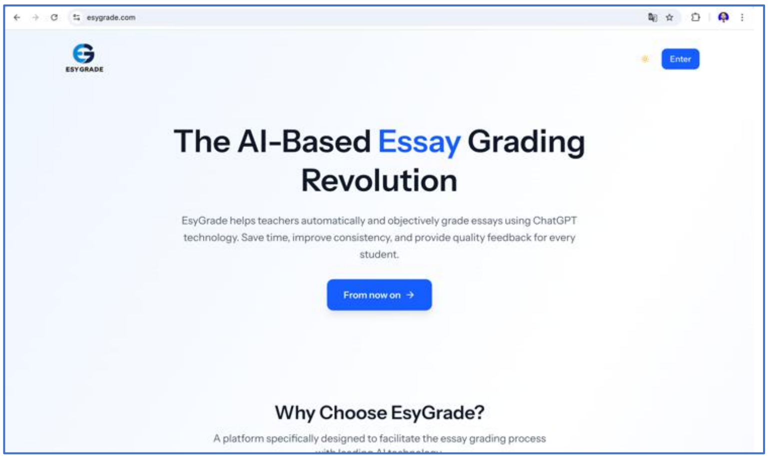


Figure 1: Early Prototype Snapshot of EsyGrade Homepage

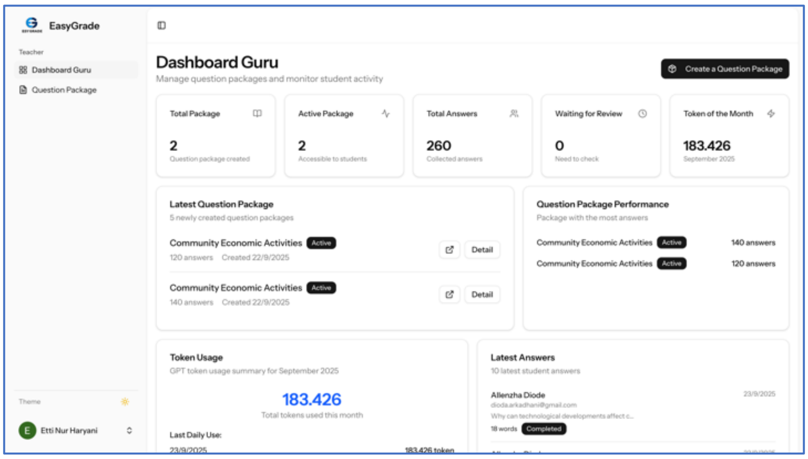


Figure: 2 Teacher Dashboard on EsyGrade

The development of this prototype demonstrates that the system can grade essays in *real-time* and present structured results based on a rubric. This capability indicates the system's potential to improve grading efficiency while providing faster and more transparent feedback. From an e-learning perspective, this feature is relevant because timely, criteria-based feedback is one of the key factors in supporting meaningful and reflective learning.

4.3 Expert Validation

The essay assessment system developed in this study went to a validation phase, involving a total of six experts to scrutinize the items' assessment and technological aspects. The validation instrument comprises 15 items that cover two main aspects: (1) assessment design quality (appropriateness of indicators, clarity of rubrics, and alignment with learning objectives) and (2) technological aspects (system functionality, ease of use, and clarity of the interface). Each aspect is rated on a 1–4 Likert scale. Example items include "The system aligns with the curriculum and applicable assessment standards" and "The system interface is easy for users to understand". Table 1 presents the content validity coefficient (V value) for each item obtained using the Aiken formula.

Table 1: Content validity coefficient (V-value) of the items

Assessment Expert		Technology Expert	
Items	V	Items	V
1	1	1	0.88888889
2	1	2	0.88888889
3	1	3	1
4	0.77777778	4	0.88888889
5	0.88888889	5	0.88888889
6	0.88888889	6	0.77777778
7	1	7	0.88888889
8	0.88888889	8	0.88888889
9	1	9	0.88888889
10	0.88888889	10	0.88888889
11	0.88888889	11	1
12	1	12	0.88888889
13	1	13	1
14	1	14	1
15	1	15	1

The validation results conducted by six experts, comprising three assessment experts and three technology experts, revealed that the Aiken's V values exhibited a range of 0.78 to 1. Notably, most of the items fulfilled the feasibility criteria, with values exceeding the V-table threshold of 0.79, thereby demonstrating high content validity. However, certain items demonstrated slightly below-threshold scores, particularly in the domains of system usability and interface clarity, suggesting a need for enhancement.

Overall, the results confirm that most items are valid and feasible for use, especially in the domains of assessment content and rubric clarity. Nevertheless, improvements were made to items with lower coefficients in response to expert suggestions. In addition to the quantitative data, experts provided written feedback focusing on rubric flexibility, and user interface enhancement. These comments were used to revise and strengthen the platform before field testing. Thus, the validation phase serves not only as an evaluation process but also as the basis for refining the system before it enters the field-testing phase.

4.4 Product Revision

Following expert validation, several items with Aiken's V values below the threshold (0.79) required revision, particularly in the aspects of system usability and interface clarity. Experts recommended improvements in two key areas: (1) rubric flexibility, to allow teachers to customise criteria and scoring according to learning objectives, and (2) user interface enhancements, to improve navigation and provide short guidance for first-time users.

These suggestions were accommodated in the revised version of EsyGrade. As illustrated in Figure 3, the rubric and scoring criteria are now fully customisable, enabling teachers to set different criteria weights (e.g., conceptual understanding, argumentation, relevance, language) and determine maximum scores per criterion. This change ensures greater alignment with classroom needs and curriculum indicators.

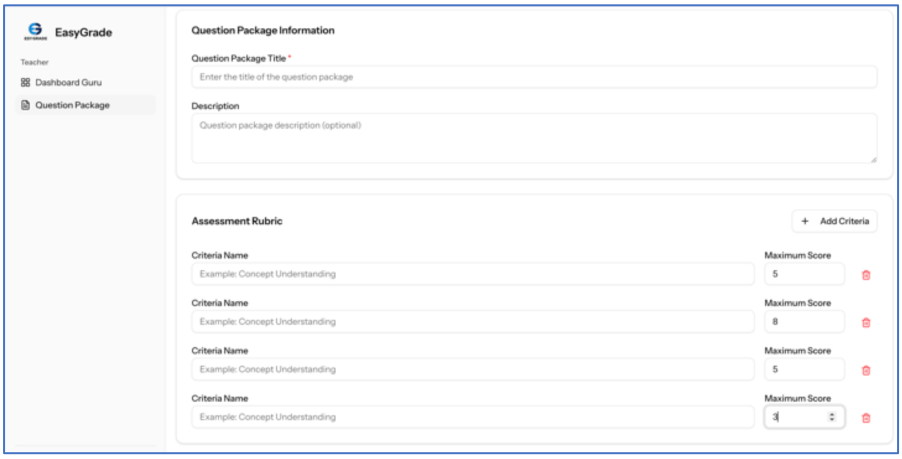


Figure 3:Revision Rubric Flexibility

Furthermore, Figure 4 demonstrates improvements in user-friendliness, where a simplified workflow (Create Question Package – Students Working – Automatic Results) is presented on the landing page, supported by a short usage guide to assist new users.

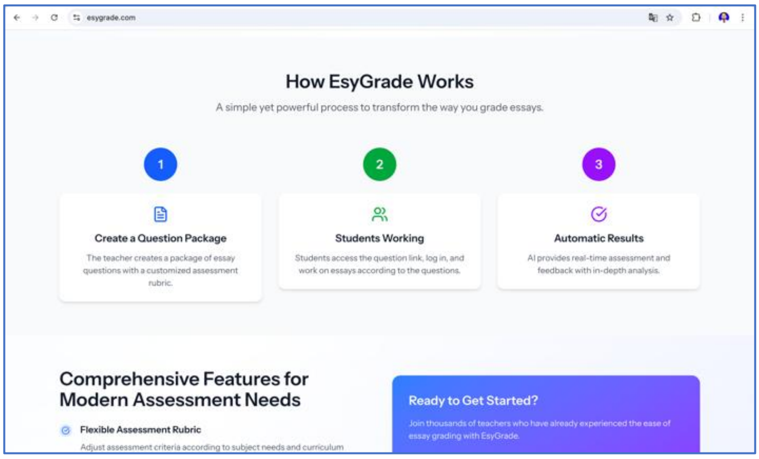


Figure 4: Revised User Interface

These changes indicate that the quality of a system is determined not only by the accuracy of its assessments but also by its ease of use and the clarity of user interactions. In the context of e-learning, the aspect of usability is crucial because it influences users' acceptance of the technology as well as the effectiveness of its implementation in learning. Consequently, the product revision phase serves as a technical enhancement and an endeavour to ensure that the developed system is pedagogically pertinent and adaptable to user requirements prior to entering the field-testing phase.

4.5 Preliminary Testing

A preliminary field test was conducted with ten students utilizing a Likert scale (1 = poor, 4 = very good) to assess the feasibility and practicality of the revised concept after expert input, as shown in Table 2.

Table 2: Preliminary Testing Results

Preliminary Testing		
Aspect	**Mean**	**Category**
Content Suitability & Validity	3.73333333	Very Good
Reliability & Fairness of Assessment	3.73333333	Very Good
Alignment with Educational Assessment Principles	3.46666667	Very Good
Clarity of Reporting Results	3.56666667	Very Good
Data Ethics & Privacy	3.55	Very Good
Integration with the Curriculum	3.9	Very Good

The questionnaire assessed the usability, interface clarity, rubric clarity, and feedback usefulness. All items as shown in Table 2, achieved the "Very Good" category (mean ≥ 3.26), with no item falling below the cut-off. These results designate strong acceptability and ease of use from the student side, confirming that the revisions flexible rubric customisation and a clear, step-by-step usage guide successfully addressed earlier expert feedback. In the context of e-learning, user acceptance is a key factor in determining the success of learning technology implementation. Systems that are easy to use and provide a clear user experience tend to be adopted more readily by users, making this initial pilot phase a crucial foundation before conducting further testing on a larger scale. Consistent "Very Good" ratings across aspects likely suggest that EsyGrade is ready to proceed to wider implementation in the main field test.

4.6 Main Field Testing

4.6.1 Normalized gain (n-gain) analysis

The main field testing involved 72 students across two grade levels: Grade 10 (36 students) who answered essay questions on Economic Activities and Grade 11 (36 students) who worked on Business and Management. Both groups completed the pre-test and post-test each of which has five essay items designed with equivalent type and weight according to Bloom's revised taxonomy. This standard is intended to ensure that changes in learning outcomes reflect improvements in conceptual understanding, rather than differences in the difficulty level of the questions.

The analysis employed the Normalised Gain (N-Gain) method, categorising improvement as *High* (≥ 0.7), *Medium* (0.3–0.69), and *Low* (< 0.3). Results showed that the average N-Gain score for both classes was in the Medium category, as shown in Table 3. These findings suggest that the use of a ChatGPT-based essay grading system contributes to an improvement in students' conceptual understanding. However, the fact that the improvement falls into the "moderate" category indicates that the resulting impact is moderate, meaning that not all students experienced optimal improvement.

Table 3: Main Field Testing

Main Field Testing							
Economic Activities (10th Grade)				**Business and Management (11th Grade)**			
Pre-Test Mean	**Post-test Mean**	**N-Gain Mean**	**Category**	**Pre-Test Mean**	**Post-Test Mean**	**N-Gain Mean**	**Category**
73.9166667	85.6388889	0.44358861	Medium	74.6111111	82.6944444	0.30379371	Medium

In this context, the observed improvement can be attributed to the role of the reflective feedback generated by the system. By using equivalent items in both the pre-test and post-test, changes in learning outcomes are more likely to result from students' ability to recognize their previous errors and correct them based on the feedback provided. Therefore, the system functions not only as an assessment tool but also as a formative learning tool that supports students' reflection on their conceptual understanding.

4.6.2 *Effect size analysis*

To complete the analysis, effect sizes were measured using paired Cohen's d. The results as shown in Table 4, show that Grade 10 had an effect size of 1.68, while Grade 11 had an effect size of 1.08, both of which fall into the "Large" category.

Table 4: Effect Size Testing

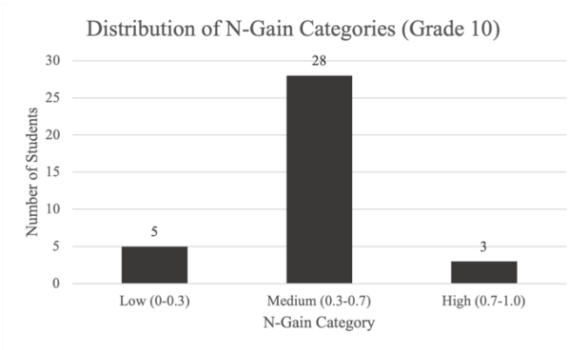
Class	Mean Difference	SD Difference	Cohen's d	Category
Economic Activities (10th Grade)	11.72	6.98	1.68	Large
Business and Management (11th Grade)	8.08	7.47	1.08	Large

*The interpretation of effect sizes follows Cohen's (1988) criteria, namely small (0.2), moderate (0.5), and large (0.8).

These findings suggest that, although the relative improvement measured using N-Gain falls into the moderate category, the magnitude of the change in learning outcomes is quite significant. This indicates that the system has a strong practical impact, while the degree of improvement varies among students.

4.6.3 *Distribution of n-gain*

To obtain a more detailed picture of the variation in improvement at the individual level, an analysis of the N-Gain distribution was conducted based on low, moderate, and high categories, as shown in Figure 5.



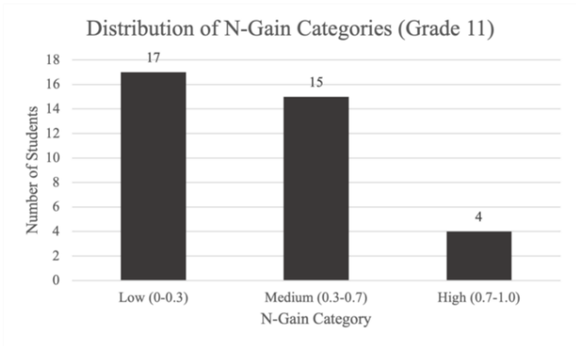


Figure 5: Distribution of N-Gain Categories

Based on the distribution of N-Gain shown in Figure 5, in 10th grade, the majority of students fell into the moderate category (28 students), with a small number of students in the low category (5 students) and the high category (3 students). This indicates that improvement occurred relatively evenly among most students. In contrast, in 11th grade, the distribution shows greater variation, with more students in the low category (17 students) than in the moderate (15 students) and high (4 students) categories. These findings suggest that the system's impact is not uniform across all students. This variation is likely influenced by factors such as initial ability, readiness to learn, and the ability to utilize the feedback provided.

4.7 Evaluation and Finalisation

The evaluation phase integrated all results from expert validation, pilot testing, and the main field trial. Overall, the results showed a moderate improvement in students' conceptual understanding based on N-Gain analysis, supported by a large effect size. These findings indicate that the system has a strong practical impact, although the relative level of improvement among students still varies. Beyond quantitative improvements, the system's primary value lies in its ability to provide reflective feedback on each student's essay response. This feedback not only presents a score but also specifically identifies the strengths and weaknesses of the response, thereby serving as a formative learning tool.

As shown in Figure 6, for partially correct answers, the system not only displays a score but also provides guidance on how to improve specific aspects such as reasoning and depth of analysis.

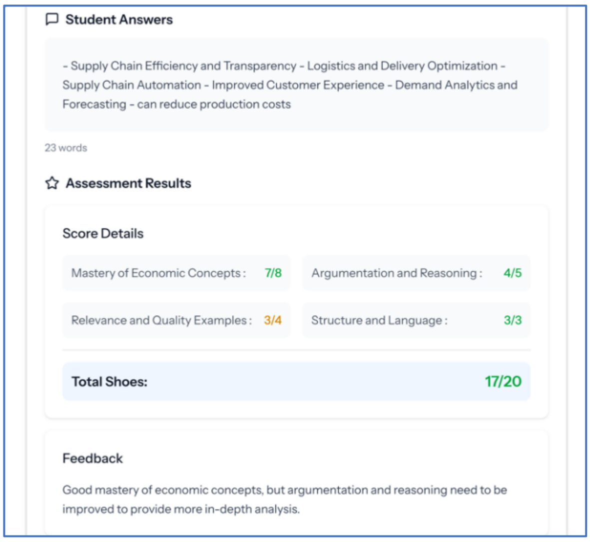


Figure 6: Feedback Partially Correct

Conversely, as shown in Figure 7, for correct and well-structured answers, the system reinforces the conceptual understanding that has been achieved.

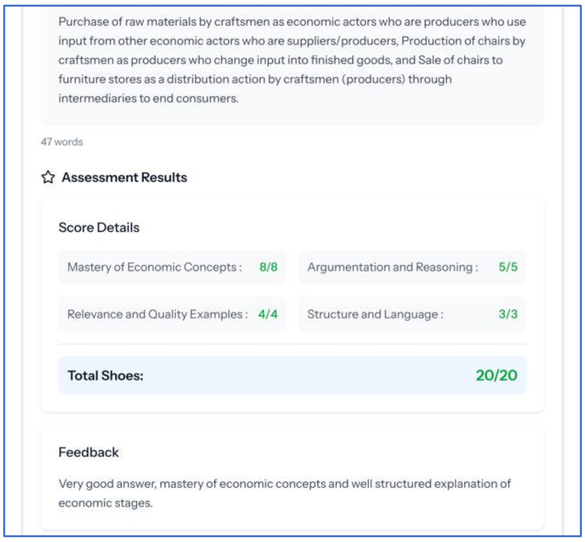


Figure 7: Feedback Completely Correct

These dual functions evaluative (scoring) and corrective (providing the right conceptual explanation) ensure that students not only recognise their errors but also learn the accurate concepts, thereby supporting formative and meaningful learning (Marton et al., 1984). The integration of reflective and corrective feedback is particularly important in economics education, where conceptual reasoning is essential. In the context of e-learning, specific and timely feedback is a crucial component in enhancing the quality of learning and student engagement.

Overall, this evaluation shows that EsyGrade functions effectively in line with its objectives: reducing teachers' workload, providing timely and structured feedback, and supporting the development of students' conceptual understanding. The system has been refined into a tool that is feasible, practical, and pedagogically sound for integrating AI-based automated essay grading into high school economics education and is ready for implementation in broader learning contexts.

5. Discussion

The results of this study indicate that the development of EsyGrade as a ChatGPT-based essay grading system can be successfully implemented and contributes positively to learning. Expert validation revealed a high level of validity, while preliminary trials indicated excellent levels of acceptance and practicality. In the main field trial phase, the analysis results showed a moderate improvement in students' conceptual understanding based on N-Gain, indicating an improvement in their abilities following the use of the system.

However, when combined with effect size results indicating a large category (Cohen's d > 1), a more comprehensive understanding emerges, indicating that the changes in learning outcomes are of substantial magnitude, even though the relative rate of improvement among students was uneven. This is reinforced by the N-Gain distribution analysis, which shows variation in improvement among individuals, particularly in 11th grade, where a larger proportion of students exhibited low improvement compared to 10th grade. Thus, the results of this study are more accurately interpreted as a moderate improvement on average, but with a strong overall impact.

These findings highlight that the primary contribution of the system lies not only in the automation of assessment, but in its ability to provide reflective feedback that supports formative learning. The feedback generated does not merely indicate right or wrong but also offers specific explanations regarding the strengths and weaknesses of students' answers. From the perspective of formative assessment theory, clear, specific, and timely feedback is a key factor in improving the quality of learning (Hattie and Timperley, 2007). Therefore, the integration of AI-based feedback in EsyGrade can be understood as a mechanism that fosters a continuous process of reflection and improvement in conceptual understanding.

These findings are consistent with previous research showing that the use of ChatGPT in essay grading can improve the quality of feedback and support students' independent learning (Mizumoto and Eguchi, 2023). Additionally, a study by Yavuz, Çelik and Yavaş Çelik, (2025) indicates that rubric-based grading assisted by ChatGPT has reliability comparable to that of human raters. This study expands on these findings by demonstrating that a similar approach can also be applied in the context of economics education, which has different conceptual and analytical characteristics compared to language learning.

In the Indonesian context, these findings are particularly relevant because assessment practices in schools still tend to be dominated by multiple-choice questions that emphasize rote memorization. This differs from trends in the broader literature, which is increasingly promoting the use of essay-based assessments and higher-order thinking as part of a more in-depth learning process (Maryani *et al.*, 2021; Zhao and Fu, 2025). Furthermore, teachers' heavy workloads and time constraints pose major obstacles to the widespread implementation of essay-based assessment. This contextual difference implies that automated grading systems like EsyGrade need to be designed adaptively, focusing not only on score automation but also on providing clear, structured, and user-friendly feedback. Therefore, features such as Bloom's taxonomy-based rubrics, flexibility in assessment customization, and a simple interface are crucial to ensuring alignment with the needs of users in Indonesia.

Nevertheless, several limitations should be noted. The system still relies on a stable internet connection, and the quality of the AI's responses sometimes requires verification by a teacher. Furthermore, this study is limited to a sample in the Surakarta region, so generalizing the findings requires caution.

Despite these limitations, this study offers valuable contributions both practically and theoretically. Practically, EsyGrade can serve as a tool to enhance the quality of formative assessment while helping to reduce teachers' workload in the essay grading process. In addition, this study contributes to e-learning research by demonstrating how AI-based assessment systems can be designed to support not only efficiency but also meaningful learning processes. Although this study was conducted in the context of economics education in Indonesia, the design principles developed such as the integration of rubrics based on Bloom's taxonomy, the use of customizable prompts, and the provision of reflective feedback have the potential to be applied to various other fields of learning. Theoretically, these findings suggest that AI-based essay grading systems can be conceptualised as tools not only to improve efficiency but also to support deep learning through structured reflective feedback. Thus, this study expands the e-learning literature by emphasizing the importance of integrating AI technology with pedagogical principles in the development of digital assessment systems.

6. Conclusion and Implications

This study aims to design, develop, and evaluate EsyGrade, a web-based automated essay grading system integrated with the ChatGPT API to assess conceptual understanding in economics education. Using a modified version of the Bennett, Borg and Gall, (1984) R&D model, the findings indicate that the developed system is feasible, user-friendly, and capable of supporting improvements in students' conceptual understanding. Expert validation indicates a high level of content validity, while preliminary trials confirm a strong level of user acceptance. Results from the main field trial show a moderate improvement based on N-Gain, supported by a large effect size, indicating a change in learning outcomes of considerable magnitude, although the level of improvement varies among students.

These findings indicate that the primary value of EsyGrade lies not only in the automation of the grading process, but also in its ability to provide structured, reflective feedback to support formative learning. Thus, the assessment is no longer viewed merely as a measurement tool, but as a means of supporting conceptual understanding and meaningful learning.

From a practical perspective, EsyGrade has the potential to serve as a tool to help teachers manage essay grading more efficiently without compromising the quality of feedback. This is particularly relevant in classes with large student populations and limited time for manual grading. From a policy perspective, these findings suggest that AI-based assessment systems can help bridge the gap between curriculum demands that emphasize higher-order thinking skills and the practical constraints faced by teachers.

More broadly, this study contributes to the development of e-learning by demonstrating how AI-based essay grading systems can be integrated with pedagogical principles such as rubric-based assessment and formative feedback. Although this system was developed in the context of economics education in Indonesia, the design principles employed such as the use of customizable prompts, alignment with rubrics, and a focus on feedback have the potential to be adapted across various fields and other educational contexts.

However, several limitations should be noted. The system still relies on a stable internet connection, and in some cases, the quality of the AI-generated feedback still requires verification by a teacher to ensure contextual appropriateness. Furthermore, this study was conducted within a limited geographical context, so the findings should be generalized with caution.

Future research could focus on several areas. First, studies with a broader and more diverse sample both in terms of geographic region and educational level are needed to enhance the generalizability of the findings. Second, further research could explore the long-term impact of AI-based feedback on the development of students' conceptual understanding. Third, integrating the system with Learning Management Systems (LMS) and evaluating its implementation over a longer period can provide a more comprehensive picture regarding the system's scalability and sustainability.

Acknowledgement

The author sincerely thanks the academic supervisors, family, and friends for their guidance and support during this study. Appreciation is also extended to the BIMA Master's Thesis Research Program of the Ministry of Higher Education, Science, and Technology of Indonesia (Fiscal Year 2025, Contract No. 105/C3/DT.05.00/PL/2025) for their valuable support.

AI Statement: No AI was used at any point in the research, writing, or creating of this paper.

Ethical Approvals: Ethical review and approval were not required for this study in accordance with the local legislation and institutional requirements. Participation in the questionnaire was entirely voluntary, and all respondents provided informed consent before participating. No personal identifying information was collected, and data were analysed anonymously to ensure participants' confidentiality.

Declaration of Conflict of Interest: All authors have read and approved the manuscript and take full responsibility for its content. The authors have no conflicts of interest regarding this research and its funding.

Data availability: Data will be made available upon request.

References

Alief, S., Irawati, I. and Mude, Muh.A. (2023) 'Impelementasi metode latent semantic analysis pada sistem informasi ujian online berbasis web', *Buletin Sistem Informasi dan Teknologi Islam*, 4(2), pp. 106–111. Available at: https://doi.org/10.33096/busiti.v4i2.1639.

Al-Obaydi, L.H., Pikhart, M. and Tawafak, R.M. (2023) 'Online Assessment in Language Teaching Environment through Essays, Oral Discussion, and Multiple-Choice Questions', *CALL-EJ*, 24(2), pp. 175–197. Available at: https://doi.org/callej.org/index.php/journal/article/view/7.

Anderson, L.W. and Krathwohl, D.R. (eds) (2001) '*A taxonomy for learning, teaching, and assessing: a revision of Bloom's taxonomy of educational objectives*'. Complete ed. New York: Longman.

Arini, Y.A.Y, Dewi, T.A. and Wibawa, F.A. (2024) 'Pengembangan e-modul ekonomi berbasis kurikulum merdeka kompetensi 4C semester ganjil kelas X SMAN 2 Metro', *EDUNOMIA: Jurnal Ilmiah Pendidikan Ekonomi*, 5(1), pp. 70–81. Available at: https://doi.org/10.24127/edunomia.v5i1.6600.

Beldar, P. (2025) Exploring the efficacy of open-ended questions in theory-based subjects', *Journal of Engineering Education Transformations*, 38(4), pp. 117–127. Available at: https://doi.org/10.16920/jeet/2024/v38i4/25101.

Bennett, N., Borg, W.R. and Gall, M.D. (1984) 'Educational research: An Introduction', *British Journal of Educational Studies*, 32(3), p. 274. Available at: https://doi.org/10.2307/3121583.

Biggs, J.B. and Tang, C.S. (2011) '*Teaching for quality learning at university: What the student does.*' 4th edition. Maidenhead: Open University Press (Society for Research into Higher Education).

Bloom, B.S. (1956) '*Taxonomy of educational objectives: The classification of educational goals*'. Longmans, Green (Taxonomy of Educational Objectives: The Classification of Educational Goals). Available at: https://books.google.co.id/books?id=1WjuAAAAMAAJ.

Cohen, J. (1988) '*Statistical power analysis for the behavioral sciences*'. 2nd ed. Hillsdale, N.J: L. Erlbaum Associates. Available at: https://doi.org/https://doi.org/10.4324/9780203771587.

Elliott, C. and Balasubramanyam, V. (2016) 'Assessing students: Real-world analyses underpinned by economic theory', *Cogent Economics & Finance*. Edited by S. Cook, 4(1), p. 1151171. Available at: https://doi.org/10.1080/23322039.2016.1151171.

Foung, D., Lin, L. and Chen, J. (2024) 'Reinventing assessments with ChatGPT and other online tools: Opportunities for GenAI-empowered assessment practices', *Computers and Education: Artificial Intelligence*, 6, p. 100250. Available at: https://doi.org/10.1016/j.caeai.2024.100250.

Gandolfi, A. (2025) 'GPT-4 in education: Evaluating aptness, reliability, and loss of coherence in solving calculus problems and grading submissions', *International Journal of Artificial Intelligence in Education*, 35(1), pp. 367–397. Available at: https://doi.org/10.1007/s40593-024-00403-3.

Hake, R.R. (1998) ' Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses', *American Journal of Physics*, 66(1), pp. 64–74. Available at: https://doi.org/10.1119/1.18809.

Hattie, J. and Timperley, H. (2007) 'The power of feedback', *Review of Educational Research*, 77(1), pp. 81–112. Available at: https://doi.org/10.3102/003465430298487.

Johnson, M. and Coleman, V. (2025) 'Teaching in uncertain times: Exploring links between the pandemic, assessment workload, and teacher wellbeing in England', *Research in Education*, 121(1), pp. 69–92. Available at: https://doi.org/10.1177/00345237231195270.

Kamalia, P.U.K., Yonisa; P. R., Fiky, A., Ghofur, M.A., Ginanjar, A.E. (2023) ' Pengembangan asesmen digital berbasis hots pada kurikulum merdeka bagi guru ekonomi, *SELAPARANG: Jurnal Pengabdian Masyarakat Berkemajuan*, (Vol 7, No 4 (2023): December), pp. 2886–2893. Available at: https://doi.org/https://journal.ummat.ac.id/index.php/jpmb/article/view/17046/8298.

Kinanti, N.L. & Qoiriah, A. (2020) ' Sistem penilaian otomatis jawaban esai Bahasa Indonesia berdasarkan kemiripan kalimat menggunakan syntactic-semantic similarity 02. Available at: https://ejournal.unesa.ac.id/index.php/jinacs/article/view/37555/33283.

Latif, E. & Zhai, X. (2024) ' Fine-tuning ChatGPT for automatic scoring', *Computers and Education: Artificial Intelligence*, 6, p. 100210. Available at: https://doi.org/10.1016/j.caeai.2024.100210.

Lee, G. G. *et al.* (2024) ' Applying large language models and chain-of-thought for automatic scoring', *Computers and Education: Artificial Intelligence*, 6, p. 100213. Available at: https://doi.org/10.1016/j.caeai.2024.100213.

Marton, F. & Hounsell, D. (eds) (1984) '*The experience of learning*'. Edinburgh: Scottish Academic Press. Available at: https://doi.org/doi.org/10.1111/b.9780631211860.1998.00025.x.

Maryani, I. *et al.* (2021) ' HOTs multiple choice and essay questions: A validated instrument to measure higher-order thinking skills of prospective teachers', *Turkish Journal of Science Education*, p. 4. Available at: https://doi.org/10.36681/tused.2021.97.

Mendonça, P.C., Quintal, F. & Mendonça, F. (2025) ' Evaluating LLMs for automated scoring in formative assessments', *Applied Sciences*, 15(5), p. 2787. Available at: https://doi.org/10.3390/app15052787.

Mizumoto, A. & Eguchi, M. (2023) ' Exploring the potential of using an AI language model for automated essay scoring, *Research Methods in Applied Linguistics*, 2(2), p. 100050. Available at: https://doi.org/10.1016/j.rmal.2023.100050.

Van Wyk, M.M. (2015) ' Teaching Economics, *International Encyclopedia of the Social & Behavioral Sciences*. Elsevier, pp. 83–88. Available at: https://doi.org/10.1016/B978-0-08-097086-8.92072-5.

Morris, C. *et al.* (2026) ' The components and implications of teacher workload: a review, *Educational Research*, pp. 1–22. Available at: https://doi.org/10.1080/00131881.2026.2629284.

Morris, R. *et al.* (2024) Can a code-based approach to marking and feedback reduce teachers' workload? An evaluation of the FLASH marking intervention, *Oxford Review of Education*, 50(4), pp. 552–569. Available at: https://doi.org/10.1080/03054985.2023.2258779.

OECD (2023) *PISA 2022 'Results (Volume I): The State of Learning and Equity in Education*. OECD Publishing (PISA). Available at: https://doi.org/10.1787/53f23881-en.

Pack, A., Barrett, A. & Escalante, J. (2024) ' Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability, *Computers and Education: Artificial Intelligence*, 6, p. 100234. Available at: https://doi.org/10.1016/j.caeai.2024.100234.

Pasaribu, N.G., Budiman, G. & Indrarini, D.I. (2024) ' Auto evaluation for essay zssessment using a 1D convolutional neural network, *IEEE Access*, 12, pp. 188217–188230. Available at: https://doi.org/10.1109/ACCESS.2024.3515837.

Permendikdasmen (2025) '*Peraturan menteri pendidikan dasar dan menengah Nomor 11 Tahun 2025 tentang Pemenuhan Beban Kerja Guru*. Available at: https://peraturan.bpk.go.id/Details/322487/permendikdasmen-no-11-tahun-2025 (Accessed: 25 November 2025).

Petersen, A., Craig, M. & Denny, P. '*ITiCSE '16: Innovation and Technology in Computer Science Education Conference 2016*, Arequipa Peru: ACM, pp. 252–253. Available at: https://doi.org/10.1145/2899415.2925503.

Plomp, T. & Nieveen, N. (2013) '*Educational design research: An introduction*. Available at: https://slo.nl/publish/pages/2904/educational-design-research-part-a.pdf (Accessed: 23 November 2025).

Poole, F.J. and Coss, M. (2023) ' Can ChatGPT reliably and accurately apply a rubric to L2 writing assessments? The devil is in the prompt(s). EdArXiv. Available at: https://doi.org/10.35542/osf.io/3r2zb.

Pratama, R.D. & Sangka, K.B. (2025) ' Assessing economic essays with ChatGPT: Systematic review and preliminary design', *International Conference on Teaching and Learning*, 1(1), pp. 337–345. Available at: https://doi.org/conference.ut.ac.id/index.php/ictl/article/view/1753/2343.

Rafi, R.R., Ramadhani, R.A. and Sanjaya, A. (2025) ' Pemanfaatan algoritma cosine similarity untuk mengkoreksi ujian esai", *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, 10(3), pp. 2242–2254. Available at: https://doi.org/10.29100/jipi.v10i3.8058.

Rittle-Johnson, B., Schneider, M. and Star, J.R. (2015) 'Not a one-way street: Bidirectional relations between procedural and conceptual knowledge of Mathematics', *Educational Psychology Review*, 27(4), pp. 587–597. Available at: https://doi.org/10.1007/s10648-015-9302-x.

Runtuwene, J.P.A. and Tangkawarow, I.R.H.T. (2020) 'The quality classification of professional teacher using fuzzy-analytical hierarchy process', *IOP Conference Series: Materials Science and Engineering*, 830(2), p. 022099. Available at: https://doi.org/10.1088/1757-899X/830/2/022099.

Sabon, S.S. (2020) 'Problematik pemenuhan beban kerja guru dan alternatif pemenuhannya (Studi kasus di kota depok provinsi jawa barat)', *Jurnal Penelitian Kebijakan Pendidikan*, 13(1), pp. 27–44. Available at: https://doi.org/10.24832/jpkp.v13i1.345.

Smerdon, D. (2024) 'AI in essay-based assessment: Student adoption, usage, and performance', *Computers and Education: Artificial Intelligence*, 7, p. 100288. Available at: https://doi.org/10.1016/j.caeai.2024.100288.

Steiss, J. *et al.* (2024) 'Comparing the quality of human and ChatGPT feedback of students' writing', *Learning and Instruction*, 91, p. 101894. Available at: https://doi.org/10.1016/j.learninstruc.2024.101894.

Tang, X. *et al.* (2024) 'Harnessing LLMs for multi-dimensional writing assessment: Reliability and alignment with human judgments', *Heliyon*, 10(14), p. e34262. Available at: https://doi.org/10.1016/j.heliyon.2024.e34262.

Walstad, W.B. (2006) 'Testing for depth of understanding in Economics using essay questions', *The Journal of Economic Education*, 37(1), pp. 38–47. Available at: https://doi.org/10.3200/JECE.37.1.38-47.

Yavuz, F., Çelik, Ö. and Yavaş Çelik, G. (2025) 'Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments', *British Journal of Educational Technology*, 56(1), pp. 150–166. Available at: https://doi.org/10.1111/bjet.13494.

Zhao, N. and Fu, Q. (2025) 'Under the framework of deep learning cognitivetheory and Bloom's Taxonomy: Investigating the Role of artificial intelligence in fostering higher-order thinking in engineering education', *2025 7th International Conference on Computer Science and Technologies in Education (CSTE)*. *2025 7th International Conference on Computer Science and Technologies in Education (CSTE)*, Wuhan, China: IEEE, pp. 892–896. Available at: https://doi.org/10.1109/CSTE64638.2025.11092073.

Appendix

This appendix presents the research instruments, including expert validation instruments, preliminary testing instruments, and pre-test and post-test questions to measure students' conceptual understanding.

A1. Instruments for educational assessment experts

	Indicator & Items
A.	**Content Relevance & Validity**
1.	The questions entered into the system can be tailored to the intended learning objectives for economics
2.	The automatic grading rubric reflects the competency indicators being assessed
3.	Test instructions, rubrics, and prompts are presented clearly and are easy for students and teachers to understand
B.	**Reliability & Fairness in Assessment**
4.	Automated essay grading is free from bias (student-neutral)
5.	The system assigns consistent scores to essay answers of the same quality
6.	The grading scale and weighting used are in accordance with educational assessment standards
C.	**Alignment with Educational Assessment Principles**
7.	The system supports the assessment of higher-order thinking skills (analysis, evaluation, synthesis)
8.	The system provides feedback in accordance with the principles of formative assessment

Indicator & Items	
9.	The system allows teachers to customize assessment rubrics and prompts as needed
D. Clarity in Reporting Results	
10.	The assessment reports are easy for teachers to read and understand
11.	Assessment results can be communicated to students in a transparent manner
12.	The system provides analysis of assessment results (such as score distributions and averages) to help teachers evaluate students
E. Data Ethics & Data Privacy	
13.	Student assessment data is stored and managed with due regard for confidentiality
14.	The system promotes student discipline and honesty (e.g., time limits)
F. Alignment with the Curriculum	
15	The system is aligned with the applicable curriculum and competency standards

A2. Instruments for educational technology experts

Indicator & Items	
A. Functional Suitability	
1.	The system's core features (question set creation, automated grading, and results reports) meet teachers' needs
2.	The automatic assessment results displayed according to the planned specifications
B. Performance Efficiency	
3.	The system's response time (for example, when processing essays) is fast and consistent
4.	The system can handle large volumes of students or questions without a significant drop in performance
C. Compatibility	
5.	The system can create custom rubrics and assessment prompts
D. Usability	
6.	The system's user interface is easy for teachers to understand and use
7.	The menu navigation (create question sets, rubrics, prompts, view results) is easy to follow
8.	The system provides sufficient guidance and help for new users
E. Reliability	
9.	The system rarely experiences errors or crashes during use
10.	User data (questions, student answers, assessment results) remains securely stored even in the event of an outage
F. Security	
11.	The system includes user data security features (secure login, protection of assessment results)
G. Maintainability	
12.	The system is easy to update (feature updates, bug fixes) without disrupting users
13.	The system's technical documentation is sufficient to support further development
H. Portability	
14.	The system can be easily moved or installed on another server, device, or operating system if necessary
15.	Assessment results can be exported as PDF or CSV

A3. Preliminary Testing Instrument for Students

Indicator & Items	
A. Ease of Access & System Navigation	
1.	It's easy for me to log in and access this system
2.	I found the menu and interface of this system easy to understand
B. Clarity of Instructions & Display	
3.	The instructions for answering the questions provided by this system are clear to me
4.	The layout of the questions and the scoring criteria in this system are engaging and easy to understand
C. Fairness & Transparency in Evaluation	
5.	The score I received feels fair based on my answers
6.	This system clearly displays the grading criteria, so I know the basis for the grading
D. Speed & Accuracy of Feedback	
7.	This system provides assessment results immediately after I finish answering the questions
8.	The feedback I received helped me understand my mistakes
E. Privacy & Security	
9.	I feel secure about my personal data when using this system
10.	My assessment results are securely stored and not misused
F. Support for Learning	
11.	This system helps me understand the economics material covered on the exam
12.	This system makes it easier for me to prepare for the next question
13.	This system motivates me to improve my academic performance
G. Overall Experience	
14.	Overall, this system is easy to use for solving problems
15.	I am satisfied with this automated grading system as a tool for evaluating the learning process

A4. Pre-test and post-test questions

Questions class X

Phase 1 (pre-test)

1. Why is production the core of a society's economic activities? (Level C4)
2. What are the four types of factors of production needed to produce a good or service? (Level C1)
3. What does distribution mean in economic activities? (Level C2)
4. A farmer grows rice. The rice is then sold to a rice merchant. The rice merchant sells it to small shops. Consumers buy rice at the shops to cook. Outline the sequence of economic activities! (Level C4)
5. Why can lifestyle influence a person's level of consumption? (Level C4)

Phase 2 (post-test)

1. Give three examples of consumption activities carried out by households on a daily basis! (Level C1)
2. Explain the difference between production and consumption activities! (Level C2)
3. What is the role of distribution institutions in delivering goods and services to consumers? (Level C2)
4. Look at the following illustration: A rattan craftsman buys raw materials from a supplier, produces chairs, and then sells them to a furniture store. Order the stages of these economic activities based on the roles of the economic actors! (Level C4)
5. Why can technological developments affect how producers distribute their goods? (Level C4)

Questions class XI

Phase 1 (pre-test)

1. Name three types of business entities in Indonesia! (Level C1)
2. Briefly explain the fundamental differences between state-owned enterprises (SOEs) and private enterprises! (Level C2)
3. A cooperative has a Surplus (SHU) of Rp100,000,000. Of this amount, 25% is distributed to members based on their deposits. If Member A's total deposit is Rp2,000,000 out of the total deposits of all members amounting to Rp20,000,000, how much of the Surplus will Member A receive? (Level C3)
4. Consider the following scenario: A new retail company is opening branches in five cities. Determine the planning and organizing steps the company should take! (Level C3)
5. Analyze the three main areas of management within a business entity (e.g., marketing, finance, and production), and explain how they are related! (Level C4)

Phase 2 (post-test)

1. Give two examples of cooperative-style business entities in Indonesia! (Level C1)
2. Briefly explain the difference between top management and middle management! (Level C2)
3. A cooperative receives a net surplus of Rp50,000,000. Thirty percent is distributed based on savings. If Member B's savings are Rp1,500,000 out of total savings of Rp15,000,000, calculate the net surplus received by Member B! (Level C3)
4. Given the following scenario: A logistics company is opening a new branch. Identify the "Actuating" and "Controlling" steps that need to be taken to ensure smooth operations! (Level C3)
5. Analyze the interrelationship between production, marketing, and finance in supporting the success of a startup! (Level C4)

Evaluating an AI-Supported Revision Module for Project-Based Research in Teacher Education

Assylzhan Yessimbekova[1], Ainash Issabekova[2], Karlygash Almenbetova[1] and Araily Shakirova[3]

[1]Department of Pedagogy and Methodology of Primary Education, Pedagogy and Psychology Faculty, Zhetysu University named after I. Zhansugurov, Taldykorgan, Kazakhstan

[2]Department of Pedagogy and Psychology, Pedagogy and Psychology Faculty, Zhetysu University named after I. Zhansugurov, Taldykorgan, Kazakhstan

[3]Department of Pedagogy and Educational Management, Farabi University, Almaty, Kazakhstan

assylzhanyessimbekova@gmail.com
ainashissabekova38@gmail.com
almenbetovakarlygash@gmail.com
shakirovaaraily@gmail.com (corresponding author)

https://doi.org/10.34190/ejel.24.3.4432

Abstract: While Generative Artificial Intelligence (GenAI) allows new methods to support students in writing processes, its incorporation into university teaching needs well-designed pedagogy to ensure academic integrity and autonomy of learners, especially in Project-Based Research (PBR) courses in teacher education, where iterative feedback is important but restricted by the size of groups and limited resources. The current study aims to address the scalability problem in the feedback process by evaluating the AI-Supported Revision Module (AIRM). It is implemented as a scaffold designed according to a rubric and embedded into Moodle, using a local version of the LLaMA-2-7B model for generation of prompts based on instructor comments for revision in four modes: polishing, restructuring, justifying, and synthesizing. The purpose of the module is to provide targeted guidance to students rather than full-text generation and support revision while preserving authorship. A mixed-methods case study approach was used involving 158 primary education student teachers, out of which 132 submitted draft and revision versions, assessed using a six-criterion rubric. A mixed-design repeated measures ANOVA test revealed interaction effects for both Literature Review and Methodology at the level of $p < .01$, with effect sizes of $d = 0.58$ and $d = 0.52$. The results suggest more improvement in structure and synthesis for the AI group than the conventional group, while no significant differences were found for Data Justification and Interpretation, suggesting a boundary between procedural support and higher-order analytical reasoning. Process data analysis revealed active involvement of learners, evidenced by the fact that on average 14.2% of draft content was changed and 39.7% of AI suggestions were rejected. Qualitative data analysis revealed that learners utilized this AI-powered module mostly to increase text coherence and clarity while staying in control of making sense of the content, whereas teachers indicated a shift from superficial to more methodological feedback practices. Thus, the findings demonstrate that GenAI may be successfully implemented in the feedback process as an additional scaffold for revision while still respecting the agency of learners.

Keywords: Generative artificial intelligence, Project-based research, Teacher education, Primary teachers, Feedback

1. Introduction

The swift advancement of Artificial Intelligence (AI) and especially Generative AI (GenAI), powered by Large Language Models (LLMs), has fundamentally shifted academic writing at universities. More broadly, AI has been adopted in higher education to support learning, assessment, and feedback. GenAI analyzes context and provides structured feedback, extending support beyond error correction toward higher-level writing processes. Unlike earlier rule-based systems, GenAI generates context-sensitive suggestions based on probabilistic language modelling, which has expanded its applicability in academic support tasks. For teacher education, this change carries a particular relevance to Project-Based Research (PBR) – a method aimed at fostering skills necessary for future primary teachers (Kokotsaki, Menzies and Wiggins, 2016). PBR's success is frequently hampered by what is known as a "feedback paradox." Although iteration is an indispensable aspect of gaining research skills, the conventional approach of providing individual feedback poses several logistic barriers within larger classes. As noted by Dean et al. (2023), without multiple feedback cycles, revisions often remain surface-level, focusing on grammar rather than methodological alignment. With insufficient feedback, students will often find it difficult to advance past purely descriptive literature reviews or to explain their conclusions transparently.

Cite this article: Yessimbekova, A.,et al. 2026. "Evaluating an AI-Supported Revision Module for Project-Based Research in Teacher Education", *Electronic Journal of e-Learning*, 24(3), pp 33-48, https://doi.org/10.34190/ejel.24.3.4432

The above limitations call for the use of GenAI as a scaffolding mechanism. While commercial applications of GenAI (such as ChatGPT) function independently to generate generic responses, this study employs a targeted approach that is tailored to a specific subject. Unlike commercial "black-box" services, which raise concerns about academic integrity and authorial voice (Zawacki-Richter et al., 2019), the AI-Supported Revision Module (AIRM) was designed as a transparent pedagogical tool. Based on the local LLaMA-2-7B model, it protects user privacy while creating prompts based on the rubric. The module processes student text together with instructor feedback to generate targeted, rubric-aligned revision suggestions at the paragraph level. Instead of producing text, it focuses on detecting gaps where improvements can be made by correlating feedback from the instructor with methods and giving suggestions for revisions. Feedback is operationalised through four modes: polish, restructure, justify, and synthesize.

AIRM is conceptualised as a pedagogical modification of the feedback cycle. The focus of this evaluation is not to claim long-term cognitive transfer, but to determine how such an AI-driven scaffold influences the quality of the final research product and the depth of the conceptual revision process. The study was conducted with third-year pre-service primary teachers at a pedagogical university in Kazakhstan, enrolled in the "Project-Based Research Activity" course. By comparing draft–revision pairs across AI-assisted and traditional cohorts, the research addresses the following questions:

RQ1: How does the AI-Supported Revision Module function within the revision cycle of project-based research in teacher education?

RQ2: Which aspects of student research writing show the greatest improvement when revisions are AI-assisted?

RQ3: How do students and instructors perceive the benefits and risks of integrating AI into revision tasks?

2. Literature Review

2.1 Project-Based Learning in Teacher Education

Project-Based Learning (PBL) encourages active, inquiry-based learning rather than passive learning. The use of projects helps the pre-service teachers learn to conduct inquiries, and also to acquire the identity of research-oriented professionals (Tsybulsky and Muchnik-Rozanov, 2023). At the same time, PBL is demanding for novices. Students often struggle to define researchable questions, formulate clear aims and narrow broad topics (Costley et al., 2023). Literature reviews tend to remain descriptive rather than synthetic, limiting the identification of gaps (Gao and Chen, 2024), while methodological reasoning is frequently weakened by misalignment between aims and methods (Helle, Tynjälä and Olkinuora, 2006).

2.2 Feedback and Revision in Project-Based Research Writing

Revision is central to PBR writing, enabling learning through iterative drafting and feedback cycles (Lu, Yao and Zhu, 2023). Feedback improves coherence and reflection, while peer input adds diverse perspectives, provided that students possess sufficient feedback literacy (Lineback and Holbrook, 2023; Wei and Liu, 2024; Carless and Boud, 2018).

However, revision practices remain inconsistent: students often focus on surface-level corrections or misinterpret feedback due to limited metacognitive skills, and instructors in large courses are unable to provide multiple feedback cycles (Hanafi et al., 2025; Yu and Xie, 2025). This creates a “feedback paradox,” where revision is essential for learning but constrained by institutional conditions. Structured feedback loops, particularly draft–feedback–resubmission cycles, have been shown to improve outcomes by narrowing the gap between current and desired performance (Bjælde, Boud and Lindberg, 2025), especially when combined with multiple feedback sources and structured debriefing (Sahlan, 2025). These limitations highlight the need for scalable approaches that operationalise feedback processes within high-enrollment contexts.

2.3 Generative Systems in Higher Education Writing

The introduction of GenAI has reframed debates about student writing, shifting from initial concerns about plagiarism, loss of originality and fabricated references toward a more balanced evidence base. Systematic reviews and meta-analyses show that, when used as a supplement rather than a substitute, generative tools can improve fluency and organisation, although effects on higher-order reasoning remain mixed and context-dependent (Deng et al., 2025; Lo, 2023). Similar patterns are observed in higher education, where improvements in readability and efficiency are accompanied by variability in output quality and the need for critical verification (Bhullar, Joshi and Chugh, 2024; O’Dea, 2024). Experimental studies further indicate gains in clarity and structure when AI is used for feedback (Mahapatra, 2024; Polakova and Ivenz, 2024) but also point to risks such as reduced creativity and shallow revisions (Niloy et al., 2024; Kim et al., 2025).

This limitation is relevant in education because research writing requires combining theoretical knowledge with practical applications; the methodology needs to be justified, and educational implications have to be discussed (Eysenbach, 2023). This indicates a limitation in human–AI collaboration between humans and artificial intelligence, since generative models are good at facilitating procedural activities in writing but cannot perform meaning-making tasks. Thus, the problem is not about how to employ the technology, but rather about how to create a proper scaffolding.

2.4 Integrity, Reliability and Reference Practices

A major limitation of generative systems is the unreliability of references, with studies showing frequent fabrication of citations in AI-generated texts (Rashidi et al., 2023), posing risks for academic writing where source credibility is essential (Chelli et al., 2024). Beyond this, overreliance on generative support may homogenise writing and weaken authorial voice, contributing to "cognitive offloading," where students uncritically adopt AI suggestions (Koo, 2023; Kooli, 2023; Nautiyal, Albrecht and Nautiyal, 2023). To mitigate these risks, students must develop evaluative judgement to critically assess automated guidance, particularly given the inherent variability of feedback processes (Ayçiçek, 2025). Recommended safeguards include DOI verification, disclosure of AI use and reflective tasks documenting how tools are applied (Walters and Wilder, 2023). However, these approaches remain largely theoretical, with limited empirical evidence on their effectiveness in authentic educational settings.

2.5 Positioning Within Technology-Integration Frameworks

Conceptual frameworks clarify how generative systems operate in education. The SAMR (Substitution–Augmentation–Modification–Redefinition) model defines technology adoption as substitution, augmentation, modification or redefinition (Hamilton, Rosenberg and Akcaoglu, 2016), with current evidence suggesting that generative tools primarily function at the levels of augmentation (improving grammar and style) and modification (supporting structural reorganisation and argument development), while redefinition remains rare in teacher education (Jiménez Sierra et al., 2023). The TPACK (Technological Pedagogical Content Knowledge) framework emphasises the alignment of technological, pedagogical and content knowledge (Mishra and Koehler, 2006), with research indicating that generative tools are most effective when embedded within pedagogical structures such as feedback cycles and aligned with disciplinary content (Lim et al., 2023). Across both frameworks, the role of agency is central: technology does not transform practice by itself, but depends on how it is integrated, perceived and critically engaged with. In this perspective, the value of AI lies in supporting a human-centred assessment transition, where technology facilitates student development and professional growth without displacing the learner's role as the primary agent in the writing process (Goode et al., 2026).

2.6 Summary of Gaps

However, PBL is crucial in teacher education and comes with constant challenges in problem identification, synthesis of literature and methodology justification. To address these challenges, feedback and revisions are needed but the process is hampered by organizational restrictions and lack of feedback literacy skills. Generative tools can enhance coherence and reflection within writing but there are concerns of invented citations, lack of originality and generalized writing.

The key limitation is the absence of scholarly studies that explain the functioning of generative tools during revision processes within teacher education. Studies have been more theoretical, focusing on either integrity and the technological potentials of the tool but not its real incorporation into feedback processes. Also, although SAMR and TPACK frameworks have been cited numerous times, very little attention has been paid to interpreting findings from the perspective of such frameworks.

3. Materials and Methods

3.1 Research Design

This study is a mixed-method case study that evaluated the effect of AIRM on a class of pre-service teachers. In a single course, students participated in two different conditions; namely, instructor-only feedback and instructor feedback with AIRM through Moodle. As randomization was not possible, equivalent baselines were set. Quantitative data included paired draft–revision submissions scored on six rubric criteria, along with process metrics (e.g., suggestions, acceptance rates, text modification, revision modes). Qualitative data included open-ended student surveys and instructor reflections, analysed thematically. Triangulation of rubric gains, system logs and thematic findings provided convergent evidence on revision behaviour, writing improvement and practical feasibility. The primary outcome of the study is the improvement in writing quality measured through rubric-based draft–revision comparisons, while interaction metrics (e.g., suggestion uptake and ratings) are used as complementary process indicators to interpret student engagement.

3.2 Context and Participants

This research was conducted at Zhetysu University named after I. Zhansugurov in Kazakhstan as part of an obligatory course during the third year of studies on PBR activities among pre-service primary school teachers. Students are required to conduct small-scale research related to primary schools. There were 158 participants in total, of which 132 provided complete pairs of drafts and revisions (Table 1). Students were allocated at the seminar group level for timetable reasons, which created a quasi-experimental design with non-random assignment but controlled instructional conditions. Sections assigned to the AI-assisted condition accessed AIRM in Moodle after receiving instructor feedback, while all other instructional conditions remained identical across groups. To minimise contamination, sections operated in partitioned Moodle spaces without cross-access to materials, while baseline comparability was established based on draft rubric scores and demographic variables (age, gender). No systematic differences between seminar groups or instructors were observed. All instructors followed a shared rubric and feedback protocol to ensure consistency.

Table 1: Participant Profile

	Participation			**Demographics**		
Delivery type	Enrolled students	Complete draft–revision pairs	Excluded submissions	Average age	Female (%)	Male (%)
Traditional (instructor comments only)	76	62	14	20.3	85.5	14.5
AI-assisted (with revision module)	82	70	12	20.5	83	17
Total	158	132	26	20.4	84	16

3.3 AI-Supported Revision Module

The AIRM was developed as an integrated extension of Moodle to support structured revision of PBR reports. It followed an iterative design process aligning LLM constraints with institutional rubric requirements, extending instructor feedback with targeted and transparent recommendations. Technically, AIRM used a locally hosted LLaMA-2-7B (temperature = 0.2, max tokens = 512) fine-tuned on institutional academic writing guides and program materials. The model was deployed locally without external API calls to ensure data privacy and full control over prompts and outputs. Predefined prompts generated concise, actionable guidance across four modes: polish (clarity and style), restructure (organisation), justify (strengthening arguments) and synthesise (connecting literature and concepts). Each prompt incorporated three elements: (1) the relevant segment of student text, (2) instructor feedback, and (3) task instructions aligned with a revision mode. Example prompt templates, including the orchestration step and mode-specific guidance, are provided in Appendix 1. The internal architecture and processing pipeline of the module are presented in Figure 1. The architecture restricts outputs to feedback-linked suggestions rather than full-text rewrites, ensuring student authorship.

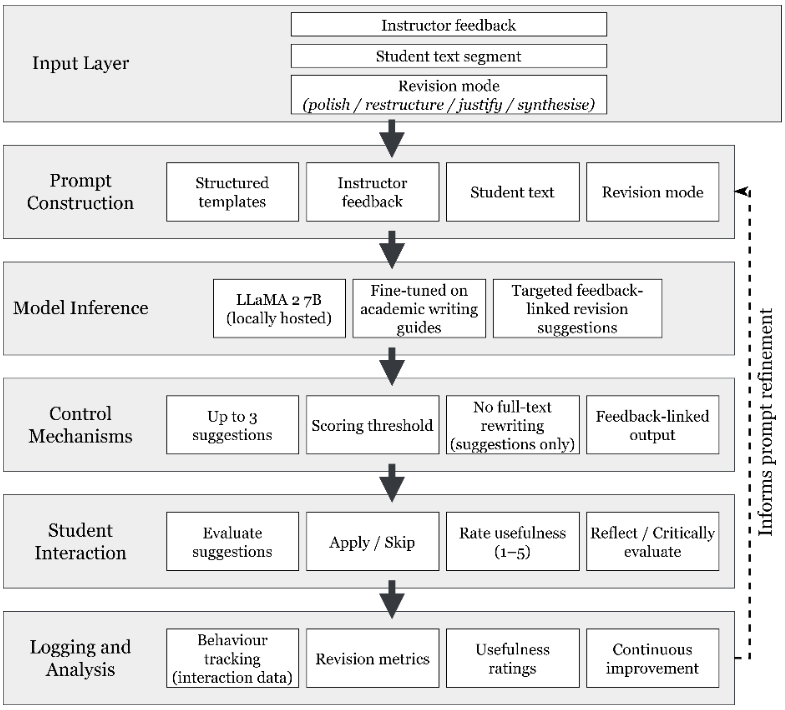


Figure 1: System Architecture of the AI-Supported Revision Module

When students accessed AIRM after receiving instructor feedback, relevant sections of the draft were identified based on instructor feedback and highlighted within the interface. Segmentation was performed at paragraph level and anchored to instructor comments to ensure consistent alignment. In cases where feedback referred to sub-paragraph elements, the nearest paragraph was used as the minimal segmentation unit to preserve contextual coherence. For each highlighted section, the system generated short, targeted suggestions indicating required improvements, without rewriting the text. Suggestions were based on the selected text, instructor feedback and instructional materials (e.g., academic writing guides) uploaded by the instructor. The overall workflow of the module within the revision cycle is illustrated in Figure 2.

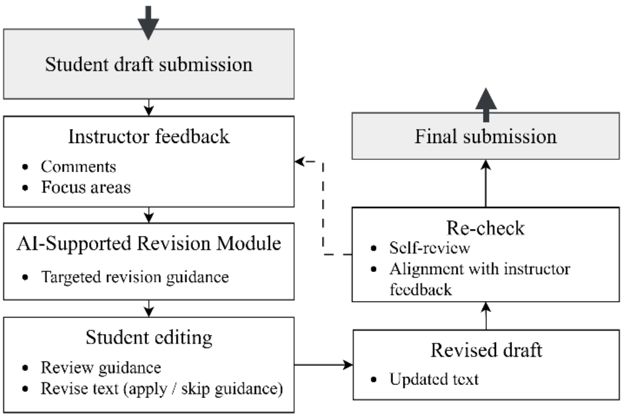


Figure 2: Workflow of the AI-Supported Revision Module Within the LMS Revision Cycle

The interface, shown in Figure 3, provided up to three suggestions for each section, allowing the user to either accept, skip, or evaluate its usefulness on a scale from one to five. Applying a suggestion indicated independent revision by the student; no automatic text replacement occurred. This interaction design was intended to support selective uptake of suggestions, ensuring active evaluation rather than passive acceptance.

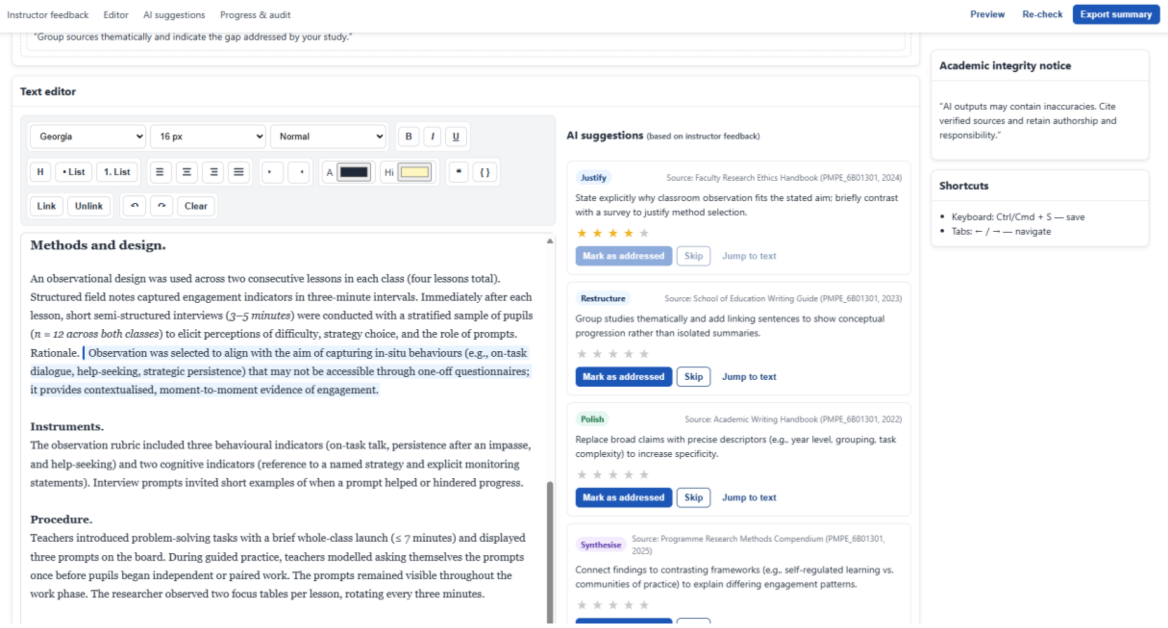


Figure 3: Student Interaction Interface and Suggestion Management in AIRM

In order to avoid cognitive overload, the number of suggestions that could be generated was restricted to three per section and determined by a threshold explicitly drawn from criteria specified within the rubric (Appendix 2). This threshold was operationalised via semantic similarity scoring between student text and rubric descriptors (embedding-based comparison), combined with rule-based heuristics (e.g., missing rationale, weak citation patterns, aim–method misalignment) to determine whether additional suggestions were required. For weaker segments, rule-based heuristics identified gaps (e.g., missing rationale, weak citation patterns, aim–method misalignment) to trigger mode-specific guidance. For example, the absence of an explicit methodological rationale triggered a "justify" suggestion, while disconnected paragraph transitions prompted a "restructure" recommendation. This ensured alignment between automated feedback and assessment standards.

Students retained full control over the revision process. Suggestions could be applied or skipped and optionally rated on a 1–5 usefulness scale. After revision, a report summarised text modifications and mode distribution. The four modes were aligned with key aspects of research writing: polish captured clarity and style, restructure targeted organisation in the literature review and methodological reasoning, justify strengthened methodological argumentation and data justification, while synthesise supported integration of sources in the literature review and interpretation. To safeguard integrity, each session began with a disclaimer reminding students to critically evaluate AI outputs and maintain authorship.

3.4 Data Collection

Multiple data sources were collected to evaluate the effects of the AI-Supported Revision Module across performance, process and perception. The core dataset consisted of 132 paired student submissions (draft and revised versions) from both cohorts.

All drafts and revisions were evaluated by two independent raters using a six-criterion rubric covering core dimensions of PBR. The raters were blinded to group allocation. Each of the 132 projects was assessed twice (draft and final), yielding 264 ratings. Inter-rater agreement was assessed with a two-way random-effects ICC (ICC(2,1), absolute agreement, single measures) with 95% confidence intervals. Overall reliability was good (ICC = .82, 95% CI [.76, .87]). Final scores were calculated as the mean of the two raters' evaluations.

Additional data were collected for the AI-assisted cohort. Module logs (N = 70 sessions) recorded suggestions generated, uptake (applied vs. skipped), percentage of text modified and changes across four modes (polish, restructure, justify, synthesise). In addition, 64 students completed voluntary surveys on benefits and limitations, and four instructors provided perspectives on feasibility and teaching implications. Survey participation was voluntary, which may introduce self-selection bias. Table 2 summarises all data sources.

Table 2: Overview of Data Sources

Data Source	Records / N	Purpose of Use
Draft–Revision Pairs	132 pairs	To measure improvement across six rubric criteria
Independent Rubric Scores	264 ratings	To ensure objective evaluation of drafts/revisions
Module Logs	70 sessions	To analyse interaction patterns (modes, % edits)
Student Surveys	64 forms	To capture perceptions of benefits and challenges
Instructor Surveys	4 forms	To assess feasibility and teaching implications

3.5 Data Analysis

Within each cohort, draft and revision scores were compared using paired t-tests across six rubric criteria. Between cohorts, differences in gain scores were analysed using independent t-tests and repeated-measures ANOVA. ANOVA tested the interaction between stage (draft vs. revision) and group (AI-assisted vs. traditional). All tests were two-tailed with the significance threshold set at $\alpha = .05$. Effect sizes (Cohen's d) were calculated using pooled Standard Deviations (SD) of draft and revision scores. For ANOVA, effect sizes are reported as partial eta squared (η^2_p), providing an indication of both statistical significance and practical importance. Assumptions of normality and homogeneity of variance were checked and met prior to analysis. All statistical analyses were conducted using JASP (version 0.95.4.0).

Qualitative data were analysed using thematic analysis to capture latent patterns and subjective experiences in student–AI interaction. Open-ended responses (n = 64 students; n = 4 instructors) were independently coded by two researchers using a hybrid framework combining deductive categories (polish, restructure, justify, synthesise) with inductive themes (e.g., cognitive load, feedback redundancy). Coding reliability was ensured through initial calibration and iterative consensus. Responses in Russian and Kazakh were translated and verified by bilingual researchers. Code frequencies and illustrative quotations are reported to support interpretation and triangulate quantitative findings.

The triangulation among rubric score improvements, log data, and qualitative themes yielded convergent and supplementary insights into how the module affected revision processes and results. The summary of the data collection process is presented in Figure 4.

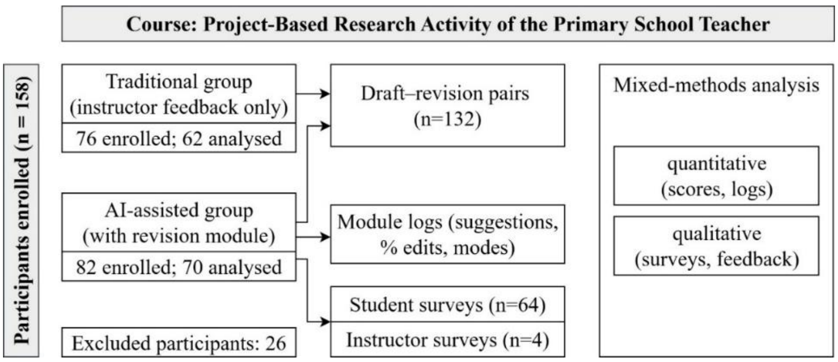


Figure 4: Study Design and Data Sources of the Research

3.6 Ethical Considerations

The study was conducted in a natural classroom environment without experimental manipulation beyond access to the AI-Supported Revision Module. All participation was voluntary and informed consent was obtained from every participant. Students could withdraw at any time without academic consequences, and participation did not affect course grades or formal assessment outcomes. No grades were assigned for the use of the module or participation in the study. All drafts and revisions were anonymised, and survey data were reported in aggregate form. The module itself was explicitly framed as a supportive tool rather than a replacement for student work. Students retained full control over revisions, including accepting or skipping suggestions, ensuring full authorship of the final text. Each session began with a disclaimer that AI outputs may be inaccurate and require critical evaluation. Students were instructed not to treat system outputs as authoritative sources. These measures ensured transparency, academic integrity and alignment with recognised ethical standards, including informed consent, data protection and voluntary participation.

4. Results and Findings

4.1 Quantitative Outcomes: Reliability and Overall Performance

The technical validity of the assessment was determined using the inter-rater reliability for 264 independent evaluations. A two-way random-effects intraclass correlation coefficient (ICC) for absolute agreement (single measures) demonstrated good reliability across the six criteria, with an overall ICC of 0.821 (95% CI [0.755, 0.869]).

Regarding overall student progress across the entire sample (N = 132), a paired samples t-test confirmed statistically significant improvements from draft to revision for all rubric criteria (Table 3). The largest effect sizes were observed in Literature Review (d = 0.58) and Methodology (d = 0.52), both indicating a medium-to-large impact. Improvements in analytical dimensions, such as Data Justification (d = 0.21) and Interpretation (d = 0.19), yielded smaller effect sizes, representing the baseline growth in higher-order research competencies (Mekheimer, 2025).

Table 3: Descriptive Statistics and Overall Growth from Draft to Revision (N = 132 Pairs)

Criterion	Draft Mean (SD)	Revision Mean (SD)	t (131)	Cohen's d	p
Problem Formulation	3.00 (1.35)	3.50 (1.64)	12.42	0.33	< .001
Literature Review	2.90 (1.12)	3.71 (1.61)	10.65	0.58	< .001
Methodology	2.80 (1.27)	3.61 (1.78)	10.44	0.52	< .001
Data Justification	3.10 (1.26)	3.40 (1.58)	2.20	0.21	.029
Interpretation	3.20 (1.29)	3.50 (1.85)	2.21	0.19	.029
Reflection	3.30 (1.39)	3.71 (1.48)	16.94	0.28	< .001

4.2 Comparative Analysis: AI-Assisted vs. Traditional Cohorts

To assess the added value of the module, a mixed-design repeated measures ANOVA was conducted to compare the performance of the AI-assisted group (n = 70) and the traditional group (n = 62). Descriptive statistics for both cohorts are presented in Table 4, including means and SD for each stage.

Table 4: Descriptive Statistics by Group and Stage (n = 132)

Group	Criterion	Draft Mean (SD)	Revision Mean (SD)	Gain (Δ)
Traditional	Problem Formulation	3.17 (1.32)	3.56 (1.64)	+0.39
	Literature Review	2.86 (1.19)	3.45 (1.73)	+0.59
	Methodology	3.03 (1.29)	3.62 (1.74)	+0.59
	Data Justification	3.13 (1.20)	3.43 (1.59)	+0.30
	Interpretation	3.36 (1.28)	3.61 (1.83)	+0.25
	Reflection	3.38 (1.36)	3.73 (1.45)	+0.35
AI-assisted	Problem Formulation	2.85 (1.37)	3.44 (1.65)	+0.59
	Literature Review	2.94 (1.05)	3.93 (1.47)	+0.99
	Methodology	2.60 (1.22)	3.59 (1.83)	+0.99
	Data Justification	3.07 (1.31)	3.37 (1.58)	+0.30
	Interpretation	3.06 (1.28)	3.41 (1.87)	+0.35
	Reflection	3.24 (1.42)	3.69 (1.52)	+0.45

The ANOVA results (Table 5) revealed a highly significant interaction effect between Stage and Group for Literature Review ($F(1, 130) = 7.506$, $p = .007$, with a partial eta squared of 0.055) and Methodology ($F(1, 130) = 7.197$, $p = .008$, with a partial eta squared of 0.052).

Table 5: Summary of Repeated Measures ANOVA (Interaction Effects: Stage x Group)

Criterion	SS	df	MS	F	p	η^2_p
Problem Formulation	0.696	1	0.696	4.718	.032	0.035
Literature Review	2.685	1	2.685	7.506	.007	0.055
Methodology	2.685	1	2.685	7.197	.008	0.052
Data Justification	0.000	1	0.000	0.000	.995	<.001
Interpretation	0.164	1	0.164	0.131	.718	0.001
Reflection	0.175	1	0.175	4.861	.029	0.036

Significant interaction effects were also found for Problem Formulation (p = .032) and Reflection (p = .029). However, no significant interaction was observed for Data Justification (F(1, 130) < 0.001) or Interpretation (F(1, 130) = 0.131, p = .718), where both cohorts demonstrated statistically equivalent progress. Figure 5 illustrates these interaction patterns, highlighting the steeper slope of improvement for the AI-assisted group in Literature Review and Methodology, contrasting with the parallel trajectories observed in analytical criteria.

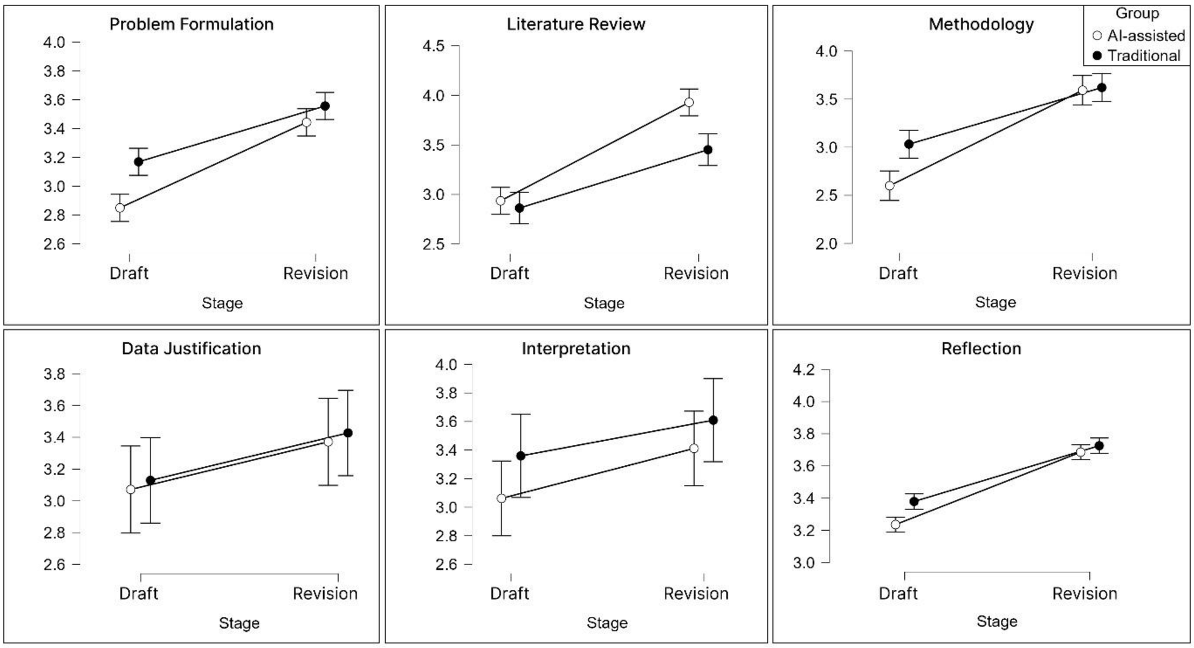


Figure 5: Interaction Effect Between Stage (Draft vs. Revision) and Condition (AI-Assisted vs. Traditional) on Mean Rubric Scores

4.3 Process Metrics: Student Interaction with AIRM

To understand the mechanisms behind the score improvements, interaction logs from the AI-assisted cohort (n = 70) were analysed. On average, the system generated 6.8 suggestions per report (SD = 2.24). A key finding was student agency: students demonstrated a selective approach, accepting 60.3% of recommendations (M = 4.1, SD = 1.45) and rejecting 39.7% (M = 2.7, SD = 0.91).

The system also captured utility ratings for each suggestion. Of the 476 generated suggestions, 412 (86.5%) received an optional rating on a 1–5 scale, yielding an overall mean of 3.9 (SD = 0.8). Perceived usefulness was highest for the Polish mode (M = 4.2, SD = 0.6), followed by Restructure (M = 3.9, SD = 0.7), Justify (M = 3.6, SD = 0.9), and Synthesise (M = 3.4, SD = 1.1). The descriptive statistics for these interaction metrics are summarised in Table 6.

Table 6: Interaction Metrics from AIRM System Logs (n = 70)

Metric	Mean (SD)	Range	Total
Suggestions Generated	6.8 (2.24)	3–12	476
Suggestions Accepted	4.1 (1.45)	2–8	287
Suggestions Rejected	2.7 (0.91)	1–5	189
Text Modified (Word Count)	281.4 (109.7)	125–519	19,700

The volume of revision was significantly higher in the AI-assisted group compared to the traditional cohort. Students using the module modified an average of 14.2% (SD = 5.02) of their draft text, whereas the traditional group, relying on peer and self-correction, modified only 8.5% (SD = 3.2) of their content. This difference suggests that the AI-assisted feedback acted as a more effective "nudge," prompting students to engage in more extensive rewriting, particularly in sections where the system identified structural or logical gaps. As shown in Figure 6, the modification rate was not uniform across the report sections. The highest percentages of modified text were observed in the Literature Review and Methodology sections, which directly correlates with the highest score gains reported in Section 4.2.

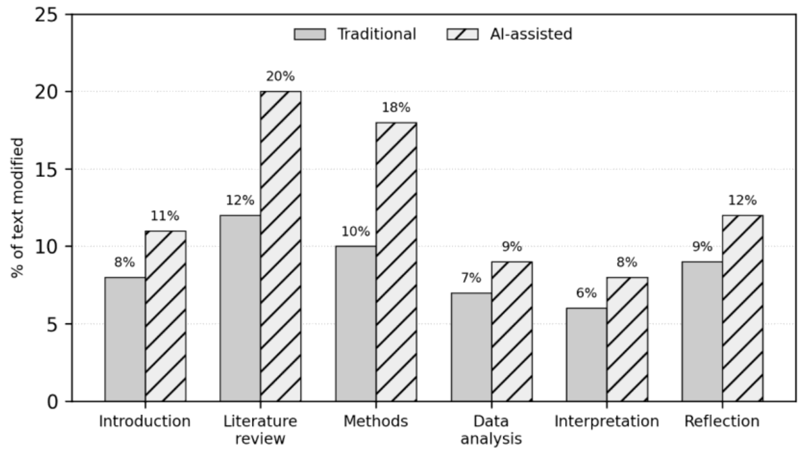


Figure 6: Percentage of Text Modified by Section and Condition

4.4 Qualitative Insights: Thematic Analysis of Feedback

To capture the subjective experiences of the participants, qualitative data from open-ended survey responses (N = 64 students) and instructor reflections (n = 4) were analysed. Thematic coding was performed by two independent researchers on a corpus of 118 individual comments. An initial inter-coder agreement of 88% was achieved, with discrepancies resolved through consensus. The analysis identified distinct patterns of engagement, with frequencies indicating the relative prominence of themes, categorised into Perceived Benefits and Reported Challenges (Table 7).

Table 7: Thematic Coding of Student and Instructor Feedback

Category	Theme	Frequency	Illustrative Quote	Source
Perceived benefits	Clarity and organisation	40	“The module helped me make my text clearer and more structured, especially in the introduction.”	Student
	Metacognitive reflection	24	“I liked the final report because it showed me what I had actually changed, not just what I thought I changed.”	Student
	Value of new perspectives	8	“The suggestion gave me a new idea for connecting literature.”	Student
	Shift in instructor role	4	“The module freed me from correcting grammar, so I could focus on methodology with students.”	Instructor
	Selective adoption of suggestions	12	“I skipped some advice because it did not fit my argument, but others were very useful.”	Student

Category	Theme	Frequency	Illustrative Quote	Source
Reported Challenges	Cognitive demand	16	“Sometimes there were too many points at once and I did not know where to start.”	Student
	Feedback redundancy	10	“This was similar to what the teacher already wrote, so I did not use it.”	Student
	Critical evaluation requirement	3	“Students must learn not to trust the AI blindly, but to use it as support.”	Instructor

Note: Frequency = number of mentions

Students consistently described the module as a tool for improving clarity and organisation, suggesting that it functioned primarily as a structural scaffold. As one student noted: “The AI didn't just fix my grammar; it highlighted where my methodology was fragmented, which made it easier to align my steps with the research aims.” Many students emphasised the role of the revision report in supporting metacognitive awareness, particularly in recognising the extent of their own revisions. Another participant reflected: “Seeing the summary of my changes at the end was eye-opening; it made me realize how much I had actually improved the text compared to my first draft.”

However, students also reported cognitive challenges, indicating that the volume of automated suggestions could at times hinder the revision process. One student remarked: “At times, receiving three different suggestions for one paragraph was overwhelming, and I struggled to decide which one to prioritize.” Finally, while Critical Evaluation (n = 3) was less frequently mentioned in surveys, the qualitative evidence confirms that student agency was active and deliberate. This was best captured by a student who explained their selective adoption: “I had to be careful; I skipped suggestions that tried to change the meaning of my results, as the AI didn’t know the specific classroom context of my project.” These insights, aligned with the 40% rejection rate in Section 4.3, demonstrate that the module fostered a critical dialogue rather than a passive acceptance of technology.

5. Discussion

5.1 Summary of Key Findings

The results show that the AIRM greatly improved students' research writing through a more stringent revision strategy. According to the mixed-design ANOVA outcomes (Section 4.2), the treatment was found to be effective in achieving significant improvement in two important dimensions – Literature Review (η^2_p = 0.055, d = 0.58) and Methodology (η^2_p = 0.052, d = 0.52). These effects suggest that the module supported idea development within a rubric framework. As argued by Bjælde, Boud and Lindberg (2025), the most effective feedback activities are those that move beyond mere information delivery and instead facilitate a continuous "draft–feedback–resubmission" cycle. In our study, the module transformed instructor comments from static advice into actionable, iterative prompts. This is evidenced by the process metrics: the AI-assisted group modified 14.2% of their draft text, which is a nearly 70% increase compared to the 8.5% modification rate observed in the traditional cohort. This discrepancy confirms that the module acted as a "catalyst," preventing the "minimalist revision trap" where students change only minor surface elements while leaving the underlying conceptual flaws untouched.

Furthermore, the significant improvements in methodological reasoning support the findings of Sahlan (2025), who emphasizes that integrating tutor feedback with structured reflection enhances the pedagogical competence of pre-service teachers. By utilizing the Restructure and Synthesise modes, students were nudged to align their research aims with their methodological choices more precisely. Operating at the SAMR Modification level, the system redesigned revision tasks to foster deeper inquiry rather than merely substituting feedback.

The concentration of gains in structural and logical dimensions (Literature Review and Methodology) highlights a strategic alignment between the inherent strengths of LLMs and the procedural needs of novice researchers. These results also suggest that the AIRM effectively managed the "procedural load" of research writing. By providing a logical framework for structural synthesis, the module enabled students to overcome the cognitive barriers of organizing complex information, allowing them to produce more coherent narratives. This alignment between pedagogy and technology (TPACK) shows that the module’s effectiveness comes from its integration into instructor-led feedback cycles.

5.2 Conceptual Boundaries and the Analytic Gap in AI Scaffolding

Results reveal an “analytic gap,” evidenced by non-significant performance gains in Data Justification and Interpretation across both cohorts. Unlike the structural improvements observed in other criteria, these analytical dimensions yielded no significant interaction effect ($p > .05$). This gap highlights a boundary between procedural support and higher-order synthesis, distinguishing AIRM’s scaffolded approach from authorship-blurring open tools.

The absence of growth in interpretation serves as empirical evidence against concerns regarding unreflective AI adoption or automated content generation. As highlighted by Tlili et al. (2023), the integration of generative tools often carries the risk of "hallucinations" that can undermine a student’s critical engagement. However, the plateau in these specific results demonstrates that the AIRM functioned strictly as a cognitive scaffold rather than a generative substitute. By restricting the model’s access to raw primary datasets, the system design ensured that students remained the sole intellectual owners of their findings. PBR requires situated knowledge in order to achieve interpretative results since there is no way for any generalized model to match the knowledge of a specific pedagogical situation existing in Kazakhstan. In this case, the "analytic gap" justifies the use of the module as an ethical tool because it successfully handles the "procedural load" required by the structure, and at the same time retains the "epistemic load" in meaning creation for the human researcher.

5.3 Selective Adoption and the Manifestation of Student Agency

The interaction logs reveal a complex pattern of engagement that contradicts the premise of unreflective AI adoption. The 39.7% rejection rate suggests that students acted as critical evaluators rather than passive recipients. This selective adoption aligns with the "active role" of students in the feedback process advocated by Bjælde, Boud and Lindberg (2025), where learning occurs through the deliberate evaluation and application of suggestions. Rejecting AI suggestions manifests evaluative judgment as a core research competence rather than scaffold failure.

The qualitative data presented in Table 7 further elucidates these thematic priorities. The high frequency of mentions regarding Clarity and Organisation (40 instances) compared to Justification and Argumentation (12 instances) reflects a hierarchical approach to revision. Students utilized the AIRM primarily to resolve "first-order" structural and linguistic barriers, thereby creating the necessary cognitive space to refine their internal arguments. Furthermore, the integration of AI guidance provides a distinct contrast to traditional peer assessment mechanisms. While Ayçiçek (2025) highlights the inherent variability and subjectivity often associated with peer-driven feedback, the AIRM offered a consistent methodological "grid." However, the 14.2% text modification rate indicates that this stability did not lead to standardisation; instead, it served as a catalyst for individualised rewriting. By positioning the student as the final authority over each suggested revision, the module ensured that the authorial voice remained intact. Consequently, the revision process was transformed into a critical dialogue, where the ability to discern and reject irrelevant advice became as pedagogically valuable as the adoption of helpful prompts.

5.4 Pedagogical Transitions and the Reconfiguration of Instructional Roles

The implementation of the AIRM facilitated a shift in instructional dynamics of the research-based course. Qualitative insights from the participating instructors indicated a transition from serving as primary providers of surface-level corrections to assuming roles as high-level methodological mentors. This redistribution of instructional effort was characterized by the delegation of repetitive structural and grammatical feedback to the AI module, which consequently permitted instructors to dedicate more time to complex, inquiry-driven dialogue.

This transition aligns with the principles of human-centered assessment articulated by Goode et al. (2026), where the focus of evaluation shifts from mere measurement toward the holistic support of student transitions and professional development. In this framework, the AI scaffold functioned as a preliminary layer of support, ensuring that drafts reached a certain level of structural maturity before human intervention. Consequently, the interaction between mentor and student was elevated, focusing on methodological nuances and pedagogical implications rather than basic clarity.

Furthermore, the integration of the module encouraged a more rigorous form of professional reflection. As suggested by Sahlan (2025), combining tutor feedback with structured revision tools is essential for enhancing the competence of pre-service teachers, as it necessitates a deliberate "debriefing" of one's own writing. The instructors observed that this model effectively addressed the logistical challenges of high-enrollment programs, allowing for a scalable feedback mechanism that does not compromise the depth of individualised guidance. This pedagogical modification suggests that the value of AI in teacher education lies in its ability to optimize human expertise, ensuring that mentors intervene at the most critical stages of the inquiry process.

5.5 Research Limitations and Future Directions

This study has several limitations. First, it evaluates scaffolded performance rather than independent skill transfer, as no post-test was conducted after removing AIRM support. Second, the quasi-experimental design without random assignment introduces potential selection bias despite baseline equivalence. Third, the findings are context-specific, as the study was conducted within a single pedagogical university in Kazakhstan with pre-service primary teachers, in a multilingual setting where academic writing is often performed in English as a Foreign Language (EFL). These contextual, cultural, and linguistic factors may have influenced revision patterns and the effectiveness of AI-supported feedback, limiting generalisability. Fourth, a possible novelty effect may have increased student engagement. Fifth, the results depend on the specific configuration of the LLaMA-2-7B model and prompt design, which may not transfer to other systems. Finally, despite implemented safeguards, the risk of cognitive offloading and over-reliance on AI remains (Tlili et al., 2023).

Future research should examine long-term skill transfer through longitudinal designs and delayed post-tests without AI support. Replication across diverse cultural, linguistic and disciplinary contexts is needed to test generalisability. Further work should also explore adaptive prompting strategies to better support higher-order analytical reasoning and address the identified “analytic gap,” while maintaining student agency and authorship.

6. Conclusion

The current research shows that it is possible to deploy the AIRM framework in a research course, thus providing an efficient example of feedback provision in the context of teacher education. By connecting a locally deployed language model with institutional rubrics and instructor intentions, this research goes further than the chatbot paradigm by creating a feedback-oriented scaffold that ensures effective revision without any creation of output text. This module has positively impacted the structure and methodology of students' essays, leaving the limits of human authorship untouched.

In relation to the research questions, the findings indicate that AIRM functioned as a structured scaffold within the revision cycle of project-based research (RQ1), leading to the most substantial improvements in literature review and methodological alignment (RQ2). At the same time, both students and instructors perceived the module as a supportive tool that enhances clarity and feedback processes while preserving learner agency and critical judgement (RQ3).

Evaluation suggests that although AIRM acts as a potent instigator of meaningful change, the extent of influence is dictated by the dynamics of a purposeful synergy and not automation per se. With the development of the “analytical gap” and the high rejection rate of AI recommendations, it is evident that students acted as the primary agents of change in the process, performing an evaluative function of judgment and “feedback literacy.” Evaluation lends credence to the human-in-the-loop model, where instructional load is partially transferred to AI, enabling educators to focus on higher-level teaching processes while informing the design of scalable, human-centred feedback systems in teacher education and beyond.

AI Statement: During the preparation of this manuscript, the authors used Grammarly for language editing and proofreading purposes only. No generative AI tool was used for content generation, data analysis, interpretation, or reference generation.

Ethics Statement: The study was conducted in accordance with institutional ethical guidelines. Formal ethical approval was not required as the study involved normal educational practice, anonymised data, and no impact on grading or student progression.

Conflict of Interest: The authors declare no conflicts of interest.

References

Ayçiçek, B. (2025). Peer assessment analysis via the many-facet Rasch model. *Mersin Üniversitesi Eğitim Fakültesi Dergisi, 21*(2), pp.631–646. https://doi.org/10.17860/mersinefd.1642676

Bhullar, P.S., Joshi, M. and Chugh, R. (2024). ChatGPT in higher education: A synthesis of the literature and a future research agenda. *Education and Information Technologies*, *29*, pp.21501–21522. https://doi.org/10.1007/s10639-024-12723-x

Bjælde, O.E., Boud, D. and Lindberg, A.B. (2025). Designing feedback activities to help low-performing students. *Active Learning in Higher Education, 26*(1), pp.123–137. https://doi.org/10.1177/14697874231212820

Carless, D. and Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. *Assessment & Evaluation in Higher Education*, *43*(8), pp.1315–1325. https://doi.org/10.1080/02602938.2018.1463354

Chelli, M., Descamps, J., Lavoué, V., Trojani, C., Azar, M., Deckert, M., Raynier, J., Clowez, G., Boileau, P. and Ruetsch-Chelli, C. (2024). Hallucination rates and reference accuracy of ChatGPT and Bard for systematic reviews: Comparative analysis. *Journal of Medical Internet Research*, *26*, e53164. https://doi.org/10.2196/53164

Costley, J., Zhang, H., Courtney, M., Shulgina, G., Baldwin, M. and Fanguy, M. (2023). Peer editing using shared online documents: The effects of comments and track changes on student L2 academic writing quality. *Computer Assisted Language Learning, 38*(4), pp.865–891. https://doi.org/10.1080/09588221.2023.2233573

Dean, C.G.P., Grossman, P., Enumah, L., Herrmann, Z. and Schneider Kavanagh, S. (2023). Core practices for project-based learning: Learning from experienced practitioners in the United States. *Teaching and Teacher Education, 133*, 104275. https://doi.org/10.1016/j.tate.2023.104275

Deng, R., Jiang, M., Yu, X., Lu, Y. and Liu, S. (2025). Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies. *Computers & Education, 227*, 105224. https://doi.org/10.1016/j.compedu.2024.105224

Eysenbach, G. (2023). The role of ChatGPT, generative language models, and artificial intelligence in medical education: A conversation with ChatGPT and a call for papers. *JMIR Medical Education, 9*, e46885. https://doi.org/10.2196/46885

Gao, J. and Chen, W. (2024). Developing culturally-situated student feedback literacy through multi-peer feedback giving: An online community-based approach. *Language Awareness, 34*(1), pp.19–42. https://doi.org/10.1080/09658416.2024.2321894

Goode, E., Valentine, J., Krautloher, A. and Anand, P. (2026). Assessment for student transitions and success: A scoping review of assessment principles in Australian enabling education. *Higher Education Research & Development, 45*(3), pp.554-571. https://doi.org/10.1080/07294360.2025.2543409

Hamilton, E.R., Rosenberg, J.M. and Akcaoglu, M. (2016). The substitution augmentation modification redefinition (SAMR) model: a critical review and suggestions for its use. *TechTrends, 60*(5), pp.433–441. https://doi.org/10.1007/s11528-016-0091-y

Hanafi, I., Kheder, K., Sabouni, R., Gorra Al Nafouri, M., Hanafi, B., Alsalkini, M., Kenjrawi, Y., Albkhetan, H. and Alhalabi, M. (2025). Improving academic writing in a low-resource country: A systematic examination of online peer-run training. *Teaching and Learning in Medicine, 37*(3), pp.388–402. https://doi.org/10.1080/10401334.2024.2332890

Helle, L., Tynjälä, P. and Olkinuora, E. (2006). Project-based learning in post-secondary education – Theory, practice and rubber sling shots. *Higher Education, 51*(2), pp.287–314. https://doi.org/10.1007/s10734-004-6386-5

Jiménez Sierra, Á.A., Ortega Iglesias, J.M., Cabero-Almenara, J. and Palacios-Rodríguez, A. (2023). Development of the teacher's technological pedagogical content knowledge (TPACK) from the Lesson Study: A systematic review. *Frontiers in Education*, *8*, 1078913. https://doi.org/10.3389/feduc.2023.1078913

Kim, J., Yu, S., Detrick, R. and Li, N. (2025). Exploring students' perspectives on generative AI-assisted academic writing. *Education and Information Technologies, 30*(1), pp.1265–1300. https://doi.org/10.1007/s10639-024-12878-7

Kokotsaki, D., Menzies, V. and Wiggins, A. (2016). Project-based learning: A review of the literature. *Improving Schools, 19*(3), pp.267–277. https://doi.org/10.1177/1365480216659733

Koo, M. (2023). The importance of proper use of ChatGPT in medical writing. *Radiology, 307*(3), e230312. https://doi.org/10.1148/radiol.230312

Kooli, C. (2023). Chatbots in education and research: A critical examination of ethical implications and solutions. *Sustainability, 15*, 5614. https://doi.org/10.3390/su15075614

Lim, W.M., Gunasekara, A., Pallant, J.L., Pallant, J.I. and Pechenkina, E. (2023). Generative AI and the future of education: Ragnarök or reformation? A paradoxical perspective from management educators. *The International Journal of Management Education, 21*(2), 100790. https://doi.org/10.1016/j.ijme.2023.100790

Lineback, J.E. and Holbrook, E. (2023). Engaging in a collaborative space: Exploring the substance and impact of peer review conversations. *Assessment & Evaluation in Higher Education, 49*(4), pp.556–571. https://doi.org/10.1080/02602938.2023.2290978

Lo, C.K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences, 13*, 410. https://doi.org/10.3390/educsci13040410

Lu, Q., Yao, Y. and Zhu, X. (2023). The relationship between peer feedback features and revision sources mediated by feedback acceptance: The effect on undergraduate students' writing performance. *Assessing Writing, 56*, 100725. https://doi.org/10.1016/j.asw.2023.100725

Mahapatra, S. (2024). Impact of ChatGPT on ESL students' academic writing skills: A mixed methods intervention study. *Smart Learning Environments, 11*, 9. https://doi.org/10.1186/s40561-024-00295-9

Mishra, P. and Koehler, M.J. (2006). Technological pedagogical content knowledge: A framework for teacher knowledge. *Teachers College Record, 108*(6), pp.1017–1054. https://doi.org/10.1111/j.1467-9620.2006.00684.x

Mekheimer, M. (2025). Generative AI-assisted feedback and EFL writing: a study on proficiency, revision frequency and writing quality. *Discover Education, 4*, 170. https://doi.org/10.1007/s44217-025-00602-7

Nautiyal, R., Albrecht, J.N. and Nautiyal, A. (2023). ChatGPT and tourism academia. *Annals of Tourism Research, 99*, 103544. https://doi.org/10.1016/j.annals.2023.103544

Niloy, A.C., Akter, S., Sultana, N., Sultana, J. and Rahman, S.I.U. (2024). Is ChatGPT a menace for creative writing ability? An experiment. *Journal of Computer Assisted Learning, 40*(2), pp.919–930. https://doi.org/10.1111/jcal.12929

O'Dea, X. (2024). Generative AI: Is it a paradigm shift for higher education? *Studies in Higher Education, 49*(5), pp.811–816. https://doi.org/10.1080/03075079.2024.2332944

Polakova, P. and Ivenz, P. (2024). The impact of ChatGPT feedback on the development of EFL students' writing skills. *Cogent Education, 11*(1), 2410101. https://doi.org/10.1080/2331186X.2024.2410101

Rashidi, H.H., Fennell, B.D., Albahra, S., Hu, B. and Gorbett, T. (2023). The ChatGPT conundrum: Human-generated scientific manuscripts misidentified as AI creations by AI text detection tool. *Journal of Pathology Informatics, 14*, 100342. https://doi.org/10.1016/j.jpi.2023.100342

Sahlan, M. (2025). Integrating self-assessment, peer assessment, and tutor feedback with structured debriefing to enhance pre-service teachers' teaching competence: Empirical evidence from Indonesia. *Qualitative Research Journal*, pp.1–17. https://doi.org/10.1108/QRJ-05-2025-0184

Tlili, A., Shehata, B., Adarkwah, M.A., Bozkurt, A., Hickey, D.T., Huang, R. and Agyemang, B. (2023). What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. *Smart Learning Environments, 10*(1), 15. https://doi.org/10.1186/s40561-023-00237-x

Tsybulsky, D. and Muchnik-Rozanov, Y. (2023). The contribution of a project-based learning course, designed as a pedagogy of practice, to the development of preservice teachers' professional identity. *Teaching and Teacher Education, 124*, 104020. https://doi.org/10.1016/j.tate.2023.104020

Walters, W.H. and Wilder, E.I. (2023). Fabrication and errors in the bibliographic citations generated by ChatGPT. *Scientific Reports, 13*, 14045. https://doi.org/10.1038/s41598-023-41032-5

Wei, Y. and Liu, D. (2024). Incorporating peer feedback in academic writing: A systematic review of benefits and challenges. *Frontiers in Psychology, 15*, 1506725. https://doi.org/10.3389/fpsyg.2024.1506725

Yu, H. and Xie, Q. (2025). Generative AI vs. teachers: Feedback quality, feedback uptake, and revision. *Language Teaching Research Quarterly, 47*, pp.113–137. https://doi.org/10.32038/ltrq.2025.47.07

Zawacki-Richter, O., Marín, V.I., Bond, M. and 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), 39. https://doi.org/10.1186/s41239-019-0171-0

Appendix 1

All prompts were executed deterministically within the same template structure to ensure consistency of outputs across users and sessions.

Table 8: Prompt Templates for AI-Supported Revision Guidance

Prompt Type	Prompt Template
Orchestrator	Task: Analyse student text in relation to instructor feedback and supporting materials, and generate structured revision guidance. Inputs: text = [TEXT]; feedback = [TEXT]; materials = [INSTITUTIONAL MATERIALS] Instructions: identify relevant segments; classify issues based on rubric-aligned criteria (clarity, coherence, argumentation, synthesis); map each issue to a revision mode (polish / restructure / justify / synthesise); generate targeted guidance per segment. Output (JSON): { "segments": [{ "text_span": "...", "mode": "...", "issue": "...", "suggestion": "..."}] } Constraints: no full-text rewriting; guidance only; strongly align with feedback and materials. Example output: { "segments": [{ "text_span": "…", "mode": "justify", "issue": "claim not supported", "suggestion": "add empirical reference linking to aim" }] }
Restructure	Task: Analyse text structure using feedback. Identify issues in flow and organisation. Suggest how to reorder or connect ideas. Constraints: no rewriting; concise guidance only.
Justify	Evaluate strength of argumentation. Identify unsupported claims or weak reasoning. Suggest how to strengthen justification and link to aims. Constraints: no rewriting; no full sentences for insertion.

Appendix 2

Table 9: Rubric for Evaluating Project-Based Research Reports

Criterion	1 – Minimal	2 – Emerging	3 – Developing	4 – Proficient	5 – Advanced
Problem Formulation	Problem absent or incomprehensible	Problem vague, aim unclear	Problem identifiable but	Problem clear and mostly	Problem precise, fully aligned with methodology,

Criterion	1 – Minimal	2 – Emerging	3 – Developing	4 – Proficient	5 – Advanced
			partly aligned with methods	aligned with aims	demonstrates originality
Literature Review	Fewer than 3 references; no academic sources	3–5 sources; descriptive, weak relevance	6–8 sources; some organisation, limited synthesis	9–12 sources; partial synthesis, mostly relevant and recent	≥ 13 sources; strong synthesis (>60% thematic discussion), clear gap identified
Methodology	Method missing or irrelevant	Method present but described superficially	Method described with some rationale, partial alignment	Method explained clearly, rationale present, aligned with aims	Method rigorously explained, rationale well supported, fully aligned with aims
Data Justification	Data absent or irrelevant	Data mentioned without justification	Data described but weakly linked to aims	Data appropriate and linked to aims with adequate justification	Data fully justified, strongly linked to aims, supported by credible references
Interpretation	No interpretation beyond restating results	Minimal interpretation, weak link to aims	Interpretation present but descriptive	Interpretation clear, with critical linkage to aims	Interpretation comprehensive, critical and connected to broader literature
Reflection	No reflection on process or outcomes	Minimal reflection, superficial	Reflection present but limited in depth	Reflection balanced, covers strengths and weaknesses	Reflection extensive, critical and linked to professional growth

Explaining Online Learning Attitudes: Community of Inquiry Presences, Learning Outcomes, and Learner Characteristics

Irit Sasson
Tel-Hai University of Kiryat Shmona and the Galilee, Israel
Shamir Research Institute, University of Haifa, Israel

iritsa@telhai.ac.il

https://doi.org/10.34190/ejel.24.3.4739

Abstract: Online learning has become a core mode of higher education, intensifying questions about what constitutes high-quality teaching and learning in digital environments and how students form evaluations of their online learning experiences. The Community of Inquiry (CoI) framework highlights teaching, social, and cognitive presence as key pedagogical conditions, yet less is known about how these conditions operate alongside learner characteristics and students' perceived learning outcomes to shape attitudes toward online learning. This quantitative study surveyed 316 undergraduates enrolled in 22 fully online undergraduate courses. Students reported demographic characteristics (including age, gender, ethnicity, faculty affiliation, learning-disability status) and self-reported cumulative Grade Point Average (GPA) range. Perceptions of CoI presences were measured using a validated CoI instrument, and perceived learning outcomes were assessed as multidimensional gains (cognitive, metacognitive, and social). Analyses included group comparisons and hierarchical regression models predicting attitudes toward online learning. Group comparisons indicated significant differences in perceived presence across demographic groups, although most effects were small. Perceptions of cognitive and social presence also varied across self-reported GPA ranges, whereas teaching presence was relatively stable. In hierarchical regression, demographic variables explained a modest portion of variance in attitudes. Adding CoI presences substantially improved prediction, with cognitive presence emerging as the primary presence-related predictor. When perceived learning outcomes were added, perceived cognitive and metacognitive gains were the strongest predictors of more positive attitudes, and the unique contribution of CoI presences was reduced, suggesting that perceived learning gains may help explain the presence–attitude link. The findings therefore point to a more integrative account of online learning quality, in which students' attitudes appear to depend less on fixed demographic differences and more on whether online courses are experienced as cognitively meaningful and supportive of reflective growth. Findings underscore the centrality of cognitive engagement and perceived learning gains for shaping students' attitudes toward online learning and point to actionable design priorities: inquiry-oriented activities, structured reflection and metacognitive scaffolds, and consistent course organization and support that promote equity across diverse learners. These results also inform institutional policy by emphasizing shared online-course quality standards and professional learning focused on evidence-informed design practices.

Keywords: Online learning, Community of inquiry, Student attitudes, Perceived learning outcomes

1. Introduction and Theoretical Background

In recent years, education systems around the world have invested considerable effort in designing innovative learning environments, with a particular emphasis on the integration of digital technologies (Collins & Halverson, 2018). The outbreak of the Covid-19 pandemic in early 2020 triggered a dramatic transformation in education, requiring a rapid and comprehensive transition to remote learning for all students (Dietrich, 2020, Fewella, 2023, McGaughey et al. 2022). Importantly, the pandemic accelerated a shift that was already underway: online learning had been steadily expanding in higher education as institutions sought to increase access and flexibility, and it has remained a central mode of provision in the post-pandemic landscape (McGaughey et al., 2022, Müller & Mildenberger, 2021, Getenet et al., 2024). Although it is difficult to equate emergency-induced online learning with programs designed intentionally for digital delivery, the pandemic nonetheless exposed persistent challenges in online learning, such as reduced interaction between students, instructors, and course content, and a noticeable decline in the sense of academic belonging for both students and faculty (Baxter & Haycock, 2014, Seifert & Bar-Tal, 2023). It also highlighted disparities in students' home learning environments, exacerbating existing inequalities (Gillis & Krull, 2020, Sasson & Yehuda, 2026). This issue is especially relevant in the Israeli higher-education context examined in the present study, which includes both Jewish majority and Arab minority students and therefore provides an opportunity to consider how online learning is experienced across a socially diverse student population. In this context, students' attitudes toward online learning have become a critical indicator of quality, shaping engagement, persistence, and willingness to enroll in future online courses (Akar, 2024, Jiang, Zhou & Yang, 2024, McIntyre et al., 2023, Paris, Lakhal & Mukamurera, 2025, Steyn et al., 2024). In addition to attitudes, students' perceived learning outcomes (Rovai et al., 2009, Yang, Wang &

Cite this article: Sasson, I. 2026. "Explaining Online Learning Attitudes: Community of Inquiry Presences, Learning Outcomes, and Learner Characteristics", *Electronic Journal of e-Learning*, 24(3), pp 49-64, https://doi.org/10.34190/ejel.24.3.4739

Zhao, 2025) provide a complementary indicator of online course quality, capturing learners' own assessment of the gains they attribute to the learning experience. In the present study, we conceptualize perceived outcomes as multidimensional, encompassing cognitive, metacognitive, and social learning gains. Examining these perceived outcomes alongside pedagogical features and learner characteristics helps clarify how pedagogical conditions translate into students' overall evaluation of online learning. Taken together, these considerations position students' attitudes toward online learning as a product of the interplay among pedagogical conditions, learner characteristics, and perceived learning gains.

Alongside these challenges, online learning also offers meaningful pedagogical opportunities. At the institutional level, flexible provision can expand access across time and place, yet its value depends on intentional design and adequate student support rather than convenience alone (Jones et al., 2025, Xie, Siau & Nah, 2020). Pedagogically, well-designed online learning can support multimodal engagement and collaborative knowledge construction (Olsson, Mozelius & Collin, 2016, Reyes, Cuenca & Martínez, 2025, Simpson, 2016). In addition, learning analytics and student-facing dashboards enable monitoring of learning processes and the provision of actionable feedback, provided they are aligned with the learning sciences and used to inform pedagogical action (Paulsen & Lindsay, 2024, Saghafi, Franz & Crowther, 2014).

The shift to remote learning during Covid-19 catalyzed a new era of pedagogical experimentation. Online learning has moved from a marginal alternative to a central and, at times, essential component of higher education. As a result, the need for thoughtful pedagogical design and institutional policy for online learning has become both urgent (Andrade & Alden-Rivers, 2019). At the same time, evidence consistently suggests that students do not benefit from online learning in the same way: learner characteristics and resources (e.g., prior academic experience and self-regulation) shape how students engage with online courses and the support they require to succeed (Xu & Jaggars, 2013, Dumford & Miller, 2018, Muljana & Luo, 2019). These differences make it especially important to clarify which pedagogical conditions foster meaningful interaction and positive attitudes for diverse learners. However, we still have limited understanding of which pedagogical conditions most strongly foster meaningful interaction with instructors, peers, and course content and how these conditions relate to students' attitudes toward online learning (Nortvig, Petersen & Balle, 2018, Rodrigues et al., 2019).

Recent scholarship has increasingly moved beyond viewing online learning as a mere technological shift and has instead highlighted the pedagogical conditions that underpin high-quality online higher education. Research evidence indicate that effectiveness is shaped by interacting factors such as clear course structure and design (e.g., coherent organization of learning activities, transparent expectations, and feasible timelines), high-quality interaction and collaborative learning processes that build a sense of community, assessment and feedback practices aligned with intended learning outcomes, and student-related capabilities such as self-regulation, metacognitive skills, and digital literacy (Broadbent & Poon, 2015, Castro & Tumibay, 2021, Muljana & Luo, 2019, Sasson & Yehuda, 2026, Sun & Chen, 2016, van Dorresteijn et al., 2025). Complementing this broader picture, work on collaborative online learning demonstrates that outcomes such as learning gains and satisfaction are closely tied to the quality of digital interaction and group-process characteristics (e.g., participation, transactivity, and group climate), underscoring that pedagogical quality is enacted through interactional processes rather than tools alone (Bach & Thiel, 2024). Existing evidence often examines these elements in relative isolation, leaving open how pedagogical features of online learning environments, learner characteristics, and perceived learning gains jointly shape students' overall experience and attitudes toward online learning (Lai & Bower, 2019). Addressing this need, the present study investigates the combined contribution of pedagogical characteristics, learner characteristics, and perceived learning outcomes to students' attitudes toward online learning in higher education.

To examine these pedagogical conditions more systematically, the present study draws on the Community of Inquiry (CoI) framework developed by Garrison, Anderson, and Archer (2000), one of the most influential models for analyzing the quality of online learning environments. This model conceptualizes meaningful online learning as the result of the dynamic interaction between three core types of presence: cognitive presence – the extent to which learners are able to construct and confirm meaning through sustained reflection and discourse, social presence – the ability of participants to present themselves as "real people" in the online environment and to engage in purposeful communication and collaboration, and teaching presence – the design, facilitation, and direction of cognitive and social processes to support learning outcomes. Together, these three dimensions form the foundation for a successful and engaging online learning experience. The CoI model provides a valuable lens for examining the pedagogical quality of online courses because it focuses directly on the interactional and instructional processes through which meaningful online learning is organized, facilitated, and experienced by students.

Recent scholarship continues to affirm the CoI framework as both a design guide and an analytic lens, while also emphasizing contextual variability in how “presence” is enacted in practice. In a systematic review of CoI applications in TESL/TEFL, Guo, Jeyaraj, and Razali (2025) show that studies increasingly use CoI not only to evaluate learners’ perceptions (often via survey-based measures) but also to inform concrete instructional design choices such as structuring interaction, scaffolding inquiry processes, and aligning activities with language-learning goals. Richardson et al. (2025) demonstrate through a multiple case study across disciplines that CoI indicators are not uniformly evident in online course designs and that instructors’ disciplinary and pedagogical goals shape which indicators are foregrounded, suggesting that CoI-informed “quality” may legitimately look different across contexts. Finally, meta-analytic evidence links CoI processes to outcomes. Yang, Wang, and Zhao (2025) report strong positive associations between cognitive presence and both perceived learning and satisfaction, while also showing that effect sizes vary by factors such as disciplinary area, course duration, and measurement scale type. Taken together, these findings motivate examining how students’ perceptions of teaching, social, and cognitive presence relate to learner characteristics and attitudes, clarifying which pedagogical features are most consequential for students’ online learning experience in higher education. At the same time, an important explanatory gap remains within this line of research. Although CoI studies have consistently linked the three presences to outcomes such as perceived learning and satisfaction, it is still unclear whether students’ attitudes toward online learning are shaped directly by teaching, social, and cognitive presence, or whether these presences matter primarily because they foster perceived cognitive and metacognitive gains.

This literature suggests that while the CoI framework offers a strong lens for conceptualizing the pedagogical processes that shape online learning, it may not by itself fully explain variation in students’ evaluations of online courses. Studies indicate that CoI indicators are enacted differently across disciplinary contexts and course designs (Lim & Richardson, 2022, Richardson et al., 2025), and meta-analytic evidence shows that CoI processes, especially cognitive presence, are closely associated with perceived learning and satisfaction (Martin et al., 2022, Yang, Wang & Zhao, 2025). At the same time, broader research on online higher education shows that students do not experience online learning uniformly, as learner characteristics and resources shape engagement and support needs (Dumford & Miller, 2018, Muljana & Luo, 2019, Xu & Jaggars, 2013). For this reason, the present study uses CoI as its primary pedagogical lens, while also incorporating learner characteristics and perceived learning outcomes to examine how pedagogical processes are experienced by different students and translated into attitudes toward online learning.

Grounded in the COI framework, the present study addresses these needs by examining (a) whether students’ perceptions of teaching, social, and cognitive presence differ across key demographic groups and self-reported achievement levels, and (b) how student characteristics, perceived three CoI presences, and perceived learning outcomes jointly explain attitudes toward online learning. By clarifying whether the association of CoI presences with attitudes is direct or is better understood through perceived cognitive and metacognitive gains, the study contributes to a more explicit account of how pedagogical quality is translated into students’ evaluations of online learning, while also offering actionable insights for improving both the effectiveness and equity of online course design and institutional policy.

1.1 Research Questions

RQ1: To what extent do students’ perceptions of the three CoI presences (teaching, cognitive, and social) differ across student characteristics, including age, gender, ethnicity, faculty affiliation, learning-disability status, and self-reported academic achievement (GPA)?

RQ2: What are the predictors of students’ attitudes toward online learning, as examined through three conceptual models: (1) student demographic characteristics, (2) perceived three CoI presences, and (3) perceived learning outcomes (cognitive, metacognitive, and social)?

2. Methodology

The quantitative research method that focuses on collecting numerical data that are analyzed using statistics in order to explain a phenomenon (Creswell & Creswell, 2018), was chosen for the present research. The goal is to identify statistical relationships between the study variables.

The research was carried out at a university in Northern Israel and examined students' experiences in fully online courses. The sample included 316 students. Notably, the Israeli context is substantively relevant because the analysis considers demographic differences between Jewish and Arab students - groups that reflect broader social and cultural diversity in Israeli higher education. Within the institution, roughly three-quarters of students are Jewish (about 75%), while the remaining quarter consists of minority students from the Arab sector. This demographic composition provides a meaningful context for interpreting patterns in students' perceptions and attitudes toward online learning.

A total of 22 undergraduate courses were represented, spanning multiple disciplines, including education, social sciences, humanities, and science and technology. The courses varied in scope and level, comprising both introductory courses and more content-specific offerings. All courses were delivered fully online. Some were conducted in a fully asynchronous format, whereas others included limited synchronous online sessions (e.g., via Zoom) or other structured online interaction components.

2.1 Instruments

A structured questionnaire was developed consisting of four parts. The first part collected demographic information from the participants, including age, gender, ethnicity (self-identified as Jewish or Arab), faculty affiliation, and whether they had been diagnosed with a learning disability. Students were also asked to report their cumulative academic grade point average (GPA) by selecting one of five predefined ranges: below 60, 60–70, 70–80, 80–90, or 90–100. This categorical measure was used to examine whether perceptions of online-course presences differed across self-reported achievement levels.

The second part assessed students' perceptions of the pedagogical components of the online course, focusing on the three types of presence - cognitive, social, and teaching presence, as defined by the ~~Community of Inquiry~~ COI framework (Garrison, Anderson & Archer, 2000). This section was based on the validated instrument developed by Arbaugh et al. (2008) and included 34 items rated on a five-point Likert scale. The third part of the questionnaire included 16 items addressing students' perceptions of learning outcomes in the online course. These items were adapted from the instruments of Rovai et al. (2009) and Alt (2017). In Alt's original questionnaire, the items were phrased to describe the pedagogical characteristics of the course, in the current study, they were modified to reflect students' perceived benefits and learning outcomes of the course. The final part of the questionnaire included five items designed to measure student satisfaction with online learning: *"I find online courses convenient," "Online courses can help me grow as a learner," "I am satisfied with online learning," "I would like to take online courses in the future,"* and *"I would recommend an online course to a friend"*. The overall reliability of this part of the questionnaire, which refers to student satisfaction with online learning, was $\alpha = 0.935$.

An Exploratory Factor Analysis (EFA) was conducted for the 34 questionnaire items, assessing students' perceptions of the pedagogical components of the online course, based on the three types of presence (cognitive, social, and teaching presence) as defined by the COI framework (Garrison, Anderson & Archer, 2000). No restriction on the number of extracted factors was made. The default criterion for factor retention was an eigenvalue greater than 1 (Eigenvalue Criterion). The analysis used the Principal Component extraction method with oblique (Oblimin) rotation, based on the assumption that the factors are correlated. The analysis yielded 34 items loading on three distinct factors (see Appendix 1), which together explained 66.2% of the total variance (KMO = 0.964, Bartlett's test < 0.001). Each item loaded on only one of the three factors, with a loading threshold set at 0.40. Table 1 presents the results of the factor analysis for the three identified factors: cognitive presence, social presence, and teaching presence.

Table 1: Factor Loadings - Exploratory Factor Analysis – Pedagogical Components

Item	Cognitive presence	Social presence	Teaching presence
Problems posed increased my interest in course issues	0.764		
Course activities piqued my curiosity	0.874		
I felt motivated to explore content related questions	0.813		
I utilized a variety of information sources to explore problems posed in this course	0.647		
Brainstorming and finding relevant information helped me resolve content related questions	0.570		

Item	Cognitive presence	Social presence	Teaching presence
Online discussions were valuable in helping me appreciate different perspectives	0.638		
Combining new information helped me answer questions raised in course activities	0.611		
Learning activities helped me construct explanations/solutions	0.696		
Reflection on course content and discussions helped me understand fundamental concepts in this class	0.616		
I can describe ways to test and apply the knowledge created in this course	0.832		
I have developed solutions to course problems that can be applied in practice	0.774		
I can apply the knowledge created in this course to my work or other non-class related activities	0.788		
Getting to know other course participants gave me a sense of belonging in the course		0.810	
I was able to form distinct impressions of some course participants		0.841	
Online or web-based communication is an excellent medium for social interaction		0.574	
I felt comfortable conversing through the online medium		0.491	
I felt comfortable participating in the course discussions		0.564	
I felt comfortable interacting with other course participants		0.795	
I felt comfortable disagreeing with other course participants while still maintaining a sense of trust		0.706	
I felt that my point of view was acknowledged by other course participants		0.718	
Online discussions help me to develop a sense of collaboration		0.520	
The instructor clearly communicated important course topics			0.626
The instructor clearly communicated important course goals			0.753
The instructor provided clear instructions on how to participate in course learning activities			0.753
The instructor clearly communicated important due dates/time frames for learning activities			0.719
The instructor was helpful in identifying areas of agreement and disagreement on course topics that helped me to learn			0.755
The instructor was helpful in guiding the class towards understanding course topics in a way that helped me clarify my thinking			0.683
The instructor helped to keep course participants engaged and participating in productive dialogue			0.667
The instructor helped keep the course participants on task in a way that helped me to learn			0.617
The instructor encouraged course participants to explore new concepts in this course			0.577

Item	Cognitive presence	Social presence	Teaching presence
Instructor actions reinforced the development of a sense of community among course participants			0.410
The instructor helped to focus discussion on relevant issues in a way that helped me to learn			0.620
The instructor provided feedback that helped me understand my strengths and weaknesses relative to the course's goals and objectives			0.506
The instructor provided feedback in a timely fashion			0.465

Cronbach's alpha reliability coefficients were calculated for each of the factors (categories) identified in this section of the questionnaire. The results indicated a reliability of $\alpha = 0.950$ for cognitive presence, $\alpha = 0.940$ for social presence, and $\alpha = 0.946$ for teaching presence. The overall reliability of this part of the questionnaire was $\alpha = 0.974$.

An Exploratory Factor Analysis (EFA) was conducted for the 16 items addressing students' perceptions of learning outcomes in the online course as well. No restriction on the number of extracted factors was made. The default criterion for factor retention was an eigenvalue greater than 1 (Eigenvalue Criterion). The analysis used the Principal Component extraction method with oblique (Oblimin) rotation, based on the assumption that the factors are correlated. The analysis yielded 15 items loading on three distinct factors (see Appendix 2), which together explained 73.82% of the total variance (KMO = 0.947, Bartlett's test < 0.001). Each item loaded on only one of the three factors, with a loading threshold set at 0.40. Table 2 presents the results of the factor analysis for the three identified factors: metacognitive learning outcomes, cognitive learning outcomes, and social learning outcomes.

Table 2: Factor Loadings - Exploratory Factor Analysis – Learning Outcomes

Item	Metacognitive outcomes	Cognitive outcomes	Social outcomes
I am able to organize the course materials in a logical way	0.680		
I am able to create a type of "study guide" that would help future students taking this course	0.943		
I am able to apply the skills I acquired in this course to other academic courses (both online and in-person)	0.856		
I am able to provide well-founded criticism of the texts used in the course	0.675		
I have expanded my skills in the subject area covered by the course	0.569		
I can demonstrate to others the skills I acquired in the subject	0.604		
I feel that I am able to think more effectively as a result of the course		0.739	
I learned how to solve real-world problems (theoretical or practical) as a result of the course		0.911	
I learned how to ask questions about complex problems as a result of the course		0.841	
I learned how to search for explanations to real-world problems as a result of the course		0.923	
In the framework of the course, I learned skills that help me investigate a topic that interests me		0.857	
In the framework of the course, I learned how to explore a selected topic in depth		0.706	
Overall, I believe I can achieve what is important to me as a result of the course		0.701	
As part of the course, I was given opportunities to work with other students and develop my teamwork skills			0.796

Item	Metacognitive outcomes	Cognitive outcomes	Social outcomes
As part of the course, I was given opportunities to engage in discussions with peers and improve my argumentation skills			0.720

Cronbach's alpha reliability coefficients were calculated for each of the factors (categories) identified in this section of the questionnaire. The results indicated a reliability of α = 0.911 for metacognitive outcomes, α = 0.828 for social outcomes, and α = 0.944 for cognitive outcomes. The overall reliability of this part of the questionnaire was α = 0.956.

2.2 Participants

The questionnaire was distributed among 578 students, of whom 316 responded (55%). Participants ranged in age from 20 to 63 years (M = 27.14, SD = 6.18, median = 26, valid age data were available for 310 students). Respondents were 58% women (182) and 42% men (134), 57% (179) majored in the social sciences and humanities, 25% (78) majored in education and teaching, and 18% (58) majored in science and technology, 73% (231) of the respondents were not diagnosed with learning disabilities, whereas 27% (85) were diagnosed. Of the respondents, 87% (276) were self-identified as Jewish, 10% (30) as Arab, and the remainder (10, 3%) did not answer the ethnicity question.

2.3 Ethical Considerations

The study received approval from the Ethics Committee of the academic institution. Participants were informed that their participation was voluntary and anonymous, and that they could withdraw from the study at any time without any consequences. The objectives of the study were clearly presented to the participants prior to data collection, and informed consent was obtained. It was explicitly stated that participation in the study would have no impact on their academic status or coursework. The questionnaire was distributed electronically via a link sent by departmental administrative staff through email communication with students.

3. Results

RQ 1: To what extent do students' perceptions of the three CoI presences (teaching, cognitive, and social) differ across student characteristics, including gender, ethnicity, faculty affiliation, learning-disability status, and self-reported academic achievement (GPA)?

To assess the normality of the dependent variables, Kolmogorov-Smirnov and Shapiro-Wilk tests were conducted for each variable separately. In all cases, the significance values (*p*) were below .05, indicating that the variables were not normally distributed. Therefore, non-parametric tests were used to address the research question. To examine differences in students' perceptions of the pedagogical design of online courses according to demographic characteristics, the Mann–Whitney U test was applied for dichotomous variables (gender, ethnicity, and learning disabilities), and the Kruskal–Wallis H test was used for the variables faculty (with three categories) and students' self-reported cumulative GPA range (categorized into five intervals).

The results revealed significant differences in students' perceptions of all types of presence based on gender (in favor of male students) and faculty (in favor of students from the Faculties of Humanities and Social Sciences, and Education). Significant differences were also found in teaching presence (in favor of students without learning disabilities) and in social presence (in favor of Arab students). In all cases, the effect sizes were relatively small. Table 3 presents the full results.

Table 3: Perceptions of the pedagogical characterization of the online course according to demographic variables

Variable		Teaching presence Mean rank	Social presence Mean rank	Cognitive presence Mean rank	Total Mean rank
Gender	**Female**	147.02	138.01	142.75	141.76
	Male	174.10	185.36	178.87	181.23

Variable		Teaching presence Mean rank	Social presence Mean rank	Cognitive presence Mean rank	Total Mean rank
	Mann–Whitney *U* test	U=10104, z=-2.607, p<.01, effect size=0.147	U=8464.5, z=-4.563, p<.01, effect size=0.257	U=9327.5, z=-3.479, p<.01, effect size=0.196	U=9148, z=-3.795, p<.01, effect size=0.214
Ethnicity	**Jews**	150.92	149.58	150.79	150.32
	Arabs	177.23	184.33	173.23	182.72
	Mann–Whitney U test	n.s.	U=3185, z=-2.052, p<.05, effect size=0.115	n.s.	n.s.
Learning disabilities	**Without diagnosis**	167.94	163.59	163.29	166.48
	With diagnosis	132.84	142.63	143.44	136.81
	Mann–Whitney *U* test	U=7636.5, z=-3.032, p<.01, effect size=0.170	n.s.	n.s.	U=7974, z=-2.560, p<.05, effect size=0.144
Faculty	**Social Sciences and Humanities**	175.90	175.58	173.50	177.15
	Education and Teaching	144.10	149.69	140.76	143.82
	Science and Technology	121.44	111.42	130.15	117.98
	Kruskal-Wallis H test	H(2)=18.12, p<.001	H(2)=22.41, p<.001	H(2)=13.40, p<.001	H(2)=21.00, p<.001
GPA	**Below 60**	198.94	213.00	199.81	210.38
	60-70	184.24	219.50	198.10	202.43
	70-80	156.80	155.89	147.17	145.43
	80-90	145.07	141.94	144.71	143.35
	90-100	167.61	162.69	169.19	166.44
	Kruskal-Wallis H test	n.s.	H(4)=16.65, p<.005	H(4)=10.74, p<.05	H(4)=12.06, p<.05

The GPA-based comparisons in Table 3 indicate that students with lower self-reported academic achievement tend to perceive stronger cognitive and social presence in online courses. In contrast, teaching presence appears to be perceived similarly across all GPA ranges. To better understand the direction of these differences, the mean ranks for each GPA group were examined. In all significant presence types—social and cognitive presence—students who reported lower cumulative GPA scores (especially those in the "< 60" and "61–70" ranges) showed higher mean ranks, indicating stronger perceptions of presence. Conversely, students in the "81–90" GPA range consistently had the lowest mean ranks, suggesting weaker perceptions of presence in the online learning environment. Post-hoc comparisons using Mann–Whitney U tests were conducted to further examine the significant differences identified in the Kruskal–Wallis tests regarding students' perceptions of cognitive and social presence across self-reported GPA ranges. For cognitive presence, students in the 61–70 GPA group reported significantly higher levels compared to those in the 71–80 and 81–90 GPA groups. Specifically, students in the 61–70 group had a higher mean rank (49.38) than those in the 71–80 group (Mean Rank = 33.10), U=302.50, Z=−2.895, p=.004, with a medium effect size (r=.337). Similarly, they reported higher cognitive presence than students in the 81–90 group (Mean Rank = 101.90 vs. 74.46), U=872.00, Z=−2.553, p=.011, with a small-to-medium effect size (r=.205).

For social presence, the 61–70 GPA group also perceived significantly higher levels than their higher-achieving peers. They scored higher than the 71–80 group (Mean Rank = 48.50 vs. 32.63), U=339.50, Z=−2.721, p=.007, with a medium effect size (r=.316), and higher than the 81–90 group (Mean Rank = 100.71 vs. 75.12), U=896.50, Z=−2.472, p=.013, with a small-to-medium effect size (r=.199). No other pairwise comparisons yielded statistically significant differences for either presence type.

RQ 2: What are the predictors of students' attitudes toward online learning, as examined through three conceptual models: (1) student demographic characteristics, (2) perceived three CoI presences, and (3) perceived learning outcomes (cognitive, metacognitive, and social)?

To identify the predictors of students' attitudes toward online courses, a hierarchical multiple regression analysis was conducted in three steps:

Model 1 – Student Demographic Characteristics: The model included five variables: gender, age, faculty, learning disability status, and ethnicity (Jewish/Arab). The model was significant, F(5, 289) = 4.731, p < .001 and explained 7.6% of the variance in attitudes (R^2 = 0.076). Among the predictors, gender (*β = -0.119, p = .044*), ethnicity (*β = 0.144, p = .014*), and faculty affiliation (*β = -0.133, p = .030*), were significant contributors. Female and Jewish students reported more positive attitudes.

Model 2 – Pedagogical Characteristics (COI Presences): The second model included the predictors from Model 1 and added pedagogical variables: teaching presence, social presence, and cognitive presence (Garrison, Anderson & Archer, 2000). The model was significant, F(8, 286) = 17.495, p < .001 and explained variance increased to 33% (R^2 = 0.329). Cognitive presence (*β = 0.338, p < .001*) was a significant contributor. Students who experienced greater cognitive engagement reported more positive attitudes. Teaching and social presence were not statistically significant when controlling for cognitive presence and demographics.

Model 3 – Perceived Learning Outcomes: The third model added variables related to perceived cognitive, metacognitive, and social learning outcomes. The final model was significant, F(11, 283) = 14.479, p < .001 and explained variance further increased to 36% (R^2 = 0.360). Among the predictors, perceived metacognitive learning outcomes (*β = 0.212, p = .033*) and perceived cognitive learning outcomes (*β = 0.205, p = .042*), were significant contributors. Greater self-reported development of metacognitive skills predicted more positive attitudes. Cognitive learning gains were also positively associated with student attitudes. None of the presences (teaching, social, or cognitive) remained significant in this model, indicating that their influence may be mediated through students' perceptions of actual learning outcomes. Table 4 presents the results of the hierarchical multiple regression analysis predicting students' attitudes toward online learning across the three conceptual models.

Table 4: Hierarchical multiple regression predicting attitudes toward online learning

Predictor	Model 1 B (SE)	β	p	Model 2 B (SE)	β	p	Model 3 B (SE)	β	p
Constant	3.683 (1.159)	-	.002	0.689 (1.037)	-	.507	0.862 (1.037)	-	.406
Gender	-0.065 (0.032)	-.119	.044	-0.015 (0.028)	-.027	.595	-0.010 (0.028)	-.019	.713
Age	-0.002 (0.010)	-.014	.809	-0.004 (0.009)	-.024	.619	-0.006 (0.009)	-.036	.453
Faculty affiliation	-0.184 (0.084)	-.133	.030	-0.027 (0.074)	-.019	.719	-0.013 (0.074)	-.009	.861
Learning disability status	0.267 (0.141)	.110	.059	0.183 (0.121)	.076	.133	0.115 (0.120)	.047	.340
Ethnicity (Jewish/Arab)	0.225 (0.091)	.144	.014	0.100 (0.079)	.064	.207	0.093 (0.077)	.060	.231
Teaching presence	-	-	-	0.147 (0.101)	.125	.147	0.048 (0.102)	.041	.637
Social presence	-	-	-	0.102 (0.074)	.108	.169	0.130 (0.085)	.138	.128
Cognitive presence	-	-	-	0.375 (0.108)	.338	<.001	0.104 (0.129)	.093	.422

Predictor	Model 1 B (SE)	β	p	Model 2 B (SE)	β	p	Model 3 B (SE)	β	p
Cognitive learning outcomes	-	-	-	-	-	-	0.209 (0.103)	.205	.042
Social learning outcomes	-	-	-	-	-	-	-0.069 (0.066)	-.080	.296
Metacognitive learning outcomes	-	-	-	-	-	-	0.230 (0.108)	.212	.033
Model statistics									
R	.275			.573			.600		
R^2	.076			.329			.360		
Adjusted R^2	.060			.310			.335		
ΔR^2	.076			.253			.032		
F for model	4.731***			17.495***			14.479***		
ΔF	4.731***			35.911***			4.651**		

**p < .01; ***p < .001.

4. Discussion

This study examined how pedagogical characteristics of online learning environments, conceptualized through the CoI presences, relate to students' attitudes toward online learning, and whether these associations are shaped by learner characteristics and students' perceived learning outcomes. Data were collected from undergraduate students enrolled in fully online courses, representing a range of disciplines. Using a quantitative survey design, teaching, social, and cognitive presence were assessed, alongside perceived learning outcomes (cognitive, metacognitive, and social) and overall attitudes toward online learning. The discussion below highlights the main findings, situates them within recent scholarship on online course quality, and outlines implications for equitable online course design and institutional policy in higher education.

Regarding group differences (RQ1), students' perceptions of CoI presences varied across several student characteristics, however, these differences should be interpreted with caution because most effect sizes were small. Rather than indicating large or stable divides between student groups, the findings point to modest variations in how different learners may experience the online learning environment under particular disciplinary and instructional conditions. In this sense, learner characteristics appear to shape online-course perceptions in limited and context-sensitive ways, while much of the overall experience may still be driven by course-level pedagogical design.

The disciplinary differences observed in the present study, where students from education and the humanities/social sciences reported stronger perceived presences than students from science and technology, are consistent with growing evidence that CoI-related processes are enacted in discipline-sensitive ways rather than as a uniform "one-size-fits-all" configuration. Research on online course design suggests that instructors' disciplinary aims and pedagogical traditions shape which CoI indicators are foregrounded in learning activities and assessment, which in turn may influence how visible teaching, social, and cognitive presence becomes to students (Richardson et al., 2025). Complementary qualitative work also indicates that while teaching presence is typically valued across fields, the salience and perceived necessity of social presence can vary by discipline, particularly in hard-applied contexts where interaction may be less central to how learning is organized and evaluated (Lim & Richardson, 2022). In line with these arguments, large-scale evidence further points to systematic disciplinary patterns in online learning experiences and presence-related dimensions, often showing comparatively weaker indicators in science/engineering relative to humanities, and highlighting that certain pedagogical components (e.g., facilitating discourse) may be especially consequential in STEM contexts (Li & Wang, 2024). Taken together, these findings suggest that the differences detected in our sample may reflect variation in how online courses operationalize interaction, facilitation, and inquiry across disciplines, underscoring the need for discipline-sensitive CoI-informed design and faculty development rather than assuming that the same presence profile will emerge, or be equally functional, across all fields. At the same time, because the effect sizes were small, these disciplinary differences are better understood as modest tendencies in students' perceptions rather than as evidence of sharply distinct online learning experiences across fields.

Gender-related differences in perceptions of online learning conditions have been reported, although findings are not always directional or consistent. Li and Wang (2024) highlight this heterogeneity by showing that specific dimensions of teaching presence relate differently to male and female students' online learning experiences, suggesting that gender effects may depend on which pedagogical processes are most salient in a given context. In the present study, male students reported higher levels of teaching, social, and cognitive presence than female students, albeit with relatively small effect sizes. This pattern may reflect differences in how students engage with and interpret online interaction and instructional support, or it may be partially shaped by contextual factors such as disciplinary enrollment patterns and course design features. Future research should therefore examine whether gender differences persist after accounting for course-level variation and whether they are mediated by participation behaviors or perceived learning gains. Given the relatively small effect sizes, this pattern should be interpreted cautiously and not as evidence of a strong or uniform gender-based divide in online learning experience.

In the current study, students' disability status was associated with a small but significant difference in teaching presence perceptions. Students without a diagnosed learning disability reported higher teaching presence, whereas no significant differences emerged for cognitive or social presence. At the same time, the effect size was small, suggesting that this result should be interpreted as a limited area of difference rather than as evidence of a broad disparity in online learning experience. This pattern is consistent with the view that, for students with special needs, the "pedagogical work" of online teaching is often experienced most strongly through instructor-led design and guidance, namely, how clearly the course is structured, how expectations are communicated, and how feedback and support are provided throughout the learning process (Öhrstedt et al., 2024). From a CoI perspective, these elements map directly onto teaching presence as the design, facilitation, and direction of learning processes, suggesting that when instructional organization and instructor support are less visible or less accessible, students with learning disabilities may evaluate teaching presence more critically even if peer interaction and cognitive engagement are perceived similarly. Practically, this finding reinforces the importance of inclusive online course design: strengthening transparency (instructions, timelines, criteria), accessibility of materials, and timely, actionable instructor feedback may be particularly consequential for reducing barriers and supporting equitable learning experiences among students with learning disabilities.

Our finding is consistent with recent evidence suggesting that, although many faculty members are genuinely motivated to support online students with disabilities, they often feel underprepared and constrained by limited time, compensation, and institutional training, conditions that can reduce the extent to which teaching presence is enacted through proactive design, facilitation, and guidance (Lehan et al., 2025). These insights offer a plausible explanation for why teaching presence, more than social or cognitive presence, may be the dimension most sensitive to disability-related differences. Students who need clearer structure, predictable communication, and timely, targeted feedback may experience variability in these "high-touch" instructional processes more strongly when inclusive design is not systematically embedded.

Ethnicity-related differences in perceived social presence should be interpreted cautiously given the small effect size, yet they may point to how online environments can reshape participation conditions for students from minoritized groups within hegemonic academic spaces. In the present study, students from the Arab minority reported higher social presence, which aligns with qualitative evidence suggesting that synchronous online learning (e.g., via Zoom) can reduce fear, self-censorship, and linguistic inhibition that sometimes characterize face-to-face participation in majority-group settings, thereby enabling more confident self-expression and interaction (Halabi, 2023). At the same time, research on Jewish–Arab undergraduates during COVID-19 indicates that minority students may concurrently experience greater academic challenges alongside lower perceived social support and institutional trust, underscoring that "feeling socially present" in online learning does not necessarily imply equal access to supportive resources or a uniformly positive overall experience (Ismail et al., 2023). Taken together, these findings suggest that social presence may be shaped not only by course design but also by perceived psychological safety, power relations, and language comfort, mechanisms that may similarly affect other underrepresented or minoritized groups across higher education systems.

Interestingly, perceptions of presence varied across self-reported cumulative GPA categories. Students in the lower GPA ranges tended to report stronger cognitive and social presence, whereas teaching presence was perceived more similarly across achievement levels. This pattern is noteworthy because the broader CoI literature typically reports positive associations between presences and learning outcomes, including actual performance and perceived learning (Martin et al., 2022, Rockinson-Szapkiw et al., 2016), and recent evidence in blended contexts likewise links teaching, social, and cognitive presence to learning performance (Li & Ye, 2025). One interpretation is a compensatory salience mechanism. Students who struggle academically may rely more on interaction, clarification, and community resources, and thus become more attuned to, or place greater value on, the social and cognitive dimensions of the online learning environment when these are present. This aligns with research framing academic help-seeking as a central self-regulated learning strategy in online courses and linking it to achievement-related outcomes (Fong et al., 2023, Won et al., 2024), as well as reviews showing that SRL processes are consistently related to academic achievement in online and blended learning (Xu, Shi & Zhao, 2023). At the same time, the observed pattern may reflect contextual confounding (e.g., disciplinary clustering, course design differences, or instructor practices that co-vary with both GPA distributions and perceived presence) and should therefore be examined in future work using multilevel models nested within courses and instructors, complemented by objective achievement and behavioral trace indicators of engagement and help-seeking.

The RQ1 findings suggest that although learner characteristics were associated with some differences in perceived presence, these differences were generally small in magnitude and context-dependent. Rather than indicating large or stable divides between student groups, the results point to modest variations in how different learners may experience the online learning environment under particular disciplinary and instructional conditions. In practical terms, these findings suggest that while student characteristics should not be ignored, online learning quality is unlikely to be determined primarily by fixed group membership alone.

In addressing Research Question 2, the hierarchical models suggest that learner demographics played a relatively limited role in explaining students' attitudes toward online learning. Although some demographic predictors reached statistical significance, Model 1 accounted for only a modest proportion of the variance. By contrast, the substantial increase in explained variance after adding CoI presences, and the further contribution of perceived learning outcomes, indicates that students' attitudes are shaped less by who they are demographically and more by how they experience the pedagogical quality and learning value of the course. More specifically, the findings suggest that cognitive presence may be important not simply as a favorable feature of online learning in itself, but because it is linked to students' sense that meaningful learning has taken place. Once perceived cognitive and metacognitive gains were introduced, the direct contribution of CoI presences was reduced, which is consistent with the interpretation that students evaluate online learning primarily through the extent to which it produces deeper understanding and reflective growth. Because the study is cross-sectional, this pattern should not be interpreted causally, however, it does support a more proximal account of student attitudes in which perceived learning gains may help explain why certain pedagogical conditions matter more than others.

This pattern aligns with accumulated evidence that cognitive presence is strongly associated with blended contexts (Martin et al., 2022), and with recent meta-analytic findings showing large correlations between cognitive presence and both perceived learning and satisfaction, alongside meaningful heterogeneity across disciplinary areas and course features (Yang, Wang & Zhao, 2025). At the same time, prior work indicates that perceived learning outcomes are a proximal determinant of satisfaction and related judgments about online learning quality (e.g., Su & Guo, 2021), suggesting that CoI presence may matter to students' attitudes to the extent that they recognize it as cognitively and metacognitively beneficial.

This study has several limitations that should be considered when interpreting the findings. First, the cross-sectional design does not allow causal conclusions about the relationships among CoI presences, perceived learning outcomes, and attitudes. Second, all variables were measured through self-report, which may introduce shared-method variance. This limitation is especially relevant for the self-reported cumulative GPA measure, as students may over-report their academic achievement due to social desirability or recall bias, because GPA was analyzed in broad intervals, even modest inaccuracies may have affected group classification and should therefore be considered when interpreting the GPA-related findings. Third, although the response rate was acceptable, non-response bias cannot be ruled out, as no formal comparison was conducted between respondents and the broader invited student cohort. Fourth, the sample was drawn from a single university, and generalizability to other institutional and national contexts should be made cautiously. Finally, students were nested within multiple online courses and instructors. Because the analyses did not explicitly model this clustering structure, some variance attributable to course-level design, instructional practices, and assessment conditions may not have been adequately separated from learner-level variance. As a result, the standard errors of some individual-level estimates may have been biased, and the magnitude of learner-level effects should therefore be interpreted with appropriate caution.

Despite these limitations, the findings suggest that improving students' attitudes toward online learning may depend less on fixed demographic differences and more on whether course design makes meaningful learning gains visible to students. The relatively modest contribution of demographic variables, compared with the stronger explanatory role of CoI presences and especially perceived cognitive and metacognitive gains, suggests that online-course quality should be designed primarily around pedagogical processes that support deeper understanding, reflective learning, and a clear sense of progress. In practical terms, this means prioritizing inquiry-oriented and application-based tasks that strengthen cognitive presence, embedding metacognitive scaffolds such as planning prompts, self-monitoring activities, and guided reflection, and ensuring transparent course organization and timely feedback, particularly as a targeted support for students who may be more sensitive to variability in teaching presence, including those with learning disabilities. At the institutional level, the findings support shared quality standards and faculty development that focus not only on technical delivery, but on designing online courses in ways that help students experience learning as cognitively meaningful and metacognitively productive.

Beyond these practical implications, the study also makes a theoretical contribution by extending CoI-based research by suggesting an integrative view in which perceived learning outcomes constitute a central link in understanding students' attitudes toward online learning, alongside the three presences. By combining process indicators (CoI presences), multidimensional perceived gains (cognitive, metacognitive, and social), and learner characteristics, the study contributes to more nuanced models of online learning quality in higher education.

Future research can build on these findings by testing formal mediation models (e.g., SEM) to examine whether perceived learning outcomes account for the association between CoI presences and attitudes. Multi-level designs that nest students within courses and instructors would help disentangle course-level pedagogical effects from learner-level differences. In addition, incorporating objective indicators, such as LMS trace data, course grades, and retention, could complement self-reports and clarify how perceived quality aligns with behavioral engagement and academic performance. Finally, comparative studies across fully online, blended, and synchronous formats would strengthen generalizability and identify which pedagogical conditions are most robust across delivery modes.

In conclusion, this study suggests that students' attitudes toward online learning are shaped less by fixed demographic differences than by how they experience the pedagogical quality and learning value of online courses. Although some subgroup differences in perceived presence were statistically significant, their generally small effect sizes indicate that these patterns should be interpreted as modest and context-dependent rather than as strong divides between student groups. By contrast, the findings point more clearly to the importance of pedagogical processes and, in particular, to the central role of perceived cognitive and metacognitive gains in shaping students' evaluations of online learning. In this sense, the study contributes to the field by advancing a more integrative account of online learning quality, one that connects CoI presences, perceived learning outcomes, and learner characteristics, while highlighting that students' judgments of online learning may depend primarily on whether they experience it as meaningful, cognitively engaging, and supportive of reflective growth.

AI Tool Use Statement: AI-assisted tools were used during manuscript preparation to support language editing (improving clarity and phrasing) and to help identify potentially relevant literature. All substantive scholarly decisions, interpretation of findings, and final wording were made by the authors.

Ethics approval: The research was approved by the Ethics Committee of the institution. Informed consent was obtained from all subjects involved in the study. The manuscript does not contain any individual person's data, therefore, consent to publish is not applicable.

Competing interests: The authors report that there are no competing interests, financial or non-financial, to declare.

Funding: Not applicable

Data Availability Statement: Data available on request from the authors.

References

Akar, S. G. M. (2024). *Why do students disengage from online courses? The Internet and Higher Education, 62,* 100948. https://doi.org/10.1016/j.iheduc.2024.100948

Alt, D. (2017). Constructivist learning and openness to diversity and challenge in higher education environments. *Learning Environments Research, 20*(1), 99-119. https://doi.org/10.1007/s10984-016-9223-8

Andrade, M. S., & Alden-Rivers, B. (2019). Developing a framework for sustainable growth of flexible learning opportunities. *Higher Education Pedagogies*, *4*(1), 1-16. https://doi.org/10.1080/23752696.2018.1564879

Arbaugh, J.B., Cleveland-Innes, M., Diaz, S.R., Garrison, D.R., Ice, P., Richardson, & Swan, K.P. (2008). Developing a community of inquiry instrument: Testing a measure of the Community of Inquiry framework using a multi-institutional sample. *The Internet and Higher Education, 11*(3-4), 133-136. https://doi.org/10.1016/j.iheduc.2008.06.003

Bach, A., & Thiel, F. (2024). Collaborative online learning in higher education - quality of digital interaction and associations with individual and group-related factors. *Frontiers in Education*, 9(1356271). https://doi.org/10.3389/feduc.2024.1356271

Baxter, J. A., & Haycock, J. (2014). International review of research in open and distance learning. *The International Review of Research in Open and Distributed Learning, 15*(1), 20-40. https://www.jstor.org/stable/44430168

Broadbent, J., & Poon, W. L. (2015). Self-regulated learning strategies & academic achievement in online higher education learning environments: A systematic review. *The Internet and Higher Education, 27*, 1–13. https://doi.org/10.1016/j.iheduc.2015.04.007

Castro, M. D. B., & Tumibay, G. M. (2021). A literature review: Efficacy of online learning courses for higher education institution using meta-analysis. *Education and Information Technologies, 26*(2), 1367–1385. https://doi.org/10.1007/s10639-019-10027-z

Collins, A., & Halverson, R. (2018). Rethinking *education in the age of technology: The digital revolution and schooling in America*. Teachers College Press.

Creswell, J.W., & Creswell, J.D. (2018). *Research design – qualitative, quantitative and mixed methods approaches* (5th edition). Sage.

Dietrich, N. (2020). Attempts, successes, and failures of distance learning in the time of COVID-19. *Journal of Chemical Education, 97(*9), 2448–2457. https://doi.org/10.1016/j.ssaho.2024.100875

Dumford, A. D., & Miller, A. L. (2018). Online learning in higher education: Exploring advantages and disadvantages for engagement. *Journal of Computing in Higher Education, 30*(3), 452-465. https://doi.org/10.1007/s12528-018-9179-z

Fewella, L. N. (2023). Impact of COVID-19 on distance learning practical design courses. *International Journal of Technology and Design Education*, *33*(5), 1703-1726. https://doi.org/10.1007/s10798-023-09806-0

Fong, C. J., Gonzales, C., Hill-Troglin Cox, C., & Shinn, H. B. (2023). Academic help-seeking and achievement of postsecondary students: A meta-analytic investigation. *Journal of Educational Psychology, 115*(1), 1–21. https://doi.org/10.1037/edu0000725

Garrison, D.R., Anderson, T., & Archer, W. (2000). Critical inquiry in a text-based environment: Computer conferencing in higher education. *The Internet and Higher Education, 2*, 87-105. https://doi.org/10.1016/S1096-7516(00)00016-6

Getenet, S., Cantle, R., Redmond, P., & Albion, P. (2024). Students' digital technology attitude, literacy and self-efficacy and their effect on online learning engagement. *International Journal of Educational Technology in Higher Education, 21*(1), 3. https://doi.org/10.1186/S41239-023-00437-Y

Gillis, A., & Krull, L.M. (2020). COVID-19 remote learning transition in spring 2020: Class structures, student perceptions, and inequality in college courses. *Teaching Sociology, 48*(4), 283-299. https://doi.org/10.1177/0092055X20954263

Guo, P., Jeyaraj, J. J., & Razali, A. B. (2025). A systematic review on the application of community of inquiry framework in teaching of English as a second/foreign language. SAGE Open, 15(3), 21582440251357697. https://doi.org/10.1177/21582440251357697

Halabi, R. (2023). Learning via Zoom during the Covid-19 pandemic: Benefits for Arab students in Hebrew academic institutes in Israel. *Multicultural Education Review, 15*(1), 28-41. https://doi.org/10.1080/2005615X.2023.2193925

Ismail, A., Schiff, M., Pat-Horenczyk, R., & Benbenishty, R. (2023). Perceived academic challenges of Jewish and Arab undergraduates during the first wave of COVID-19. *International Journal of Psychology, 58*(1), 7-15. https://doi.org/10.1002/ijop.12873

Jiang, L., Zhou, N., & Yang, Y. (2024). Student motivation and engagement in online language learning using virtual classrooms: Interrelationships with support, attitude and learner readiness. *Education and Information Technologies, 29*(13), 17119–17143. https://doi.org/10.1007/s10639-024-12514-4

Jones, I., Pauli, R., Brewer, G., & Peat, J. (2025, June 20). Is flexibility in higher education just a matter of convenience? A critical literature review. *Quality Assurance Agency for Higher Education*. https://www.qaa.ac.uk/docs/qaa/members/is-flexibility-in-higher-education-just-a-matter-of-convenience---a-critical-literature-review.pdf

Lai, J. W., & Bower, M. (2019). How is the use of technology in education evaluated? A systematic review. *Computers & Education, 133*, 27-42. https://doi.org/10.1016/j.compedu.2019.01.010

Lehan, T., Murphy, Z., Dolan, V., & St. Louis, L. (2025). Faculty members' perceptions and experiences of working with online students with disabilities. *Open Praxis*, *17*(3), 529-543. https://doi.org/10.55982/openpraxis.17.3.859

Li, W., & Wang, W. (2024). The impact of teaching presence on students' online learning experience: Evidence from 334 Chinese universities during the pandemic. *Frontiers in Psychology*, *15*, 1291341. https://doi.org/10.3389/fpsyg.2024.1291341

Li, X., & Ye, Y. (2025). Mediating role of online academic emotions between online presence and learning performance in blended learning environments. *Scientific Reports, 15*, 29875. https://doi.org/10.1038/s41598-025-15947-0

Lim, J., & Richardson, J. C. (2022). Considering how disciplinary differences matter for successful online learning through the Community of Inquiry lens. *Computers & Education*, *187*, 104551. https://doi.org/10.1016/j.compedu.2022.104551

Martin, F., Wu, T., Wan, L., & Xie, K. (2022). A meta-analysis of the Community of Inquiry presences and learning outcomes in online and blended learning environments. *Online Learning, 26*(1), 325–359. https://doi.org/10.24059/olj.v26i1.2604

McGaughey, F., Watermeyer, R., Shankar, K., Suri, V. R., Knight, C., Crick, T., ... & Chung, R. (2022). 'This can't be the new norm': academics' perspectives on the COVID-19 crisis for the Australian university sector. *Higher Education Research & Development, 41*(7), 2231–2246. 10.1080/07294360.2021.1973384

McIntyre, K. P., Kashi, K., Ringer, A. M., Ferrer, J., Ren, B., Chaw, J., Masuch, S., & Saville, K. (2023). Understanding student intentions to take online courses: A theory-based understanding. *Education and Information Technologies, 28*, 16067–16093. https://doi.org/10.1007/s10639-023-11823-4

Muljana, P. S., & Luo, T. (2019). Factors contributing to student retention in online learning and recommended strategies for improvement: A systematic literature review. *Journal of Information Technology Education: Research, 18,* 19–57. https://doi.org/10.28945/4182

Müller, C., & Mildenberger, T. (2021). Facilitating flexible learning by replacing classroom time with an online learning environment: A systematic review of blended learning in higher education. *Educational Research Review, 34*, 100394. https://doi.org/10.1016/j.edurev.2021.100394

Nortvig, A.M., Petersen, A.K., & Balle, S. H. (2018). A literature review of the factors influencing e-learning and blended learning in relation to learning outcome, student satisfaction and engagement. *Electronic Journal of e-Learning*, *16*(1), 46-55. https://files.eric.ed.gov/fulltext/EJ1175336.pdf

Öhrstedt, M., Käck, A., Reierstam, H., & Ghilagaber, G. (2024). Studying online with special needs: A student perspective. *Journal of Research in Special Educational Needs, 24*(3), 771-785. https://doi.org/10.1111/1471-3802.12670

Olsson, M., Mozelius, P., & Collin, J. (2016). Visualisation and gamification of e-learning and programming education. *Electronic Journal of e-Learning, 13*(6), 441-454.

Paris, J., Lakhal, S., & Mukamurera, J. (2025). Relationships between presence, satisfaction, and persistence in higher education online courses. *Educational Technology & Society, 28*(1), 178–193. https://doi.org/10.30191/ETS.202501_28(1).RP10

Paulsen, L., & Lindsay, E. (2024). Learning analytics dashboards are increasingly becoming about learning and not just analytics-A systematic review. *Education and Information Technologies*, *29*(11), 14279-14308. https://doi.org/10.1007/s10639-023-12401-4

Reyes, E. R., Cuenca, A. A., & Martínez, A. I. M. (2025). Recommendations for designing collaborative activities in online higher education: a systematic review. *Journal of Technology and Science Education, 15*(1), 4-17. https://doi.org/10.3926/jotse.2818

Richardson, J. C., Koehler, A. A., Tagare, D., Urena-Rodriguez, L., Xu, Q., Zhang, Z., ... & Duha, M. S. U. (2025). Community of Inquiry design decisions across disciplines. *Journal of Computing in Higher Education*, *37*(3), 1119-1160. https://doi.org/10.1007/s12528-024-09417-1

Rockinson-Szapkiw, A. J., Wendt, J., Wighting, M., & Nisbet, D. (2016). The predictive relationship among the Community of Inquiry framework, perceived learning, and online graduate students' course grades in online synchronous and asynchronous courses. *The International Review of Research in Open and Distributed Learning, 17*(3), 18-35. https://doi.org/10.19173/irrodl.v17i3.2203 Copied

Rodrigues, H., Almeida, F., Figueiredo, V., & Lopes, S. L. (2019). Tracking e-learning through published papers: A systematic review. *Computers & Education, 136*, 87-98. https://doi.org/10.1016/j.compedu.2019.03.007

Rovai, A.P., Wighting, M.J., Baker, J.D., & Grooms, L.D. (2009). Development of an instrument to measure perceived cognitive, affective, and psychomotor learning in traditional and virtual classroom higher education settings. *The Internet and Higher Education, 12*(1),7-13. https://doi.org/10.1016/j.iheduc.2008.10.002

Saghafi, M.R., Franz, J., & Crowther, P.H. (2014). A holistic model for blended learning. *Journal of Interactive Learning Research, 25*(4), 531-549.

Sasson, I., & Yehuda, I. (2025). Cognitive demands and course pedagogical design: Understanding success in online learning environments. Accepted to *Online Learning.* https://doi.org/10.24059/olj.v30i2.4936

Seifert, T., & Bar-Tal, S. (2023). Student-teachers' sense of belonging in collaborative online learning. *Education and Information Technologies*, *28*(7), 7797-7826. https://doi.org/10.1007/s10639-022-11498-3

Simpson, A. (2016). Designing pedagogic strategies for dialogic learning in higher education. *Technology, Pedagogy and Education, 25*(2), 135–151. https://doi.org/10.1080/1475939X.2015.1038580

Steyn, A. A., van Slyke, C., Dick, G., Twinomurinzi, H., & Amusa, L. B. (2024). Student intentions to continue with distance learning post-COVID: An empirical analysis. *PLOS ONE, 19*(1), e0293065. https://doi.org/10.1371/journal.pone.0293065

Su, C. Y., & Guo, Y. (2021). Factors impacting university students' online learning experiences during the COVID-19 epidemic. *Journal of Computer Assisted Learning, 37*(6), 1578-1590. https://doi.org/10.1111/jcal.12555

Sun, A., & Chen, X. (2016). Online education and its effective practice: A research review. *Journal of Information Technology Education: Research, 15*, 157–190. https://doi.org/10.28945/3502

van Dorresteijn, C., Fajardo-Tovar, D., Pareja Roblin, N., Cornelissen, F., Meij, M., Voogt, J., & Volman, M. (2025). What factors contribute to effective online higher education? A meta-review. *Technology, Knowledge and Learning*, *30*(1), 171-202. https://doi.org/10.1007/s10758-024-09750-5

Won, S., & Chang, Y. (2024). Antecedents and consequences of academic help-seeking in online STEM courses. *Frontiers in Psychology, 15*, 1438299. https://doi.org/10.3389/fpsyg.2024.1438299

Xie, X., Siau, K., & Nah, F. F.H. (2020). COVID-19 pandemic – Online education in the new normal and the next normal. *Journal of Information Technology Case and Application, 22*(3), 175-187. https://doi.org/10.1080/15228053.2020.1824884

Xu, D., & Jaggars, S. (2013). *Adaptability to online learning: Differences across types of students and academic subject areas* (CCRC Working Paper). New York, NY: Teachers College, Columbia University. Retrieved from http://ccrc.tc.columbia.edu/publications/adaptability-to-online-learning.html.

Xu, R., Shi, R., & Zhao, P. (2023). Self-regulated learning and academic achievement in blended learning: A scoping review. *Frontiers in Psychology, 14*, 1154780. https://doi.org/10.3389/fpsyg.2023.1154780

Yang, D., Wang, S., & Zhao, L. (2025). Relationships between cognitive presence and students' learning outcomes in online higher education: A meta-analysis. *Distance Education*, *46*(4), 669-690. https://doi.org/10.1007/s12528-024-09417-1

Appendix 1

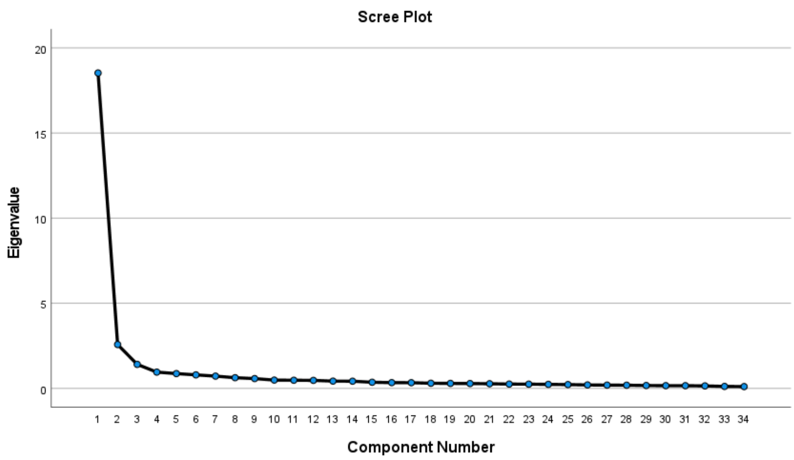


Figure 1: Scree plot for the exploratory factor analysis of students' perceptions of the three types of presence (cognitive, social, and teaching presence) as defined by the Community of Inquiry framework

Appendix 2

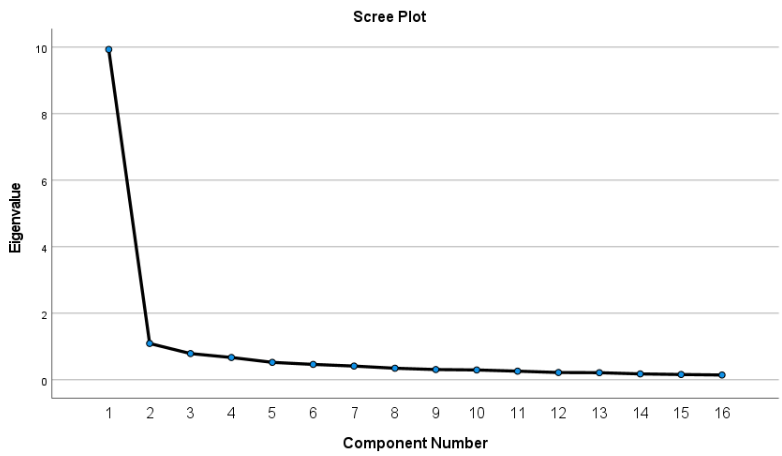


Figure 2: Scree plot for the exploratory factor analysis of students' perceptions of learning outcomes in the online course

Transforming the Ethnochemistry Technology Lecture Model: Integrating Digital Literacy and Character Education

Florida Doloksaribu, Lusia Narsia Amsad and Wigati Yektiningtias
Cenderawasih University, Indonesia

floridadoloks@gmail.com
lusianarsiaams ad@gmail.com
wigati_y@yahoo.com

https://doi.org/10.34190/ejel.24.3.4483

Abstract: Papua, Indonesia, has a rich ethnochemical heritage, including plant-based dyeing and starch processing, that is rarely represented in tertiary chemistry curricula. At the same time, uneven internet connectivity constrains students' opportunities to develop robust digital literacy. These dual challenges highlight a need for e-learning models that are culturally grounded, technologically adaptive, and pedagogically transformative. This study redesigned an Ethnochemistry Technology lecture into a low-bandwidth, culturally embedded, and virtue-infused e-learning model to improve students' digital literacy and character outcomes measurably. Using a semester-long Design-Based Research approach at a public university in Jayapura, the course integrated Papuan case vignettes, flipped and HyFlex delivery, downloadable H5P interactives, short micro-lectures, an augmented-reality dye-extraction lab with offline alternatives, communal reflection journals, peer mentoring, and a service-learning partnership with a village craft cooperative. Cultural consultants, including community elders and a dye artisan, ensured epistemic authenticity. Learning tasks were aligned with Kurikulum Merdeka virtues, respect for local wisdom, collaboration, and environmental stewardship, and mapped to UNESCO digital-literacy domains. Findings indicate substantial and educationally meaningful gains in digital literacy and character mastery. Students demonstrated great improvement across all digital-literacy domains, alongside marked increases in demonstrated respect for local wisdom, collaborative engagement, and environmental responsibility. Engagement indicators also rose consistently throughout the semester, particularly during immersive and community-linked activities. Qualitative feedback revealed heightened confidence in evaluating digital sources and a deeper recognition of chemistry within local cultural practices. Beyond local impact, this study contributes to e-learning practice by presenting a scalable blueprint for culturally responsive, low-bandwidth digital pedagogy in resource-constrained contexts. It advances the field of e-learning in Papua by validating indigenous knowledge as a central epistemic resource in technology-enhanced higher education. The model demonstrates that digital transformation need not marginalize local culture; instead, it can amplify it. Future research should explore multi-campus replication, long-term retention effects, and adaptation for additional Indigenous language communities.

Keywords: Ethnochemistry, Digital literacy, Character education, Flipped learning, Design-based research

1. Introduction

Authors in Indonesia repeatedly note that integration of local wisdom into chemistry curricula, learning tools, and teaching materials remains rare, creating a persistent implementation gap despite strong conceptual justification. Ethnochemistry research consistently argues that chemistry instruction becomes more meaningful when chemical concepts are grounded in learners' lived cultural contexts and "home-cultures," thereby improving engagement and conceptual accessibility (Irfandi, 2022). This gap is also documented in Indonesia, where ethnochemistry is positioned as a means to legitimize indigenous cultural knowledge within formal chemistry education and to connect classroom chemistry to culturally situated practices (Wahyudiati and Qurniati, 2023).

Within the existing ethnochemistry literature, "digital mediation" has mainly been operationalized as technology-based learning media rather than as a systematically engineered lecture model that simultaneously measures digital literacy competencies and character outcomes. Nevertheless, multiple studies demonstrate that ethnochemistry content can be embedded in web-based or multimedia formats, such as Weebly-based websites (Irnawati and Rahmawan, 2024), and microblog-based learning resources (Anggreni, Hadiarti and Fadhilah, 2023). These studies collectively provide evidence that ethnochemistry can be designed for digitally mediated delivery. Still, they do not, as a body, establish an integrated higher-education lecture architecture that jointly targets (i) digital literacy as an explicit competence, (ii) indigenous epistemic legitimacy, and (iii) measurable character outcomes, reinforcing the fragmentation highlighted in the task (Leny Heliawati *et al.*, 2022). Incorporating Papua's ethnochemical heritage into STEM education and addressing digital literacy gaps offer significant opportunities to enhance the relevance and effectiveness of education in the region.

Cite this article: Doloksaribu, F., Narsia Amsad, L. and Yektiningtias, W. 2026. "Transforming the Ethnochemistry Technology Lecture Model: Integrating Digital Literacy and Character Education", *Electronic Journal of e-Learning*, 24(3), pp 65-78, https://doi.org/10.34190/ejel.24.3.4483

Furthermore, aligning educational practices with character education mandates can ensure that students emerge as informed, responsible, and culturally aware individuals who can contribute positively to their communities.

Current ethnochemistry technology courses at the host public university in Jayapura still follow a teacher-centred slide-and-chalk routine. Formative evaluations in 2024 revealed that fewer than 40 % of students could (i) describe a locally sourced dye-extraction pathway or (ii) pass a basic digital-information verification task. Moreover, observational rubrics found limited evidence of the Merdeka virtues during laboratory teamwork.

1.1 Research Question

Against this background, this study asks the following questions:

RQ1: How can a culturally grounded, digitally mediated lecture model be engineered to lift students' digital-literacy competencies while legitimising Papuan ethnochemistry?

RQ2: To what extent does that model foster measurable growth in the targeted character values of respect, collaboration, and environmental stewardship?

The remainder of this article proceeds as follows. Section 2 synthesises literature on ethnochemistry pedagogy, technology-enhanced lecture models, digital-literacy frameworks for developing regions, and character-education theory, culminating in a visual conceptual model. Section 3 details the Papuan context and needs analysis that informed the design specification. Section 4 outlines the mixed-methods Design-Based Research protocol, instruments, and ethical safeguards. Section 5 documents the iterative development of the Transformative Ethnochemistry Lecture Model, including interactive H5P content and augmented-reality dye labs co-facilitated by community elders. Section 6 reports quantitative and qualitative results; Section 7 interprets those findings in light of prior work and identifies limitations. Section 8 concludes with practical recommendations and directions for multi-campus replication.

By foregrounding ethnochemical heritage, scaffolding digital literacies, and embedding *Merdeka*-aligned virtues in every task, the proposed model seeks to demonstrate that chemistry education in Papua can be simultaneously scientific, digital, and deeply human.

2. Literature Review

2.1 Ethnochemistry in Tertiary Chemistry Education

Ethnochemistry, as a subfield of ethnoscience, is commonly defined as culturally specific knowledge and practices that can be interpreted through chemistry concepts (i.e., “specific cultural behavior related to chemistry”) and therefore can function as a legitimate learning source rather than mere anecdote (Leny Heliawati *et al.*, 2022). Empirical studies in Indonesia largely operationalize ethnochemistry by identifying, inventorying, and mapping local traditions to chemistry topics (e.g., aligning cultural practices with atomic-structure content), indicating its potential to contextualize abstract theory through culturally familiar phenomena (Irawati *et al.*, 2023). Although many implementations are reported in school-oriented settings, they explicitly argue for curriculum-level integration of indigenous knowledge to support scientific literacy and innovation tied to regional cultural identity, rationales that are directly transferable to tertiary chemistry programs (Leny Heliawati *et al.*, 2022).

In regions like Papua, where knowledge of traditional practices such as sago fermentation is rich yet underrepresented in academic curricula, there is an acute need to redesign educational models that do not compromisescientific rigor (S. Rahmawati et al., 2025).One promising approach involves intertwining ethnochemistry with modern scientific teaching methods, thereby creating a curriculum that not only deconstructs indigenous knowledge but also reconstructs it within a scientific framework. This would empower students as custodians of their cultural heritage while promoting scientific literacy (Smith, Avraamidou and Adams, 2022). Implementing education models that support cultural sustainability will be crucial in bridging existing gaps between traditional knowledge and modern scientific methodologies (Yanti *et al.*, 2025a). Ultimately, integrating ethnochemical principles into tertiary chemistry education in Indonesia offers a dual opportunity: fostering deeper engagement with chemistry through cultural relevance. It enhances chemical literacy through innovative pedagogical strategies. This approach not only helps preserve indigenous knowledge but also prepares students to navigate and contribute meaningfully to modern scientific landscapes (Torres, Mouro and Mesquita, 2024).

2.2 Technology-Enhanced Lecture Models

Technology-enhanced lecture models in higher-education science are framed as ICT-enhanced teaching that leverages LMS, mobile devices, and virtual/online tools for flexible, blended, and distance learning rather than simple digitization of lectures (Asli, Basheer and Hugerat, 2023). Design-oriented frameworks emphasize structuring online learning for inquiry and discourse (e.g., Community of Inquiry) (Redfors *et al.*, 2022) and recognizing that "mobile learning" outcomes depend on integrated device, learner social dimensions, not tools alone (Sezen-Barrie *et al.*, 2023). Empirical implementations often operationalize enhancement through scalable interactivity (e.g., rapid response/feedback, and peer learning), which is associated with improved conceptual and quantitative problem-solving compared with traditional lecturing (Whipple *et al.*, 2021). However, syntheses caution that educational technology adoption can remain under-theorized (e.g., documented theoretical gaps in gamification research), thereby limiting epistemic transformation (Tok, 2025). In parallel, culturally responsive, dialogic, technology-enabled curricula and culturally sustaining STEM pedagogy argue that digital mediation should be designed to sustain learners' cultures and address power/identity, not merely deliver content (Parsons *et al.*, 2023). Anti-racist science education further warns that culture-based frameworks require substantive curriculum shifts rather than "buzzwords" (Mshayisa, 2020).

2.3 Digital-literacy Frameworks for Developing Regions

In developing regions such as Indonesia, digital literacy frameworks are crucial for enhancing educational quality and equity, particularly in underrepresented areas like West Papua. Three noteworthy frameworks can significantly impact these initiatives. Digital literacy (often framed as digital competence) is commonly operationalized through international reference frameworks such as the European DigComp and UNESCO's Digital Literacy Global Framework (DLGF), which specify assessable components for education and training initiatives(Heaney, 2025). DigComp delineates competence across information/data literacy (supporting cognitive evaluation), communication/collaboration, digital content creation (technical production), safety (socio-ethical participation), and problem solving, and it includes knowledge, skills, and attitudes (Çebi and Reisoğlu, 2020). Reviews in higher education further show that digital literacy research clusters around measuring levels, identifying influencing factors, relating competence to achievement, and using pedagogical approaches, supporting the need for contextualized evaluation rubrics and frameworks across settings (Rosli *et al.*, 2023). Importantly, educator/learner digital-competence needs are reported to vary with local conditions and region-specific constraints, implying that "developing region" implementations require adaptation rather than direct transplantation of global standards (Schumann, Lehmann and Peters, 2025). However, these mainstream frameworks do not explicitly foreground epistemic sensitivity to indigenous knowledge systems, suggesting a conceptual gap when digital literacy is intended to support knowledge pluralism in culturally diverse contexts (Heaney, 2025).

2.4 Character Education Theories

Character education is critical to students' moral and civic development, and several theories and frameworks guide its implementation in educational systems. In Indonesia, the *Kurikulum Merdeka* operationalizes Lickona's principles by outlining six core virtues: faith, global citizenship, independence, critical-creative thinking, collaboration, and environmental stewardship (Sitanggang *et al.*, 2025). This curriculum mandates integrating these virtues across all disciplines, ensuring that character education is not siloed but an integral component of the educational experience. Research supports the idea that character education fosters social competencies and emotional attachments to moral values. Sitanggang et al. (2025) conducted a systematic review that highlights the varying cultural contexts in character education, emphasizing the integration of local values and digital literacy, thereby further enrichinghe educational experience for students (Lewis *et al.*, 2024). This integration encourages empathy and a civic-minded approach to learning. Moreover, the implementation of interactive teaching strategies enhances the effectiveness of character education by fostering dialogue and engagement among students, enabling them to internalize moral and civic values actively (Maloney *et al.*, 2022).

Furthermore, the role of culturally responsive teaching intertwined with character education is significant. By integrating local cultural contexts, such as Papua culture or local wisdom, into character education frameworks, educators can enhance the relevance and applicability of character values in students' lives (Suswandari, 2017).This approach promotes a connection to heritage and nurtures a sense of identity and belonging, which is essential for the development of robust civic characters.

2.5 Synthesised Conceptual Model

The integration of indigenous knowledge into modern pedagogy requires a framework that honours cultural heritage while leveraging contemporary educational innovations. The Transformative Ethnochemistry Lecture Model positions Papuan ethnochemistry at the centre of a dynamic, multi-layered learning ecosystem. By embedding traditional chemical knowledge and artifacts into technology-supported instruction, the model creates culturally relevant learning experiences while simultaneously cultivating 21st-century competencies. It theorises that sustainable improvement in student learning outcomes arises not from isolated strategies, but from the interplay between content, process, skills, and values, iteratively refined through Design-Based Research (DBR).

The model further highlights the reciprocal reinforcement between these systems: digital tools amplify the visibility of ethnochemistry, indigenous cases contextualise digital practices, and character education shapes the ethical and collaborative use of technology. Collectively, this synthesised conceptual model offers a pathway for culturally responsive, technology-mediated education that advances both digital literacy and character formation, while safeguarding indigenous chemical heritage, as in Figure 1.

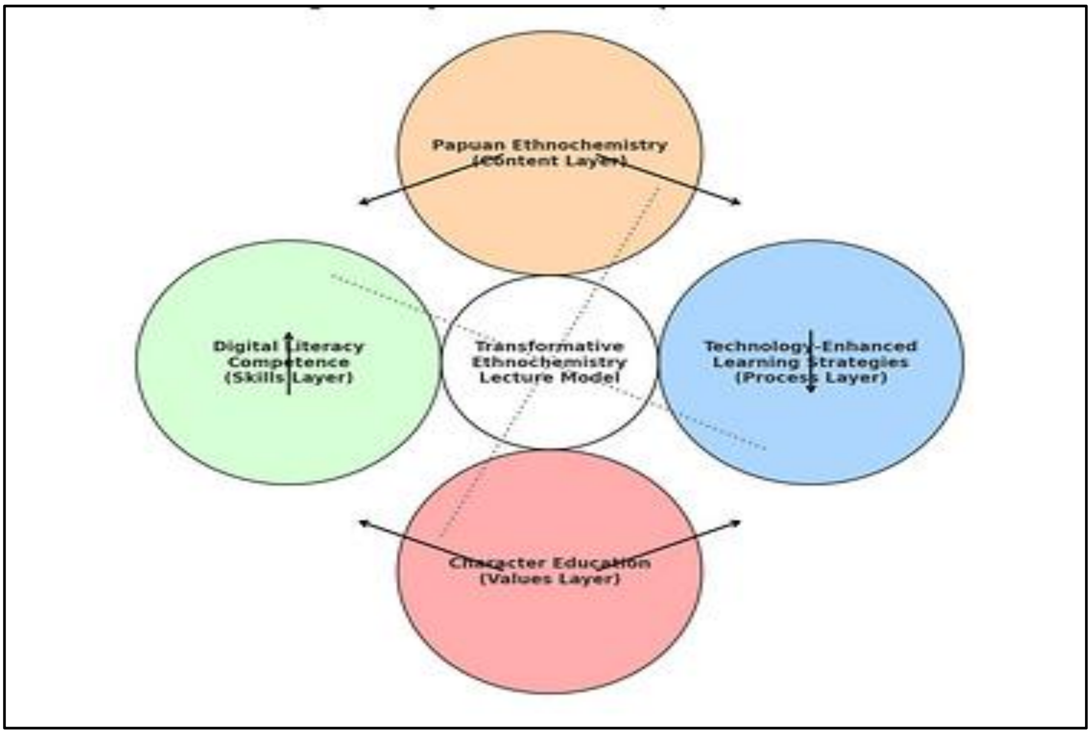


Figure 1: Conceptual Model

Figure 1 shows the four layers, the Transformative Ethnochemistry Lecture Model at the centre, cyclic DBR arrows, and dotted reciprocal reinforcement lines. It is a conceptual illustration of the "Transformative Ethnochemistry Lecture Model" at the intersection of four layers: content, Papuan ethnochemistry; process; technology-enhanced learning strategies; and skills, digital literacy competence, and values, character education. Arrows indicate cyclic DBR iterations, while dotted lines show reciprocal reinforcement between layers.

3. Methods

3.1 Research Design

Design-Based Research (DBR) is a methodological approach that integrates the design and study of educational interventions in real-world settings. It is characterized by iterative cycles of analysis, design, development, implementation, and evaluation that allow refinement of both the intervention and the underlying design principles. This approach is efficient in creating validated educational products and transferable design principles, as demonstrated in various studies across different educational contexts. The analysis phase involves identifying the problem and understanding the context. For instance, in the study of mobile social media tools for heutagogical learning, the analysis phase helped in establishing draft design principles in collaboration with teachers and literature (Narayan, Herrington and Cochrane, 2019).

Design-based research (DBR) enables quick adjustments to interventions based on real-time feedback, as seen in studiesf technology integration courses, where community of practice models were iteratively refined (Sadik, 2021). DBR's iterative nature leads to the development of robust educational tools, such as mobile virtual reality learning environments designed to enhance learner-centered pedagogies (Cochrane *et al.*, 2017). DBR generates principles that can be adapted to similar contexts, as demonstrated in the study of online instructor training, which identified principles applicable to other online education settings (Shattuck and Anderson, 2013).

3.2 Sampling Techniques

The purposive sampling strategy described in the study aligns with the instructional intervention by ensuring that all participants, including students, lecturers, and community partners, are integral to the learning environment. This approach enhances ecological validity by embedding the intervention within a naturally occurring classroom setting, thus facilitating a comprehensive educational experience that integrates academic, technological, and cultural dimensions. Selecting participants based on their roles and expertise ensures the learning model is both inclusive and representative of the diverse knowledge systems involved.

The purposive sampling strategy effectively integrates academic and cultural dimensions by involving community partners recognized for their cultural authority and expertise in traditional practices. This aligns with the concept of knowledge coevolution, which emphasizes integrating Western and Indigenous knowledge systems to create mutually beneficial new knowledge (Chapman and Schott, 2020). The involvement of lecturers with ongoing teaching responsibilities ensures that the academic component of the course is robust and aligned with the instructional goals. This is similar to the approach in food-systems education, where collective agency is scaffolded across curricula to enhance learning outcomes (Jordan *et al.*, 2023).

The study was conducted at a public university in Jayapura, Papua, Indonesia, within the instructional context of the Ethnochemistry Technology course. Participants comprised three distinct groups representing both academic and community-based perspectives. The primary group consisted of 120 students in second-year chemistry majors (68% female; mean age = 20.7 years) who were officially enrolled in the course. Instruction was facilitated by four faculty members who regularly co-taught the subject and contributed expertise in chemistry education and digital pedagogy. In addition, three community partners, two respected village elders, and one local craft-dye artisan were engaged as cultural consultants to provide authentic ethnochemical cases and artifacts. Their participation ensured that the course design was grounded in local traditions while bridging academic and indigenous knowledge systems.

3.3 Intervention Phases and Learning Environment

The intervention was structured around a five-phase design-based research (DBR) cycle, integrating both pedagogical and technological innovations within a culturally grounded learning environment. Each phase was carefully sequenced to ensure alignment between needs assessment, collaborative design, digital resource development, classroom implementation, and rigorous evaluation. The learning environment combined HyFlex delivery with contextually responsive tools and strategies to accommodate diverse student needs, including those with limited internet bandwidth. Table 1 outlines the phases, weeks, key activities, and the researcher's role.

Table 1: Intervention Phases and Learning Environment

Phase	Weeks	Key Activities	Researcher Role
Analyse	1-3	Needs analysis, baseline surveys, syllabus audit	Collect data; map gaps
Design	4-5	Co-design workshops with lecturers & elders; storyboard flipped materials; align tasks to UNESCO-DLGF and *Merdeka* virtues	Facilitate; document decisions.
Develop	6-8	Produce low-bandwidth H5P interactives, AR dye-lab app (Blippar), and character-reflection journals; pilot usability tests.	Build & iterate
Implement	9-14	HyFlex delivery: asynchronous pre-class videos + in-class problem-based sessions + field trip to local dye site; live data capture via LMS	Observe; provide tech support.
Evaluate	15-16	Post-tests, focus groups, lecturer reflections, and member-checking with elders	Analyse; validate findings

The intervention not only addressed technological and pedagogical dimensions but also cultivated a culturally sustaining learning environment, adaptable to bandwidth constraints and responsive to local values, as shown in the flowchart-style diagram of the intervention phases in Figure 2.

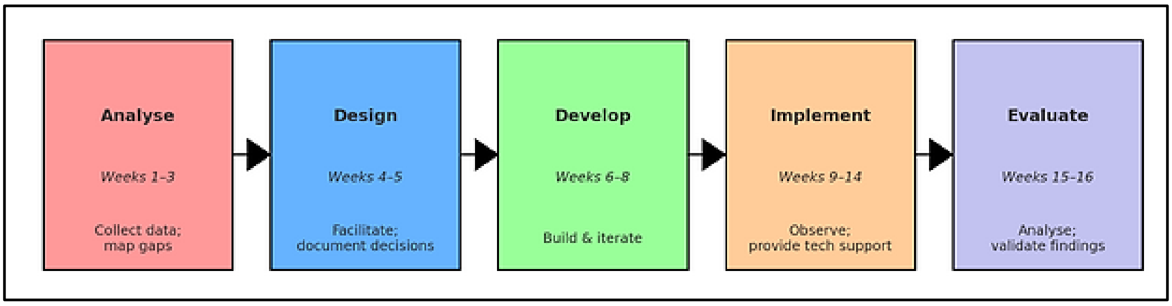


Figure 2: Flowchart of the Intervention Phase

A timeline diagram that visually represents the sequence of intervention sessions across the five phases. It shows how **Analyse → Design → Develop → Implement → Evaluate** unfolded over the 16 weeks of the study. It shows each phase in sequence, the corresponding weeks, and the researcher's role, connected with directional arrows to illustrate progression.

3.4 Instruments and Measures

To evaluate outcomes aligned with the course's learning goals, we combined validated self-report scales, structured observations, platform analytics, and qualitative inquiry. Where adaptations were necessary, we used best-practice translation procedures and inter-rater checks to preserve measurement quality, as in Table 2.

Table 2: Instruments and Measures

Construct	Instrument	Reliability	Sample Item
Digital literacy (5 UNESCO domains)	30-item Likert Digital Literacy Scale (adapted & back-translated)	Cronbach α = 0.87	"I can verify the credibility of online chemistry sources."
Character education (respect, collaboration, stewardship)	Character Values Observation Rubric (5-point, 12 indicators) rated by two observers	Inter-rater κ = 0.81	Shares equipment responsibly during lab."
Engagement	Moodle analytics (time-on-task, clickstream)	System-logged	—
Qualitative insights	Semi-structured focus-group protocol (students n = 24; lecturers n = 4)	Member-checked summaries	"Describe a moment the course connected with your community knowledge."

3.5 Data Collection and Analysis Procedures

Quantitative tests and learning analytics provide measurable evidence of the magnitude and significance of educational interventions. For instance, learning analytics can track student engagement and performance, offering data-driven insights into learning processes (Hilliger *et al.*, 2020). In the context of online learning, quantitative surveys have been used to assess student satisfaction and identify challenges, such as a lack of community and concerns about well-being (Brown *et al.*, 2023). Learning analytics frameworks, such as those discussed in the literature, emphasize the importance of data quality and technological readiness for effective implementation (Clark, Liu and Isaias, 2020).

Focus groups and thematic analysis are instrumental in uncovering the underlying mechanisms and contextual factors influencing educational outcomes. These methods enable exploration of student and staff perceptions, providing a deeper understanding of their experiences (Brown *et al.*, 2023). Thematic analysis has been used to identify key themes in students' experiences, such as facilitators and barriers to learning, which inform practice recommendations(Brown *et al.*, 2023).

Qualitative data from interviews and focus groups can reveal diverse stakeholder needs and priorities, as seen in studies on learning analytics adoption (Hilliger *et al.*, 2020). The evaluation of the design artefact is crucial for assessing the quality of implementation. This involves examining the educational design principles and their impact on learning outcomes (Müller, Mildenberger and Steingruber, 2023a). Studies have highlighted the importance of an adequate course structure, activating learning tasks, and timely feedback in enhancing the effectiveness of blended learning programs (Müller, Mildenberger and Steingruber, 2023b). The evaluation process can also provide critical feedback for refining educational approaches, such as flipped learning, by considering stakeholder perspectives and experiences (Simmons *et al.*, 2020).

Pre- and post-test scores were computed for each participant, screened for normality using the Shapiro–Wilk test, and analyzed with paired-samples t-tests at $\alpha = .05$ to determine the significance of gains; effect sizes were reported as Cohen's d. Learning management system (LMS) traces were exported as CSV files, cleaned (e.g., removal of idle-time outliers), and correlated with individual learning gains using Pearson's r. Focus-group recordings were transcribed verbatim and imported into NVivo 14 for thematic analysis using a hybrid approach that combined deductive codes (drawn from the Digital Literacy Global Framework and the target virtue categories) with inductive, data-driven codes; credibility was strengthened through analyst triangulation and participant member-checking of summary interpretations. Model fidelity was appraised with a 15-item usability/fidelity checklist, yielding a mean score of M = 4.4/5, indicating high alignment between the intended design and its enacted implementation.

4. Findings and Discussion

4.1 Learning-Management-System Analytics

We analyzed Moodle trace data over Weeks 9–14 to quantify behavioral engagement and relate it to a key design feature (the AR dye-extraction lab). Metrics include interaction volume, time-on-task trends, and feature uptake, with inferential testing via repeated-measures ANOVA to assess change over time as in Table 3.

Table 3: Summary of LMS Engagement Metrics (Weeks 9–14)

Metric	Value	How to read it
Total interactions (page views + quiz attempts)	14,678	Aggregate activity across all students and weeks
Analytic sample size	N = 120	From ANOVA df: $F(1, 119)$
Interactions per student (6 weeks)	≈122	14,678 / 120 ≈ 122 → ≈20/week
Mean time-on-task (Week 9 → Week 14)	31 → 52 min/week	+21 min total change (+67.7%)
Average weekly increase in time-on-task	+4.2 min/week	Linear rise across Weeks 9–14
Linear trend (RM-ANOVA)	$F(1,119) = 38.7$, $p < .001$	partial $\eta^2 = .25$ (large effect)
AR lab uptake (Week 12)	92% accessed	≈110 of 120 students
AR activity dwell time (median)	11 min	+46% vs course-wide median (~7.5 min)

Engagement increased significantly during implementation ($F(1,119) = 38.7$, $p < .001$; partial $\eta^2 = .25$). Time-on-task rose from 31 to 52 minutes per week (+67.7%), with about 20 interactions per student weekly. The Week 12 AR lab boosted engagement (92% participation; 11-minute median). Interactive activities mid-semester can sustain momentum. Engagement data should be carefully measured and linked to learning outcomes.

4.2 Digital-literacy Gain Scores

Pre– and post-administration of the 30-item Digital Literacy Scale (DLS) met normality assumptions (Shapiro–Wilk *ps* > .10), permitting paired-samples *t* tests. Results show substantial improvements across all UNESCO domains and overall DLS, with consistently significant effects as in Table 4.

Table 4: Digital Literacy Pre–Post Comparisons

DLS Domain	MPre (SD)	MPost (SD)	Gain (Δ)	% Change	*t*(119)	*p*	Cohen's *d*
Information/Data	11.9 (3.0)	17.4 (2.6)	5.5	46.2%	15.80	< .001	1.44
Communication	10.8 (2.7)	15.0 (2.5)	4.2	38.9%	14.23	< .001	1.30
Content Creation	12.4 (3.1)	17.6 (2.8)	5.2	41.9%	16.12	< .001	1.47
Safety	10.5 (2.9)	13.8 (2.4)	3.3	31.4%	11.77	< .001	1.07
Problem-Solving	10.7 (2.8)	14.0 (2.3)	3.3	30.8%	12.89	< .001	1.18
Overall	56.3 (9.7)	77.8 (8.1)	21.5	38.2%	14.62	< .001	1.34

The findings show great improvement in students' digital literacy across all UNESCO domains within one semester. The largest gains were in Content Creation (d = 1.47) and Information/Data (d = 1.44). Communication (d = 1.30), Safety (d = 1.07), and Problem-Solving (d = 1.18) also improved significantly. All effect sizes were large (d > 1.0), indicating a meaningful impact. Integrating culturally responsive pedagogy with accessible technology effectively strengthened digital competence. Figure 3 to visually represent the pre- and post-test gains in digital literacy, and an effect size (Cohen's d) visualization to emphasize the magnitude of improvements across domains.

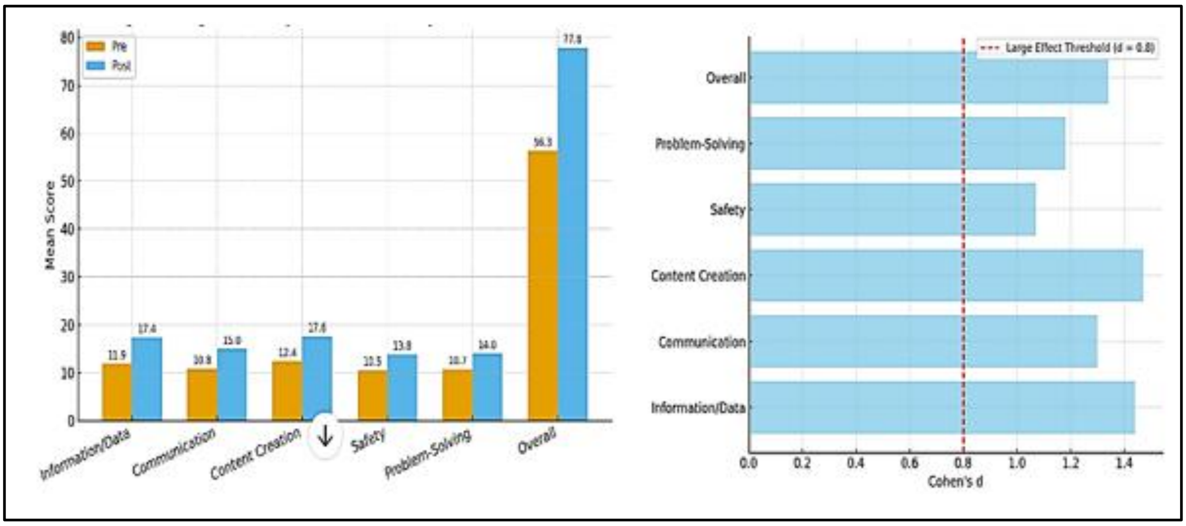


Figure 3: Digital Literacy Pre–Post Means by Domain

The chart shows consistent pre→post gains across all UNESCO domains and the overall DLS score. Visually, the "Post" bars exceed "Pre" bars for every domain, mirroring the significant paired *t*-tests and large effects reported earlier (all $d \geq 1.07$). The steepest lifts appear in Content Creation and Information/Data, aligning with coursework that emphasized producing artifacts and evaluating sources. Safety improved the least (though still strongly), suggesting an opportunity to bolster privacy, security, and digital well-being activities in future iterations. The overall increase substantiates broad competence growth within a single semester, reinforcing that the instructional model was not domain-specific but systematically elevated digital literacy. The dashed red line marks the conventional threshold for a "large" effect size ($d = 0.8$), illustrating that all observed improvements exceeded this benchmark substantially.

4.3 Character-Education Outcomes

Character education was assessed using a 12-indicator rubric that captured students' observable behaviors across five in-class collaborative tasks. Two trained observers made ratings on a 0–4 scale, with strong inter-rater reliability (κ = 0.81). For analysis, mastery was defined as scores of 3 or higher. Results indicate significant post-intervention improvements across all domains, reflecting the integration of Merdeka Belajar values and local cultural knowledge into the HyFlex model, as in Table 5.

Table 5: Mastery Rates on Character Domains (≥3 on 0–4 rubric)

Character Domain	Baseline %	Post %	Δ (pp)	Relative Increase	Risk Ratio	Cohen's *h*
Respect for Local Wisdom	42%	68%	+26	+61.9%	1.62	0.53
Collaboration	47%	71%	+24	+51.1%	1.51	0.49
Environmental Stewardship	35%	54%	+19	+54.3%	1.54	0.39

Notes. McNemar tests confirmed significant paired gains in all domains (χ^2 range **22.3–29.7**, *ps* < .001). Cohen's *h* provides a standardized magnitude for proportion shifts; values ~0.2/0.5/0.8 ≈ small/medium/large. Effects were medium to large. In a cohort of about 120 students, gains equal roughly 31 more students reaching mastery in Respect, 29 in collaboration, and 23 in stewardship. All domains improved significantly (McNemar χ^2 < .001). The largest gain was respect (+26 pp), followed by collaboration (+24 pp) and stewardship (+19 pp). Reliability was strong (κ = 0.81). Adding targeted environmental roles may further strengthen stewardship outcomes.

4.4 Qualitative Insights

To complement quantitative measures, qualitative insights were gathered through semi-structured focus groups with 24 students and four lecturers. Transcripts were coded inductively, and thematic analysis identified three high-salience themes that reveal how the intervention influenced participants' perceptions, cultural connections, and digital learning behaviors. These themes provide contextual depth to the observed gains in digital literacy and character education.

The first theme, **"Seeing Chemistry in Our Culture,"** illustrates how the AR dye-lab experience bridged disciplinary knowledge with local traditions. Students' realization that "chemistry already lives in my village" reflects the success of situating scientific learning within cultural heritage, fostering both relevance and pride.

The second theme, **"Learning by Teaching Elders,"** demonstrates the reciprocal nature of the service-learning activities. By creating infographic posters for community dye practices, students not only disseminated knowledge but also developed humility and respect through intergenerational dialogue. This aligns directly with gains observed in respect for local wisdom and collaboration.

The third theme, **"Digital Confidence,"** highlights students' growing ability to evaluate online information critically. Statements such as "Now I check who wrote the article before I copy anything" indicate that the intervention cultivated transferable critical-search skills, supporting the quantitative evidence of substantial improvements in the Information/Data and Communication domains of digital literacy.

These themes portray a virtuous cycle: culturally situated activities sparked relevance, service-learning cemented social purpose and collaborative norms, and explicit search/credibility scaffolds matured into agency. The convergence with quantitative results (engagement lift, significant DLS effects, gains in character domains) strengthens confidence that the observed improvements reflect genuine learning rather than test familiarity.

4.5 Model Fidelity and Usability

To assess the quality and consistency of implementation, external observers applied a 15-item fidelity checklist, and lecturers completed usability surveys regarding integration and student engagement. These measures were selected to evaluate not only whether the intervention was delivered as designed but also how it was experienced by instructors responsible for sustaining the model. Results indicate strong fidelity and positive perceptions of usability, with some noted challenges in workload management during the production phase, as in Table 6.

Table 6: Fidelity and Usability Summary

Measure	Scale	Result	Interpretation	Notes
External fidelity checklist (15 items)	1–5 per item	M = 4.4, SD = 0.3	High adherence to design intentions	Observers noted precise flipped sequencing, explicit virtue prompts, and materials optimized for low bandwidth.
Usability: ease of integration into syllabus	1–7 Likert	M = 6.1	Very usable for course embedding	Lecturers reported smooth mapping to existing weeks and assessments.
Usability: perceived student engagement	1–7 Likert	M = 6.4	High perceived engagement	Perceptions align with LMS analytics peaking around the AR lab.
Implementation burden (qualitative)	—	Workload spikes	Targeted constraint rather than systemic barrier	Spikes concentrated during video production and curation for flipped sessions.

Fidelity scores near the ceiling (4.4/5) indicate the model was enacted as intended in a resource-constrained Papuan context, preserving its core pillars (flipped flow, virtue alignment, low-bandwidth access). High lecturer ratings for ease of integration (6.1/7) and perceived engagement (6.4/7) suggest strong practical viability and face validity, echoing quantitative engagement lifts and qualitative reports of relevance.

The primary caveat is production workload—notably for video creation—implying that the model's most substantial returns occur when media assets are front-loaded or shared. Pragmatic next steps include (a) a reusable asset library (short, modular clips), (b) co-creation with students to distribute effort, and (c) low-bandwidth-first templates (scripts, slide skeletons) to minimize editing time.

Bottom line: Taken with the mixed-method results reported elsewhere (large digital-literacy gains, significant increases in character-value mastery, and upward LMS trends), these fidelity and usability findings show the model is both practical and implementable—with a clear, solvable pinch point in media production.

5. Discussion

5.1 Answering the Research Questions

The Transformative Ethnochemistry Lecture Model significantly addresses gaps in students' digital literacy and character development within educational environments with limited technological resources. The model uses a flipped HyFlex teaching approach, which research indicates can enhance student engagement and motivation in learning, particularly in contexts focused on digital literacy. A study highlights that implementing flipped classroom strategies can substantially improve students' awareness and responsibility in their learning journey (Torres, Mouro and Mesquita, 2024). Furthermore, incorporating local ethnochemical cases alongside relevant digital tasks is essential for fostering digital literacy, especially in environments constrained by bandwidth limitations, as the literature on ethnochemistry education's effectiveness suggests (Ridwan, Hadi and Jailani, 2022).

Quantitative assessments show that students' digital literacy skills saw considerable enhancement, with effect sizes reported around $d \approx 1.3$ to 1.5 across various domains, corroborating findings in studies focused on the impact of integrated digital tasks in educational models (Tour *et al.*, 2023). This improvement can be attributed to the interactive and participatory nature of the flipped HyFlex model, which aligns with findings from digital literacy frameworks emphasizing engagement through contemporary pedagogical strategies (Saputri, Muchtarom and Triyanto, 2019). In addition to digital skills, character outcomes, including respect for local wisdom, collaboration, and environmental stewardship, exhibited measurable improvements, increasing by approximately 19-26 percentage points among students (Murphy and Arciuli, 2024). Integrating community-driven initiatives, such as field-trip service-learning, allows students to engage meaningfully with their environments, resulting in authentic experiences that translate moral reasoning into actionable steps (Rahmawati and Agustini, 2023). Reflective practices through journaling reinforce this growth by encouraging students to connect their experiences with character education principles, thereby significantly transforming educational outcomes when embedded directly into task rubrics (Suroso and Husin, 2024).

Key factors contributing to these enhancements include authentic engagement with local communities through an educational lens, documented as vital for character development and for establishing students as active participants in sustainability efforts (Rahmawati and Atmojo, 2023). By embedding local wisdom and context-specific knowledge into the learning process, students gain a deeper understanding of their cultural identities and their relationship with environmental sustainability, which is essential for character education in contemporary contexts (L. Heliawati *et al.*, 2022). Thus, the Transformative Ethnochemistry Lecture Model not only bridges digital literacy gaps for students in resource-limited environments but also cultivates a profound respect for local wisdom and environmental responsibilities, indicative of a robust character education framework.

5.2 Positioning Within Prior Literature

The findings from the Transformative Ethnochemistry Lecture Model reveal advances in both digital literacy and character education. The observed improvements in digital literacy suggest that integrating culturally relevant content enhances the motivational aspects of technology-enhanced learning. Effect sizes reported in recent flipped-STEM meta-analyses indicate improvements in student engagement and digital competency when culturally situated content is applied (Colque-Quispe *et al.*, 2025).

Furthermore, the model’s impact on character education supports Lickona’s assertion that virtues thrive when embedded in community-linked, project-based learning environments. Evidence shows that character outcomes improve significantly when moral education is connected to practical contexts rather than relying solely on traditional, abstract storytelling. Engaging students in community involvement and hands-on experiences can enhance moral development and deepen their engagement with the subject matter, corroborated by research on effective community-oriented educational strategies (Suroso & Husin, 2024; Sitanggang et al., 2025).

The current study builds upon previous educational initiatives in Indonesia, emphasizing the significance of cultural relevance in enhancing student engagement and performance in ethnochemistry. Notably, it is the first to empirically connect cultural relevance with scaffolding in digital skills and systematic virtue assessment (Rahmawati et al., 2023). This approach aligns with advocacy for integrating digital literacy into culturally responsive teaching practices and supports sustainable educational paradigms that foster both digital competencies and ethical character development (Torres, 2024). This study highlights the importance of a holistic educational approach that blends technological engagement with culturally relevant pedagogies, ultimately promoting both digital literacy and character development in students.

5.3 Implications for Practice and Policy

Utilizing the Transformative Ethnochemistry Lecture Model has implications beyond initial educational results. It influences curricular design, faculty development, and policy leverage in STEM education, particularly in peripheral regions.

Chemistry departments in similar peripheral regions can effectively adopt the design principles derived from Design-Based Research (DBR) that emphasize cultural anchoring, low-bandwidth technology, and virtue-aligned assessment. This approach is expected to revitalize other STEM disciplines, such as ethnobotany and mineral processing, by creating content that resonates with local cultural contexts. Ethnochemistry enhances engagement and understanding in chemistry by linking abstract concepts to students' cultural practices (Rahmawati et al., 2023). Incorporating culturally relevant materials not only makes learning more relatable but also enhances comprehension and retention of complex scientific ideas (Yanti *et al.*, 2025b).

Given the necessity for innovation in instructional design, faculty members require structured support in multimedia authoring and integrating virtual-infused task design strategies. Establishing a micro-credential program focused on digital competencies could institutionalize essential skills for educators. Such initiatives promote a proactive stance in addressing educational gaps in technology and in character education conformity, which are essential for preparing students for the demands of contemporary educational environments (Rathod and Kämppi, 2025).

The findings of this study provide empirical evidence supporting character education objectives that can coexist synergistically with rigorous, digitally enhanced science instruction. This dual focus could serve as a compelling argument to policymakers for securing funding for rural broadband upgrades, which are critical to expanding access to digital resources and innovative teaching methods in underserved areas. Additionally, demonstrating the feasibility of integrating virtue and character education into STEM fields is pivotal in advocating for more inclusive educational policies (Hafina, Nur and Malik, 2022). The evidence presented underscores the importance of community integration in educational practices, which can enhance student engagement and achievement (Phelan, Maguire and Finnegan, 2025).

By embracing culturally responsive pedagogies and digital literacy frameworks, policymakers and educational institutions can bolster the educational ecosystem in peripheral regions, ensuring that students are not only technologically adept but also well-rounded individuals grounded in their cultural values.

6. Conclusion

This study set out to re-envision an Ethnochemistry Technology course for Papua Indonesia by weaving together three strands that are seldom integrated in STEM higher education: indigenous chemical heritage, explicit digital-literacy scaffolding, and character-education mandates from the *Kurikulum Merdeka*. Through a semester-long Design-Based Research cycle, the resulting Transformative Lecture Model achieved three salient outcomes. First, it delivered large, statistically significant gains in all five UNESCO Digital Literacy domains—demonstrating that low-bandwidth, flipped HyFlex design can close digital-skills gaps even in infrastructure-limited regions. Second, rubric-based observations showed marked growth in respect for local wisdom, collaboration, and environmental stewardship, validating the potency of virtue-aligned tasks such as service-learning with village elders. Third, strong LMS engagement metrics and high fidelity scores indicate that the model is both pedagogically robust and practically implementable.

Strengths include the mixed-methods DBR approach, strong fidelity monitoring, and genuine community partnership—all of which bolster ecological validity. *Limitations* revolve around scope (single university), duration (one semester), and potential observer bias in character-rubric scoring despite high κ. Replication across multiple Papuan and non-Papuan campuses, using blinded assessors and longitudinal follow-ups, would test generalisability and persistence of effects. Scalability studies should also explore automated content-translation pipelines to accommodate additional indigenous languages.

References

Anggreni, U.D., Hadiarti, D. and Fadhilah, R. (2023) "Development of the Acid-Base Microblogs Based on Malay Ethnochemistry to Preserve Culture," *Jurnal Penelitian Pendidikan Ipa*, 9(8), pp. 6067–6075. Available at: https://doi.org/10.29303/jppipa.v9i8.4300.

Asli, S., Basheer, A. and Hugerat, M. (2023) "Acid-Base Chemistry and Societal, Individual, and Vocational Aspects of Students' Learning," *Creative Education*, 14(05), pp. 943–960. Available at: https://doi.org/10.4236/ce.2023.145060.

Brown, M. *et al.* (2023) "A pragmatic evaluation of university student experience of remote digital learning during the COVID-19 pandemic, focusing on lessons learned for future practice," *PLoS ONE*, 18(5 May). Available at: https://doi.org/10.1371/journal.pone.0283742.

Çebi, A. and Reisoğlu, İ. (2020) "Digital Competence: A Study From the Perspective of Pre-Service Teachers in Turkey," *Journal of New Approaches in Educational Research*, 9(2), pp. 294–308. Available at: https://doi.org/10.7821/naer.2020.7.583.

Chapman, J.M. and Schott, S. (2020) "Knowledge coevolution: generating new understanding through bridging and strengthening distinct knowledge systems and empowering local knowledge holders," *Sustainability Science*, 15(3), pp. 931–943. Available at: https://doi.org/10.1007/s11625-020-00781-2.

Clark, J.-A., Liu, Y. and Isaias, P. (2020) "Critical success factors for implementing learning analytics in higher education: A mixed-method inquiry," *Australasian Journal of Educational Technology*, 36(6), pp. 89–107. Available at: https://doi.org/10.14742/ajet.6164.

Cochrane, T. *et al.* (2017) *A DBR framework for designing mobile virtual reality learning environments*, *Australasian Journal of Educational Technology*. Available at: https://twitter.com/hashtag/FOAMed?src=hash.

Colque-Quispe, L.W. *et al.* (2025) "Enhancing autonomous learning through flipped classroom and virtual education: A strategic approach to SDG 4," *Multidisciplinary Science Journal*, 7(12). Available at: https://doi.org/10.31893/multiscience.2025618.

Hafina, A., Nur, L. and Malik, A.A. (2022) "The development and validation of a character education model through traditional games based on the Socratic method in an elementary school," *Cakrawala Pendidikan*, 41(2), pp. 404–415. Available at: https://doi.org/10.21831/cp.v41i2.46125.

Heaney, A. (2025) "Early Adopters: Older Teens' Conceptions of Their Significant Social Media Practices," *Reading Research Quarterly*, 60(3). Available at: https://doi.org/10.1002/rrq.70026.

Heliawati, Leny *et al.* (2022) "Ethnochemistry-Based Adobe Flash Learning Media Using Indigenous Knowledge to Improve Students' Scientific Literacy," *Jurnal Pendidikan Ipa Indonesia*, 11(2), pp. 271–281. Available at: https://doi.org/10.15294/jpii.v11i2.34859.

Heliawati, L. *et al.* (2022) "ETHNOCHEMISTRY-BASED ADOBE FLASH LEARNING MEDIA USING INDIGENOUS KNOWLEDGE TO IMPROVE STUDENTS' SCIENTIFIC LITERACY," *Jurnal Pendidikan IPA Indonesia*, 11(2), pp. 271–281. Available at: https://doi.org/10.15294/jpii.v11i2.34859.

Hilliger, I. *et al.* (2020) "Identifying needs for learning analytics adoption in Latin American universities: A mixed-methods approach," *Internet and Higher Education*, 45. Available at: https://doi.org/10.1016/j.iheduc.2020.100726.

Irawati, R.K. *et al.* (2023) "Exploration and Inventory of Banjar Etnochemistry as a Learning Source in Indonesia Senior High School Chemistry Context," *JTK (Jurnal Tadris Kimiya)*, 8(1), pp. 42–58. Available at: https://doi.org/10.15575/jtk.v8i1.22380.

Irfandi, I. (2022) "Etnochemistry: Analysis Relevance of Cultural 'Pacu Jalur' in Chemical Materials as Learning Sources," *InternationalConferenceUNIKS*, pp. 56–61. Available at: https://doi.org/10.36378/internationalconferenceuniks.v0i0.2824.

Irnawati, I. and Rahmawan, S. (2024) "Development of Weebly-Based Website Learning Media Containing Ethnochemical Acid-Base Material," *Prisma Sains Jurnal Pengkajian Ilmu Dan Pembelajaran Matematika Dan Ipa Ikip Mataram*, 12(2), p. 316. Available at: https://doi.org/10.33394/j-ps.v12i2.10279.

Jordan, N.R. *et al.* (2023) "Scaffolding collective agency curriculum within food-systems education programs," *Frontiers in Sustainable Food Systems*, 7. Available at: https://doi.org/10.3389/fsufs.2023.1119459.

Lewis, C.C. *et al.* (2024) "Developing an Educational Resource Aimed at Improving Adolescent Digital Health Literacy: Using Co-Design as Research Methodology," *Journal of Medical Internet Research*, 26. Available at: https://doi.org/10.2196/49453.

Maloney, S. *et al.* (2022) "Using LMS Log Data to Explore Student Engagement with Coursework Videos," *Online Learning Journal*, 26(4), pp. 399–423. Available at: https://doi.org/10.24059/olj.v26i4.2998.

Mshayisa, V. V (2020) "Students' Perceptions of Plickers and Crossword Puzzles in Undergraduate Studies," *Journal of Food Science Education*, 19(2), pp. 49–58. Available at: https://doi.org/10.1111/1541-4329.12179.

Müller, C., Mildenberger, T. and Steingruber, D. (2023a) "Learning effectiveness of a flexible learning study programme in a blended learning design: why are some courses more effective than others?," *International Journal of Educational Technology in Higher Education*, 20(1). Available at: https://doi.org/10.1186/s41239-022-00379-x.

Müller, C., Mildenberger, T. and Steingruber, D. (2023b) "Learning effectiveness of a flexible learning study programme in a blended learning design: why are some courses more effective than others?," *International Journal of Educational Technology in Higher Education*, 20(1). Available at: https://doi.org/10.1186/s41239-022-00379-x.

Murphy, A. and Arciuli, J. (2024) "Digital reading comprehension instruction in English for children with English as an additional language: A systematic review," *Journal of Research in Reading*, 47(3), pp. 348–394. Available at: https://doi.org/10.1111/1467-9817.12448.

Narayan, V., Herrington, J. and Cochrane, T. (2019) *Design principles for heutagogical learning: Implementing student-determined learning with mobile and social media tools*, *Australasian Journal of Educational Technology*.

Parsons, D. *et al.* (2023) "Mobile Learning Frameworks and Pedagogy: A Systematic Review," *European Journal of Education*, 59(2). Available at: https://doi.org/10.1111/ejed.12601.

Phelan, D., Maguire, H. and Finnegan, C. (2025) "Scoping Review of Universal Design for Learning Principles Embedded in Subjects in Secondary Education," *European Journal of Education*. John Wiley and Sons Inc. Available at: https://doi.org/10.1111/ejed.70016.

Rahmawati, A.N. and Agustini, R. (2023) "Science Process Skills Oriented E-LKPD Assisted by Liveworksheets on the Subject of Shifting Factors in Chemical Equilibrium," *Hydrogen Jurnal Kependidikan Kimia*, 11(4), p. 535. Available at: https://doi.org/10.33394/hjkk.v11i4.8422.

Rahmawati, N. and Atmojo, I.R.W. (2023) "The Difference in the Effect of Teacher's Learning Model in TPACK Approach," in N.Y. Indriyanti and M.W. Sari (eds.) *AIP Conference Proceedings*. American Institute of Physics Inc. Available at: https://doi.org/10.1063/5.0111317.

Rahmawati, S. *et al.* (2025) "Unpacking the digital competence challenge in vocational education: A case from Indonesia," *Social Sciences and Humanities Open*, 12. Available at: https://doi.org/10.1016/j.ssaho.2025.101803.

Rathod, P. and Kämppi, P. (2025) "Improving Online and E-Learning Education Through Train-the-Trainer Model: A Case Study of European Innovation Project ECOLHE," *European Journal of Education*, 60(1). Available at: https://doi.org/10.1111/ejed.12900.

Redfors, A. *et al.* (2022) "Early Years Physics Teaching of Abstract Phenomena in Preschool—Supported by Children's Production of Tablet Videos," *Education Sciences*, 12(7), p. 427. Available at: https://doi.org/10.3390/educsci12070427.

Ridwan, M.R., Hadi, S. and Jailani, J. (2022) "A meta-analysis study on the effectiveness of a cooperative learning model on vocational high school students' mathematics learning outcomes," *Participatory Educational Research*, 9(4), pp. 396–421. Available at: https://doi.org/10.17275/per.22.97.9.4.

Rosli, N.S. *et al.* (2023) "A Theoretical Framework on Exploring the Implementation of Digital Entrepreneurship Education in Malaysian Polytechnics' Business Incubation Program," *International Journal of Academic Research in Business and Social Sciences*, 13(12). Available at: https://doi.org/10.6007/ijarbss/v13-i12/20333.

Sadik, O. (2021) "Exploring a community of practice to improve quality of a technology integration course in a teacher education institution," *Contemporary Educational Technology*, 13(1), pp. 1–16. Available at: https://doi.org/10.30935/cedtech/8710.

Saputri, R.A., Muchtarom and Triyanto, W. (2019) "Reinforcing civics literacy in sustaining students' learning in the industrial era 4.0," *Universal Journal of Educational Research*, 7(9 A), pp. 36–43. Available at: https://doi.org/10.13189/ujer.2019.071605.

Schumann, M., Lehmann, M. and Peters, H. (2025) "Beyond Carrots and Sticks. Exploring Faculty Motivation to Join a Digital Health Professions Educator Program," *Frontiers in Medicine*, 12. Available at: https://doi.org/10.3389/fmed.2025.1554011.

Sezen-Barrie, A. *et al.* (2023) "Epistemic Discourses and Conceptual Coherence in Students' Explanatory Models: The Case of Ocean Acidification and Its Impacts on Oysters," *Education Sciences*, 13(5), p. 496. Available at: https://doi.org/10.3390/educsci13050496.

Shattuck, J. and Anderson, T. (2013) "Using a design-based research study to identify principles for training instructors to teach online," *International Review of Research in Open and Distance Learning*, 14(5), pp. 186–210. Available at: https://doi.org/10.19173/irrodl.v14i5.1626.

Simmons, M. *et al.* (2020) "Introducing the flip: A mixed method approach to gauge student and staff perceptions on the introduction of flipped pedagogy in pre-clinical medical education," *Australasian Journal of Educational Technology*, 2020(3), p. 36. Available at: https://doi.org/10.14742/ajet.5600.

Sitanggang, A.O. *et al.* (2025) "A systematic literature review: Character education to build tolerance," *Multidisciplinary Reviews*, 8(10). Available at: https://doi.org/10.31893/multirev.2025201.

Smith, T., Avraamidou, L. and Adams, J.D. (2022) "Culturally relevant/responsive and sustaining pedagogies in science education: theoretical perspectives and curriculum implications," *Cultural Studies of Science Education*, 17(3), pp. 637–660. Available at: https://doi.org/10.1007/s11422-021-10082-4.

Suroso, S. and Husin, F. (2024) "Analyzing Thomas Lickona's Ideas in Character Education (A Library Research)," *Proceedings of the 7th FIRST 2023 International Conference on Global Innovations (FIRST-T3 2023)*, pp. 39–47. Available at: https://doi.org/10.2991/978-2-38476-220-0_5.

Suswandari (2017) "Incorporating beliefs, values and local wisdom of Betawi culture in a character-based education through a design-based research," *European Journal of Contemporary Education*, 6(3), pp. 574–585. Available at: https://doi.org/10.13187/ejced.2017.3.574.

Tok, H.H. (2025) "The Effect of Forum Theatre on Nursing Students' Ageist Attitudes Towards Older Adults," *Australasian Journal on Ageing*, 44(4). Available at: https://doi.org/10.1111/ajag.70107.

Torres, M., Mouro, J. and Mesquita, M. (2024) "EDUCATION, SUSTAINABILITY AND LOCAL ECOLOGICAL KNOWLEDGE," *Janus.net*, 15(1 TD1), pp. 81–103. Available at: https://doi.org/10.26619/1647-7251.DT0224.5.

Tour, E. *et al.* (2023) "Investigating the efficacy of the AMEP Digital Literacies Framework and Guide for adult EAL settings," *Journal of Adolescent and Adult Literacy*, 67(3), pp. 150–161. Available at: https://doi.org/10.1002/jaal.1309.

Wahyudiati, D. and Qurniati, D. (2023) "Ethnochemistry: Exploring the Potential of Sasak and Javanese Local Wisdom as a Source of Chemistry Learning to Improve the Learning Outcomes of Pre-Service Teachers," *Jurnal Pendidikan Sains Indonesia*, 11(1), pp. 12–24. Available at: https://doi.org/10.24815/jpsi.v11i1.26790.

Whipple, S. *et al.* (2021) "The Field Experience as a Potential Barrier to Underrepresented Minority Student Participation in Ecological Sciences," *Bulletin of the Ecological Society of America*, 102(4). Available at: https://doi.org/10.1002/bes2.1928.

Yanti, F. *et al.* (2025a) "Development of Technopreneurship-Based E-Modules for Ethnochemistry, Redox, and Science Literacy," *APTISI Transactions on Technopreneurship*, 7(2), pp. 469–480. Available at: https://doi.org/10.34306/att.v7i2.517.

Yanti, F. *et al.* (2025b) "Development of Technopreneurship-Based E-Modules for Ethnochemistry, Redox, and Science Literacy," *APTISI Transactions on Technopreneurship*, 7(2), pp. 469–480. Available at: https://doi.org/10.34306/att.v7i2.517

Simulation-based Learning and Project Management Certification Outcomes: Evidence from IPMA Level D Training

Marcin Opas and Małgorzata Ćwil
Department of Quantitative Methods and Information Technology, Kozminski University, Warsaw, Poland

mopas@kozminski.edu.pl (corresponding author)
mcwil@kozminski.edu.pl

https://doi.org/10.34190/ejel.24.3.4740

Abstract: The development and assessment of project management competencies remain a persistent challenge in professional education. While competence frameworks emphasise behavioural and contextual judgement, assessment practices in entry-level certification contexts continue to rely predominantly on standardised knowledge-based examinations. At the same time, simulation-based and game-based learning approaches are increasingly used to support experiential learning, yet evidence linking such interventions to externally validated assessment outcomes remains limited. This study examines whether the inclusion of a simulation-based learning intervention in a preparatory course for the International Project Management Association (IPMA) Level D certification is associated with differences in certification examination performance. Using a quasi-experimental design, examination results for participants who completed simulation-supported training (n = 178) were compared with those of a reference cohort who completed the same course without the simulation component (n = 455). Certification exam scores were analysed across competence areas defined in the IPMA Individual Competence Baseline (ICB 4.0). The results indicate that participants in the simulation-supported group achieved higher overall examination scores, with statistically significant differences concentrated mainly in selected People and Practice competence elements. Mean scores in the Perspective area were also higher in the simulation-supported group, but the differences were not statistically significant. While the non-randomised design does not allow causal conclusions, the findings suggest an association between competence-oriented simulation-based learning and higher performance in a standardised, externally administered certification examination. The analysis focused not only on overall examination performance but also on the distribution of differences across individual competence elements. This made it possible to examine whether observed differences corresponded to the areas most directly activated by the simulation scenario. The study therefore provides evidence on the alignment between simulation design, competence frameworks, and externally administered assessment outcomes. The study contributes to e-learning research by demonstrating how simulation-based learning can be examined in relation to formal assessment outcomes within a professional certification context. It highlights the importance of aligning experiential learning design with competence frameworks while maintaining independence from existing assessment formats.

Keywords: Project management, Simulation-based learning, Simulation games, IPMA certification, Quasi-experiment

1. Introduction

Project management is a dynamic and complex discipline requiring diverse competencies, ranging from technical and strategic skills to interpersonal and behavioural capabilities. As organisations increasingly rely on structured project management methodologies to enhance efficiency and competitiveness (Martinelli and Milosevic, 2016), the role of competent project managers has become more critical than ever (Maylor, 2017). Ensuring that project managers are well-equipped with the necessary skills is a central concern for certification bodies and training institutions. Among the recognised competency frameworks, the IPMA ICB 4.0 provides a structured assessment of project management competencies through its four-level certification system (IPMA, 2015).

Traditional project management education relies heavily on theoretical instruction, case studies, and structured training programs. However, these approaches may not fully prepare individuals for the complexities and uncertainties of real-world project environments. To address this gap, simulation-based learning (SBL) has gained traction as an innovative educational tool that offers hands-on, experiential learning in a risk-free environment. Despite the growing adoption of SBL in project management training (Rumeser and Emsley, 2019), the effectiveness of such methods remains a subject of academic and practical debate. Some studies highlight the benefits of SBL in enhancing cognitive and behavioural competencies, whereas others raise concerns regarding methodological inconsistencies and the difficulty of measuring learning outcomes. The current study aims to contribute to this discourse by examining whether participation in a simulation-supported preparatory course is associated with differences in IPMA Level D certification examination outcomes.

Cite this article: Opas, M. and Ćwil, M. 2026. "Simulation-based Learning and Project Management Certification Outcomes: Evidence from IPMA Level D Training", *Electronic Journal of e-Learning*, 24(3), pp 79-91,
https://doi.org/10.34190/ejel.24.3.4740

The structure of the paper is as follows. Section 2 provides an overview of project manager competency models, certification procedures, and the role of IPMA Level D certification. It then discusses simulation-based learning in project management education, clarifying its relationship with serious games and game-based learning, and explaining how it can be aligned with competence-oriented educational frameworks. Section 3 outlines the research methodology, including the quasi-experimental design, participant selection, data collection process, and statistical methods used to analyse associations between simulation-supported learning and certification examination outcomes. Section 4 presents the research findings, comparing exam performance between participants who followed conventional project management training and those who engaged in SBL. The results highlight the extent to which simulation-supported learning is associated with competence-oriented learning outcomes as reflected in certification examination results. Section 5 discusses the practical and theoretical implications of the findings and explores their relevance for project management training institutions, certification bodies, and professional development programs. The paper concludes with recommendations for integrating simulation games into project management education, emphasising the need for future research on long-term competency retention and adaptive learning technologies.

This study addresses the following research question:

RQ1: Is participation in a simulation-supported IPMA Level D preparatory course associated with higher certification examination performance compared with participation in the same course without the simulation component?

Two supplementary questions are also considered:

RQ2: In which IPMA ICB 4.0 competence areas and competence elements are between-group differences most pronounced?

RQ3: To what extent do the observed differences correspond to the competence areas and competence elements activated by the SBL intervention?

By investigating the role of simulation games in supporting competence-oriented learning, as reflected in certification examination outcomes, this research seeks to provide empirical insights into SBL's role in professional project management certification training. The findings of this study will contribute to the ongoing discussion of innovative learning methodologies in project management education and offer practical recommendations for training institutions and certification bodies seeking to enhance learning outcomes through digital and experiential learning.

2. Review of Literature

2.1 Project Manager Competencies

The concept of project manager competencies has evolved from prescriptive lists of tasks and knowledge areas to integrative constructs that emphasise performance, behaviour, and contextual judgment. Early approaches to project management competence were predominantly grounded in bodies of knowledge and method-oriented standards, focusing on procedural correctness and technical proficiency (PMI, 2002; Crawford, 2005). While these approaches contributed to the professionalisation of project management, they have proven insufficient to explain variations in performance among practitioners with comparable formal qualifications.

Contemporary competency research, therefore, adopts a broader and more dynamic perspective, recognising that effective project management depends not only on technical expertise but also on behavioural capabilities and the ability to interpret and respond to organisational and environmental conditions (Koi-Akrofi et al., 2024; Turner, Huemann and Keegan, 2014; Le Deist and Winterton, 2005). From this perspective, competence is understood as a context-dependent integration of knowledge, skills, and attitudes that becomes visible through action rather than declarative understanding alone. In project environments characterised by uncertainty and competing constraints, competencies are revealed through decision-making, stakeholder interactions, and adaptive behaviour (Alvarenga et al., 2020; Crawford and Pollack, 2007).

In project management, several competency frameworks have been developed to operationalise this holistic view. Comparative analyses of widely recognised standards, including those proposed by the Project Management Institute, IPMA, Association for Project Management, and Australian Institute of Project Management, indicate a recurring tripartite structure encompassing task-oriented, person-oriented, and context-oriented dimensions of competence (Koi-Akrofi et al., 2024; Ochoa Pacheco et al., 2023; Omidvar et al., 2011). Task-oriented competencies relate to planning, monitoring, and controlling project activities. Person-oriented competencies capture leadership, communication, self-regulation, and collaboration, while context-oriented competencies address governance, strategic alignment, and stakeholder environments. Despite differences in terminology, this conceptual convergence suggests a shared understanding of competent project management practice.

Among existing frameworks, the IPMA ICB 4.0 explicitly positions competence as a personal capability demonstrated through performance in real or simulated project situations (IPMA, 2015). Unlike knowledge-based certification models, the IPMA approach places stronger emphasis on behavioural indicators and contextual judgment, supported by evidence gathered through portfolios, interviews, and observed performance. This person-centred logic aligns with contemporary competence theory and is particularly relevant in educational and certification contexts where learning outcomes extend beyond the acquisition of factual knowledge (Omidvar et al., 2011).

Despite the conceptual maturity of competency frameworks, a persistent challenge remains in translating abstract competency constructs into reliable and scalable assessment mechanisms. While frameworks define what project managers should be capable of, they provide limited guidance on how to consistently measure these capabilities, especially at entry or early certification levels (Leigh et al., 2007). Standardised examinations, commonly used in certification schemes, tend to privilege knowledge recall and conceptual understanding, raising concerns about their ability to capture behavioural and contextual competencies that are central to contemporary models (PMI, 2002).

Empirical studies further indicate that not all competencies contribute equally to perceived effectiveness. Research applying the Competing Values Framework demonstrates that behavioural competencies such as creativity, engagement and motivation, efficiency, and value-based judgement differentiate higher- and lower-performing project managers across multiple dimensions of organisational effectiveness (Trivellas and Drimoussis, 2010). Similar conclusions have been drawn in international project contexts, where leadership, communication, stakeholder engagement, and adaptability consistently emerge as critical across project phases (Bashir, Maqsood and Jeong, 2021). These findings reinforce the argument that competence manifests through adaptive behaviour in the face of competing demands rather than adherence to predefined procedures.

In response to these limitations, scholars increasingly call for assessment approaches that combine standardisation with experiential evidence (Leigh et al., 2007; Omidvar et al., 2011). Simulation-based environments and other experiential learning tools are frequently proposed as mechanisms for eliciting observable competence-oriented behaviours in conditions that approximate the objective project complexity. However, empirical evidence linking such learning interventions with externally validated assessment outcomes, such as certification examinations, remains limited.

Against this backdrop, this study adopts the IPMA ICB 4.0 framework as its conceptual reference point. It examines whether simulation-based learning can support the acquisition of competencies that are subsequently reflected in standardised certification outcomes. By focusing on certification exam results as an external performance indicator, the study addresses a gap between competence-oriented educational design and conventional assessment practices, contributing to ongoing discussions on how project manager competencies can be meaningfully developed and evaluated in higher education and professional training contexts.

2.2 Simulation-based Learning in Project Management Education

Although simulation-based learning, serious games and game-based learning partly overlap in the literature, this study conceptualises the intervention primarily as simulation-based learning because its core educational mechanism is the modelling of project management decisions under dynamic constraints. In this sense, game-based learning and serious games provide a broader conceptual context, whereas simulation-based learning is the focal category used to describe the educational intervention examined in this study.

SBL has become an established approach in technology-enhanced education, particularly in domains requiring the development of complex, non-routine skills. In contrast to traditional instructional methods, simulation-based environments place learners in realistic decision situations in which they must manage resources, respond to dynamic feedback, and adapt their actions under conditions of uncertainty (Michael and Chen, 2006; Bellotti et al., 2019). This experiential character makes SBL especially relevant for professional education contexts where learning outcomes extend beyond factual knowledge to include judgment, behaviour, and situational awareness (Yaman et al., 2024).

Simulation-based games have been widely adopted in project management education to support the development of decision-making, communication, and coordination skills. Empirical and review studies indicate that project management simulations support active engagement in learning and can help learners develop and apply both technical project management competencies and behavioural skills in realistic, risk-free project scenarios (Geithner and Menzel, 2016; Jaccard, Bonnier and Hellström, 2022; Hellström, Jaccard and Bonnier, 2023). Unlike case studies or static exercises, simulations expose learners to the temporal consequences of their decisions, making explicit trade-offs among time, cost, scope, and stakeholder satisfaction.

Beyond their instructional value, serious games are increasingly seen as potential tools for formative and summative assessment. Assessment-oriented research in education highlights that complex competencies are best evaluated through integrated performance tasks rather than isolated knowledge tests (Leigh et al., 2007). In this respect, games generate rich performance data such as decision paths, resource allocation patterns, and outcome trajectories that can serve as evidence of competence-oriented behaviour. When combined with structured debriefing and reflection, these data support both learner self-assessment and instructor-led evaluation (Bellotti et al., 2019).

Despite this potential, most studies on game-based learning in project management focus primarily on learning effectiveness, learner satisfaction, or perceived competence development (Elenany and Ahmed, 2024). Far fewer investigations examine whether outcomes from SBL translate into externally validated assessment results, such as standardised examinations or certification outcomes. This gap is evident in entry-level professional certification contexts, where assessment practices remain predominantly knowledge-based despite competence-oriented learning designs.

Recent research in e-learning and professional education, therefore, calls for greater alignment between learning interventions and assessment mechanisms. Simulation-based environments are increasingly viewed not only as pedagogical tools but also as boundary objects that can connect educational practice with formal evaluation systems (Walker and Engelhard, 2014; Hellström, Jaccard and Bonnier, 2023; Özsoy, T., 2025). When learning activities are explicitly mapped to recognised competence frameworks, games can serve as a bridge between experiential learning and standardised assessment, without reducing competence to mere declarative knowledge.

Building on the view that simulation-based environments can serve as boundary objects connecting educational practice with formal evaluation systems, a central challenge in SBL research is to articulate how game experiences translate into valid, interpretable assessment evidence (Plass, Homer and Kinzer, 2015). In other words, the credibility of SBL claims depends not only on whether learners appear to improve, but on the explicit relationship between (a) intended learning objectives, (b) the game mechanics and learning processes they instantiate, and (c) the assessment evidence used to substantiate outcomes. Comprehensive accounts of SBL describe learning as emerging from integrated cognitive, motivational, affective, and sociocultural processes, suggesting that evaluation strategies should align with the underlying learning mechanisms rather than rely exclusively on post hoc declarative tests (Plass, Homer, and Kinzer, 2015).

A well-established response to this methodological challenge is embedded or stealth assessment, in which gameplay yields observable evidence of target constructs (e.g., decision quality, action sequences, problem-solving trajectories), and these indicators are mapped to competency models through transparent inference rules (Shute and Ke, 2012). This approach is particularly relevant in competency-oriented education, where intended outcomes include behavioural and judgment components that are difficult to capture through traditional examinations alone. From this standpoint, robust SBL evaluation typically requires an explicit validity argument that links design features to constructs and constructs to interpretable indicators (Shute and Ke, 2012; Tettegah, McCreery and Blumberg, 2015).

In practice, however, many professional education contexts rely on external standardised assessments (e.g., certification examinations) that cannot be readily adapted to capture nuanced in-game evidence. In such settings, the evaluation problem becomes one of construct alignment: the extent to which external measures capture the exact competency domains and cognitive processes that the simulation is designed to develop, and the extent to which test formats may underrepresent key competencies. Applied work on simulation games in blended formats highlights that credible claims require deliberate integration of learning activities, feedback, and assessment, including an explicit discussion of what counts as evidence and why (Wall and Ahmed, 2008; de Freitas, 2006).

Recent synthesis work also indicates that "assessment literacy" remains a practical barrier to wider SBL adoption, as educators often struggle to justify SBL when the chain from learning experience to defensible evidence of an outcome is not explicit (Yaman et al., 2024). Therefore, when SBL is evaluated using external examinations, methodological quality is strengthened when authors make alignment assumptions explicit, discuss likely construct underrepresentation, and position external test performance as one evidence source within a broader validity argument, an approach that directly informs how SBL can support evidence-based assessment of competence development in project management education (Shute and Ke, 2012; Yaman et al., 2024).

In the context of project management education, this perspective suggests that the value of SBL lies not only in enhancing learning experiences but also in its capacity to support evidence-based assessment of competence development. Building on the competence-oriented framework discussed in Section 2.1, the present study examines SBL as an educational intervention whose effects can be analysed using certification examination results as an external reference point. By doing so, the study contributes to ongoing discussions in e-learning research on how technology-enhanced learning designs can be meaningfully aligned with formal assessment practices.

3. Methodology

3.1 Research Design and Educational Content

The study was designed to examine whether participation in a simulation-supported preparatory course was associated with differences in IPMA Level D certification examination outcomes. Rather than testing the causal effectiveness of the intervention, the study analyses between-group differences in externally assessed certification results within an authentic professional training context.

The study adopts a quasi-experimental, intervention-based research design conducted within a formal project management training program preparing participants for the IPMA Level D certification. The research is positioned as applied educational research embedded in an authentic professional learning environment. Its purpose is to examine whether the inclusion of an SBL intervention is associated with differences in externally assessed learning outcomes.

Participants were not randomly assigned to groups, as they were naturally grouped by course enrolment. Certification examination outcomes for participants who completed the simulation-supported training were compared with those of a reference cohort that followed the same preparatory program, but without the simulation component, during the same period. This design reflects organisational and ethical constraints typical of certification-oriented education while allowing meaningful comparison of learning outcomes under comparable conditions.

3.2 Participants

The experimental group consisted of 178 participants who attended IPMA Level D preparatory courses between December 2022 and March 2024, during which the SBL intervention was implemented. The control group comprised 455 participants who attended the same course in the same time period but did not participate in the simulation activity. The comparison group was restricted to the same period as the simulation-supported group to reduce the risk of historical cohort effects and temporal changes in training or examination conditions.

All participants were candidates for entry-level project management certification and did not hold an IPMA certificate before enrolment in the course. The course content, instructional approach, and certification examination format were identical for both groups to ensure comparability. Anonymised examination results for both groups were obtained from the IPMA Poland certification examinations database after the course was completed.

3.3 Simulation-based Learning Intervention

The SBL intervention was implemented using the PM2PM computer-based project management simulation game, developed by PM2PM Ltd. This Polish company offers, among other consulting services, project management training courses, primarily to prepare applicants for the IPMA project management certification (PM2PM, Ltd., 2025).

At the start of the simulation, participants assumed the role of project managers responsible for delivering a newly assigned project. The initial planning phase involved defining project scope and timeline, allocating resources based on competency matrices, estimating costs, setting budget constraints, and identifying potential risks and mitigation strategies. Participants were provided with project documentation and competency matrices to support decision-making.

The simulation operated through an interactive digital dashboard that enabled participants to:

- assign human resources using a competency matrix (see Figure 1),
- schedule and adjust project timelines using Gantt charts (see Figure 2),
- monitor budget allocation and financial constraints, respond to emerging risks, delays, skill shortages, and cost fluctuations.

Figure 1: PM2PM game competency matrix dashboard (PM2PM, Ltd., 2025)

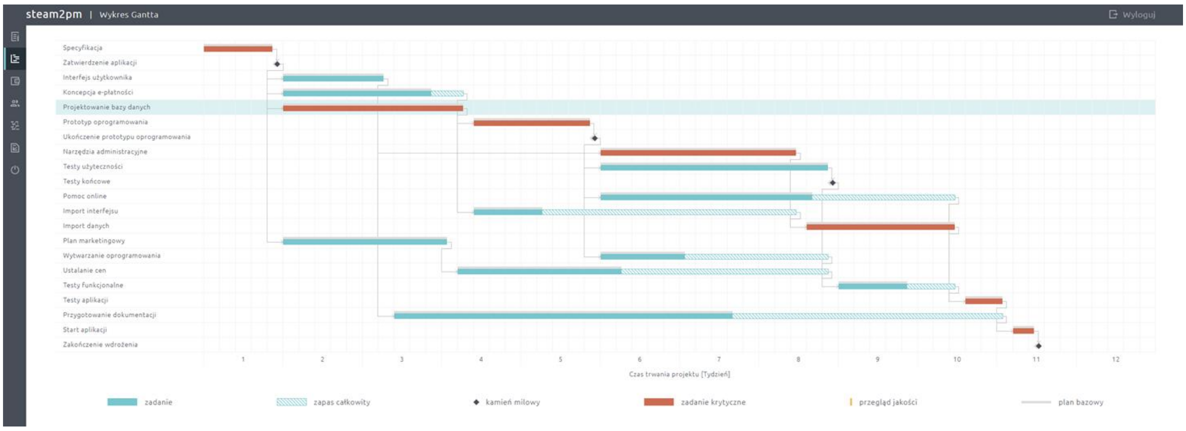


Figure 2: PM2PM game Gantt dashboard (PM2PM, Ltd., 2025)

The simulation session followed a structured timeline consisting of:

- a planning phase (approx. 45 minutes) focused on project setup, resource allocation, and budgeting,
- an execution phase (approx. 60 minutes) involving dynamic decision-making and project adjustments,
- a debriefing and evaluation phase (approx. 30 minutes) dedicated to performance analysis and reflective discussion.

During execution, the simulation dynamically modified project conditions, requiring participants to trade off cost, schedule, quality, and resource availability. Real-time feedback allowed participants to observe the consequences of their decisions and adjust strategies accordingly. At the end of the simulation, participants received performance reports summarising schedule efficiency, budget adherence, resource utilisation, risk exposure, and overall project quality.

3.4 Alignment of the Simulation with the IPMA ICB 4.0

The design of the PM2PM simulation reflects the competence-oriented logic of the IPMA ICB 4.0. Rather than addressing individual competence elements in isolation, the simulation presents integrated project scenarios that require the simultaneous application of competencies from the Perspective, People, and Practice areas.

Decision situations embedded in the simulation engage competence elements related to strategy, governance, and compliance (Perspective); teamwork, communication, and conflict management (People); and scope, time, cost, risk, and change management (Practice) (Table 1). These competences are activated through observable decision-making and interaction patterns rather than through explicit instruction or examination-style assessment.

Table 1: Learning objectives of the PM2PM game aligned with the IPMA ICB 4.0

Area	ICB Competence Element (CE)	PM2PM Game Learning Objective
4.3 Perspective	4.3.1. Strategy	
	4.3.2. Governance, structures, and processes	
	4.3.3. Compliance, standards, and regulations	
	4.3.4. Power and interest	
	4.3.5. Culture and values	
4.4 People	4.4.1. Self-reflection and self-management	
	4.4.2. Personal integrity and reliability	
	4.4.3. Personal communication	X
	4.4.4. Relationships and engagement	
	4.4.5. Leadership	X
	4.4.6. Teamwork	X
	4.4.7. Conflict and crisis	
	4.4.8. Resourcefulness	
	4.4.9. Negotiation	
	4.4.10. Result orientation	
4.5 Practice	4.5.1. Project design	
	4.5.2. Requirements and objectives	X
	4.5.3. Scope	X
	4.5.4. Time	X
	4.5.5. Organisation and information	
	4.5.6. Quality	X
	4.5.7. Finance	X
	4.5.8. Resources	X
	4.5.9. Procurement	
	4.5.10. Plan and control	X
	4.5.11. Risk and opportunity	X
	4.5.12. Stakeholders	X
	4.5.13. Change and transformation	

Importantly, the simulation was not designed to replicate the structure or content of the IPMA certification examination and did not function as direct exam preparation. Its purpose was to operationalise competence development by creating conditions in which conceptual knowledge acquired during the course had to be applied in dynamic, uncertain project contexts. The alignment with IPMA ICB 4.0, therefore, exists at the level of competence domains and behavioural manifestations rather than at the level of exam items or scoring criteria.

3.5 Assessment Measures and Data Collection

The primary outcome measure in the study was performance on the IPMA Level D certification examination, a standardised, externally administered assessment aligned with IPMA ICB 4.0. The examination is conducted independently of the training provider and researchers. For both experimental and control groups, anonymised examination results were collected after course completion. The dataset included overall certification exam scores and competence element scores aggregated within the Perspective, People, and Practice areas, as reported in official examination records. No in-game performance metrics or self-reported learning measures were used as dependent variables. The study deliberately focused on externally validated assessment outcomes to ensure comparability across cohorts and to avoid bias associated with internally generated performance indicators.

3.6 Ethical Considerations

Participation in the study was voluntary and did not affect access to training or certification. Examination data were analysed in an anonymised form and used exclusively for research purposes. The study adhered to institutional guidelines for educational research and data protection.

The authors declare no commercial affiliation with the simulation provider.

4. Research Results

All data were analysed using SPSS v. 29. The data were first examined for skewness and kurtosis, and no problematic issues were noted (skewness and kurtosis in between –1 and 1). An independent-samples t-test was performed for each competence element to examine differences in exam scores between the experimental ($n = 178$) and control ($n = 455$) groups. Cohen's d was used to assess the magnitude of observed between-group differences. Given the exploratory nature of competence-element-level comparisons, p-values are interpreted descriptively and reported alongside effect sizes. The emphasis is placed on the overall pattern of differences rather than on isolated statistically significant results.

4.1 Perspective Area

Across all five competence elements in the perspective area, the average scores were higher among individuals who completed the simulation game; however, no statistically significant differences were observed for any of these elements (Table 2). The negative values of Cohen's d indicate that the mean results in the experimental group were higher than in the control group. The higher the Cohen's d values, the bigger the differences between the two analysed groups.

Table 2: Scores for the competence elements in the perspective area

ICB CE	Mean 1	Mean 2	SD 1	SD 2	Sig.		Cohen's *d*	RANK
Power and interest	3.84	3.97	1.325	1.244	0.140		-0.096	
Strategy	3.40	3.52	1.272	1.272	0.145		-0.094	
Governance, structures, and processes	3.97	4.04	1.119	1.005	0.226		-0.066	
Culture and values	4.38	4.39	1.020	0.946	0.422		-0.017	
Compliance, standards, and regulations	3.72	3.74	1.149	1.020	0.426		-0.017	

4.2 People Area

The statistically significant difference was observed in three out of 10 competence elements belonging to the people area, with the highest differences occurring for teamwork ($M1 = 3.70$; $M2 = 4.09$; $p < 0.001$, $d = -0.324$), followed by relationships and engagement ($M1 = 4.23$; $M2 = 4.49$; $p = 0.001$, $d = -0.240$), result orientation ($M1 = 3.94$; $M2 = 4.10$; $p = 0.041$, $d = -0.144$). The analysis results are summarised in Table 3.

Table 3: Scores for the competence elements in the people area

ICB CE	Mean 1	Mean 2	SD 1	SD 2	Sig.		Cohen's *d*	RANK
Teamwork	3.70	4.09	1.137	1.075	<0.001	***	-0.324	1
Relationships and engagement	4.23	4.49	1.155	0.916	0.001	***	-0.240	2
Result orientation	3.94	4.10	1.153	0.978	0.041	*	-0.144	3
Personal communication	4.34	4.48	1.005	0.891	0.104		-0.144	
Conflict and crisis	3.97	4.01	1.046	1.031	0.316		-0.042	
Leadership	3.68	3.74	1.301	1.307	0.325		-0.040	
Resourcefulness	4.37	4.40	1.000	0.929	0.352		-0.034	
Self-reflection and self-management	4.47	4.50	0.947	0.885	0.379		-0.027	
Negotiation	4.02	4.04	1.120	0.965	0.410		-0.020	
Personal integrity and reliability	3.91	3.80	1.186	1.147	0.136		0.097	

4.3 Practice Area

In the practice area, a statistically significant difference was observed in 5 of 13 cases, with the largest effect observed for change and transformation (M1 = 3.40; M2 = 3.81; p < 0.001, d = -0.311), followed by scope (M1 = 3.77; M2 = 4.08; p = 0.002, d = -0.241) and resources (M1 = 4.01; M2 = 4.26; p = 0.001, d = -0.225). The analysis results are presented in Table 4.

Table 4: Scores for the competence elements in the practice area

ICB CE	Mean 1	Mean 2	SD 1	SD 2	Sig.		Cohen's *d*	RANK
Change and transformation	3.40	3.81	1.367	1.264	<0.001	***	-0.311	1
Scope	3.77	4.08	1.319	1.152	0.002	**	-0.241	2
Resources	4.01	4.26	1.150	0.992	0.001	**	-0.225	3
Finance	3.93	4.12	1.282	1.126	0.006	**	-0.153	4
Project design	3.62	3.82	1.308	1.254	0.042	*	-0.153	5
Procurement	4.00	4.12	1.215	1.168	0.114		-0.107	
Time	3.48	3.63	1.232	1.158	0.071		-0.130	
Risk and opportunity	2.89	3.03	1.471	1.424	0.135		-0.098	
Plan and control	3.85	3.96	1.278	1.234	0.165		-0.086	
Requirements and objectives	3.20	3.26	1.304	1.272	0.301		-0.046	
Organisation and information	4.15	4.17	1.131	1.028	0.414		-0.019	
Stakeholders	3.63	3.64	1.403	1.321	0.468		-0.007	
Quality	3.48	3.42	1.282	1.330	0.284		0.051	

4.4 Overall IPMA Level D Exam Scores

The last step of the analysis was to compare the overall exam scores of the two groups. According to the independent-samples t-test results, there is a significant difference in exam scores between the two analysed groups (N1 = 455, N2 = 178). The simulation-supported group achieved significantly higher overall exam scores than the comparison group (M1 = 107.34; M2 = 110.54; p = 0.02, d = -0.181). The analysis results are presented in Table 5.

Table 5: Overall exam scores for the two analysed groups

Mean 1	Mean 2	SD 1	SD 2	Sig.		Cohen's *d*
107.34	110.54	18.614	15.116	0.02	**	-0.181

Thus, the quasi-experimental analysis indicates an association between participation in simulation-supported learning and higher performance in the IPMA Level D certification examination. Across the IPMA Level D examination results, statistically significant differences between the simulation-supported and comparison cohorts were observed for a substantial subset of competence elements. The significant differences were identified in the People and Practice competence areas, particularly in elements related to teamwork, relationships and engagement, change, scope, resource and finance, with effect sizes ranging from small to moderate (highest Cohen's d ≈ 0.3-0.35). Differences within the Perspective area were generally weaker and not significant. At the overall examination performance level, the simulation-supported cohort achieved significantly higher scores with reduced score dispersion, indicating more homogeneous outcomes.

5. Discussion and Conclusions

5.1 Discussion

This study aimed to examine whether the inclusion of an SBL activity within a certification-oriented project management course is associated with differences in externally assessed learning outcomes. By focusing on certification examination results rather than self-reported learning gains, the study addresses a significant gap between competence-oriented educational design and predominantly knowledge-based assessment practices in professional education.

The results indicate that participants who completed the simulation-supported training achieved higher overall certification examination scores, with statistically significant differences concentrated in selected People and Practice competence elements. Although mean scores in the Perspective area were also higher in the simulation-supported group, these differences were not statistically significant. This pattern suggests that the association between simulation-supported learning and examination performance was not evenly distributed across all IPMA ICB 4.0 competence areas. Instead, the biggest differences appeared in competence elements that were more directly activated by the simulation scenario, particularly those related to teamwork, relationships and engagement, change and transformation, scope, resources, finance, and project design.

This distribution of results is consistent with the internal logic of the simulation-based learning intervention. During the simulation, participants were required to define the project scope, allocate resources, estimate and monitor costs, respond to emerging constraints, and adjust project decisions as conditions changed. These activities correspond closely to selected Practice competences, especially scope, resources, finance, project design, and change and transformation. At the same time, the collaborative and decision-oriented nature of the simulation may explain the observed differences in selected People competences, particularly teamwork, relationships, and engagement. By contrast, Perspective competences concern broader organisational, strategic, and contextual aspects of project management, such as governance, compliance, power structures, and culture. These elements were less directly represented in the simulated project environment, which may explain why differences in this area were weaker and not statistically significant.

The findings therefore suggest that simulation-based learning may be particularly useful for supporting competence-oriented learning in areas where learners must integrate technical project management decisions with interpersonal coordination and adaptive judgement. However, the results should not be interpreted as evidence that participation in the simulation-supported course was associated with higher performance across all competence areas equally. Rather, they indicate a selective association between participation in the simulation-supported course and higher performance on the competence elements most closely aligned with the decision-making tasks embedded in the simulation.

From an e-learning perspective, the study supports the view that simulation-based environments can function as a bridge between experiential learning and formal assessment systems. Although certification examinations are limited in their ability to directly capture behavioural and contextual competence, the results indicate that learning designs grounded in dynamic decision-making, feedback, and reflection may also be reflected in externally administered assessment outcomes. This is particularly relevant in professional education contexts where external certification bodies define assessment formats and cannot be easily modified by educators or training providers.

The magnitude of the observed differences should be interpreted cautiously. The overall difference in examination scores was statistically significant, but the effect size was small. This is consistent with expectations for applied educational interventions implemented in authentic professional training settings, where participants differ in prior experience, motivation, learning strategies, and professional background. The relatively small effect sizes also reinforce the need to interpret the findings as evidence of association rather than as strong evidence of causal impact.

The study contributes to the existing body of knowledge in three main ways. First, it extends research on simulation-based learning by examining externally validated certification examination outcomes rather than relying solely on learner satisfaction, perceived competence development, or in-game performance data. Second, it contributes to project management education by demonstrating how a simulation-based intervention can align with a recognised professional competence framework, namely IPMA ICB 4.0. Third, it contributes to e-learning assessment research by illustrating both the potential and the limitations of evaluating experiential learning interventions through standardised assessment systems that were not originally designed to measure behavioural competence directly.

Taken together, the findings contribute to ongoing discussions on competence development and assessment in project management education. While competence frameworks such as IPMA ICB 4.0 emphasise demonstrated ability in action, entry-level certification assessments remain largely knowledge-based. The present study illustrates a pragmatic approach to this tension by examining whether competence-oriented learning designs can be reflected, at least partially, in existing assessment systems without altering certification formats. At the same time, the selective pattern of significant results highlights the importance of carefully aligning simulation design, intended learning outcomes, and assessment evidence when evaluating simulation-based learning in professional certification contexts.

5.2 Limitations and Directions for Future Research

Several limitations of the study should be acknowledged. First, the quasi-experimental design without random assignment limits causal interpretation. Although the comparison group was restricted to the same period as the simulation-supported group to reduce the risk of historical cohort effects and temporal changes in training or examination conditions, unmeasured differences in prior project management experience, motivation, learning strategies, or professional background may have influenced the results. Therefore, the findings should be interpreted as evidence of association rather than as definitive evidence of causal impact.

Second, the study did not include a pre-test, which restricts the ability to control for baseline differences between groups. Without pre-intervention measures of project management knowledge or competence, it is not possible to determine whether the observed between-group differences resulted from the simulation-supported learning activity or from pre-existing differences between participants.

Third, the groups differed in size. Although unequal group sizes do not invalidate the analysis, they may affect the precision of estimates and reinforce the need to interpret statistical significance together with effect sizes. In this study, the overall effect size was small, and the competence-element-level effects ranged from small to moderate. These results should therefore be interpreted cautiously, with attention to the overall pattern of differences rather than to isolated statistically significant results.

Fourth, certification examination results were used as the sole outcome measure. While this choice strengthens external validity and comparability across groups, it does not capture behavioural performance during the simulation, reflective learning processes, learner engagement, or longer-term transfer of competence to professional practice. Moreover, because the IPMA Level D examination is an externally administered standardised assessment, it may underrepresent some behavioural and contextual aspects of competence that simulation-based learning is designed to develop.

Finally, because multiple competence-element-level comparisons were conducted, the results should be interpreted as exploratory and pattern-oriented rather than as confirmatory evidence for isolated competence elements. Future research could address these limitations by combining externally validated assessment outcomes with in-simulation performance data, reflective artefacts, learner feedback, or longitudinal follow-up measures. Where feasible, randomised, matched-group, or crossover designs would further strengthen causal inference. More fine-grained analysis of the relationship between specific simulation design features and particular IPMA ICB 4.0 competence elements may also provide deeper insight into the mechanisms through which simulation-based learning supports competence development.

5.3 Conclusions

This study contributes to e-learning research by demonstrating how SBL can be examined in relation to externally validated assessment outcomes within a professional certification context. By explicitly aligning simulation design with a recognised competence framework while maintaining independence from the certification assessment itself, the study offers a methodological approach for investigating experiential learning in environments governed by external evaluation standards.

The findings indicate that simulation-supported learning is associated with higher and more consistent performance on certification examinations, particularly in competence areas that require integrative judgement and interaction. While these results do not establish causality, they provide empirical support for aligning competence-oriented learning designs with formal assessment systems.

For educators and program designers, the study highlights SBL's potential to enhance certification-oriented training without modifying existing assessment frameworks. For researchers, it underscores the value of analysing technology-enhanced learning interventions through the combined lenses of learning design, competence frameworks, and assessment alignment.

AI Statement: During the preparation of this manuscript, the authors used Grammarly to enhance readability and language. The authors also used ChatGPT to support language refinement, structural revision, and preparation of responses to reviewer comments. All AI-generated suggestions were critically reviewed, edited, and approved by the authors. No AI tool was used to generate or analyse the study data. The authors remain fully responsible for the content of the manuscript.

Ethics Statement: Ethical approval was not required because the study used anonymised archival examination data. The simulation-based activity was part of regular course delivery and was not introduced solely for research purposes. The analysis did not involve identifiable personal data and had no impact on participants' certification outcomes.

References

Alvarenga, J.C., Bronco, R.R., Guedes, A.L.A., Soares, C.A. and Silva, W.S. (2020). The project manager core competencies to project success. *International Journal of Managing Projects in Business*, 13(2). Available at: https://doi.org/10.1108/IJMPB-12-2018-0274 [Accessed 30 April 2026]

Bashir, H., Maqsood, T. and Jeong, K.Y. (2021). Project managers' competencies in international development projects: A Delphi study. *International Journal of Project Management*, 39(8), pp 891–904. Available at: https://doi.org/10.1177/21582440211058188 [Accessed 20 December 2025]

Bellotti, F., Kapralos, B., Lee, K., Moreno-Ger, P. and Berta, R. (2019). Assessment in and of serious games: An overview. *Advances in Human-Computer Interaction*, pp 1–12. Available at: https://doi.org/10.1155/2013/136864 [Accessed 7 May 2025]

Crawford, L. (2005). Senior management perceptions of project management competence. *International Journal of Project Management*, 23(1), pp 7–16. Available at: https://doi.org/10.1016/j.ijproman.2004.06.005 [Accessed 8 December 2025]

Crawford, L. and Pollack, J. (2007). How generic are project management knowledge and practice? *Project Management Journal*, 38(1), pp 87–96. Available at: https://doi.org/10.1177/875697280703800109 [Accessed 8 December 2025]

de Freitas, S. (2006). Using games and simulations for supporting learning. Learning, Media and Technology, 31(4), pp 343–358. Available at: https://doi.org/10.1080/17439880601021967 [Accessed 7 May 2025]

Elenany, Y. and Ahmed, V. (2024). A Digital Simulation Game for Resource Management in Construction Projects", *Electronic Journal of e-Learning*, 22(10), pp 01-18. https://doi.org/10.34190/ejel.22.10.3495 [Accessed 10 January 2026]

Gee, J.P. (2003). *What video games have to teach us about learning and literacy*. Palgrave Macmillan

Geithner, S. and Menzel, D. (2016). Effectiveness of learning through experience and reflection in a project management simulation, *Simulation & Gaming*, 47(2), pp. 228–256. Available at: https://doi.org/10.1177/1046878115624312 [Accessed 24 February 2026]

Hellström, M.M., Jaccard, D. and Bonnier, K.E. (2023). A systematic review on the use of serious games in project management education, International Journal of Serious Games, 10(2), pp. 3–24. Available at: https://doi.org/10.17083/ijsg.v10i2.630 [Accessed 24 February 2026]

International Project Management Association (IPMA) (2015). *Individual Competence Baseline for Project, Programme and Portfolio Management (ICB 4.0)*. IPMA.

Jaccard, D., Bonnier, K.E. and Hellström, M.M. (2022). How might serious games trigger a transformation in project management education? Lessons learned from 10 years of experimentations, *Project Leadership and Society*, 3, 100047. Available at: https://doi.org/10.1016/j.plas.2022.100047 [Accessed 24 February 2026]

Koi-Akrofi, G.Y., Aboagye-Darko, D., Koi-Akrofi, J. and Boateng, D. (2024). Project management competencies: A literature-based analysis. *Journal of Business and Professional Studies*, 15(1), Available at: https://www.researchgate.net/publication/385205723_PROJECT_MANAGEMENT_COMPETENCIES_A_LITERATURE-BASED_ANALYSIS [Accessed 30 April 2026]

Le Deist, F.D. and Winterton, J. (2005). What is competence? *Human Resource Development International*, 8(1), pp 27–46. Available at: https://doi.org/10.1080/1367886042000338227 [Accessed 12 February 2026]

Leigh, I. W., Smith, I. L., Bebeau, M. J., Lichtenberg, J. W., Nelson, P. D., Portnoy, S., Rubin, N. J., and Kaslow, N. J. (2007). Models of competency assessment. *Professional Psychology: Research and Practice,* 38(5), 463–473. Available at: https://psycnet.apa.org/doi/10.1037/0735-7028.38.5.463 [Accessed 10 February 2026]

Martinelli, R. J., & Milosevic, D. Z. (2016). *Project management toolbox.* John Wiley & Sons, Inc.

Maylor, H. (2017). Understand, reduce, respond: Project complexity management. *International Journal of Operations & Production Management, 37*(8), pp 1076–1093. Available at: https://doi.org/10.1108/IJOPM-05-2016-0263 [Accessed 10 February 2026]

Michael, D.R. and Chen, S.L. (2006). *Serious games: Games that educate, train, and inform*. Thomson.

Omidvar, G., Jaryani, F., Samad, Z.B.A., Zafarghandi, S.F. and Nasab, S.S. (2011). A proposed framework for project managers' competencies and the role of e-portfolio. *International Journal of e-Education, e-Business, e-Management and e-Learning*, 1(4), pp 311–321.

Ochoa Pacheco, P., Coello-Montecel, D., Tello, M., Lasio, V. and Armijos, A. (2023). How do project managers' competencies impact project success? A systematic literature review. *PLoS ONE*, 18(12) https://doi.org/10.1371/journal.pone.0295417 [Accessed 28 April 2026]

Özsoy, T. (2025). Unveiling the potential of metaverse in project management education. *International Journal of Information Systems and Project Management*, 13(4), pp 1-27. Available at: https://doi.org/10.12821/ijispm130404 [Accessed 28 April 2026]

Plass, J. L., Homer, B. D., & Kinzer, C. K. (2015). Foundations of game-based learning. *Educational Psychologist*, 50(4), pp 258–283. Available at: https://doi.org/10.1080/00461520.2015.1122533 [Accessed 10 February 2026]

PM2PM, Ltd. (2025). Available at: https://pm2pm.pl/o-nas/ [Accessed 20 December 2025]

Project Management Institute (PMI) (2002). *Project Manager Competency Development Framework*. PMI

Rumeser, D., and Emsley, M. (2019). Can serious games improve project management decision making under complexity? *Project Management Journal*, 50(1), pp 23–39. Available at: https://doi.org/10.1177/8756972818808982 [Accessed 10 December 2025]

Shute, V. J., & Ke, F. (2012). Games, learning, and assessment. *Assessment in Education: Principles, Policy & Practice*, 19(2), pp. 105–130. Available at: https://doi.org/10.1007/978-1-4614-3546-4_4 [Accessed 10 February 2026]

Tettegah, S. Y., McCreery, M. P. and Blumberg, F. C. (2015). Designing games for learning: Cognitive, emotional, and social considerations. In R. L. Rosas & M. F. Blumberg (Eds.), *Handbook of game-based learning*. MIT Press, pp. 135–158

Trivellas, P. and Drimoussis, C. (2010). Skills and competences of project managers. In: *Proceedings of CIB 2010 – Advancing Project Management for the 21st Century*. Heraklion, pp.735–743.

Turner, J.R., Huemann, M. and Keegan, A. (2008). *Human resource management in the project-oriented organisation*. Project Management Institute

Wall, J. and Ahmed, V. (2008). Lessons learned from a case study in deploying a simulation tool for teaching project management. *Journal of Information Technology Education*, 7, pp 31–51. Available at: https://doi.org/10.1108/09699980810852691 [Accessed 10 February 2026]

Walker, A.A., Engelhard Jr., G. (2014). Game-Based Assessments: A Promising Way to Create Idiographic Perspectives. *Measurement: Interdisciplinary Research and Perspectives*, 12(1-2), pp 57-91. Available at: http://dx.doi.org/10.1080/15366367.2014.921037 [Accessed 28 April 2026]

Yaman, H., Sousa, C., Neves. P.P. and Luz F. (2024). Implementation of Game-Based Learning in Educational Contexts: Challenges and Intervention Strategies. *Electronic Journal of e-Learning*, 22(10), pp 19-36. Available at: https://doi.org/10.34190/ejel.22.10.3480 [Accessed 15 January 2026]

Early Childhood Computational Thinking through Tangible Floor-Robot Programming in an eTwinning Community of Practice

Paraskevi Foti and Tharrenos Bratitsis
Early Childhood Education Department, University of Western Macedonia, Greece

vivifoti@gmail.com (Corresponding Author)

https://doi.org/10.34190/ejel.24.3.4601

Abstract: This study explores how early-years teachers evaluate and implement Computational Thinking (CT) through tangible floor-robot activities within an eTwinning Community of Practice (CoP), with attention to gender-related participation patterns**.** Although CT is increasingly recognised as an essential dimension of Early Childhood Education (ECE - referring to ages 4-6 according to the Greek Education System), there is still limited empirical evidence on how teachers transform CT concepts into developmentally appropriate practice and how gender may shape children's engagement in such activities. Addressing these gaps, the study focuses on teachers' evaluations and classroom implementation of CT, their experiences with tangible programming tasks, and their perceptions of gender-related patterns in participation, support needs, and CT performance. The research was conducted over a 24-week asynchronous professional-learning programme hosted on Moodle and involved one national cohort of Greek early-years educators (N = 473). During the professional-learning programme, participants engaged in weekly STEAM-oriented CT activities and collaboratively designed classroom learning scenarios using floor robots (e.g., Bee-Bot). Survey instruments captured teachers' perceptions of CT, educational robotics, STEAM pedagogy, and gender-related classroom observations, while focus-group discussions provided complementary insights into classroom enactment. Following a DBR-informed exploratory mixed-methods approach, the research combined iterative engagement in the CoP with descriptive analysis of survey and focus-group data to identify patterns in teachers' experiences and reported classroom practices. Findings indicate consistently high levels of participation for both girls and boys, with only small gender differences. Boys appeared slightly more often in the highest engagement band but were also more likely to require ongoing support, whereas girls were more frequently described as working independently and showed a modest descriptive advantage in problem solving, sequencing, and simple algorithm design. In addition, the learning scenarios were rated as useful or very useful for cultivating CT and for fostering collaboration, communication, and problem solving in early-years classrooms. Qualitative findings identified three design features as particularly effective in supporting children's engagement: explicit sequencing supports, structured testing and debugging cycles, and the use of cooperative roles. Taken together, these findings underpin a set of practical learning-design principles for implementing CT through floor-robot activities in early childhood. More broadly, the research illustrates how an online CoP can support the adaptation of CT-focused designs into everyday classroom practice. It also contributes to e-learning research by illustrating how developmentally appropriate robotics activities within a Community of Practice may support equitable opportunities for young children to engage with foundational CT practices from the earliest years of schooling.

Keywords: Computational thinking, Early childhood, Educational robotics, Floor-robot programming, Tangible programming, eTwinning, Communities of practice

1. Introduction

Computational Thinking (CT) comprises analytical and algorithmic ways of thinking used to formulate, analyze, and solve problems (Aho, 2012). Since the mid-2000s, CT has gained prominence as a key twenty-first-century literacy, emphasized by international bodies such as the OECD and UNESCO (Bocconi et al., 2016; Organization for Economic Cooperation and Development, 2018; Scott, 2015). Policy frameworks position CT among essential science and engineering practices (National Research Council [NRC], 2012) and recommend systematic integration across compulsory education so that all children can participate fully in a digital society (Bocconi et al., 2016).

Parallel research strands connect CT with STEAM—often discussed as computational pedagogy (Yaşar et al., 2016). Beyond computer science, CT is treated as a transferable constellation of knowledge, skills, and dispositions that support designing solutions, creating systems, and interpreting human behavior (Hsu, Chang and Hung, 2018; Selby & Woollard, 2013; Wing, 2011). Cultivating CT from the early years is therefore a pedagogical and societal priority (Battelle for Kids, 2015; European Communities, 2007; Schola Europaea, 2018). Despite this momentum, integrating CT in ECE (4-6 y.o.) settings remains challenging and under-researched. This study addresses this gap by analyzing ECE teachers' attitudes and perceptions within an eTwinning Community of Practice (CoP) during a 24-week asynchronous training program. The study reports on teachers' perspectives on CT teaching via tangible floor-roaming robotic devices. This article forms part of a broader research

Cite this article: Foti, P. and Bratitsis, T. 2026. "Early Childhood Computational Thinking through Tangible Floor-Robot Programming in an eTwinning Community of Practice", *Electronic Journal of e-Learning*, 24(3), pp 92-105, https://doi.org/10.34190/ejel.24.3.4601

programme and focuses specifically on the gender-related dimensions of children's participation and CT performance as reported by ECE educators. The objective of this study is to examine how ECE educators evaluate and implement CT-oriented floor-roaming robotics activities within an eTwinning Community of Practice and to explore teacher-reported gender-related differences in children's participation and CT performance.

2. Literature Review

Early conceptualizations of CT frame it as the set of thought processes used to formulate problems and express solutions executable by humans or machines (Wing, 2006). Subsequent work casts CT as reasoning about information processes and algorithms (Denning, 2011) and, more broadly, as a lens on how people think about computation (Guzdial, 2008). Definitions remain plural (CSTA & ISTE, 2011), with recent reviews identifying many variants (Foti & Bratitsis, 2025; Foti, 2025). Despite this plurality, the literature converges on three complementary views: (a) CT as a way of thinking for developing solutions executable by computers or robots (Corradini, Lodi and Nardelli, 2017b; Eickelmann et al., 2019); (b) CT as a problem-solving process (Grover & Pea, 2018; Hazzan, Ragonis and Lapidot, 2020; Zhang & Nouri, 2019); and (c) CT as a transferable thinking skill applicable to real-world problems across disciplines via algorithmic methods (Israel-Fishelson et al., 2021; Román-González et al., 2019; Shute, Sun and Asbell-Clarke, 2017).

Across definitions, a recurring **Core of attributes** includes abstraction, decomposition, algorithmic thinking, automation, and generalization supported by practices such as creating computational artifacts, testing/debugging, collaboration, creativity, and working with open-ended problems (Curzon et al., 2019; Grover & Pea, 2018; Selby & Woollard, 2014). In early-years' contexts—such as those fostered within the teachers' eTwinning CoP of this study these ideas are operationalized through developmentally appropriate, play-based designs; in this study, the focus is on tangible (on-robot) programming with floor robots embedded in STEAM-oriented activities. These core CT attributes and representative definitions are summarized in Table 1.

Table 1: Core CT concepts and representative definitions from the literature

CT Attribute	Description and corresponding literature
Problem Abstraction	The process of making a problem more understandable by reducing unnecessary details and focusing on essential aspects, leading to simpler solutions (Barr & Stephenson, 2011; Lee et al., 2011; Grover & Pea, 2013; Selby & Woollard, 2013; Angeli et al., 2016; Cansu & Cansu, 2019; Huang & Looi, 2020).
Algorithmic Thinking	Thought processes for creating, expressing, executing, and evaluating step-by-step procedures to solve a problem (Barr & Stephenson, 2011; Lee et al., 2011; Grover & Pea, 2013; Selby & Woollard, 2013; Shute, Sun and Asbell-Clarke, 2017; Komm et al., 2020; Oyelere, Agbo and Oyelere, 2023).
Automation	Labor-saving execution of repetitive tasks by instructing a computer/robot to perform them quickly and efficiently (Wing, 2008; Barr & Stephenson, 2011; Lee et al., 2011; Bocconi et al., 2016).
Problem Decomposition	Breaking down a complex problem/system into smaller, manageable parts to aid understanding and solution design (Wing, 2006; Barr & Stephenson, 2011; Grover & Pea, 2013; Selby & Woollard, 2013; Angeli et al., 2016; Bocconi et al., 2016; Fried et al., 2018; Curzon et al., 2019; Upadhyaya, McGill and Decker, 2020).
Debugging	Systematic testing, tracing, and logical reasoning to predict, verify, and improve outcomes (Grover & Pea, 2013; Csizmadia et al., 2015; Angeli et al., 2016).
Generalization	Identifying patterns, similarities, and connections, and transferring solutions across related problems/contexts (Selby & Woollard, 2013; Csizmadia et al., 2015; Angeli et al., 2016; Grover & Pea, 2018; Zhang & Nouri, 2019; Hazzan, Ragonis and Lapidot, 2020).

In ECE classrooms, tangible (on-robot) programming with floor roamers offers a developmentally appropriate pathway into CT: children do not merely talk about algorithms; they enact them with their and materials. In this study, such learning designs are embedded within an eTwinning CoP and examined with attention to gender-related participation patterns. Brennan and Resnick's (2012) framework of concepts, practices, and perspectives—applied with preschoolers by Dietz et al. (2019)—helps explain how CT grows through iterative task refinement informed by evidence and experience. Within this view, testing, debugging, and feedback are not add-ons but core practices through which children appropriate CT attributes via problem solving. Complementary strands in the literature emphasize developmentally appropriate ICT integration (Bers & Flannery, 2014) and the role of play and social interaction in CT development (Grover & Pea, 2013). From a cognitive standpoint, the information-processing tradition models thinking as the reception, transformation, storage, and retrieval of information (Eysenck & Keane, 2015), a perspective that aligns well with classroom work on sequencing, monitoring actions, and revising plans. Recent systematic reviews of preschool robotics interventions similarly emphasize the importance of age-appropriate, embodied, and play-based approaches for fostering CT in early childhood settings (Bers, González-González and Armas-Torres, 2021).

A growing body of research suggests that technology-enhanced learning environments can foster CT (Manches & Plowman, 2017; González & Armas, 2019), with educational robotics highlighted as a particularly effective approach for young learners (Bers, 2019). Teachers typically employ playful and experiential strategies, emphasising collaboration, learning from errors, and the introduction of foundational CT attributes through floor roamers (e.g., Bee-Bot) or block-based platforms (Lin et al., 2020; Foti, 2023). As children program virtual or physical artefacts to perform sequences of actions, they move from passive technology use to active creation; dialogue with teachers and peers scaffolds understanding (Baker & Clark, 2010; Blatchford et al., 2003).

Gender has been another focal dimension in CT research, though findings remain mixed. Several studies attribute potential gender differences to attitudes toward technology and related self-beliefs (Stein & Nickerson, 2004), while others report associations among gender, CT proficiency, and programming self-efficacy (Lee et al., 2014; Sarıtepeci & Durak, 2017). Results vary across contexts and specific CT components: some studies report minimal or no gender effects (Aşkar & Davenport, 2009; Werner et al., 2012), whereas others identify disparities that widen with age (Román-González, Pérez-González and Jiménez-Fernández, 2017). Subskill-level analyses also present variability: boys may show stronger decomposition tendencies (Tsai, Liang and Hsu, 2020), while girls sometimes outperform boys in abstraction or computational creativity (Rijke et al., 2018; Israel-Fishelson et al., 2021). Broader cognitive and socio-cultural factors intersect with these patterns, including girls' strengths in language and memory (Duckworth & Seligman, 2006) and boys' visuospatial tendencies (Gurian & Ballew, 2003). Stereotype threat, self-efficacy, and limited teacher preparation further shape outcomes (Yücel & Rızvanoğlu, 2019; Cateté et al., 2020). These findings highlight the importance of equity-oriented pedagogy, especially in early-years CT activities.

Within this context, e-learning Communities of Practice (CoPs) serve as structured environments that support teacher learning, resource exchange, and iterative refinement of CT activities. Asynchronous participation allows flexibility and accommodates diverse schedules (Nichols, 2003; Papachristos et al., 2010). In this study, an eTwinning CoP hosted on a Moodle platform facilitated collaboration among ECE teachers through discussion forums, shared artefacts, and iterative classroom enactments. The CoP followed a 24-week STE(A)M-oriented cycle that introduces tangible programming with floor robots and incorporated collaborative eTwinning project work. Deliverables—including lesson plans, grids, and photographs—are curated into shared repositories such as a collective ebook (Foti, Tzimopoulos, & Bratitsis, 2023). Table 2 presents an exemplary Social Studies learning scenario ("What will we celebrate today?") illustrating how cultural-diversity goals are integrated with CT tasks such as sequencing, testing, and debugging. This is an exemplary learning scenario for an ECE classroom, which was designed by ECE teachers while participating training activities within the eTwinning CoP. The teachers were presented with and were obliged to design their own learning scenarios for CT, utilizing floor roamers (among other materials).

Supplementary examples of classroom materials and task designs (including sequencing activities and grid-based navigation tasks) can be found in the teaching resources associated with the CoP but are not reproduced here for brevity.

Table 2: Example Social Studies learning scenario: "What will we celebrate today?"

Element	Content
Subject	Social Studies
Subject / Study Module	Holiday traditions and cultural diversity
Objective	Recognition of festive and cultural symbols; discussion of meanings and traditions.
Summary	The scenario supports classroom discussions on holidays, traditions, and cultural diversity across seasons. Children connect holiday names to visual symbols and reflect on similarities/differences.
Procedure	(a) Class discussion on holidays throughout the school year in relation to seasons. (b) Present holiday cards; children identify and name them with teacher support. (c) Place holiday images on a card mat or under a transparent grid. (d) Children receive a holiday name, locate a matching image, trace the path with their bodies on paper, then program Bee-Bot to navigate the path, incorporating CT tasks (sequencing, testing, debugging).

Element	Content
Differentiated teaching	For advanced children, include multiple symbols and/or holidays from other countries; for children needing support, provide step cues, reduced grid size, or pair programming.
Collaborate	Children work together in small groups.
Time distribution	Multiple short iterations as needed (≈ 10–20 minutes per iteration).
Resources	Card mat or transparent grid/cardboard; holiday photos or images from magazines and catalogs; "Month Cards."
Evaluation (K–W–L–H)	Children record what they Know, what they Want to know, what they Learned, and How they will learn more (Foti & Rellia, 2023).

3. Community of Practice Context and Learning Design

The study was situated within an eTwinning CoP that supported the design, implementation, and evaluation of CT activities in ECE classrooms. The learning scenarios developed within the CoP aimed to support effective teaching and learning through tangible Floor Roamer programming activities and developmentally appropriate pedagogical practices. The pedagogical stance draws on project-based, child-centered, and collaborative learning within well-organized environments (Misirli & Komis, 2014; Foti, 2022).

The learning scenarios used in the Community of Practice were informed by the TPACK framework (Mishra & Koehler, 2006), which served as a general guide for aligning computational thinking content, developmentally appropriate pedagogy, and Floor Roamer technologies. In this study, TPACK functioned as a design reference rather than as an analytical framework.

A learning scenario in this study addresses both teacher practice and classroom activity: it specifies objectives, sequences activities, outlines forms of assessment, and offers implementation guidance. As summarised in Table 2, each scenario follows a common outline that supports coherence, differentiation, and alignment with classroom expectations.

Throughout the 24-week programme, participating ECE educators engaged in a cyclical process of implementation, reflection, adaptation, and redesign. Educators implemented CT-oriented floor roamer activities in their classrooms, shared experiences and classroom evidence through Moodle discussion forums, received peer and trainer feedback, and progressively refined or developed learning scenarios. Many of the resulting classroom scenarios were subsequently compiled into a shared collection of educational resources (Foti, Tzimopoulos, & Bratitsis, 2023).

Across the CoP cycle, participating educators designed, adapted, and implemented a range of CT-oriented classroom activities using floor robots. Typical activities included grid-based navigation tasks, route planning with arrow cards, sequencing and prediction exercises, debugging tasks, collaborative role-based programming, and movement-based path tracing. Children were encouraged to predict robot movement, test programmed sequences, identify errors, and revise commands collaboratively. These activities aimed to support foundational CT concepts such as sequencing, decomposition, problem solving, and debugging through developmentally appropriate and play-based learning experiences.

4. Methodology

4.1 Research Design

The study followed a DBR-informed exploratory mixed-methods design aimed at examining teachers' evaluations and classroom implementation of CT-oriented floor-robot activities within an eTwinning Community of Practice (CoP). The work was conducted in collaboration with practitioners participating in the CoP. This design and these instruments directly addressed RQ1 (teachers' evaluations and implementation of CT-oriented floor-robot learning scenarios) and RQ2 (teacher-reported gender-related differences in children's participation, support needs, problem solving, and sequencing/algorithmic design).

4.2 Triangulation and Data Sources

Consistent with methodological triangulation (Robson, 2007), multiple sources and methods were combined:

- a targeted literature review on STE(A)M-oriented CT and educational robotics;
- a nationwide questionnaire administered at the completion of the CoP cycle and

- focus-group discussions with eTwinning trainers.

This blend supports convergent evidence on teachers' attitudes, perceptions, and evaluations of the CoP experience.

4.3 Community of Practice Context

The CoP was implemented as a 24-week asynchronous professional-learning programme hosted on Moodle. Each week, participating ECE educators were provided with CT-oriented learning scenarios, supporting materials, and implementation guidelines. Educators implemented the activities in their classrooms with children aged 4–6 years, shared reflections and classroom evidence through discussion forums, and completed weekly evaluation activities. The forum environment supported peer exchange, collaborative reflection, and feedback on classroom implementation.

Throughout the programme, educators were encouraged to adapt existing scenarios and develop new classroom activities aligned with the programme objectives. The resulting learning scenarios incorporated Floor Roamer programming activities across different curriculum areas and emphasized sequencing, route planning, prediction, testing, debugging, and collaborative problem solving. Through iterative cycles of implementation, reflection, and feedback, educators progressively refined their classroom practices and contributed learning scenarios that were later compiled into a shared collection of educational resources (Foti, Tzimopoulos, & Bratitsis, 2023).

4.4 Instruments and Data Collection

This study forms part of a broader research programme conducted within an CoP between January and June 2022. While the wider project involved multiple strands of data collection, this paper focuses specifically on the gender-related dimensions of teachers' evaluations of floor-robot CT activities.

The primary dataset analysed in this study derives from a nationwide electronic questionnaire administered at the completion of the 24-week CoP cycle. The questionnaire examined teachers' evaluations of innovative learning designs for cultivating computational thinking through floor-robot programming activities in ECE settings. Responses were obtained from N = 473 ECE educators.

The questionnaire included 5-point Likert-type items addressing CT, educational robotics, STEAM pedagogy, children's participation, support needs, problem-solving performance, sequencing and algorithmic design, perceived CT cultivation, and teachers' evaluations of the learning scenarios. Additional items explored broader issues related to educational robotics, STEAM pedagogy, and the integration of robotics activities within eTwinning projects. For the purposes of the study, analyses focused on items related to children's participation, support needs, problem-solving performance, sequencing and algorithmic design, perceived CT cultivation, and teachers' evaluations of the learning scenarios. Items examining gender-related classroom observations were included to explore possible differences in participation and performance between girls and boys.

To complement the survey findings, two focus-group meetings were conducted in January and April 2022 with eTwinning seminar trainers (N = 39). The discussions explored classroom implementation experiences, perceived challenges during floor-robot activities, scaffolding strategies, children's participation patterns, and the contribution of collaborative design within the CoP. The focus-group data were used to contextualize, interpret, and triangulate the survey findings.

The instruments used in the study, together with their purpose and response formats, are summarized in Table 3.

Table 3: Overview of research instruments and response formats

Instrument	Purpose	Main focus areas	Response format
Questionnaire (N = 473)	End-of-cycle evaluation of classroom implementation	Participation, support needs, problem solving, sequencing and algorithmic design, usefulness of learning scenarios, perceived CT cultivation, gender-related classroom observations	5-point Likert-type items and percentage-band ratings
Focus groups (N = 39)	Qualitative exploration of implementation practices	Classroom enactment, facilitation, scaffolding, debugging, cooperative roles, participation patterns	Semi-structured group discussions

Participation- and support-related items used 5-point response scales ranging from "not at all" to "to a very high extent." Additional questionnaire items asked teachers to estimate children's observed performance using percentage bands (e.g., 0–19%, 20–39%, 40–59%, 60–79%, and 80–100%). Prior to administration, the questionnaires were reviewed by experienced eTwinning trainers to ensure clarity and relevance to ECE educational practice.

4.5 Timing

Data collection followed the DBR cycle: Questionnaire at end-of-cycle., focus groups mid-course (January and April 2022).

4.6 Participants

The second (nationwide) questionnaire was disseminated across regions to ensure broad coverage beyond the CoP's immediate locales. The highest participation rates were from Attica (21.8%) and Central Macedonia (22.5%). Regarding teaching experience, the largest subgroup reported 19–25 years (38.2%), followed by 12–18 years (31.7%). By age, most respondents were 50–59 (44.8%) and 40–49 (34.8%). In terms of education, most held a bachelor's degree (42.8%) or a master's degree (33.6%). For ICT competence, 72.3% reported Level B certification (advanced). Detailed distributions appear in Table 4.

Table 4: Baseline characteristics of the study sample (N = 473)

Variable	Category	N	%
Gender	Female	465	98.3
	Male	8	1.7
Job title	Early Childhood Education and Care (ECEC) educator	434	91.8
	Special education teacher	5	1.1
	Primary school teacher	34	7.2
School district	Region of Eastern Macedonia and Thrace	49	10.4
	Region of Attica	93	19.7
	Region of North Aegean	13	2.7
	Region of Western Greece	33	7.0
	Region of Western Macedonia	24	5.1
	Region of Thessaly	22	4.7
	Region of Ionian Islands	36	7.6
	Region of Central Macedonia	96	20.3
	Region of Central Greece	34	7.2
	Region of Southern Aegean	29	6.1
	Region of Peloponnese	21	4.4
	Region of Crete	23	4.9
Teaching experience (years)	1–5	30	6.3
	6–11	34	7.2
	12–18	150	31.7
	19–25	181	38.2
	26–30	78	16.5
Age (years)	20–29	12	2.5
	30–39	71	15.0
	40–49	165	34.9
	50–59	212	44.8
	60–65	13	2.7

Variable	Category	N	%
Highest education	Doctoral degree	13	2.7
	Bachelor's degree	198	41.9
	Second bachelor's degree	35	7.4
	Master's degree	159	33.6
	Training courses	69	14.6
ICT level	A	81	17.1
	B	127	26.8
	B1	139	29.4
	B2	77	16.3
	Training courses	49	10.4

5. Data Analysis

Survey data were exported to spreadsheets and analyzed in SPSS v. 28.0. Descriptive statistics were used to summarize distributions for most variables, and the significance level was set at $\alpha = .05$ (two-tailed) for statistical decision making. Where appropriate, associations were examined using Pearson/Spearman correlations and cross-tabulations with χ^2 tests to explore relationships among teaching experience, educational background, ICT knowledge, and attitudes toward STEAM and CT. Missing data were <5% and handled via listwise deletion.

Qualitative data from focus-group discussions were analyzed using reflexive thematic analysis. Codes were iteratively refined, and discrepancies were resolved through discussion to enhance the trustworthiness of the analysis and support triangulation with quantitative findings.

6. Findings

Overview of evidence

Quantitative results from the two questionnaires were complemented by notes from study-group meetings within the e-learning eTwinning Community of Practice (CoP). The statistics reported below are descriptive and summarize teachers' observations from classroom enactments of tangible (on-robot) CT activities with floor robots (e.g., Bee-Bot).

6.1 Participation in CT Activities (RQ2)

Participation referred to children's observable engagement in sequencing, robot navigation, prediction, testing, and debugging activities during classroom floor-robot tasks. Teachers evaluated participation using a 5-point scale ranging from "not at all" to "to a very high extent." Combining the two highest response categories, 74.9% of boys and 71.2% of girls were rated as highly engaged. Smaller shares fell into the moderate or slight bands, and very few were observed not to participate at all (boys 5.3%, girls 1.1%). Overall, the pattern suggests broadly comparable—and generally strong—participation across genders (Figure 1).

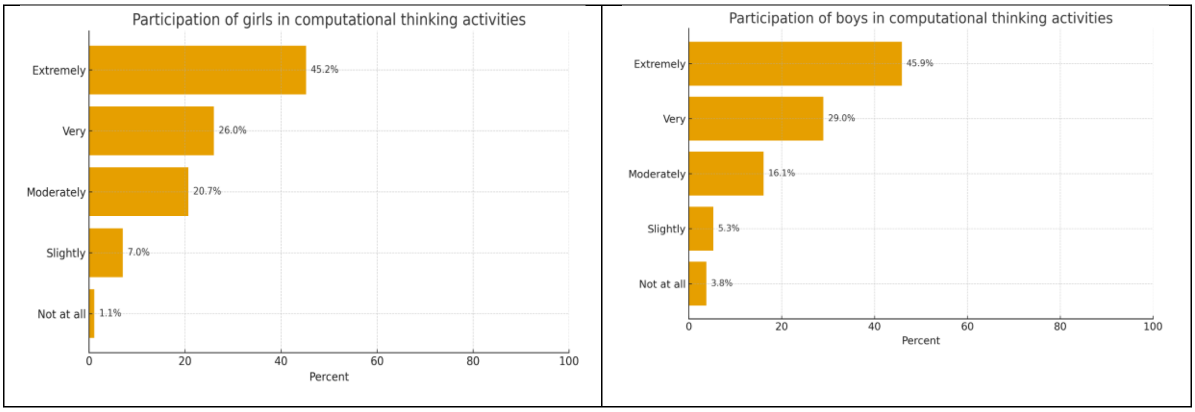


Figure 1: Teacher ratings of girls' and boys' participation in floor-robot CT activities (N = 473)

6.2 Need for Constant Support During CT Tasks (RQ2)

Support needs referred to the extent to which children required continuous teacher guidance during sequencing, navigation, testing, and debugging activities. Support included verbal prompts, modelling, step-by-step guidance, peer collaboration, and visual scaffolds. Teachers evaluated support needs using a 5-point scale ranging from "not at all" to "to a very high extent." Ratings of ongoing support needs also showed similar profiles across genders, with most children in the moderate or slight ranges. Two differences stand out:

- a larger share of girls were judged able to work without constant support (18.4% vs. 11.2% for boys),
- whereas boys were slightly more represented at the highest support level (9.1% vs. 5.9% for girls).

Taken together, these results indicate that girls were descriptively more independent in sustaining multi-step CT work, though differences were modest (Figure 2).

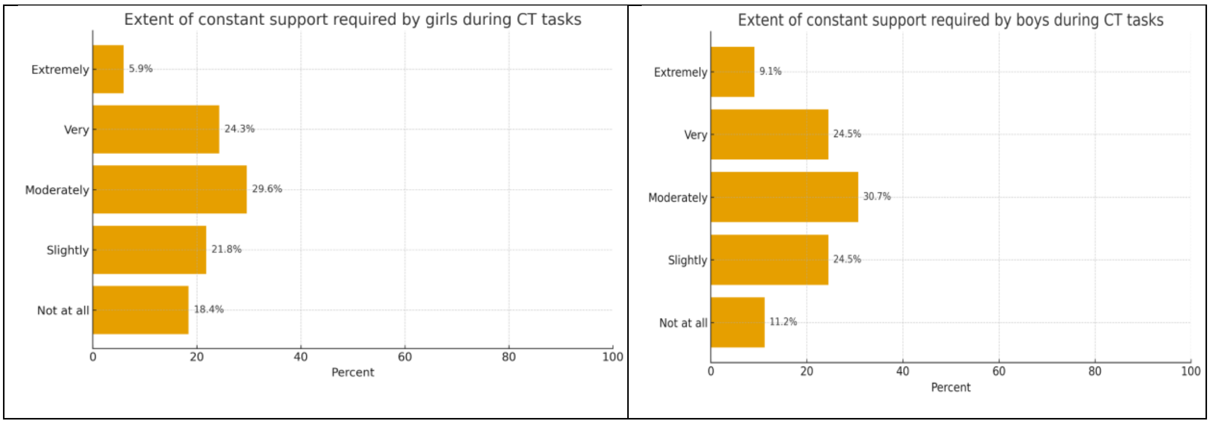


Figure 2: Teacher ratings of girls' and boys' support needs during floor-robot CT activities (N = 473)

6.3 Problem-solving Performance in CT (RQ2)

Problem-solving performance referred to children's observed ability to complete floor-robot tasks involving sequencing, prediction, testing, and correction of movement paths. Teachers estimated observed performance using percentage-band categories ranging from 0–19% to 80–100%. Distributions for computational thinking and problem-solving performance were highly similar for girls and boys. The majority in both groups clustered in the 60–79% band (boys 44.4%, girls 43.7%), with progressively fewer in lower bands. A small difference emerged at the upper end: 23.6% of girls were rated 80–100%, compared with 17.5% of boys. Descriptively, this suggests a slight edge for girls among the highest performers (Figure 3).

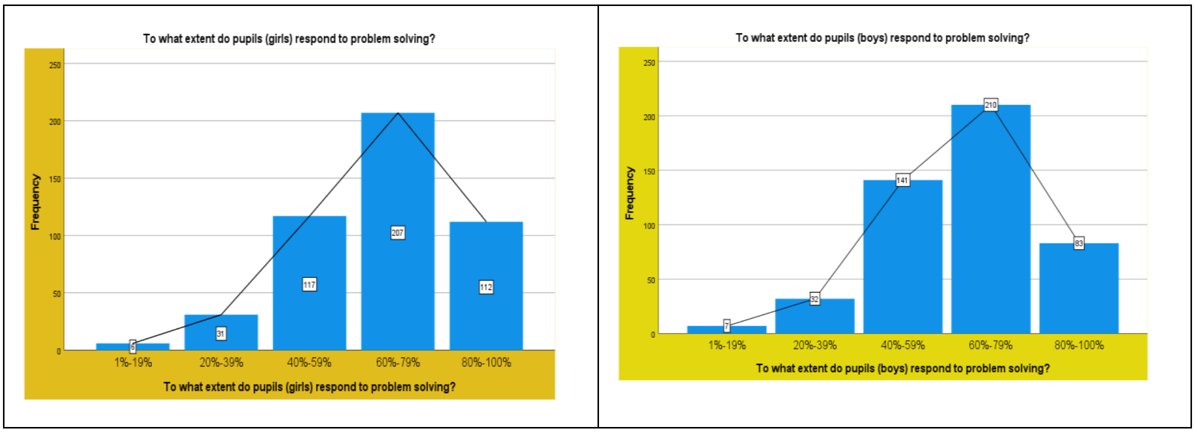


Figure 3: Teacher ratings of girls' and boys' problem-solving performance in floor-robot CT activities (N = 473)

6.4 Sequence Perception and Algorithmic Design (RQ2)

Sequence perception and algorithmic design referred to children's ability to recognize ordered steps, plan movement sequences, and revise simple command structures during floor-robot activities. Teachers evaluated these skills using percentage-band performance categories. A similar pattern appears for sequence perception and simple algorithm design. Most children again fell in the 60–79% range (boys 41.4%, girls 44.4%). At the upper end, 23.6% of girls and 19.8% of boys reached 80–100%. Overall, differences were small, with girls descriptively outperforming boys in planning, ordering steps, and refining simple algorithms (Figure 4).

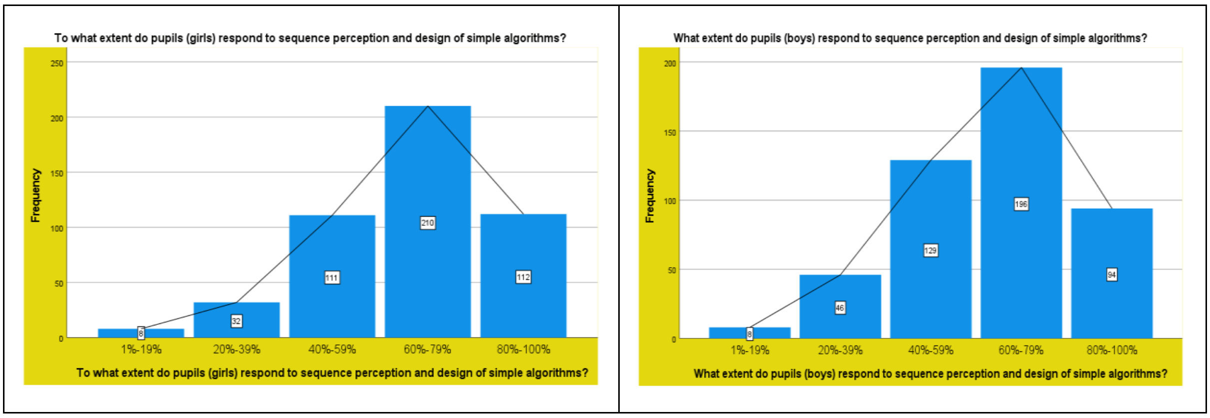


Figure 4: Teacher ratings of girls' and boys' sequencing and algorithmic-design performance in floor-robot CT activities (N = 473)

6.5 Perceived Usefulness and CT Cultivation (RQ1)

Teachers evaluated the usefulness of the learning scenarios in relation to classroom implementation, children's engagement, and perceived support for CT development. Usefulness ratings were based on 5-point Likert-type response categories ranging from "not useful at all" to "very useful." Evaluative items regarding the usefulness of the eTwinning CoP learning scenarios and their classroom impact were strongly positive. Across N = 473 teachers: 93.2% judged the scenarios quite/very useful, 96.0% reported that the floor roamer activities effectively cultivated CT skills. Only 6.8% selected "somewhat useful," and 4.0% indicated only moderate/slight cultivation. These converging ratings suggest that CoP-based designs were perceived as useful and supportive for CT development (Figures 5-6).

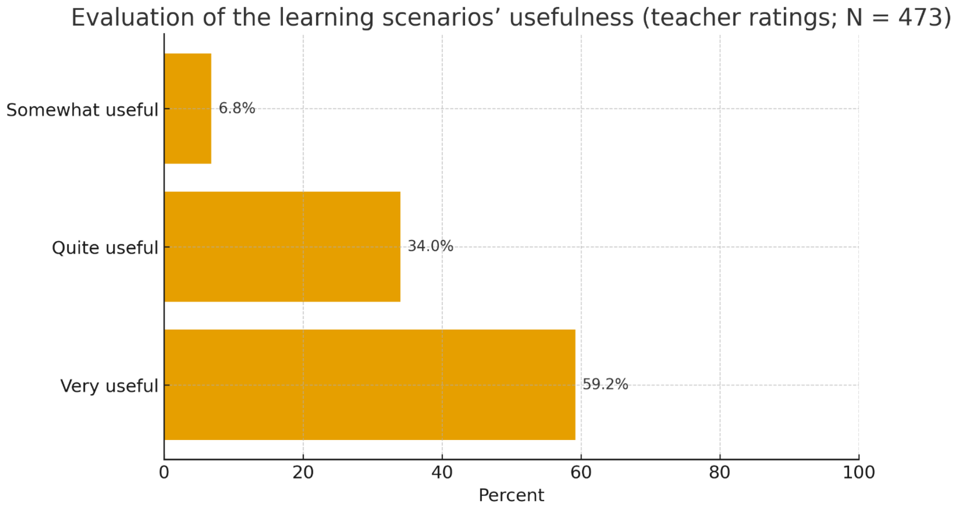


Figure 5: Teacher evaluations of the usefulness of floor-robot CT learning scenarios (N = 473)

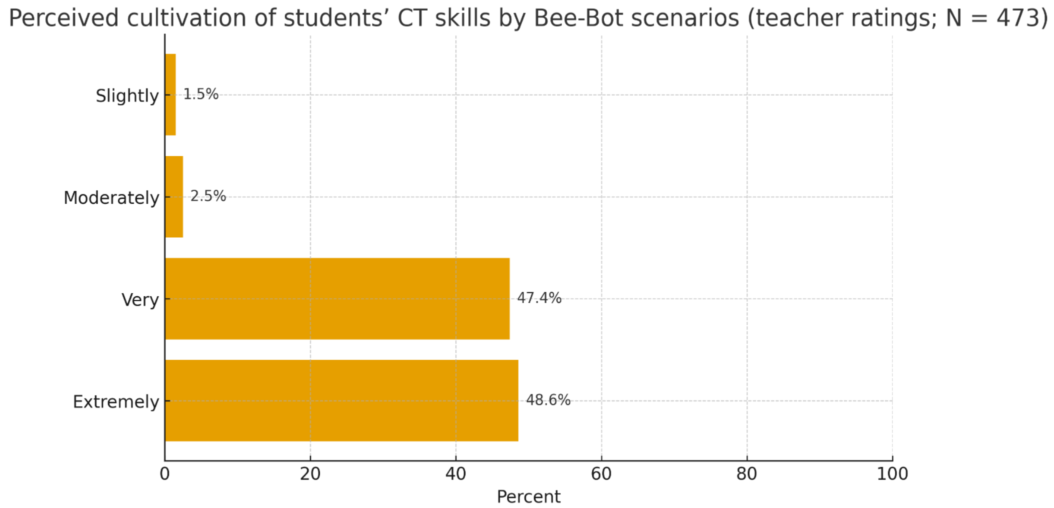


Figure 6: Teacher perceptions of children's CT cultivation through floor-robot learning scenarios (N = 473)

6.6 Qualitative Synthesis (RQ1 & RQ2)

Qualitative findings from the focus-group discussions (N = 39) complemented the questionnaire results by providing additional insights into classroom implementation practices and teacher facilitation strategies. Study-group discussions converged on three recurring design strategies that appeared to sustain engagement and reduce reliance on constant adult support: explicit sequencing scaffolds (e.g., step cards and path previews), structured testing and debugging cycles embedded within activities, and cooperative role assignment (e.g., navigator/programmer/checker).

Participants reported that sequencing supports were particularly valuable for children who initially experienced difficulty organizing movement commands and planning routes. Visual prompts and route-planning activities appeared to reduce uncertainty and facilitate participation during Floor Roamer tasks.

Participants also emphasized the importance of testing and debugging activities. Opportunities to test commands, identify errors, and revise routes encouraged children to reflect on their decisions and engage in problem-solving processes rather than simply seeking correct answers.

Finally, participants highlighted the value of cooperative role assignment. Structured responsibilities appeared to increase participation opportunities, promote collaboration, and reduce the likelihood that individual children would dominate the activity.

Teachers further emphasized the CT concepts most salient in practice—sequencing, decomposition, and debugging—informing the learning-design principles discussed later. Descriptively, these findings address RQ2 by showing small gender differences, with a modest advantage for girls in the upper performance bands and greater independence during CT work. They address RQ1 by indicating that CoP-designed tangible floor-robot activities were perceived as useful and effective for cultivating CT in ECE settings.

7. Discussion

This study investigated an e-learning eTwinning Community of Practice (CoP) in which early-years teachers co-designed and enacted classroom-ready learning scenarios to cultivate computational thinking (CT) through tangible floor-robot activities (Bee-Bot). Across classroom enactments, teachers consistently rated the scenarios as usable and educationally valuable and reported that the blend of unplugged and on-robot tasks supported core CT aims while engaging collaboration, communication, critical thinking, and creativity in routine practice. These findings reinforce existing evidence that developmentally appropriate robotics activities, when embedded in structured learning designs, can meaningfully support CT in early childhood settings.

The results also align with prior research on CoP-based professional learning, indicating that iterative plan–teach–reflect cycles and shared design templates (e.g., sequencing, debugging, and cooperative roles) enable teachers to translate pedagogical intentions into classroom action. Within this CoP, teachers appeared to leverage these design elements in ways that increased task clarity and reduced cognitive load for young children, which may explain the high engagement levels reported across gender groups. Similar findings regarding the importance of structured professional learning for supporting early-years teachers' implementation of robotics and CT activities have also been reported in recent teacher-development research (Yang, 2025). These observations are also consistent with research emphasizing the importance of teacher perspectives and pedagogical support structures in educational robotics implementation (Chevalier et al., 2021).

Descriptively, gender differences were small across outcomes derived from teacher ratings. Girls and boys exhibited high participation overall; boys appeared somewhat more represented at the highest engagement band yet simultaneously more likely to require ongoing support, whereas girls were more often described as working independently and showed a modest advantage in the upper performance bands for problem solving and for sequence/algorithm design. These results cohere with previous reports of slight female advantages in specific CT dimensions in early primary years and with evidence that sustained engagement functions as a lever for opportunity to learn in CT-rich tasks. While not inferential, these descriptive trends suggest that gender differences, where present, may be context-sensitive and shaped by scaffolding, task structure, and opportunities for supported autonomy.

Perceived usefulness of the learning scenarios and perceived cultivation of CT moved together in the expected direction—most teachers rated both highly—suggesting a clear descriptive association between classroom readiness of designs and their perceived impact on CT within the CoP. In practical terms, classroom-tested scenarios that include explicit sequencing supports, structured testing/debugging cycles, and cooperative role assignment appear to broaden participation while reducing reliance on constant adult guidance. This echoes the broader literature showing that well-scaffolded, embodied, and collaborative CT activities can mitigate barriers to participation, including gender-linked differences in confidence or perceived competence.

7.1 Limitations and Implications

The evidence synthesized here is teacher-reported and descriptive; participants were volunteers in a national CoP and may therefore represent a more motivated or pedagogically progressive subset. No direct child assessments or inferential statistics are presented, and the reliance on self-report may introduce expectancy or confirmation biases. Future work should triangulate teacher ratings with learner performance measures, analyze classroom interactions in situ, examine persistence of learning over time, and experimentally compare alternative design variants (e.g., different scaffolding intensities or cooperation structures).

Even so, the present results provide actionable direction for early-years practice and professional learning. First, embedding CT development within e-learning CoPs appears to support teacher implementation fidelity and collective capacity-building. Second, retaining explicit sequencing and debugging scaffolds in tangible floor-robot tasks helps stabilize task demands for young children. Third, structuring cooperative roles can balance participation and encourage sustained engagement. Finally, routine monitoring of engagement and support needs by gender may serve as an important formative assessment practice for ensuring equitable CT learning opportunities.

8. Conclusion

This study offers evidence that an e-learning eTwinning Community of Practice (CoP) can serve as a valuable mechanism for supporting early-years teachers in designing and enacting developmentally appropriate computational thinking (CT) activities with tangible floor robots. Through iterative cycles of co-design, classroom enactment, and reflection, teachers produced learning scenarios that were consistently perceived as usable, pedagogically appropriate, and effective in cultivating foundational CT skills. The strong alignment between scenario design features—explicit sequencing supports, structured testing/debugging cycles, and cooperative roles—and teachers' positive evaluations underscores the importance of intentional scaffolding when introducing CT in early childhood settings.

Across gender, participation and performance patterns were broadly similar, with only modest differences favoring girls in independence and upper-band problem-solving outcomes. While descriptive, these results suggest that well-structured, play-based CT activities may help neutralize gender disparities often reported in later educational stages, reinforcing the value of early and equitable exposure to CT-rich learning environments.

The findings also highlight the broader contribution of e-learning CoPs to professional learning in early childhood education. By providing a sustained, collaborative, and practice-centered space, the CoP enabled teachers to refine their pedagogical reasoning, deepen their understanding of CT, and translate technological tools into meaningful classroom action. This model holds promise for scaling CT integration across diverse early-years contexts.

Future work should complement teacher-reported data with direct learner assessments, longitudinal analyses, and experimental comparisons of alternative design elements. Nonetheless, the present study offers a clear and actionable message: when early-years educators are supported through structured, collaborative professional learning, computational thinking can be cultivated effectively, equitably, and playfully—from the very first years of schooling.

AI Statement: AI tools were used only for language editing, proofreading, and final compliance-checking support. The authors reviewed, verified, and take full responsibility for the final content of the paper.

Ethics Statement: Ethical approval was not required for this study. All participants provided informed consent by proceeding after reading a statement on the study's purpose, anonymity, and voluntary nature of participation. The study followed the principles of voluntary participation, anonymity, confidentiality, and responsible data handling.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

Aho, A.V. (2012). 'Computation and Computational Thinking.' *The Computer Journal*, 55:832–835.

Angeli, C., Voogt, J., Fluck, A., Webb, M., Cox, M., Malyn-Smith, J. and Zagami, J., (2016). 'A K-6 Computational Thinking Curriculum Framework: Implications for Teacher Knowledge.' *Educational Technology & Society*, 19(3):47–57.

Askar, P. and Davenport, D. (2009). 'An Investigation of Factors Related to Self-Efficacy for Java Programming among Engineering Students.' *Online Submission*, 8(1).

Baker, T. and Clark, J. (2010). 'Cooperative Learning—A Double-Edged Sword: A Cooperative Learning Model for Use with Diverse Student Groups.' *Intercultural Education*, 21:257–268.

Barr, V. and Stephenson, C. (2011). 'Bringing Computational Thinking to K–12: What Is Involved and What Is the Role of the Computer Science Education Community?' *ACM Inroads*, 2(1):48–54.

Bers, M.U., González-González, C. and Armas-Torres, M.B. (2021). 'Preschool Children, Robots, and Computational Thinking: A Systematic Review.' *Computers & Education*, 164:104123. https://doi.org/10.1016/j.compedu.2020.104123

Blatchford, P., Kutnick, P., Baines, E. and Galton, M. (2003). 'Toward a Social Pedagogy of Classroom Group Work.' *International Journal of Educational Research*, 39:153–172.

Bocconi, S., Chioccariello, A., Dettori, G., Ferrari, A., Engelhardt, K., Kampylis, P. and Punie, Y. (2016). *Developing Computational Thinking in Compulsory Education*. JRC Science for Policy Report. Luxembourg: Publications Office of the European Union.

Bocconi, S., Chioccariello, A. and Kampylis, P. (2022). *Reviewing Computational Thinking in Compulsory Education: State of Play and Practices from Computing Education*. Luxembourg: Publications Office of the European Union. Available at: https://data.europa.eu/doi/10.2760/126955

Cansu, F.K. and Cansu, S.K. (2019). 'An Overview of Computational Thinking.' *International Journal of Computer Science Education in Schools*, 3(1):17–30.

Carr, E. and Ogle, D. (1987). 'KWL Plus: A Strategy for Comprehension and Summarization.' *Journal of Reading*, 30:626–631.

Cateté, V., Alvarez, L., Isvik, A., Milliken, A., Hill, M. and Barnes, T. (2020). 'Aligning Theory and Practice in Teacher Professional Development for Computer Science.' In *Koli Calling '20: Proceedings of the 20th Koli Calling International Conference on Computing Education Research*, pp. 1–11. https://doi.org/10.1145/3428029.3428560

Chevalier, M., El-Hamamsy, L., Giang, C., Bruno, B. and Mondada, F. (2021). 'Teachers' Perspective on Fostering Computational Thinking through Educational Robotics.' *Education and Information Technologies*, 26:1–22. https://doi.org/10.1007/s10639-021-10455-7

Corradini, I., Lodi, M., & Nardelli, E. (2017b). Conceptions and Misconceptions about Computational Thinking among Italian Primary School Teachers. Proceedings of the 2017 ACM Conference on International Computing Education Research, 136–144. https://doi.org/10.1145/3105726.3106194

Csizmadia, A., Curzon, P., Dorling, M., Humphreys, S., Ng, T., Selby, C. and Woollard, J. (2015). *Computational Thinking: A Guide for Teachers*. Swindon, UK: Computing at School.

CSTA and ISTE (2011). *Operational Definition of Computational Thinking for K-12 Education*. Available at: http://www.iste.org/docs/pdfs/Operational-Definition-of-Computational-Thinking.pdf

Curzon, P., Bell, T., Waite, J. and Dorling, M. (2019). 'Computational Thinking.' In Fincher, S.A. and Robins, A.V. (eds.) *The Cambridge Handbook of Computing Education Research*. Cambridge: Cambridge University Press, pp. 513–546. https://doi.org/10.1017/9781108654555.018

Durak, H. and Sarıtepeci, M. (2017). 'Investigating the Effect of Technology Use in Education on Classroom Management within the Scope of the FATİH Project.' *Çukurova Üniversitesi Eğitim Fakültesi Dergisi*, 46(2):441–457. https://doi.org/10.14812/cuefd.303511

Duckworth, A.L. and Seligman, M.E.P. (2006). 'Self-Discipline Gives Girls the Edge: Gender in Self-Discipline, Grades, and Achievement Test Scores.' *Journal of Educational Psychology*, 98(1):198–208.

Eickelmann, B., Bos, W., Gerick, J., Goldhammer, F., Schaumburg, H., Schwippert, K., Senkbeil, M., & Vahrenhold, J. (2019). *ICILS 2018 #Deutschland: Computer- und informationsbezogene Kompetenzen von Schülerinnen und Schülern im zweiten internationalen Vergleich und Kompetenzen im Bereich Computational Thinking.* Münster: Waxmann Verlag.

Foti, P., Tzimopoulos, N. and Bratitsis, T. (2023). *Educational Robotics in Kindergarten with Bee-Bot* [E-book]. eTwinning Seminars. ISBN 978-618-85916-4-6. Available at: https://www.etwinning.gr/etwinning/publications/1318-beebot

Foti, P. and Bratitsis, T. (2025). 'The Importance of Learning and Practice Communities in Education: Strengthening Collaboration, Professional Growth, and Innovation: Educators' Perspectives in Greece.' In *Proceedings of the 19th International Technology, Education and Development Conference*. Valencia, Spain, pp. 6465–6474. https://doi.org/10.21125/inted.2025.1668

Foti, P. (2025). 'Using Bee-Bot to Foster Computational Thinking: A Model Lesson Plan for the Pedagogical Integration of Educational Robotics.' In Foti, P. and Tzimopoulos, N. (eds.) *Proceedings of the eTwinning Online Conference: Using Bee-Bot in the Classroom and in Collaborative School Projects in Primary and Secondary Education*. Athens: Hellenic National eTwinning – ITYE Diofantos, pp. 7–17. ISBN 978-618-87865-1-8.

Foti, P. (2023). 'Educational Robotics and Computational Thinking in Early Childhood—Linking Theory to Practice with ST(R)EAM Learning Scenarios.' *European Journal of Open Education and E-Learning Studies*. https://doi.org/10.46827/ejoe.v8i1.4677

Foti, P. (2022). 'Integrating Educational Robotics with the Bee-Bot into the Development of ST(R)EAM Pedagogical Frameworks for Early Childhood Education.' In *Proceedings of the 8th Online eTwinning Conference*. Athens: Hellenic National eTwinning – ITYE Diofantos, pp. 550–559. ISBN 978-618-85916-3-9.

Foti, P. (2021). 'Exploring Kindergarten Teachers' Views on STEAM Education and Educational Robotics: Dilemmas, Possibilities, Limitations.' *Advances in Mobile Learning Educational Research*, 1(2):82–95. https://doi.org/10.25082/AMLER.2021.02.004

Foti, P., Kosyvas, G., Katopodi, M., Pantelopoulou, S., Papazissi, C., Patrinopoulos, M. and Zografou, E. (2021). *Robogirls: Comprehensive Guide for Educators*. Regional Directorate for Primary and Secondary Education.

Fried, D., Legay, A., Ouaknine, J. and Vardi, M.Y. (2018). 'Sequential Relational Decomposition.' In *Proceedings of the 33rd Annual ACM/IEEE Symposium on Logic in Computer Science (LICS)*, pp. 432–441.

Grover, S. and Pea, R. (2013). 'Computational Thinking in K–12: A Review of the State of the Field.' *Educational Researcher*, 42(1):38–43.

Gurian, M. and Ballew, A. (2003). *The Boys and Girls Learn Differently: Action Guide for Teachers*. San Francisco: Jossey-Bass.

Halpern, D.F., Benbow, C.P., Geary, D.C., Gur, R.C., Hyde, J.S. and Gernsbacher, M.A. (2007). 'The Science of Sex Differences in Science and Mathematics.' *Psychological Science in the Public Interest*, 8(1):1–51.

Hazzan, O., Ragonis, N. and Lapidot, T. (2020). 'Computational Thinking.' In *Guide to Teaching Computer Science*. 2nd ed. Cham: Springer. https://doi.org/10.1007/978-3-030-39360-1_4

Hsu, T.-C., Chang, S.-C. and Hung, Y.-T. (2018). 'How to Learn and How to Teach Computational Thinking: Suggestions Based on a Review of the Literature.' *Computers & Education*, 126:296–310. https://doi.org/10.1016/j.compedu.2018.07.004

Huang, W. and Looi, C.-K. (2020). 'A Critical Review of Literature on "Unplugged" Pedagogies in K–12 Computer Science and Computational Thinking Education.' *Computer Science Education*, pp. 1–29. https://doi.org/10.1080/08993408.2020.1789411

Hyde, J.S., Fennema, E. and Lamon, S.J. (1990). 'Gender Differences in Mathematics Performance: A Meta-Analysis.' *Psychological Bulletin*, 107(2):139–155.

Israel-Fishelson, R., Hershkovitz, A., Eguíluz, A., Garaizar, P. and Guenaga, M. (2021). 'The Associations between Computational Thinking and Creativity: The Role of Personal Characteristics.' *Journal of Educational Computing Research*, 58(8):1415–1447. https://doi.org/10.1177/0735633120940954

Komis, V., Tzavara, A., Karsenti, T., Collin, S. and Simard, S. (2013). 'Educational Scenarios with ICT: An Operational Design and Implementation Framework.' In McBride, R. and Searson, M. (eds.) *Proceedings of SITE 2013 – Society for Information Technology & Teacher Education International Conference*. New Orleans, Louisiana, United States: Association for the Advancement of Computing in Education, pp. 3244–3251. Available at: https://www.learntechlib.org/primary/p/48594/

Komm, D., Hauser, U., Matter, B., Staub, J. and Trachsler, N. (2020). 'Computational Thinking in Small Packages.' In Kori, K. and Laanpere, M. (eds.) *Lecture Notes in Computer Science*, 12518. Cham: Springer, pp. 170–181. https://doi.org/10.1007/978-3-030-63212-0_14.

Lee, I., Martin, F., Denner, J., Coulter, B., Allan, W., Erickson, J. and Werner, L. (2011). 'Computational Thinking for Youth in Practice.' *ACM Inroads*, 2(1):32–37.

Li, Q., Jiang, Q., Liang, J.-C., Xiong, W. and Zhao, W. (2024). 'Engagement Predicts Computational Thinking Skills in Unplugged Activity: Analysis of Gender Differences.' *Thinking Skills and Creativity*, 52:101537.

Lin, S.-Y., Chien, S.-Y., Hsiao, C.-L., Hsia, C.-H. and Chao, K.-M. (2020). 'Enhancing Computational Thinking Capability of Preschool Children by Game-Based Smart Toys.' *Electronic Commerce Research and Applications*, 44:101011.

Marzano, R.J., Pickering, D.J. and Pollock, J.E. (2005). *Classroom Instruction That Works*. Upper Saddle River, NJ: Pearson Education.

Mishra, P. and Koehler, M.J. (2006). 'Technological Pedagogical Content Knowledge: A Framework for Teacher Knowledge.' *Teachers College Record*, 108(6):1017–1054.

Misirli, A. and Komis, V. (2014). 'Robotics and Programming Concepts in Early Childhood Education: A Conceptual Framework for Designing Educational Scenarios.' In *Research on e-Learning and ICT in Education*. New York: Springer, pp. 99–118.

National Research Council (2012). *A Framework for K–12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*. Washington, DC: National Academies Press.

Nichols, M.A. (2003). 'A Theory for eLearning.' *Educational Technology & Society*, 6:1–10.

OECD (2018). *The Future of Education and Skills: Education 2030*. Paris: OECD.

Oyelere, S.A., Agbo, F.J. and Oyelere, S.S. (2023). 'Formative Evaluation of Immersive Virtual Reality Expedition Mini-Games to Facilitate Computational Thinking.' *Computers & Education: X Reality*, 2:100016.

Papachristos, D., Alafodimos, N., Arvanitis, K., Vassilakis, K., Kalogiannakis, M., Kikilias, P. and Zafeiri, E. (2010). 'An Educational Model for Asynchronous e-Learning: A Case Study in Higher Technology Education.' *International Journal of Advanced Corporate Learning*, 3(1):32–36.

Prottsman, C.L.L. (2011). 'Computational Thinking and Women in Computer Science.' PhD thesis, University of Oregon.

Rijke, W.J., Bollen, L., Eysink, T.H.S. and Tolboom, J.L.J. (2018). 'Computational Thinking in Primary School: An Examination of Abstraction and Decomposition in Different Age Groups.' *Informatics in Education*, 17(1):77–92.

Robson, C. (2007). *How to Do a Research Project: A Guide for Undergraduate Students*. Oxford: Blackwell.

Román-González, M., Pérez-González, J.C. and Jiménez-Fernández, C. (2017). 'Which Cognitive Abilities Underlie Computational Thinking? Criterion Validity of the Computational Thinking Test.' *Computers in Human Behavior*, 72:678–691. https://doi.org/10.1016/j.chb.2016.08.047

Rose, S.A., Feldman, J.F., Jankowski, J.J. and Van Rossem, R. (2012). 'Information Processing from Infancy to 11 Years: Continuities and Prediction of IQ.' *Intelligence*, 40(5):445–457. https://doi.org/10.1016/j.intell.2012.05.007

Saad, A., Elbashir, A., Abdou, R., Alkhair, S., Ali, R., Parangusan, H., Ahmad, Z. and Al-Thani, N.J. (2024). 'Exploring the Gender Variations in 4Cs Skills among Primary Students.' *Thinking Skills and Creativity*, 52:101510.

Scott, C.L. (2015). *The Futures of Learning 2: What Kind of Learning for the 21st Century?* Education Research and Foresight Working Papers 3. Paris: UNESCO.

Selby, C.C. and Woollard, J. (2013). *Computational Thinking: The Developing Definition*. Southampton: University of Southampton.

Shute, V.J., Sun, C. and Asbell-Clarke, J. (2017). 'Demystifying Computational Thinking.' *Educational Research Review*, 22:142–158. https://doi.org/10.1016/j.edurev.2017.09.003

Tsai, M.-J., Liang, J.-C. and Hsu, C.-Y. (2020). 'The Computational Thinking Scale for Computer Literacy Education.' *Journal of Educational Computing Research*, 59(4):579–602. https://doi.org/10.1177/0735633120972356

Upadhyaya, B., McGill, M.M. and Decker, A. (2020). 'A Longitudinal Analysis of K–12 Computing Education Research in the United States: Implications and Recommendations for Change.' In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education (SIGCSE '20)*. New York: Association for Computing Machinery, pp. 605–611. https://doi.org/10.1145/3328778.3366809

Wang, F. and Hannafin, M.J. (2005). 'Design-Based Research and Technology-Enhanced Learning Environments.' *Educational Technology Research and Development*, 53(4):5–23.

Werner, L., Denner, J., Campe, S. and Kawamoto, D.C. (2012). 'The Fairy Performance Assessment: Measuring Computational Thinking in Middle School.' In *Proceedings of the 43rd ACM Technical Symposium on Computer Science Education (SIGCSE)*, pp. 215–220.

Wing, J.M. (2006). 'Computational Thinking.' *Communications of the ACM*, 49(3):33–35.

Wing, J.M. (2008). 'Computational Thinking and Thinking about Computing.' *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 366(1881):3717–3725.

Wu, B. and Su, H.-H. (2021). 'Gender Differences in Primary School Students' Computational Thinking Skills in Unplugged Programming Activities.' *Journal of Educational Computing Research*, 59(2):303–326.

Yang, W. (2025). 'A Three-Phase Professional Development Approach to Introducing Robot Programming and Computational Thinking in Early Childhood Education.' *Computers & Education*, 216:105040. https://doi.org/10.1016/j.compedu.2025.105040

Yasar, O., Veronesi, P., Maliekal, J., Little, L.J., Vattana, S.E. and Yeter, I.H. (2016). 'Computational Pedagogy: Fostering a New Method of Teaching.' Presented at: *ASEE Annual Conference and Exposition*, June 2016.

Yücel, Y. and Rizvanoğlu, K. (2019). 'Battling Gender Stereotypes: A User Study of a Code-Learning Game, "Code Combat," with Middle School Children.' *Computers in Human Behavior*, 99:352–365.

Zhang, L. and Nouri, J. (2019). 'A Systematic Review of Learning Computational Thinking through Scratch in K–9.' *Computers & Education*, 141:103607. https://doi.org/10.1016/j.compedu.2019.103607

Quantile-based e-Learning Student Engagement Classification

Aditya Galih Sulaksono[1,2], Syaad Patmanthara[1] and Harits Ar Rosyid[1]
[1]State University of Malang, Indonesia
[2]Universitas Merdeka Malang, Indonesia

aditya.galih.2205349@students.um.ac.id
syaad.ft@um.ac.id
harits.ar.ft@um.ac.id

https://doi.org/10.34190/ejel.24.3.4678

Abstract: Classifying student engagement accurately is critical for timely academic intervention; however, most existing approaches rely on arbitrarily defined thresholds that lack statistical grounding and are difficult to transfer across institutional contexts. This limitation reduces the practical applicability of engagement analytics in diverse educational settings. This study evaluates a quantile-based engagement classification framework across two contrasting datasets to assess its validity, transferability, and consistency of predictive features. Unlike threshold-based approaches, the proposed framework derives engagement categories directly from dataset-specific interaction distributions. The Open University Learning Analytics Dataset (OULAD) represents large-scale fully online learning, while the Unistudium dataset reflects a smaller blended learning context. The two datasets differ substantially in size and delivery mode, with a student ratio of approximately 17.4 to 1. This contrast provides a rigorous basis for assessing method transferability. Engagement categories (passive, moderate, and active) are derived using dataset-specific quartile thresholds (Q1 and Q3). This strategy adapts automatically to local interaction distributions and avoids manual parameter tuning. Five temporal behavioural features were extracted, including active days, unique actions, and learning consistency. Random Forest was employed as the proposed model, while a Decision Tree classifier was included as a baseline for comparative evaluation. The results indicate that the proposed framework remains effective across different educational contexts. In the OULAD dataset, the model achieved an accuracy of 92.04% with a Cohen's κ of 0.87. In the Unistudium dataset, accuracy reached 72.50% with a Cohen's κ of 0.59. Although performance differed between datasets, variance remained low. Feature importance analysis further revealed strong consistency across contexts, with a Spearman correlation of 0.90. Active days and unique actions were the most influential predictors in both cases. The baseline comparison further confirmed the superiority of Random Forest over the Decision Tree baseline across both datasets. These findings support e-learning practice by offering institutions a statistically grounded and automated method for engagement classification. The approach removes the need for arbitrary thresholds and reduces operational overhead in analytics deployment. From a research perspective, the study establishes realistic performance benchmarks for engagement analytics at different institutional scales, demonstrates the applicability of quantile-based engagement classification across heterogeneous datasets, and confirms that key behavioural engagement indicators transfer reliably across online and blended learning environments.

Keywords: Learning analytics, Student engagement, Quantile classification, Cross-dataset validation, Random forest, Educational data mining

1. Introduction

When engagement classification fails, the cost is borne by students. Misclassifying a disengaged student as moderately active delays academic intervention, while misclassifying an active student as struggling wastes limited tutoring resources. At institutional scale, even modest error rates translate into hundreds of missed support opportunities each semester. Digital learning platforms generate extensive interaction data that can inform student support, intervention, and learning analytics applications (Xu *et al.*, 2025). However, converting these logs into meaningful engagement categories remains methodologically problematic. A central issue lies in how engagement thresholds are defined and validated. Many studies continue to rely on fixed interaction counts that lack clear statistical grounding (Conijn *et al.*, 2017). As a result, engagement definitions vary widely across the literature, making comparison and replication difficult (Howard, Meehan and Parnell, 2018). These rigid thresholds are particularly fragile in cross-institutional settings. A boundary that performs well in a large research-intensive university may be inappropriate for a smaller, teaching-oriented institution.

Alternative analytical approaches introduce different limitations. Unsupervised clustering methods often produce results that are difficult to interpret and insufficiently stable for operational use (Howard, Meehan and Parnell, 2018). Supervised classification methods, by contrast, require extensive manual labelling and tend to perform poorly when applied beyond their original context (Brinton and Chiang, 2015). Both approaches therefore present challenges for institutions seeking scalable and transferable engagement analytics.

Cite this article: Sulaksono, A.G., Patmanthara, S. and Ar Rosyid, H. 2026. "Quantile-based e-Learning Student Engagement Classification", *Electronic Journal of e-Learning*, 24(3), pp 106-122, https://doi.org/10.34190/ejel.24.3.4678

Quantile-based classification offers a promising alternative. By defining engagement categories using distribution-derived thresholds calculated automatically from each dataset, this approach provides a statistically grounded basis for classification (Xu *et al.*, 2025). Thresholds are computed directly from local interaction patterns, allowing the method to adapt automatically to different datasets. This property reduces the need for manual tuning and supports consistent interpretation across contexts.

Despite these advantages, evidence for the cross-context validity of quantile-based engagement classification remains limited. Most existing studies evaluate performance within a single dataset or institutional setting (Rizvi *et al.*, 2022). As a result, it is unclear whether reported performance levels can be expected to transfer to institutions with different scales, delivery modes, or student populations.

To address this gap, the present study evaluates quantile-based engagement classification using two contrasting datasets. The Open University Learning Analytics Dataset (OULAD) represents large-scale distance education, while the Unistudium dataset reflects a smaller blended learning environment. The two contexts differ markedly in institutional setting, delivery model, and scale, with a student ratio of approximately 17.4 to 1. This contrast provides a rigorous basis for examining method generalisability.

The quantile-based approach is particularly suited to this validation task. Quartile thresholds are calculated independently for each dataset, eliminating the need to transfer thresholds between contexts. This automatic adaptation is a central methodological strength. If comparable classification performance and feature importance patterns emerge despite large differences in interaction volume and institutional scale, this would indicate that the underlying engagement constructs are robust.

Accordingly, this study addresses three research questions. First, can quantile-based engagement classification achieve acceptable accuracy across different educational contexts when thresholds adapt automatically to local data distributions? Second, do feature importance patterns remain stable across datasets with substantial differences in scale and institutional characteristics? Third, how do engagement category characteristics differ between large-scale distance education and smaller blended learning environments?

Answering these questions contributes to both theory and practice in learning analytics. First, the study provides empirical evidence regarding the cross-context validity of quantile-based engagement classification. Second, it identifies behavioural features that remain informative across educational settings with substantially different scales and delivery modes. Third, it establishes practical performance benchmarks that can guide institutions seeking to implement transferable engagement analytics frameworks. Demonstrating cross-context validity would support wider institutional adoption of quantile-based methods. Identifying stable behavioural predictors would guide feature selection in new implementations. Finally, clarifying how performance varies with dataset scale would help institutions form realistic expectations when deploying engagement analytics systems.

2. Literature Review

This literature review establishes the theoretical and methodological foundation for the proposed engagement classification framework. The review is organised into four interconnected areas. First, it examines the concept of student engagement in digital learning and its behavioural manifestations within learning management systems. Second, it reviews existing approaches to engagement classification, highlighting the limitations of arbitrary threshold-based, clustering-based, and supervised methods. Third, it discusses quantile-based classification as a statistically grounded alternative for defining engagement categories. Finally, it examines the role of cross-dataset validation in learning analytics research and the challenges associated with transferring analytical models across educational contexts. Together, these strands of literature provide the rationale for evaluating a quantile-based engagement classification framework across heterogeneous learning environments.

2.1 Student Engagement in Digital Learning

Student engagement encompasses behavioural, emotional, and cognitive dimensions (Fredricks et al., 2004). Behavioural engagement manifests through participation in learning activities, attendance, and time-on-task. Emotional engagement reflects students' affective responses to learning experiences. Cognitive engagement involves self-regulation and strategic learning approaches. While all three dimensions matter, digital learning environments most readily capture behavioural indicators through interaction logs.

Recent work has reinforced this framing. (Saqr, Fors and Nouri, 2018), in their longitudinal analysis of online engagement across a full programme, confirm that behavioural engagement signals from interaction logs reliably anticipate downstream academic outcomes when measured consistently over time. (Bergdahl *et al.*,

2024) further argue that the trustworthiness of any engagement instrument depends on the methodological transparency of how its categories are defined.

Learning management systems record every student action: accessing content, submitting assignments, participating in forums, watching videos, and taking assessments (Kuzilek, Hlosta and Zdrahal, 2017). These interaction logs provide granular data about when, how often, and which resources students engage with. Temporal patterns emerge from these logs. Some students access materials consistently throughout the course, while others concentrate activity around assessment deadlines.

The relationship between interaction volume and academic success remains complex. More interactions generally correlate with better outcomes, but the relationship is not linear (Conijn *et al.*, 2017). A student with 1,000 interactions distributed across diverse activities and consistent time periods demonstrates qualitatively different engagement than a student with 1,000 interactions concentrated in a single cramming session before the final exam. This distinction motivates the development of temporal-behavioural features that capture engagement quality rather than mere quantity. These behavioural differences motivate the need for classification approaches that can distinguish meaningful engagement patterns while remaining interpretable and transferable across educational contexts.

2.2 Classification Approaches for Engagement Analysis

Learning analytics researchers employ several strategies for engagement classification. Expert-defined thresholds represent the simplest approach. Researchers or practitioners set boundaries based on domain knowledge or intuition (Howard et al., 2018). For example, students performing fewer than 50 interactions might be labelled passive, those with 50-150 interactions moderate, and those exceeding 150 interactions active. This approach offers interpretability and ease of implementation. However, threshold choices remain arbitrary and context specific. What constitutes 'low' engagement at one institution may represent average behaviour at another.

Unsupervised clustering methods discover natural groupings in behavioural data without requiring predefined thresholds. K-means clustering partitions students based on similarity metrics, while hierarchical approaches build dendrograms revealing nested engagement structures (Kovanović et al., 2016). These methods adapt to data characteristics automatically. Yet they introduce new challenges. Cluster assignments vary with algorithm selection, distance metrics, and predetermined cluster counts. Moreover, clusters may not align with educationally meaningful categories. A cluster might group students by temporal pattern rather than engagement level, complicating interpretation.

Supervised machine learning trains models on pre-labelled examples. Random Forest, support vector machines, and neural networks can learn complex decision boundaries from expert-classified training data (Fincham et al., 2019). These approaches achieve high accuracy when training labels are reliable and datasets are large. The limitation lies in label acquisition. Instructors must manually review student behaviours and assign engagement categories before model training. For large courses or institution-wide analytics, this requirement becomes impractical. Furthermore, models trained in one context may not transfer to others without retraining on locally labelled data.

2.3 Quantile-based Classification Methods

Quantile-based approaches offer methodological alternatives that balance simplicity and statistical rigour. Quartiles divide distributions into four equal parts, with Q1 marking the 25th percentile, Q2 the median, and Q3 the 75th percentile (Xu *et al.*, 2025). Using Q1 and Q3 as category boundaries produces three groups of unequal size but clear interpretation: the bottom quarter (passive), the middle half (moderate), and the top quarter (active). This approach requires no arbitrary threshold selection and adapts automatically to data characteristics.

The statistical properties of quartiles make them particularly suitable for educational contexts. Unlike clustering algorithms, quartile boundaries remain stable across repeated analyses with identical data. They are robust to outliers compared to mean-based approaches. Most importantly, quartiles maintain constant theoretical properties across contexts. The 25th percentile always represents the value below which 25% of observations fall, regardless of dataset size or distribution shape. This mathematical consistency suggests that quartile-based categories might generalise better than arbitrary thresholds.

Table 1 compares the three main approaches used for student engagement classification. The comparison highlights differences in statistical foundation, adaptability, reproducibility, interpretability, and computational requirements. The proposed quantile-based framework is positioned as a statistically grounded and reproducible alternative that preserves interpretability while adapting automatically to local data distributions.

Table 1: Comparison of Student Engagement Classification Approaches

Characteristic	Arbitrary Thresholds	Unsupervised Clustering	Quantile-Based (This Study)
Statistical Basis	None	Data-driven	Distribution percentiles
Adaptability	Fixed across contexts	Emergent patterns	Context-adaptive
Reproducibility	Low	Moderate	High
Interpretability	High	Low	High
Computational Cost	Minimal	Moderate	Minimal
Main Advantage	Simple implementation	No labels required	Statistical grounding + reproducibility
Main Limitation	Arbitrary boundaries	Unstable, subjective k	Relative categories

2.4 Cross-dataset Validation in Learning Analytics

Learning analytics research increasingly recognises the importance of external validation (Tempelaar, Nguyen and Rienties, 2020). Models trained and tested on single datasets may overfit to context-specific patterns that do not generalise. Cross-dataset validation tests whether methods work across different institutions, course types, and student populations. This validation approach provides stronger evidence for practical utility than single-dataset studies.

Few learning analytics studies conduct rigorous cross-dataset validation. Consequently, evidence regarding the transferability of engagement classification frameworks across heterogeneous educational contexts remains limited. Most researchers validate on subsets of single datasets using cross-validation or train-test splits (Rizvi *et al.*, 2022). While these techniques prevent overfitting, they cannot assess generalisability across contexts. Studies that do attempt cross-validation often use datasets from similar institutions or course types, limiting the diversity of validation contexts.

Domain shift presents a major challenge in cross-dataset validation. Different institutions employ different learning management systems with varying interaction possibilities. Student demographics, course structures, and assessment methods differ. These contextual factors influence both the volume and patterns of student interactions. A classification method that works well in one context might fail when interaction volumes, temporal patterns, or feature distributions shift substantially.

Quantile-based classification may exhibit greater robustness to domain shift than alternative approaches. By calculating thresholds from local data distributions rather than importing fixed values, quartile methods automatically adjust to new contexts. However, this theoretical advantage requires empirical validation. Testing quantile-based classification across datasets with substantial differences in scale and characteristics would provide evidence for or against cross-context generalisability.

Recent cross-institutional learning analytics studies make the same point even more forcefully. (Kaliisa *et al.*, 2024) demonstrate that engagement patterns identified within a single institution often fail to replicate when the same analytical pipeline is applied to a structurally different setting. (Xing *et al.*, 2023) similarly find that process feedback interventions grounded in learning analytics produce uneven effects across student subgroups, suggesting that single-context validation systematically overstates the practical utility of any given method. (Tempelaar, Nguyen and Rienties, 2020) extend this concern to the design of analytics-driven feedback systems, arguing that classroom-level robustness is achievable only when classification methods are tested against the heterogeneity of real institutional environments.

3. Methodology

This section describes the methodological procedures used to evaluate the proposed quantile-based engagement classification framework. The methodology integrates cross-dataset validation, quantile-based label construction, temporal-behavioural feature engineering, and machine learning classification within a unified analytical pipeline. Two datasets representing substantially different educational contexts were processed using identical procedures to ensure methodological consistency and enable meaningful comparison. The section outlines the research design, dataset characteristics, engagement classification process, feature extraction strategy, model development, validation procedures, and performance evaluation metrics employed throughout the study.

3.1 Research Design

Single-dataset evaluations cannot establish whether a classification method generalises beyond its original context. This study therefore adopts a cross-dataset design, applying quantile-based classification to two markedly different settings: a large distance-education platform (OULAD) and a smaller blended-learning system (Unistudium). The contrast across scale, delivery mode, and geographic setting provides a rigorous basis for evaluating methodological generalisability. Both datasets are processed through an identical pipeline, so that any observed differences reflect genuine contextual variation rather than methodological inconsistency.

The design was intentionally structured to isolate methodological robustness from contextual influences. Rather than transferring engagement thresholds or predictive models directly between institutions, each dataset was processed independently using the same analytical procedures. This approach allows the study to evaluate whether quantile-based classification preserves its effectiveness when applied to educational environments that differ substantially in scale, delivery mode, and learner behaviour. Consistent performance across these settings would provide stronger evidence of methodological generalisability than validation within a single institutional context.

Figure 1 illustrates the overall research framework and cross-dataset validation process. OULAD includes interaction data from 21,420 students across seven fully online modules in a distance education context. The Unistudium dataset captures learning behaviour from 1,234 students enrolled in blended courses at the University of Perugia. Together, these datasets enable evaluation across substantially different educational conditions.

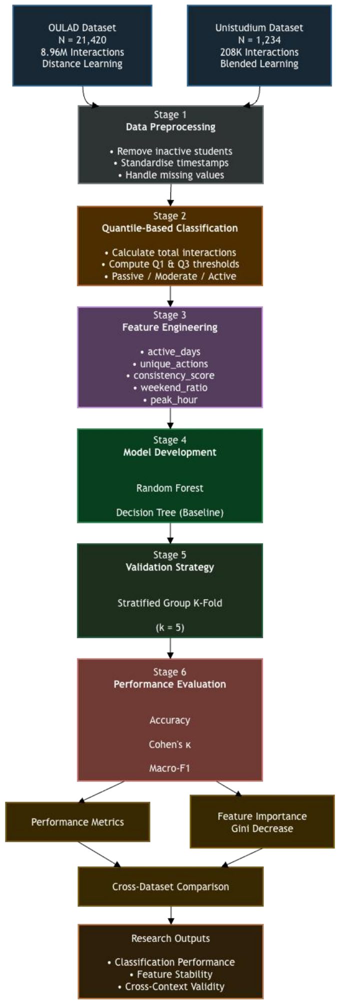


Figure 1: Cross-dataset validation workflow for quantile-based engagement classification

3.2 Dataset Descriptions

The Open University Learning Analytics Dataset contains interaction logs from seven undergraduate modules offered by The Open University UK during 2013 and 2014 (Kuzilek, Hlosta and Zdrahal, 2017). The dataset includes 21,420 students generating 8.96 million interactions across fully online distance education courses. Students accessed materials through the university's virtual learning environment, which recorded all clicks, resource views, forum posts, and assessment submissions.

OULAD represents a typical massive open online learning context. Students came from diverse backgrounds, studied independently, and interacted primarily through digital platforms. Course presentations ran for approximately 269 days, providing substantial temporal data for engagement analysis. The dataset's large size and comprehensive logging make it well-suited for training and validating analytics methods.

The Unistudium dataset contains interaction logs from the University of Perugia's Moodle-based e-learning platform during one semester (Milani, Biondi and Franzoni, 2024). This dataset includes 1,234 students generating 208,109 interactions in blended learning courses that combine face-to-face instruction with online activities. Students attended physical classes while also accessing digital resources, submitting assignments online, and participating in discussion forums.

Unistudium represents a more typical institutional scale and pedagogical approach. The smaller student population and blended delivery model create different interaction patterns compared to fully online distance education. Lower overall interaction volumes reflect the supplementary role of digital platforms in campus-based education. This contextual difference provides a strong test of cross-validation robustness.

Comparison of OULAD and Unistudium dataset characteristics showing substantial differences in scale (17.4× student ratio), educational context (distance vs blended learning), platform (custom VLE vs Moodle), and geographic location (UK vs Italy). Despite these differences, quartile-based thresholds automatically adapt to each context (Q3: 594 vs 309), yielding balanced class distributions (25%-50%-25%) in both datasets. The diversity between datasets provides a rigorous test of cross-context generalisability.

Table 2 presents detailed characteristics of both datasets. The substantial differences in scale (17.4× student ratio), educational model, and temporal scope provide a rigorous test of methodological generalisability. Despite these contextual differences, the quartile-based approach automatically adapts thresholds (OULAD Q3=594, Unistudium Q3=309) while maintaining balanced class distributions.

Table 2: Comparison of dataset characteristics

Characteristic	OULAD	Unistudium
Institution	Open University, UK	University of Perugia, Italy
Educational Model	Distance learning (fully online)	Blended learning (hybrid)
Platform	Custom VLE	Moodle 2.5
Time Period	2013-2014 academic year	Sept 1 - Dec 31, 2022
Duration	Full academic year (7 presentations)	One semester (4 courses)
Number of Students	21,420	1,234
Number of Interactions	8,964,036	208,228
Dataset Size Ratio	17.4× larger	Baseline
Courses/Modules	7 modules	4 courses
Student Anonymization	Pseudonymized IDs	Fully anonymized
Data Source	Kuzilek et al. (2017)	University of Perugia README

3.3 Quantile-based Engagement Classification

The classification follows a simple principle. Each student's total interaction count is mapped to one of three engagement categories using percentile thresholds derived from the dataset itself. Students below the 25th percentile are classified as Passive, students above the 75th percentile as Active, and the remainder as Moderate. Because the threshold values are computed independently for each dataset, they adapt automatically to local interaction volumes without manual calibration. The procedure comprises three steps:

computing total interactions per student, calculating dataset-specific quartile thresholds, and assigning each student to a category on the basis of those thresholds.

Total interactions for student *s* were calculated by summing all recorded actions:

$$total\_action = \sum_{i=1}^{n} interaction$$

This metric aggregates all platform activities including content views, forum posts, assignment submissions, quiz attempts, and resource downloads. Each interaction receives equal weight regardless of type or duration. Conceptually, the thresholds divide a list of students ranked by activity level at the 25th percentile (Q1) and the 75th percentile (Q3). Students whose activity falls below Q1 are classified as Passive, those whose activity exceeds Q3 as Active, and those between Q1 and Q3 as Moderate. Formally, these thresholds were calculated using the Type 7 quantile algorithm, which produces unbiased estimates for continuous distributions. Three thresholds divide the distribution.

Distribution quartiles were calculated using the Type 7 quantile algorithm, which provides unbiased estimates for continuous distributions. Three thresholds divided the distribution. Q_1 represents the 25th percentile of the total_actions distribution. Q_2 represents the 50th percentile (median). Q_3 represents the 75th percentile. For OULAD, the calculated thresholds were Q_1 = 105 interactions, Q_2 = 278 interactions, and Q_3 = 594 interactions. For Unistudium, the thresholds were Q_1 = 37 interactions, Q_2 = 125 interactions, and Q_3 = 309 interactions. These substantial differences illustrate how quantile-based classification automatically adapts to context.

Students were assigned to engagement categories based on their total interactions relative to the quartile thresholds. Passive students fell below Q_1, moderate students ranged from Q_1 to Q_3, and active students exceeded Q_3. This scheme produces a balanced distribution with 25% passive, 50% moderate, and 25% active by mathematical definition. This three-category classification provides interpretable groupings that align with educational practice. Passive students demonstrate minimal platform usage and may require intervention. Moderate students show typical engagement levels. Active students exhibit extensive platform usage and exploratory behaviour.

Figure 2 visualises the distribution of total interactions in both datasets using box plots. The figure highlights the quartile thresholds used to define passive, moderate, and active engagement categories and illustrates how threshold values adapt automatically to local interaction distributions.

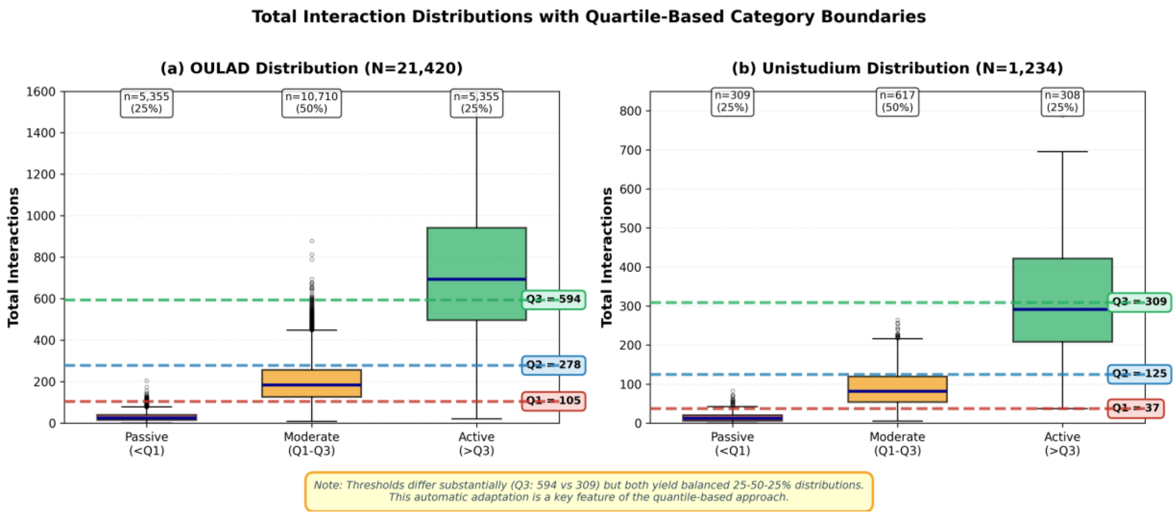


Figure 2: Box plots showing total interaction distributions with quartile thresholds for OULAD and Unistudium

3.4 Feature Engineering

Once the engagement labels are assigned, the prediction task requires a separate feature set. Using total interactions for both labelling and prediction would introduce circular reasoning, since the variable used to define the categories would also serve as the strongest predictor. Five temporal-behavioural features were therefore derived from the raw logs. Drawing on Fredricks et al.'s multidimensional framework, these features operationalise behavioural engagement through three theoretically distinct dimensions: consistency of access, diversity of activity, and scheduling patterns. None of the five features uses total interactions directly.

Active days represents the number of unique calendar dates on which a student performed at least one interaction.

$$active\_days_s = |\{d \in D : interactions_s(d) \geq 1\}|$$

where D represents the set of all dates in the course period. This feature captures engagement consistency across time. Students who access the platform regularly across many days demonstrate different patterns than those with sporadic access, even if total interaction counts are similar.

Unique actions represents the count of distinct activity types a student performed.

$$unique\_action_s = |\{a \in A : student_s \; performed \; action \; a\}|$$

where A represents the set of all possible activity types in the platform. This feature captures behavioural diversity. Students who combine multiple activity types (viewing, submitting, posting, downloading) demonstrate broader engagement than those focused on single activities.

Consistency score measures regularity of daily interaction patterns using the coefficient of variation.

$$consistency\_score_s = 1 - min\left(1, \frac{\sigma(daily\_interactions_s)}{\mu\,(daily\_interactions_s)}\right)$$

where σ represents standard deviation and μ represents mean of daily interaction counts across active days. Scores near 1 indicate consistent daily patterns while scores near 0 indicate erratic behaviour. The minimum operator prevents negative values when standard deviation exceeds the mean.

Weekend ratio measures the proportion of interactions occurring on Saturdays or Sundays.

$$weekend\_ratio_s = \frac{interactions \; on \; weekends}{total\_actions_s}$$

This feature captures temporal preferences that may relate to employment status or study habits. Students integrating learning into weekend schedules demonstrate different patterns than those confining study to weekdays.

Peak hour represents the hour of day (0-23) during which a student performed the most interactions.

$$peak\_hour_s = \underset{h \in [0,23]}{\arg max} \sum_{i \in I_h} interaction$$

where I_h represents the set of interactions occurring during hour h. This feature captures diurnal preferences. Some students engage primarily during morning hours while others prefer evening or night-time study. Definitions and examples of five temporal-behavioral features extracted from interaction logs for engagement classification. Features capture temporal consistency (active_days), behavioral diversity (unique_actions), regularity (consistency_score), and scheduling patterns (weekend_ratio, peak_hour). Example values from OULAD show clear separation between passive, moderate, and active categories, particularly for active_days (9.7 vs 145.1 days) and unique_actions (18.2 vs 163.8 types) which account for 86% of combined predictive power. All features avoid circular reasoning by excluding total_interactions from the feature set despite using it for quartile-based labeling.

Table 3 presents detailed definitions and example values for all five features. The features capture complementary aspects of digital engagement: temporal consistency (active_days: 9.7 vs 145.1 days for passive vs active), behavioral breadth (unique_actions: 18.2 vs 163.8 types), and regularity (consistency_score: 0.68 vs 0.82). Together, active_days and unique_actions account for 86% of predictive power, validating the theoretical premise that temporal consistency and behavioral diversity represent fundamental engagement dimensions.

Table 3: Feature definitions with example calculations

Feature	Passive	Moderate	Active	Separation
active_days	9.7	56.5	145.1	15× difference
unique_actions	18.2	70.2	163.8	9× difference
consistency	0.68	0.83	0.82	Moderate separation
weekend_ratio	0.25	0.24	0.26	Minimal separation
peak_hour	-	-	-	No clear pattern

3.5 Classification Model

A Random Forest operates as a panel of decision trees, each casting a vote for the final prediction. Each tree is trained on a slightly different random sample of the data and considers a random subset of features at every split. Each tree then assigns one of the three engagement categories (Active, Moderate, or Passive) to a given student, and the majority vote across all trees determines the final classification. Because no single tree dominates the outcome, the method is robust to noise and resists overfitting. Random Forest was selected specifically because it handles nonlinear relationships effectively, does not require feature scaling, and produces interpretable feature importance scores directly from the training process.

The Random Forest classifier employed 100 decision trees (n_estimators=100), with each split considering the square root of total features (max_features='sqrt'). Maximum tree depth remained unlimited (max_depth=None), allowing trees to grow until leaves contained pure classes or reached minimum sample thresholds. The minimum samples required for splitting was two (min_samples_split=2), and minimum samples per leaf was one (min_samples_leaf=1). A fixed random seed (random_state=42) ensured reproducibility across runs.

These hyperparameters represent standard Random Forest defaults without tuning. Using default settings tests the classifier's inherent ability to generalise without optimisation for specific datasets. This conservative approach provides realistic estimates of performance that institutions could expect when implementing the method without extensive tuning.

Random Forest provides feature importance scores through mean decrease in impurity. When a feature is used to split a node, the algorithm calculates the reduction in Gini impurity. Features that consistently produce large impurity reductions across trees receive high importance scores. Scores were normalised to sum to 1.0, enabling direct comparison across datasets.

3.6 Cross-validation Strategy

Stratified Group K-Fold cross-validation was employed to ensure robust performance estimation while preventing data leakage. Standard K-Fold cross-validation randomly distributes samples across folds, which risks placing students from the same course presentation in both training and test sets. This temporal leakage inflates performance estimates artificially.

Students were grouped by their enrolment cohort or course presentation. For OULAD, this corresponded to module-presentation combinations (e.g., Module AAA 2013B). For Unistudium, this corresponded to course-semester combinations. All students from the same group remained together in the same fold, preventing temporal information from training data leaking into test predictions.

The dataset was divided into five folds (k=5) with stratification by engagement category. This ensures each fold maintains approximately the same proportion of passive, moderate, and active students as the full dataset. Stratification prevents situations where some folds contain predominantly one category, which could bias performance estimates.

Each fold served as the test set exactly once, with the remaining four folds forming the training set. Models trained on the training folds predicted engagement categories for students in the held-out test fold. This process repeated five times, producing five independent performance estimates. Final results report the mean and standard deviation across all five folds.

3.7 Evaluation Metrics

Three complementary metrics were calculated to assess classification performance comprehensively. Each metric captures different aspects of classifier quality, providing a balanced assessment.

Overall accuracy measures the proportion of correctly classified students. This metric provides an intuitive assessment of classifier performance but can be misleading with imbalanced classes. The balanced category distribution from quantile-based labelling makes accuracy a reliable metric in this context.

$$Accuracy = \frac{Correct\ Predictions}{Total\ Predictions}$$

Cohen's κ adjusts for chance agreement, providing a more robust measure when classes are imbalanced (Cohen, 1960). Κ ranges from -1 to 1, where 1 indicates perfect agreement, 0 indicates agreement equivalent to chance, and negative values indicate systematic disagreement.

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

where P_o is observed agreement (accuracy) and P_e is expected agreement by chance. Landis and Koch (1977) suggest that κ values range from -1 (perfect disagreement) to +1 (perfect agreement), with 0 indicating chance-level performance. Interpretation guidelines suggest: κ < 0.20 (slight), 0.21-0.40 (fair), 0.41-0.60 (moderate), 0.61-0.80 (substantial), 0.81-1.00 (almost perfect).

F1-score balances precision and recall for each class. Precision measures the proportion of predicted category members that truly belong to that category. Recall measures the proportion of true category members correctly identified. F1-score represents the harmonic mean of precision and recall. Macro-averaging calculates F1-score independently for each category then averages across categories, giving equal weight to all classes regardless of size.

$$F1_c = \frac{2 \times Precision_c \times Recall_c}{Precision_c + Recall_c}$$

4. Results

This section presents the empirical findings obtained from applying the proposed quantile-based engagement classification framework to both datasets. The results are organised into three parts. First, classification performance is evaluated using accuracy, Cohen's κ, and macro-averaged F1-score to assess predictive effectiveness across educational contexts. Second, feature importance analysis is conducted to examine the consistency of behavioural predictors between datasets. Third, descriptive analyses of engagement categories are presented to characterise behavioural differences among passive, moderate, and active learners. Together, these results provide evidence regarding the validity, transferability, and practical applicability of the proposed framework across heterogeneous learning environments.

4.1 Classification Performance

Classification performance comparison between OULAD (N=21,420) and Unistudium (N=1,234) across three evaluation metrics. OULAD achieves 92.04% accuracy with almost perfect agreement (κ=0.8725), while Unistudium achieves 72.50% accuracy with moderate agreement (κ=0.5875). The 19.54 percentage point gap (26.95% relative difference) primarily reflects dataset size differences (17.4× ratio) and behavioral pattern distinctiveness. Both datasets show low cross-validation standard deviations, indicating stable model behavior. Results demonstrate that quantile-based classification achieves high accuracy on large datasets while maintaining practically useful performance on medium-sized institutional datasets.

Table 4 presents classification performance across both datasets. The Random Forest classifier achieved 92.04% accuracy (κ=0.8725) on OULAD and 72.50% accuracy (κ=0.5875) on Unistudium. The 19.54 percentage point difference reflects the substantial dataset size gap (17.4× student ratio) and differences in behavioral pattern distinctiveness.

Table 4: Classification performance metrics

Metric	OULAD	Unistudium	Gap	Relative Diff
Accuracy	92.04% ± 0.28%	72.50% ± 1.45%	**+19.54 pp**	+26.95%
Cohen's κ	0.8725 ± 0.0039	0.5875 ± 0.0218	**+0.2850**	+48.51%
F1-Score	0.9203 ± 0.0024	0.7180 ± 0.0156	**+0.2023**	+28.18%

Table 5 compares the performance of the proposed Random Forest model with the Decision Tree baseline. The comparison provides a benchmark for assessing the contribution of ensemble learning to engagement classification.

Table 5: Baseline comparison between Decision Tree and Random Forest

Dataset	Metric	Decision Tree	Random Forest
OULAD	Accuracy	81.62%	**92.04%**
OULAD	Cohen's κ	0.6977	**0.8725**
Unistudium	Accuracy	63.45%	**72.50%**
Unistudium	Cohen's κ	0.3918	**0.5875**

To provide a contextual performance benchmark, a Decision Tree classifier was evaluated as a baseline model using the same engagement labels and behavioural features. Random Forest consistently outperformed Decision Tree across both datasets. In OULAD, Random Forest improved accuracy from 81.62% to 92.04% and increased Cohen's κ from 0.6977 to 0.8725. Similar improvements were observed in Unistudium, where accuracy increased from 63.45% to 72.50% and Cohen's κ improved from 0.3918 to 0.5875. These findings indicate that ensemble learning provides a more robust representation of student engagement patterns than a single-tree classifier.

The Random Forest classifier achieved 92.04% accuracy (SD = 0.28%) on OULAD. This low standard deviation demonstrates highly consistent predictions across cross-validation folds. Cohen's κ reached 0.8725 (SD = 0.0039), falling in the 'almost perfect agreement' range according to Landis and Koch (1977). The macro-averaged F1-score of 0.9203 (SD = 0.0024) confirms balanced performance across all three engagement categories.

The minimal standard deviations across metrics indicate robust model stability. Cross-validation fold performance varied by less than 0.3 percentage points for accuracy and 0.004 for Cohen's κ. This stability suggests the classifier learned generalizable patterns rather than overfitting to specific course presentations.

Unistudium achieved 72.50% accuracy (SD = 1.45%), representing moderate classification performance. Cohen's κ of 0.5875 (SD = 0.0218) falls in the 'moderate agreement' range. The macro F1-score reached 0.7180 (SD = 0.0156). While these metrics trail OULAD performance, they represent statistically significant classification accuracy well above chance levels.

The larger standard deviations compared to OULAD reflect smaller dataset size and fewer cross-validation groups. With only 1,234 students versus 21,420 in OULAD, individual fold composition exerts greater influence on performance estimates. Nevertheless, the consistency of results across folds demonstrates reliable classification even at this smaller scale.

Figure 3 compares classification performance between OULAD and Unistudium across the three evaluation metrics. Error bars represent variation across cross-validation folds and provide an indication of model stability.

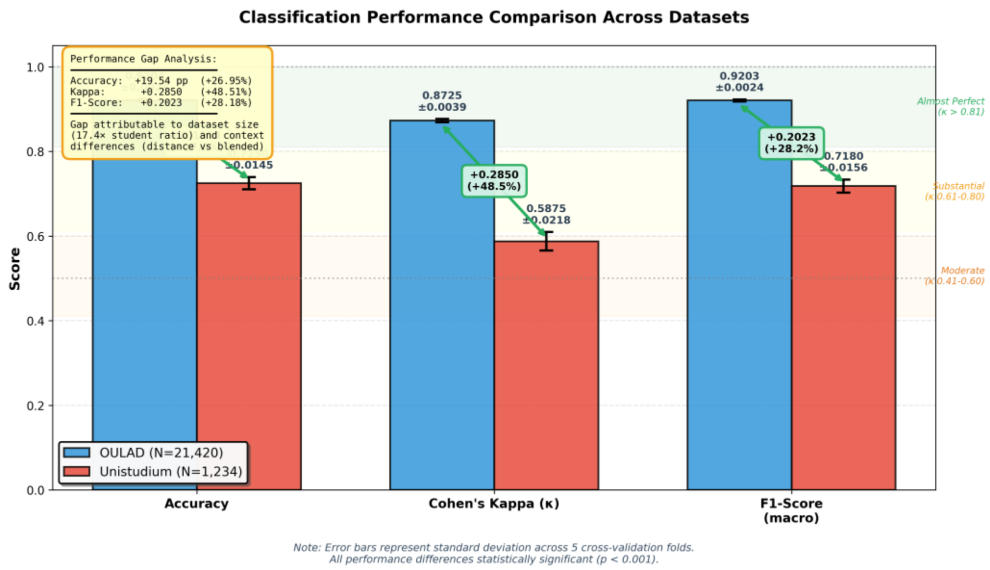


Figure 3: Bar chart comparing accuracy, κ, and F1-score between datasets with error bars

The 19.54 percentage point accuracy gap between datasets (92.04% vs 72.50%) primarily reflects scale differences rather than methodological limitations. Larger datasets provide more training examples and clearer separation between categories. The 17.4× student ratio between OULAD and Unistudium explains much of the performance differential. Studies across machine learning domains consistently show performance improvements with increased training data.

Behavioural pattern distinctiveness also contributes to performance differences. OULAD students engaged exclusively through digital platforms, creating clearer behavioural signatures. Unistudium students balanced online and face-to-face activities, potentially diluting digital engagement signals. Despite these contextual differences, both datasets achieved statistically significant classification performance, supporting the generalisability of quantile-based approaches.

4.2 Feature Importance Analysis

Feature importance scores from Random Forest models showing mean Gini importance for each feature across both datasets. Active days (48-53%) and unique actions (33-35%) dominate predictions, collectively accounting for >80% of predictive power in both datasets. High rank correlation (Spearman's ρ=0.90) demonstrates remarkable consistency despite substantial contextual differences. Only minor rank swap between weekend ratio and consistency score (ranks 3-4), while top two features and peak hour maintain identical rankings. Results validate temporal consistency and behavioral diversity as fundamental engagement dimensions that generalize across educational contexts.

Table 6 presents feature importance scores from both datasets. Active days emerged as the dominant predictor (OULAD: 52.74%, Unistudium: 48.23%), followed by unique actions (OULAD: 33.39%, Unistudium: 34.56%). Together, these two features account for 86.13% and 82.79% of predictive power respectively. The high Spearman rank correlation (ρ=0.90) indicates remarkable consistency in feature importance patterns across datasets.

Table 6: Feature importance scores for both datasets

Rank	Feature	OULAD	Unistudium	Agreement	Average
1	**active_days**	**52.74%**	**48.23%**	✓	**50.49%**
2	**unique_actions**	**33.39%**	**34.56%**	✓	**33.98%**
3	weekend_ratio	6.62%	6.78%	~ (rank 4)	6.70%
4	consistency_score	5.87%	8.91%	~ (rank 3)	7.39%
5	peak_hour	1.38%	1.52%	✓	1.45%
TOTAL		**100.00%**	**100.00%**		**100.00%**

Active days emerged as the most important feature in both datasets, accounting for 52.74% of predictive power in OULAD and 48.23% in Unistudium. This dominance confirms that temporal consistency represents the primary dimension distinguishing engagement levels. Students who access the platform regularly across many days demonstrate fundamentally different engagement than those concentrating activity in brief periods.

Unique actions ranked second in both datasets, contributing 33.39% in OULAD and 34.56% in Unistudium. This consistency validates behavioural diversity as a universal engagement indicator. Students exploring many different activity types show broader engagement than those repeatedly using only a few platform features. Together, active days and unique actions account for over 80% of predictive power in both contexts.

Consistency score, weekend ratio, and peak hour contributed smaller but measurable importance. These features ranked 3-5 in both datasets, though their relative ordering differed slightly. OULAD placed weekend ratio (6.62%) ahead of consistency score (5.87%), while Unistudium reversed this ordering (consistency 8.91%, weekend 6.78%). Peak hour contributed least in both datasets (1.38% and 1.52% respectively).

The minor role of temporal preference features (weekend ratio, peak hour) suggests that when students engage matters less than how consistently and diversely they engage. Students achieve high engagement through regular, diverse platform use regardless of whether they primarily study on weekends or weekdays, mornings or evenings.

Feature importance rankings remained remarkably stable across datasets. The rank correlation (Spearman's ρ) between OULAD and Unistudium importance scores reached 0.90, indicating very high consistency. Only one minor rank swap occurred: consistency score and weekend ratio exchanged positions 3 and 4. The top two features (active days, unique actions) and bottom feature (peak hour) maintained identical rankings.

This cross-dataset consistency provides strong evidence that temporal consistency and behavioural diversity represent fundamental engagement dimensions. These patterns generalise across different scales, delivery models, and institutional contexts. Institutions implementing engagement analytics can prioritise these features with confidence that they will capture meaningful engagement signals in their local context.

4.3 Engagement Category Characteristics

Descriptive statistics for behavioral features across three engagement categories in the OULAD dataset (N=21,420). Clear separation evident across all features, with active students demonstrating 15× more active days (145.1 vs 9.7), 22× more total interactions (1,027.2 vs 46.4), and 9× more unique actions (163.8 vs 18.2) compared to passive students. Moderate and active categories show similar consistency scores (0.83 vs 0.82), both substantially higher than passive students (0.68). All between-group differences statistically significant (p<0.001) with large effect sizes for temporal and behavioral features. Passive students show highest coefficient of variation, indicating heterogeneous group with diverse engagement patterns.

Table 7 summarises descriptive statistics for each engagement category within the OULAD dataset. The results provide insight into behavioural differences between passive, moderate, and active learners across all extracted features.

Table 7: Descriptive statistics by engagement category

Feature (Active/Passive)	Passive (N=5,355) Mean ± SD	Moderate (N=10,710) Mean ± SD	Active (N=5,355) Mean ± SD	Ratio
active_days (days)	9.7 ± 7.3	56.5 ± 27.5	145.1 ± 43.8	15.0×
total_actions (count)	46.4 ± 31.1	300.4 ± 135.9	1,027.2 ± 437.8	22.1×
unique_actions (types)	18.2 ± 11.2	70.2 ± 32.2	163.8 ± 61.3	9.0×
consistency_score (0-1 scale)	0.68 ± 0.33	0.83 ± 0.19	0.82 ± 0.17	1.2×
weekend_ratio (proportion)	0.25 ± 0.25	0.24 ± 0.10	0.26 ± 0.06	1.04×
peak_hour (hour, 0-23)	13.2 ± 5.8	14.1 ± 4.9	14.5 ± 4.2	1.1×

Active days showed dramatic separation between categories in OULAD. Passive students accessed the platform on only 9.7 days on average (SD = 7.3), compared to 56.5 days for moderate students (SD = 27.5) and 145.1 days for active students (SD = 43.8). This represents a 15-fold difference between passive and active categories, demonstrating clear behavioural boundaries.

The substantial standard deviations within categories indicate heterogeneous subgroups. Some passive students accessed the platform on only a few days, while others approached the moderate range. This variability suggests that category boundaries represent pragmatic groupings rather than discrete behavioural types. Nevertheless, the clear mean differences support the validity of the three-category scheme.

Total interactions showed even larger category separation, as expected given their role in category definition. Passive students performed 46.4 interactions on average (SD = 31.1), moderate students 300.4 (SD = 135.9), and active students 1,027.2 (SD = 437.8). The 22-fold difference between passive and active categories confirms substantial behavioural variation across the engagement spectrum.

Unique actions followed similar patterns, with a 9-fold difference between passive (18.2 types, SD = 11.2) and active students (163.8 types, SD = 61.3). Moderate students fell clearly between these extremes (70.2 types, SD = 32.2). This progression demonstrates that higher engagement involves both more interactions and greater behavioural diversity.

Consistency scores revealed unexpected patterns. Moderate and active students showed similar consistency (0.83 and 0.82 respectively), both substantially higher than passive students (0.68). This suggests a threshold effect. Once students achieve moderate engagement levels, they tend to maintain regular patterns. Passive students, conversely, show erratic behaviour with high variability in daily interaction counts.

The high coefficient of variation for passive students (0.49) compared to moderate (0.23) and active students (0.21) confirms greater heterogeneity in this category. Some passive students rarely accessed the platform at all, while others showed sporadic bursts of activity. This diversity suggests that passive classification might encompass several distinct subgroups warranting different intervention strategies.

5. Discussion

This section interprets the findings in relation to the study objectives, research questions, and existing literature on learning analytics and student engagement. Beyond reporting classification performance, the discussion examines the implications of the results for understanding engagement behaviour across different educational contexts. Particular attention is given to the effectiveness of quantile-based classification, the consistency of behavioural predictors across datasets, and the practical significance of the observed performance differences between large-scale and medium-scale learning environments. The section also considers how the findings contribute to current knowledge on transferable engagement analytics and identifies implications for future research and institutional practice.

5.1 Interpretation of Findings

Quantile-based engagement classification showed strong performance across both datasets. Accuracy was high in the large-scale OULAD dataset (92.04%) and remained acceptable in the Unistudium dataset (72.50%). These results confirm that the method functions across different institutional contexts. Automatic threshold adaptation preserved the statistical structure of engagement categories while accommodating large differences in interaction volume. The accuracy gap of 19.54 percentage points requires careful interpretation. Dataset scale is the primary factor. OULAD includes a student population 17.4 times larger than Unistudium, which provides more training data and clearer behavioural patterns. Behavioural expression also differs between contexts. Students in fully online programmes generate richer digital traces than those in blended courses, where learning activity is partly offline. Temporal coverage further contributes to the difference. OULAD spans 269 days, whereas Unistudium covers approximately 120 days. The longer observation window allows more stable engagement patterns to emerge. Taken together, these factors explain why performance differs without indicating instability in the classification method itself.

The baseline comparison provides additional evidence regarding the robustness of the proposed approach. Across both datasets, Random Forest consistently outperformed the Decision Tree baseline. The improvement was particularly evident in Unistudium, where behavioural patterns were more heterogeneous and training data were more limited. This finding suggests that ensemble learning is better able to capture complex engagement patterns than a single-tree classifier, particularly when engagement behaviours exhibit substantial variation.

5.2 Comparison with Prior Research

Performance on OULAD exceeds that reported in earlier studies using the same dataset. The classification performance observed in this study is comparable to, and in some cases exceeds, results reported in prior learning analytics research. Previous studies have demonstrated the predictive value of behavioural interaction data for identifying student outcomes and engagement-related patterns (Howard, Meehan and Parnell, 2018; Fincham *et al.*, 2019). The strong performance observed here may be attributed to the use of quantile-based category construction and temporal-behavioural features that capture engagement quality rather than simple activity volume. First, quartile-based labelling produces balanced engagement categories and avoids class imbalance. Second, temporal behavioural features capture engagement quality rather than simple activity volume.

More importantly, this study addresses a limitation highlighted in prior work. (Rizvi *et al.*, 2022) noted that most engagement studies rely on validation within a single institutional dataset, which restricts claims of generalisability. By applying the same method to datasets that differ substantially in scale, context, and delivery mode, this study provides stronger evidence of cross-context validity. The high consistency in feature importance across datasets ($\rho = 0.90$) further supports the stability of engagement constructs beyond a single setting.

5.3 Practical Implications for Educational Institutions

Institutions can form realistic expectations based on dataset size. The findings suggest that larger datasets generally support stronger classification performance, whereas smaller institutional datasets may exhibit greater variability. Institutions should therefore interpret expected performance in relation to dataset size, interaction density, and local learning practices.

Automatic threshold adaptation is a key practical advantage of the proposed approach. Institutions do not need to import thresholds from prior studies or conduct extensive calibration. Quartiles calculated from local interaction data generate boundaries that reflect the institution's own learning environment. This simplicity lowers technical barriers and makes the method accessible to institutions with limited analytics expertise.

Feature importance analysis provides guidance for intervention design. Active days and unique actions accounted for most of the predictive power in both datasets. This suggests that engagement is driven more by consistency and behavioural diversity than by raw activity counts. Institutions may therefore prioritise interventions that promote regular platform access and varied resource use. Examples include scheduled reminders that encourage daily engagement, learning activities that require exploration of different materials, and early alerts for students with irregular access patterns.

5.4 Methodological Contributions

The cross-dataset validation framework used in this study offers a practical template for future learning analytics research. Testing methods across datasets that differ in scale, institutional context, and delivery model provides stronger evidence of generalisability than single-dataset evaluation. The consistent results observed across OULAD and Unistudium demonstrate that such validation is both feasible and informative.

The inclusion of a Decision Tree baseline further strengthens the methodological contribution of the study by demonstrating that the observed performance gains are attributable to the proposed modelling approach rather than to the engagement labelling scheme alone.

This study also addresses a common methodological issue in engagement research. Engagement labels are often defined using the same interaction metrics later used for prediction, resulting in circular reasoning. Here, total interactions were used only for label construction. Classification relied on independent temporal and behavioural features. This separation enables a more meaningful evaluation of whether engagement patterns beyond simple activity volume distinguish engagement categories.

5.5 Limitations

Validation on two datasets limits the breadth of generalisation. Additional datasets would strengthen the conclusions. Future work should include institutions from different geographical regions, educational levels, subject domains, and learning platforms. In addition, only two institutional contexts were examined, limiting the breadth of cross-context validation.

Engagement categories were validated against quartile-based labels rather than external learning outcomes. While statistically grounded, these categories have not yet been linked to grades, completion rates, or retention. Examining these relationships is an important direction for future research.

Engagement was treated as static over the course duration. In practice, engagement may fluctuate as motivation and workload change. Tracking transitions between engagement categories over time could support earlier intervention. Such analysis would require dynamic modelling approaches, such as time-series classification or state-transition models.

6. Conclusion

This study demonstrates that quantile-based engagement classification can achieve robust performance across diverse educational contexts without requiring manual threshold calibration. Cross-dataset validation using OULAD and Unistudium revealed consistent accuracy and stable feature importance patterns despite a 17.4-fold difference in dataset scale and fundamental differences in delivery mode.

Three findings are particularly important. First, automatic threshold adaptation through quartile calculation successfully accommodates large differences in interaction volume while maintaining balanced engagement categories. Second, temporal consistency and behavioural diversity emerged as dominant engagement indicators in both datasets. Third, Random Forest consistently outperformed the Decision Tree baseline, indicating that ensemble learning provides additional robustness for engagement classification across different educational contexts.

From a practical perspective, the proposed approach offers a statistically grounded and interpretable solution for institutions seeking to implement engagement analytics. Quartiles can be calculated directly from local data, a small set of temporal behavioural features can be extracted, and classification models can be trained using standard tools. The demonstrated cross-context validity suggests that this method is suitable for deployment beyond the setting in which it was developed.

Future research should extend validation to additional institutional contexts, examine links between engagement categories and learning outcomes, and explore temporal changes in engagement over time. Such work would further strengthen the practical relevance of engagement analytics.

More broadly, this study highlights the importance of cross-dataset validation in learning analytics research. Methods evaluated on a single dataset cannot demonstrate generalisability. Testing across multiple contexts provides stronger evidence of practical utility and supports wider adoption by educational institutions.

Ethics Statement: This study used anonymised educational datasets obtained from publicly available and institutional learning management system records. No direct interaction with human participants was conducted, and no personally identifiable information was accessed or processed. All analyses were performed on anonymised data in accordance with applicable ethical principles for educational data research.

AI Statement: Generative AI tools were used exclusively for editorial assistance, including language refinement and manuscript formatting. No AI system was used to generate research data, perform data analysis, interpret findings, or draw scientific conclusions. All intellectual contributions and final decisions remain the responsibility of the authors.

References

Bergdahl, N. *et al.* (2024) "Unpacking student engagement in higher education learning analytics: a systematic review," *International Journal of Educational Technology in Higher Education*, 21(1), pp. 1–33. Available at: https://doi.org/10.1186/S41239-024-00493-Y/TABLES/6.

Brinton, C.G. and Chiang, M. (2015) "MOOC performance prediction via clickstream data and social learning networks," *Proceedings - IEEE INFOCOM*, 26, pp. 2299–2307. Available at: https://doi.org/10.1109/INFOCOM.2015.7218617.

Conijn, R. *et al.* (2017) "Predicting student performance from LMS data: A comparison of 17 blended courses using moodle LMS," *IEEE Transactions on Learning Technologies*, 10(1), pp. 17–29. Available at: https://doi.org/10.1109/TLT.2016.2616312.

Fincham, E. *et al.* (2019) "From Study Tactics to Learning Strategies: An Analytical Method for Extracting Interpretable Representations," *IEEE Transactions on Learning Technologies*, 12(1), pp. 59–72. Available at: https://doi.org/10.1109/TLT.2018.2823317.

Howard, E., Meehan, M. and Parnell, A. (2018) "Contrasting prediction methods for early warning systems at undergraduate level," *The Internet and Higher Education*, 37, pp. 66–75. Available at: https://doi.org/10.1016/J.IHEDUC.2018.02.001.

Kaliisa, R. *et al.* (2024) "Have Learning Analytics Dashboards Lived Up to the Hype? A Systematic Review of Impact on Students' Achievement, Motivation, Participation and Attitude," *ACM International Conference Proceeding Series*. Association for Computing Machinery, pp. 295–304. Available at: https://doi.org/10.1145/3636555.3636884.

Kuzilek, J., Hlosta, M. and Zdrahal, Z. (2017) "Data Descriptor: Open University Learning Analytics dataset," *Scientific Data*, 4(1), pp. 170171-. Available at: https://doi.org/10.1038/SDATA.2017.171;SUBJMETA.

Milani, A., Biondi, G. and Franzoni, V. (2024) "Unistudium 2022: students groups logs on moodle." IEEE Dataport. Available at: https://doi.org/10.21227/3k76-z181.

Rizvi, S. *et al.* (2022) "Beyond one-size-fits-all in MOOCs: Variation in learning design and persistence of learners in different cultural and socioeconomic contexts," *Computers in Human Behavior*, 126. Available at: https://doi.org/10.1016/j.chb.2021.106973.

Saqr, M., Fors, U. and Nouri, J. (2018) "Using social network analysis to understand online problem-based learning and predict performance," *PLoS ONE*, 13. Available at: https://doi.org/10.1371/journal.pone.0203590.

Tempelaar, D., Nguyen, Q. and Rienties, B. (2020) "Learning Analytics and the Measurement of Learning Engagement," pp. 159–176. Available at: https://doi.org/10.1007/978-3-030-47392-1_9.

Xing, W. *et al.* (2023) "Using learning analytics to explore the multifaceted engagement in collaborative learning," *Journal of Computing in Higher Education*, 35, pp. 633–662. Available at: https://doi.org/10.1007/s12528-022-09343-0.

Xu, Z. *et al.* (2025) "Leveraging Learning Analytics to Model Student Engagement in Graduate Statistics: A Problem-Based Learning Approach in Agricultural Education †," *Behavioral Sciences*, 15. Available at: https://doi.org/10.3390/bs15101360.

Microlearning in Teacher Education: Effects on Pre-service Teachers' Digital Self-efficacy and Digital Competence

Tomas Javorcik, Tatiana Havlaskova, Magdalena Zavodna and Katerina Kostolanyova
Department of Information and Communication Technologies, Faculty of Education, University of Ostrava, Czechia

tomas.javorcik@osu.cz
tatiana.havlaskova@osu.cz
magdalena.zavodna@osu.cz
katerina.kostolanyova@osu.cz

https://doi.org/10.34190/ejel.24.3.4790

Abstract: The rapid digital transformation of education has increased demands on teachers' digital competences and their ability to meaningfully integrate digital technologies into teaching practice. Developing these competences in pre-service teachers therefore represents a key challenge for contemporary teacher education. Microlearning has emerged as a promising instructional approach that enables flexible, targeted and time-efficient development of digital skills through short, focused learning units. However, empirical evidence regarding the impact of microlearning on the development of digital competences among future teachers remains limited. This study therefore examines the effects of a microlearning course on pre-service teachers' perceived self-efficacy in using digital technologies for teaching. The study employed a quantitative pre-post research design conducted over four academic years (2021–2025). Data were collected through an online questionnaire administered before and after completion of a microlearning course designed to support the development of selected digital skills relevant to teaching practice. The analysis included responses from 1,437 students in the pre-survey and 871 students in the post-survey. The questionnaire measured students' perceived competence across multiple domains of digital skills, including activities such as creating presentations, designing interactive worksheets, editing video materials and preparing electronic tests. The collected data were analysed using descriptive and inferential statistical methods to evaluate changes in perceived competence and examine differences across selected demographic variables. The findings demonstrate statistically significant improvements across all examined domains of digital skills following completion of the microlearning course. The most pronounced improvements were observed in areas that students initially perceived as the most challenging, particularly in creating interactive worksheets, editing video content and designing electronic tests. The results also indicate a partial reduction in differences between student groups defined by factors such as gender, age and mode of study, suggesting a homogenising effect of the microlearning intervention on students' perceived digital competence. Overall, the study shows that microlearning can effectively support future teachers' digital competences and strengthen their confidence in using digital technologies for teaching.

Keywords: Microlearning, Pre-service teachers, Digital competences, Self-efficacy, Teacher education, Digital literacy

1. Introduction

The accelerating digital transformation of education has intensified expectations that future teachers will integrate digital technologies meaningfully into pedagogical practice. Although pre-service teachers commonly use digital tools in everyday life, their ability to apply them for instructional purposes remains uneven. Developing technological–pedagogical competence has therefore become a central objective of teacher education, commonly framed through models such as TPACK (Koehler and Mishra, 2005) and European competence frameworks including DigCompEdu (Redecker, 2017).

Despite these frameworks, the integration of digital competences into teacher education remains inconsistent. TALIS 2018 data show that only 56% of teachers reported that digital technologies were included in their initial preparation and fewer than 40% felt prepared to use them effectively (OECD, 2019). In the Czech context, DigCompEdu provides a framework for competence development (Ministry of Education, Youth and Sports, n.d.), yet the use of digital technologies in teaching remains below the OECD average (OECD, 2023). Teacher-related barriers, including insufficient skills and negative attitudes, continue to constrain effective integration (Schmitz et al., 2022), while systematic preparation for technology integration in initial teacher education often remains limited and is frequently supported through modelling and mentoring during teaching practice (Heine et al., 2024).

Cite this article: Javorcik, T. et al. 2026. "Microlearning in Teacher Education: Effects on Pre-service Teachers' Digital Self-Efficacy and Digital Competence", *Electronic Journal of e-Learning*, 24(3), pp 123-136, https://doi.org/10.34190/ejel.24.3.4790

These challenges became particularly visible during the COVID-19 pandemic, which accelerated educational digitalisation and increased expectations regarding teachers' digital readiness (Weigold and Weigold, 2021; Zancajo, Verger and Bolea, 2022). At the same time, disparities in digital literacy persisted among teachers and across the wider population (Deursen, 2020; Tejedor et al., 2020). European initiatives such as the Digital Education Action Plan 2021–2027 therefore emphasize the need to strengthen digital skills and educational ecosystems, although their success depends largely on teacher preparedness (European Commission, 2020).

A key determinant of successful technology integration is teachers' self-efficacy, which influences their willingness to experiment with digital tools and adopt innovative pedagogical practices. Even in environments with adequate technological infrastructure, low confidence may inhibit pedagogical innovation (Ertmer and Ottenbreit-Leftwich, 2010). Teachers with lower technological self-efficacy are less likely to implement innovative digital practices (Fransson et al., 2019). Self-confidence and motivation are widely recognised as important determinants of learning effectiveness (Hoz et al., 2024; Ormrod, Anderman and Anderman, 2020), while professional knowledge is closely related to instructional quality (Shulman, 1987). Teacher professional identity evolves through experience and reflection (Vanassche and Kelchtermans, 2014), and digitalisation may either strengthen or weaken teachers' confidence depending on their level of competence (Fransson et al., 2019).

Professional development—formal, informal, and self-directed—plays an important role in strengthening technological self-efficacy (Barton and Dexter, 2020). Meta-analytic evidence confirms a strong positive relationship between technological self-efficacy and the TPACK framework (Zeng, Wang and Li, 2022), while structured preparation and continuous professional development positively influence teachers' confidence and technology integration practices (Amponsah et al., 2024). Higher self-efficacy is also associated with greater willingness to adopt innovative tools and overcome barriers (Prestridge, 2012; Tondeur et al., 2017). Consequently, teacher characteristics interact dynamically, with professional self-confidence representing a significant secondary barrier after the availability of technologies and opportunities for further training (Abedi and Ackah-Jnr, 2023).

Taken together, these findings indicate that teacher education needs flexible, structured, and practice-oriented approaches that can support the gradual development of digital competences while strengthening pre-service teachers' confidence in using technologies for teaching. In response to these challenges, new forms of professional learning have emerged within technology-enhanced education, among which microlearning has recently attracted increasing attention.

2. Literature Review

Microlearning represents a promising instructional approach. Although no single, universally accepted definition exists (Mohammed, Wakil and Nawroly, 2018), microlearning is typically characterized by short, learner-centered, and predominantly asynchronous instructional segments (Major and Calandrino, 2018), delivered through modular multimedia units that can be accessed anytime and anywhere (Cronin and Durham, 2024). Individual learning activities usually last between two and ten minutes (Buchem and Hamelmann, 2010) and support distributed practice, which enhances retention (Nowak, Speed and Vuk, 2023), although their effectiveness may vary depending on learners' individual characteristics (Rof, Bikfalvi and Marques, 2024).

The potential effectiveness of microlearning can also be explained through established cognitive and instructional design theories. From the perspective of Cognitive Load Theory, short and clearly structured units may reduce extraneous cognitive load by limiting the amount of information processed at one time and focusing attention on a specific task or concept (Paas, Renkl and Sweller, 2003). Microlearning is also consistent with distributed practice and retrieval-based learning, which emphasize repeated engagement and recall as mechanisms supporting retention (Dunlosky et al., 2013). It should therefore be understood not merely as fragmented content, but as an instructional design approach combining concise content, focused activity, feedback, and repeated practice.

Microlearning aligns well with learning preferences frequently associated with Generation Z. Research suggests that these learners prefer concise multimedia content, mobile accessibility, interactive activities, and rapid feedback (Aivaz and Teodorescu, 2022; Jayathilake et al., 2021; Manzoni et al., 2021; Navarrete et al., 2025; Wang et al., 2023; Zhang et al., 2025). These characteristics correspond closely with core microlearning principles such as short modular units, multimedia segmentation, mobile delivery, and gamified or socially supported learning activities (Arnab et al., 2021; Buchem and Hamelmann, 2010; Göschlberger, 2017). The relationship between learning characteristics frequently attributed to Generation Z and the design principles of microlearning is summarized in Table 1.

Empirical research across higher education, medical education, and professional development indicates that microlearning can enhance short-term learning outcomes, knowledge retention, learner satisfaction, and practical skill development (Dennen, Arslan and Bong, 2024; Gorham, Majumdar and Ogata, 2023; Hlazunova et al., 2024; Javorcik, Kostolanyova and Havlaskova, 2023; Kohnke et al., 2024; Richardson et al., 2023; Román-Sánchez et al., 2023; Waldia et al., 2023;). At the same time, common methodological limitations persist, including small sample sizes, short intervention periods, and a strong reliance on self-report instruments (Al-Zahrani, 2024; Richardson et al., 2023; Román-Sánchez et al., 2023).

Table 1: Structural alignment between Generation Z learning characteristics and microlearning design principles

Generation Z Learning Characteristic	**Corresponding Microlearning Design Principle**
Preference for concise and brief content	Modular micro-lessons (typically 2–10 minutes)
Strong orientation toward multimedia (visual/auditory)	Segmented multimedia instructional design
High level of mobile device usage	Mobile-first approach with anytime, anywhere accessibility
Affinity for gamification and interactive formats	Game-oriented and activity-driven micro-tasks
Preference for social interaction and collaboration	Integration of social media and peer interaction elements
Demand for personalization and autonomy	Adaptive learning systems and individually paced instruction
Expectation of immediacy and rapid feedback	Immediate feedback mechanisms embedded within micro-lessons
Multitasking and fragmented attention	Short, focused learning segments supporting distributed practice

It should be noted, however, that multitasking should not be interpreted as an inherently beneficial learning strategy. Evidence from educational research suggests that media multitasking may interfere with attention, working memory, comprehension, recall, and academic performance (May and Elder, 2018). In the present study, microlearning is therefore not understood as a response that encourages multitasking, but rather as a design strategy intended to reduce cognitive overload, focus attention on clearly defined tasks, and support distributed practice through short, structured learning sequences.

Although microlearning has been associated with increased self-efficacy in several educational contexts, including medical education (Sozmen, 2022; Zarshenas et al., 2022), patient education (Janssen et al., 2023; Rahbar, Zarifsanaiey and Mehrabi, 2024), language education (Prasittichok and Smithsarakarn, 2024), and corporate training (Karlsen, Balsvik and Rønnevik, 2023), its impact on the digital self-efficacy of pre-service teachers remains underexplored. Digital literacy and self-efficacy constitute significant predictors of innovative thinking (Al-Hattami, 2025), and teacher self-efficacy develops primarily through practical experience rather than as a function of gender differences (Symes, Lazarides and Hußner, 2023).

In response to this research gap, the present study aims to examine the impact of a microlearning course on pre-service teachers' perceived self-efficacy in the use of digital technologies for instructional purposes within the Czech educational context. The study addresses the following research questions:

RQ1: *What differences in pre-service teachers' self-efficacy across specific domains of digital skills can be identified before and after completing the microlearning course?*

RQ2: *To what extent did participation in the microlearning course influence pre-service teachers' self-efficacy in individual digital skills, and are these changes statistically significant?*

RQ3: *What is the effect of factors such as gender, age, and mode of study on pre-service teachers' self-efficacy in digital skills before and after completing the course?*

RQ4: *Are there differences in self-efficacy across academic cohorts, and how did the microlearning course affect these differences?*

3. Methods

The study employed a quantitative pre–post design to examine changes in pre-service teachers' perceived self-efficacy in digital competences following completion of a microlearning course. This approach allows comparison between baseline and post-intervention levels of perceived professional confidence. Given that self-efficacy represents a subjective belief in one's capability to perform specific tasks (Bandura, Freeman and Lightsey, 1999), a pre–post measurement approach is considered an appropriate method for identifying changes resulting from mastery experiences (Ertmer and Ottenbreit-Leftwich, 2010; Jahnke et al., 2020; Schmitz et al., 2022).

The research was carried out over four consecutive academic years (2021–2025) at a Czech university (Faculty of Education, University of Ostrava) within a compulsory course focused on the development of pre-service teachers' digital competences and the pedagogical use of digital technologies. The pre-survey was completed by 1,437 respondents and the post-survey by 871 respondents. The discrepancy in the number of responses reflects voluntary participation in the research component and the anonymous nature of data collection. The demographic characteristics of the sample (gender, age categories, mode of study—full-time/part-time) are presented in Table 2. Gender was self-reported and coded as male or female. The inclusion of multiple cohorts made it possible not only to assess the immediate effect of the intervention but also to examine the stability of findings across academic years and to account for potential contextual influences.

Table 2: Structure of the research sample

Variable	Category	2021/2022	2022/2023	2023/2024	2024/2025	Total
		Pre/Post	Pre/Post	Pre/Post	Pre/Post	Pre/Post
Gender	Male	116/61	96/40	86/54	82/40	380/195
	Female	238/159	249/173	300/192	270/152	1057/676
Age	Up to 20	226/127	237/138	250/144	226/116	939/525
	21–30 years	100/78	83/60	93/73	82/57	358/268
	31–40 years	20/13	12/9	26/17	30/10	88/49
	41–50 years	8/2	12/6	15/12	13/8	48/28
	Over 50 years	0/0	1/0	2/0	1/1	4/1
Focus of pedagogical education	Kindergartens	15/9	21/11	11/10	23/10	70/40
	Primary schools (1st stage)	70/51	90/21	122/70	102/14	384/156
	Primary schools (2nd stage)	141/87	121/116	149/95	119/93	530/391
	Secondary schools	128/73	113/65	104/71	108/75	453/284
Form of study	Regular learning	287/183	285/175	303/191	268/150	1143/699
	Distance learning	67/37	60/38	83/55	84/42	294/172

The intervention consisted of a fully online microlearning course structured into modular units composed of short, task-oriented segments (2–10 minutes), in accordance with the defining characteristics of microlearning described in the literature (Buchem and Hamelmann, 2010). The microlearning course *Information Technology in Education* was developed and delivered within the Moodle learning management system (LMS) as a compulsory component of all teacher education programmes at the Faculty of Education, University of Ostrava. Cílem předmětu bylo poskytnout budoucím učitelům obecný základ pro praktické a pedagogické využívání digitálních technologií ve vzdělávání, zatímco oborově specifické aplikace jsou rozvíjeny v navazujících didakticky orientovaných předmětech. The individual modules focused on specific domains of digital competences relevant to instructional practice: **(1)** Text processing and worksheet design, **(2)** Image and graphic editing, **(3)** Multimedia production, **(4)** Online tools and digital sharing, **(5)** Mobile technologies and applications. The course was not designed as discipline-specific training, but as a shared foundation of transversal digital competences relevant to pre-service teachers across educational levels and subject areas.

To ensure that the microlearning units did not function merely as short fragments of instructional content, each module followed a recurring instructional sequence. A typical unit began with a concise introduction to the tool, concept, or pedagogical problem, followed by a short demonstration or guided example. Students then completed an application task in which they created a concrete educational output, such as a worksheet, digital poster, video material, or electronic questionnaire. These tasks were embedded in pedagogically situated scenarios to connect technical skill acquisition with instructional use.

The modules also incorporated elements supporting recall, reflection, consolidation, and feedback. After selected activities, students checked their understanding through short verification tasks, compared their outputs with stated criteria, or reflected on how the created digital material could be used in teaching practice. Feedback was provided through automated Moodle activities, task completion criteria, model examples, or teacher comments on submitted outputs. Through this sequence, the course combined content segmentation, modular progression, self-paced learning, active application, and immediate feedback, thereby reinforcing mastery experiences as a core mechanism underlying the development of perceived self-efficacy (Bandura, Freeman and Lightsey, 1999).

Perceived self-efficacy was measured using a structured questionnaire targeting the specific domains of digital competences aligned with the course objectives. The use of a self-report instrument is grounded in the theoretical conceptualization of self-efficacy as an individual's subjective belief in their capability to perform specific tasks (Bandura, Freeman and Lightsey, 1999). Items were rated on a Likert scale and included the response option "N – unable to assess," enabling the identification of areas in which respondents lacked prior experience. The questionnaire was administratively identical in both the pre-survey and post-survey, ensuring full comparability of measurements. The internal consistency of the instrument across cohorts was high (Cronbach's α = .83–.95; see Table 3), confirming its reliability.

Table 3: Cronbach's alpha for pre-survey and post-survey questionnaires in 2021/2022 to 2024/2025

Academic year	Pre-survey	Post-survey
2021/2022	0.88	0.95
2022/2023	0.85	0.95
2023/2024	0.83	0.94
2024/2025	0.86	0.92

The pre-survey was administered prior to the commencement of instruction, and the post-survey immediately after its completion. Due to the anonymous nature of data collection, it was not possible to match individual responses; consequently, a paired (dependent) statistical model could not be employed. The analysis was therefore conducted at the level of independent samples. This approach is methodologically appropriate in contexts where anonymity precludes longitudinal tracking of individuals, while still permitting the assessment of change at the group level. Non-parametric tests for independent samples were employed due to the ordinal nature of Likert-scale data.

The data were analyzed using descriptive and inferential statistics, with pre- and post-survey differences tested at α = .05 in NCSS 2023 statistical software (64-bit), version 23.0.2. Responses to the two open-ended items were analyzed using Braun and Clarke's (2012) open-ended response coding and analysis methodology. Participation was voluntary and anonymous, no personally identifiable data were collected, and the study followed institutional ethical standards.

4. Results

The results are presented according to the research questions: changes in self-efficacy across digital skill areas, their statistical significance, the role of demographic variables, and cohort effects across four academic years.

4.1 Differences in the Self-efficacy Ratings of Student Teachers Before and After Completing the Microlearning Course

Aggregated data across all observed academic years provide an overview of students' perceived self-efficacy in specific areas of digital technology use before and after completing the microlearning course (Table 4). Three indicators were used in the analysis: mean (M), median (Mdn), and standard deviation (SD).

Before completing the course, students reported the highest levels of confidence in word processing (M = 1.76) and creating presentations (M = 1.73). These results indicate that commonly used office tools were already familiar to most participants prior to the intervention. In contrast, creating an interactive worksheet (M = 3.03) and editing a video (M = 3.07) were evaluated as the most demanding activities. Higher variability of responses in these areas (SD > 1.2) suggests considerable differences in students' prior experience with more advanced digital tasks.

Following completion of the microlearning course, mean values decreased in almost all skill domains, indicating higher perceived competence. The most visible improvements were observed in creating a worksheet (M = 2.31 to M = 1.67) and creating a poster or flyer (M = 2.38 to M = 1.62). Although editing a video and creating an interactive worksheet remained among the more challenging activities, both domains showed substantial improvement (e.g., 3.07 to 1.98). In addition, slightly lower SD values across most areas indicate more consistent levels of confidence among students after completing the course. Changes in the relative ranking of individual digital skills before and after the course are summarised in Table 5.

A further indicator of change is the frequency of the response option "N – unable to assess." In the pre-survey this option was selected 1,445 times, whereas in the post-survey it appeared only 42 times. This substantial reduction suggests that students developed greater awareness of the assessed competence domains and were better able to evaluate their own skills after completing the course.

Table 4: Results of the self-efficacy assessment in each area of digital skills before and after the course for the entire observation period

Digital skill area	Pre-survey			Post-survey		
	M	Mdn	SD	M	Mdn	SD
Word processing in a word processor	1.76	2	0.74	1.58	1	0.95
Data processing in a spreadsheet editor	2.80	3	0.99	2.48	2	0.97
Creating a presentation	1.74	2	0.76	1.53	1	0.97
Creating a worksheet	2.31	2	0.95	1.67	1	0.95
Creating an interactive worksheet	3.03	3	1.04	1.85	2	0.94
Creating an educational infographic	2.60	3	1.04	1.91	2	0.98
Creating a poster/flyer	2.38	2	1.08	1.62	1	0.99
Editing a video	3.07	3	1.23	1.98	2	1.00
Using shared documents	2.15	2	1.01	1.65	1	0.97
Creating an electronic questionnaire or test	2.40	2	1.08	1.53	1	0.97
Navigating an educational application	2.44	2	1.00	1.92	2	0.93

M – Mean, Mdn – Median, SD – Standard deviation

Table 5: Changes in the ranking of digital skills before and after the microlearning course

Digital skill area	Pre-survey M	Pre-survey rank	Post-survey M	Post-survey rank	Rank change
Creating a presentation	1.74	1	1.53	1	0
Word processing in a word processor	1.76	2	1.58	3	+1
Using shared documents	2.15	3	1.65	5	+2
Creating a worksheet	2.31	4	1.67	6	+2
Creating a poster/flyer	2.38	5	1.62	4	−1
Creating an electronic questionnaire or test	2.40	6	1.53	2	−4
Navigating an educational application	2.44	7	1.92	9	+2
Creating an educational infographic	2.60	8	1.91	8	0
Data processing in a spreadsheet editor	2.79	9	2.48	11	+2
Creating an interactive worksheet	3.03	10	1.85	7	−3
Editing a video	3.07	11	1.98	10	−1

M – Mean

4.2 Impact of the Microlearning Course on Students' Self-efficacy in Digital Skills

To determine whether participation in the microlearning course was associated with changes in students' self-efficacy, responses from the pre-survey and post-survey were compared using the Mann–Whitney U test (Table 6). Statistically significant differences were identified in all examined areas ($p < .001$).

The most pronounced changes were observed in creating an interactive worksheet ($U = 186,812$; $Z = –23.51$), creating an electronic questionnaire or test ($U = 278,438$; $Z = –20.78$), and editing a video ($U = 268,990$; $Z = –19.84$). These results indicate that the microlearning course was particularly associated with improvements in areas involving the creation of interactive and multimedia educational materials, which students initially perceived as more demanding tasks.

Smaller, although still statistically significant, differences were observed in word processing ($U = 499,477$; $Z = –9.25$) and data processing in spreadsheets ($U = 509,125$; $Z = –7.16$). These domains already showed relatively high levels of confidence in the pre-survey, leaving less room for substantial improvement.

Overall, the results indicate that participation in the microlearning course was associated with improvements in self-efficacy across all examined digital skill domains, although the magnitude of change varied depending on the complexity of the activity and students' prior experience.

Table 6: Analysis of differences in students' self-efficacy before and after the microlearning course using Mann-Whitney U test

Digital skill area	Pre-survey		Post-survey		U	Z	p-value
	M	Mdn	M	Mdn			
Word processing in a word processor	1.76	2	1.58	1	499477	-9.2543	p < .001
Data processing in a spreadsheet editor	2.80	3	2.48	2	509125	-7.1593	p < .001
Creating a presentation	1.74	2	1.53	1	486593.5	-10.141	p < .001
Creating a worksheet	2.31	2	1.67	1	329173	-17.13	p < .001
Creating an interactive worksheet	3.03	3	1.85	2	186812	-23.507	p < .001
Creating an educational infographic	2.60	3	1.91	2	333552.5	-15.414	p < .001
Creating a poster/flyer	2.38	2	1.62	1	330940.5	-18.057	p < .001
Editing a video	3.07	3	1.98	2	268990.5	-19.843	p < .001
Using shared documents	2.15	2	1.65	1	411659	-13.429	p < .001
Creating an electronic questionnaire or test	2.40	2	1.53	1	278438	-20.779	p < .001
Navigating an educational application	2.44	2	1.92	2	398576	-12.696	p < .001

M – Mean, Mdn – Median, U – U-value of the statistic, Z – Z-value of the statistic, Significance level $\alpha = 0.01$

4.3 Impact of Gender, Age, and Form of Study on the Development of Self-Efficacy

The subsequent analysis examined potential differences in self-efficacy development in relation to gender, age, and form of study (regular or distance learning). Statistical testing using the Kruskal–Wallis One-Way ANOVA identified several statistically significant differences between groups (Table 7).

Table 7: Statistically significant differences in self-efficacy of student teachers by gender, age and form of study (Kruskal-Wallis One-Way ANOVA)

	Pre-survey			Post-survey		
Digital skill area	Gender[1]	Age[2]	Form of study[3]	Gender	Age	Form of study
Word processing in a word processor	.00876 (F)	—	—	.00002 (F)	—	—
Data processing in a spreadsheet editor	.00000 (M)	.00088 (31-40)	.00000 (DL)	—	.02599 (41-50)	.00159 (DL)
Creating a presentation	.00001 (F)	.00000 (<20)	.00000 (RL)	.00001 (F)	.00001 (<20)	.00001 (RL)

	Pre-survey			Post-survey		
Digital skill area	**Gender[1]**	**Age[2]**	**Form of study[3]**	**Gender**	**Age**	**Form of study**
Creating a worksheet	.00061 (F)	—	—	.00000 (F)	—	—
Creating an interactive worksheet	—	—	—	.00184 (F)	.02112 (<20)	.01969 (RL)
Developing an educational infographic	—	.002479 (<20)	—	.00507 (F)	—	—
Creating a poster/flyer	.00085 (F)	—	—	.00001 (F)	—	—
Editing a video	.00281 (M)	.00281 (31-40)	—	.03784 (F)	—	—
Using shared documents	—	—	.00179 (RL)	.01132 (F)	—	—
Creating an electronic questionnaire or test	—	.00099 (<20)	.00131 (RL)	.02348 (F)	—	—
Navigating an educational application	—	.00001 (<20)	.00358 (RL)	.01175 (F)	—	—

[1]Female, male (self-reported)

[2]Up to 20 years of age, 21-30 years of age, 31-40 years of age, 41-50 years of age, over 50 years of age

[3]Regular learning (RL), distance learning (DL)

In the pre-survey, women reported higher self-efficacy in tasks related to commonly used productivity tools, including word processing, creating presentations, creating worksheets, and creating posters or flyers. In contrast, men reported higher self-efficacy in spreadsheet processing and video editing, which may reflect different prior technological experiences.

After completing the course, women showed significant increases in confidence particularly in creating an interactive worksheet and editing a video, suggesting that gender-related differences in these domains were partially reduced during the course. This pattern indicates that structured training in digital competences may contribute to reducing gender disparities in perceived technological abilities.

Age differences were primarily observed in the pre-survey phase. Younger students (below 20 years of age) reported higher self-efficacy in creating presentations, developing infographics, and navigating educational applications, whereas students aged 31–40 years showed higher confidence in spreadsheet processing and video editing. After the course, the number of statistically significant differences between age groups decreased, suggesting a partial levelling of self-efficacy across generations.

A similar pattern was observed in the comparison between forms of study. Before the course, regular students reported higher confidence in presentation creation, whereas distance students reported higher confidence in spreadsheet processing. After course completion, these differences became less pronounced, which may reflect the homogenising influence of exposure to the same instructional content.

4.4 Four-Year Comparison of Student Teachers' Self-efficacy in the Use of Digital Technologies

To examine potential cohort effects, self-efficacy levels were compared across four academic years (2021/2022–2024/2025). Differences between cohorts were analysed using the Kruskal–Wallis One-Way ANOVA test.

In the pre-survey, statistically significant differences between academic years were identified in two skill areas: data processing in spreadsheets ($p = 0.02$) and creating posters or flyers ($p = 0.004$). Students in the academic year 2021/2022 reported higher self-efficacy in spreadsheet processing, while students in 2024/2025 reported higher confidence in poster and flyer creation.

However, no statistically significant differences between cohorts were observed in the post-survey results. This finding suggests that the structure and design of the microlearning course contributed to reducing differences between student cohorts and resulted in more uniform levels of perceived self-efficacy across groups.

Overall, the four-year comparison indicates that the microlearning course produced consistent outcomes across different student cohorts, suggesting that the instructional design maintained comparable effectiveness over time.

4.5 Open-ended Responses

Responses to the open-ended questions complemented the quantitative findings and provided deeper insight into students' perceived strengths and weaknesses in the use of digital technologies.

In the pre-survey, many respondents indicated limited confidence in their digital abilities and were often unable to identify specific strengths. When strengths were mentioned, they most frequently related to commonly used tools such as MS Word, PowerPoint, Canva, or basic graphic editing, reflecting a relatively basic level of digital competence prior to the course.

In contrast, post-survey responses demonstrated both an increase in the number of reported strengths and greater diversity in the technologies mentioned. Students frequently referred to multimedia creation, interactive educational tools, and the pedagogical integration of digital technologies, including the use of AI tools, virtual reality, augmented reality, and digital content creation for personalised learning.

Despite the overall improvement, several areas remained challenging. Students frequently mentioned databases, programming, video editing, and web design as weaker areas both before and after the course. These topics were only marginally represented in the course structure, which may explain their continued presence among perceived weaknesses. Artificial intelligence also appeared as a newly mentioned area in the post-survey responses, indicating increased awareness of the topic while simultaneously reflecting uncertainty regarding its pedagogical application.

Overall, the qualitative findings support the quantitative results by illustrating a shift from basic operational familiarity with digital tools toward a broader and more pedagogically oriented understanding of digital technology use in education.

5. Discussion

This study found statistically significant improvements in pre-service teachers' self-efficacy across all examined digital skill domains after completing a microlearning course.

The most pronounced improvements were observed in areas initially perceived as the most challenging, particularly in the creation of interactive worksheets, instructional video materials, and electronic assessments. These tasks are complex pedagogical–technological activities because they require students to use digital tools to create instructional artefacts with a clear pedagogical function. Comparable findings have been reported in microlearning research across educational and professional contexts, where participation in short, structured modules resulted in improvements in both learning outcomes and self-efficacy (Rahbar, Zarifsanaiey and Mehrabi, 2024; Richardson et al., 2023; Zarshenas et al., 2022;). Similar effects have been documented in medical and professional education, including the training of healthcare professionals (Janssen et al., 2023; Zarshenas et al., 2022), supporting the theoretical assumption that mastery experiences constitute a central mechanism in the development of self-efficacy (Bandura, Freeman and Lightsey, 1999).

The substantial reduction in responses marked "N – unable to assess" further indicates enhanced metacognitive awareness and greater confidence in evaluating one's own competences. At the same time, isolated decreases in post-survey self-efficacy scores may be interpreted as a process of recalibration rather than regression. As students develop a deeper understanding of task complexity, they may reassess previously overestimated perceptions of their own competence. This interpretation aligns with research highlighting discrepancies between perceived and externally evaluated competence among pre-service teachers (Dassa and Nichols, 2019) and with models of reflective professional development in digital competence (Ertmer and Ottenbreit-Leftwich, 2010; Tondeur et al., 2017).

With respect to demographic variables, gender differences were evident in the pre-survey but became partially attenuated following course completion. Women initially reported higher self-efficacy in traditional productive tasks, whereas men demonstrated greater confidence in more technically demanding domains. The equalizing effect observed after the intervention suggests that structured development of digital competences may contribute to reducing gender disparities. The literature on academic self-efficacy presents mixed findings: some studies report higher confidence among men, while others indicate higher academic self-efficacy among women or emphasize women's stronger emotional competences supporting performance (Kurniawan et al., 2022; Salavera, Usán and Jarie, 2017; Webb-Williams, 2014). These inconsistent results reinforce the view that self-efficacy is context-dependent and shaped by learning experiences rather than by fixed personal characteristics.

Age and mode of study also influenced baseline levels of self-efficacy. Older students and those enrolled in part-time programmes reported higher confidence in certain domains in the pre-survey, possibly reflecting accumulated experience and more developed coping strategies (Bausch, Michel and Sonntag, 2014; Doba et al., 2016). Conversely, younger individuals—particularly women—may exhibit lower initial self-efficacy (Szota et al., 2024). Post-survey findings, however, revealed an equalizing effect across age groups and modes of study, suggesting that appropriately designed online educational interventions can significantly strengthen self-efficacy (Blazer et al., 2012). Nevertheless, research in practice-based training contexts (e.g., cardiopulmonary resuscitation instruction) indicates that face-to-face instruction may, in some cases, produce stronger increases in confidence for highly procedural skills (Ko, Kim and Cho, 2023), underscoring the importance of instructional design.

Longitudinal comparisons across four academic years further support the homogenizing effect of the microlearning course. Although differences in baseline self-efficacy were observed among cohorts in the pre-survey—likely influenced by prior educational quality, learning environments, and other contextual factors (Duraku et al., 2022)—these differences were no longer apparent in the post-survey. This finding suggests that the systematic implementation of microlearning may stabilize competence development across diverse student populations.

Several limitations should be acknowledged. Because the surveys were anonymous, individual responses could not be paired across measurement points, limiting causal interpretation at the individual level. Differences in the number of respondents between the pre- and post-survey may also indicate attrition bias. Furthermore, the study measured perceived self-efficacy rather than objectively assessed digital performance; the findings should therefore be interpreted as evidence of changes in students' confidence, not as direct evidence of demonstrated mastery of digital competences. Since the post-survey was administered immediately after course completion, the study also does not allow conclusions about the long-term stability of the observed gains. Future research should therefore incorporate performance-based measures, analysis of student-created digital artefacts, delayed post-tests, and evidence of transfer into authentic pedagogical practice.

Overall, the results confirm that microlearning can substantially enhance the digital self-efficacy of pre-service teachers while simultaneously contributing to the reduction of disparities associated with gender, age, mode of study, and cohort differences. When systematically integrated into teacher education, microlearning represents a promising strategy for fostering both competence development and the formation of professional confidence within digitally enriched educational environments.

6. Conclusion

This study provides longitudinal empirical evidence that a systematically designed microlearning course can significantly enhance the digital self-efficacy of pre-service teachers across multiple domains of digital skills. The most substantial gains were observed in areas initially perceived as the most challenging, suggesting that microlearning may be particularly effective in fostering complex pedagogical–technological competences.

The findings further indicate a homogenizing effect of the intervention, as differences associated with gender, age, mode of study, and cohort affiliation were reduced or fully eliminated following course completion. This suggests that a structured, modular approach to digital competence development can contribute to stabilizing and equalizing competence levels across heterogeneous groups of pre-service teachers.

The results support the theoretical assumptions of the social-cognitive conceptualization of self-efficacy and confirm the importance of mastery experiences as a central mechanism in the development of professional confidence. Microlearning thus emerges not only as an effective tool for the acquisition of discrete skills but also as a strategy that supports the formation of professional identity within a digitally transformed educational environment.

Despite limitations arising from the use of self-report instruments and the absence of paired individual-level tracking, the study indicates that the systematic integration of microlearning into teacher education represents a promising pathway for strengthening the digital readiness of future teachers. Future research should examine the long-term transfer of perceived competences into authentic pedagogical practice and combine self-report data with performance-based indicators.

AI Statement: ChatGPT was used solely for language editing and grammatical refinement. The ideas, structure, interpretations, and conclusions presented in this manuscript are entirely the authors' original work.

Ethics Statement: The study was approved by the Ethics Committee of the Faculty of Education, University of Ostrava. Participants were informed about the purpose of the study and their inclusion in the research and participated voluntarily. All data were collected and analysed anonymously.

References

Abedi, E.A. and Ackah-Jnr, F.R. (2023) 'First-order barriers still matter in teachers' use of technology: An Exploratory study of multi-stakeholder perspectives of technology integration barriers', *International Journal of Education and Development using Information and Communication Technology*, 19(2), pp. 148–165.

Aivaz, K.A. and Teodorescu, D. (2022) 'College Students' Distractions from Learning Caused by Multitasking in Online vs. Face-to-Face Classes: A Case Study at a Public University in Romania', *International Journal of Environmental Research and Public Health*, 19(18), p. 11188. Available at: https://doi.org/10.3390/ijerph191811188.

Al-Hattami, H.M. (2025) 'Understanding how digital accounting education fosters innovation: The moderating roles of technological self-efficacy and digital literacy', *The International Journal of Management Education*, 23(2), p. 101131. Available at: https://doi.org/10.1016/j.ijme.2025.101131.

Al-Zahrani, A.M. (2024) 'Enhancing postgraduate students' learning outcomes through Flipped Mobile-Based Microlearning', *Research in Learning Technology*, 32. Available at: https://doi.org/10.25304/rlt.v32.3110.

Amponsah, K.D., Adu-Gyamfi, K., Awoniyi, F.C. and Commey-Mintah, P. (2024) 'Navigating academic performance: Unravelling the relationship between emotional intelligence, learning styles, and science and technology self-efficacy among preservice science teachers', *Heliyon*, 10(9), p. e29474. Available at: https://doi.org/10.1016/j.heliyon.2024.e29474.

Arnab, S., Walaszczyk, L., Lewis, M. and Kernaghan-Andrews, S. (2021) 'Designing Mini-Games as Micro-Learning Resources for Professional Development in Multi-Cultural Organisations', *Electronic Journal of e-Learning*, 19(2), pp. 44–58. Available at: https://doi.org/10.34190/ejel.19.2.2141.

Bandura, A., Freeman, W.H. and Lightsey, R. (1999) 'Self-Efficacy: The Exercise of Control', *Journal of Cognitive Psychotherapy*, 13(2), pp. 158–166. Available at: https://doi.org/10.1891/0889-8391.13.2.158.

Barton, E.A. and Dexter, S. (2020) 'Sources of teachers' self-efficacy for technology integration from formal, informal, and independent professional learning', *Educational Technology Research and Development*, 68(1), pp. 89–108. Available at: https://doi.org/10.1007/s11423-019-09671-6.

Bausch, S., Michel, A. and Sonntag, K. (2014) 'How gender influences the effect of age on self-efficacy and training success', *International Journal of Training and Development*, 18(3), pp. 171–187. Available at: https://doi.org/10.1111/ijtd.12027.

Blazer, K.R., Christie, C., Uman, G. and Weitzel, J.N. (2012) 'Impact of Web-Based Case Conferencing on Cancer Genetics Training Outcomes for Community-Based Clinicians', *Journal of Cancer Education*, 27(2), pp. 217–225. Available at: https://doi.org/10.1007/s13187-012-0313-8.

Braun, V. and Clarke, V. (2012) 'Thematic analysis.', in H. Cooper, P.M. Camic, D.L. Long, A.T. Panter, D. Rindskopf, and K.J. Sher (eds) *APA handbook of research methods in psychology, Vol 2: Research designs: Quantitative, qualitative, neuropsychological, and biological.* Washington: American Psychological Association, pp. 57–71. Available at: https://doi.org/10.1037/13620-004.

Buchem, I. and Hamelmann, H. (2010) 'Microlearning: a strategy for ongoing professional development', *eLearning Papers*, 21(7), pp. 1–15.

Cronin, J. and Durham, M.L. (2024) 'Microlearning: A Concept Analysis', *CIN: Computers, Informatics, Nursing*, 42(6), pp. 413–420. Available at: https://doi.org/10.1097/CIN.0000000000001122.

Dassa, L. and Nichols, B. (2019) 'Self-Efficacy or Overconfidence? Comparing Preservice Teacher Self-Perceptions of Their Content Knowledge and Teaching Abilities to the Perceptions of Their Supervisors', *The New Educator*, 15(2), pp. 156–174. Available at: https://doi.org/10.1080/1547688X.2019.1578447.

Dennen, V.P., Arslan, Ö. and Bong, J. (2024) 'Optional embedded microlearning challenges: Promoting Self-Directed Learning and Extension in a Higher Education Course', *Educational Technology & Society*, 27(1), pp. 166–182.

Deursen, A.J.V. (2020) 'Digital Inequality During a Pandemic: Quantitative Study of Differences in COVID-19–Related Internet Uses and Outcomes Among the General Population', *Journal of Medical Internet Research*, 22(8), p. e20073. Available at: https://doi.org/10.2196/20073.

Doba, N., Tokuda, Y., Saiki, K., Kushiro, T., Hirano, M., Matsubara, Y. and Hinohara, S. (2016) 'Assessment of Self-Efficacy and its Relationship with Frailty in the Elderly', *Internal Medicine*, 55(19), pp. 2785–2792. Available at: https://doi.org/10.2169/internalmedicine.55.6924.

Dunlosky, J., Rawson, K.A., Marsh, E.J., Nathan, M.J. and 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), pp. 4–58. Available at: https://doi.org/10.1177/1529100612453266.

Duraku, Z.H., Blakaj, V., Shllaku Likaj, E., Boci, L. and Shtylla, H. (2022) 'Professional training improves early education teachers' knowledge, skills, motivation, and self-efficacy', *Frontiers in Education*, 7, p. 980254. Available at: https://doi.org/10.3389/feduc.2022.980254.

Ertmer, P.A. and Ottenbreit-Leftwich, A.T. (2010) 'Teacher Technology Change: How Knowledge, Confidence, Beliefs, and Culture Intersect', *Journal of Research on Technology in Education*, 42(3), pp. 255–284. Available at: https://doi.org/10.1080/15391523.2010.10782551.

European Commission (2020) 'Digital Education Action Plan (2021–2027)'. European Commission. Available at: https://education.ec.europa.eu/focus-topics/digital-education/actions/plan.

Fransson, G., Holmberg, J., Lindberg, O.J. and Olofsson, A.D. (2019) 'Digitalise and capitalise? Teachers' self-understanding in 21st-century teaching contexts', *Oxford Review of Education*, 45(1), pp. 102–118. Available at: https://doi.org/10.1080/03054985.2018.1500357.

Gorham, T., Majumdar, R. and Ogata, H. (2023) 'Analyzing learner profiles in a microlearning app for training language learning peer feedback skills', *Journal of Computers in Education*, 10(3), pp. 549–574. Available at: https://doi.org/10.1007/s40692-023-00264-0.

Göschlberger, B. (2017) 'Social Microlearning Motivates Learners to Pursue Higher-Level Cognitive Objectives', in G. Vincenti, A. Bucciero, M. Helfert, and M. Glowatz (eds) *E-Learning, E-Education, and Online Training*. Cham: Springer International Publishing (Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering), pp. 201–208. Available at: https://doi.org/10.1007/978-3-319-49625-2_24.

Heine, S., Krepf, M., Jäger-Biela, D.J., Gerhard, K., Stollenwerk, R. and König, J. (2024) 'Preservice teachers' professional knowledge for ICT integration in the classroom: Analysing its structure and its link to teacher education', *Education and Information Technologies*, 29(9), pp. 11043–11075. Available at: https://doi.org/10.1007/s10639-023-12212-7.

Hlazunova, O.H., Schlauderer, R., Korolchuk, V.I., Voloshyna, T.V. and Saiapina, T.P. (2024) 'Microlearning technology based on video content: advantages, methodology and quality factors', *Journal of Physics: Conference Series*, 2871(1), p. 012028. Available at: https://doi.org/10.1088/1742-6596/2871/1/012028.

Hoz, A.D.L., Melo, L., Cañada, F. and Cubero, J. (2024) 'Educational robotics for science and mathematics teaching: Analysis of pre-service teachers' perceptions and self-confidence', *Heliyon*, 10(21), p. e40032. Available at: https://doi.org/10.1016/j.heliyon.2024.e40032.

Jahnke, I., Lee, Y.-M., Pham, M., He, H. and Austin, L. (2020) 'Unpacking the Inherent Design Principles of Mobile Microlearning', *Technology, Knowledge and Learning*, 25(3), pp. 585–619. Available at: https://doi.org/10.1007/s10758-019-09413-w.

Janssen, A., Shah, K., Rabbets, M., Nagrial, A., Pene, C., Zachulski, C., Phillips, J.L., Harnett, P. and Shaw, T. (2023) 'Feasibility of Microlearning for Improving the Self-Efficacy of Cancer Patients Managing Side Effects of Chemotherapy', *Journal of Cancer Education*, 38(5), pp. 1697–1709. Available at: https://doi.org/10.1007/s13187-023-02324-6.

Javorcik, T., Kostolanyova, K. and Havlaskova, T. (2023) 'Microlearning in the Education of Future Teachers: Monitoring and Evaluating Students' Activity in a Microlearning Course', *Electronic Journal of e-Learning*, 21(1), pp. 13–25. Available at: https://doi.org/10.34190/ejel.21.1.2623.

Jayathilake, H.D., Daud, D., Eaw, H.C. and Annuar, N. (2021) 'Employee development and retention of Generation-Z employees in the post-COVID-19 workplace: a conceptual framework', *Benchmarking: An International Journal*, 28(7), pp. 2343–2364. Available at: https://doi.org/10.1108/BIJ-06-2020-0311.

Karlsen, J.T., Balsvik, E. and Rønnevik, M. (2023) 'A study of employees' utilization of microlearning platforms in organizations', *The Learning Organization*, 30(6), pp. 760–776. Available at: https://doi.org/10.1108/TLO-07-2022-0080.

Ko, J.-S., Kim, S.-R. and Cho, B.-J. (2023) 'The Effect of Cardiopulmonary Resuscitation (CPR) Education on the CPR Knowledge, Attitudes, Self-Efficacy, and Confidence in Performing CPR among Elementary School Students in Korea', *Healthcare*, 11(14), p. 2047. Available at: https://doi.org/10.3390/healthcare11142047.

Koehler, M.J. and Mishra, P. (2005) 'What Happens When Teachers Design Educational Technology? The Development of Technological Pedagogical Content Knowledge', *Journal of Educational Computing Research*, 32(2), pp. 131–152. Available at: https://doi.org/10.2190/0EW7-01WB-BKHL-QDYV.

Kohnke, L., Foung, D., Zou, D. and Jiang, M. (2024) 'Creating the conditions for professional digital competence through microlearning', *Educational Technology & Society*, 27(1). Available at: https://doi.org/10.30191/ETS.202401_27(1).SP05.

Kurniawan, C., Soepriyanto, Y., Zakaria, Z. and Aulia, F. (2022) 'Gender Differences in E-Learning Self-Efficacy during Pandemic Covid-19': *2nd World Conference on Gender Studies (WCGS 2021)*, Malang, Indonesia. Available at: https://doi.org/10.2991/assehr.k.220304.011.

Major, A. and Calandrino, T. (2018) 'Beyond Chunking: Microlearning Secrets for Effective Online Design', *Distance Learning*, 15(2), pp. 27–30.

Manzoni, B., Caporarello, L., Cirulli, F. and Magni, F. (2021) 'The Preferred Learning Styles of Generation Z: Do They Differ from the Ones of Previous Generations?', in C. Metallo, M. Ferrara, A. Lazazzara, and S. Za (eds) *Digital Transformation and Human Behavior*. Cham: Springer International Publishing (Lecture Notes in Information Systems and Organisation), pp. 55–67. Available at: https://doi.org/10.1007/978-3-030-47539-0_5.

May, K.E. and Elder, A.D. (2018) 'Efficient, helpful, or distracting? A literature review of media multitasking in relation to academic performance', *International Journal of Educational Technology in Higher Education*, 15(1), p. 13. Available at: https://doi.org/10.1186/s41239-018-0096-z.

Ministry of Education, Youth and Sports (n.d.) 'Framework of digital competences for teachers'. Ministry of Education, Youth and Sports. Available at: https://msmt.gov.cz/vzdelavani/ramec-digitalnich-kompetenci-ucitele.

Mohammed, G.S., Wakil, K. and Nawroly, S.S. (2018) 'The Effectiveness of Microlearning to Improve Students' Learning Ability', *International Journal of Educational Research Review*, 3(3), pp. 32–38. Available at: https://doi.org/10.24331/ijere.415824.

Navarrete, E., Nehring, A., Schanze, S., Ewerth, R. and Hoppe, A. (2025) 'A Closer Look into Recent Video-based Learning Research: A Comprehensive Review of Video Characteristics, Tools, Technologies, and Learning Effectiveness', *International Journal of Artificial Intelligence in Education*, 35(4), pp. 1631–1694. Available at: https://doi.org/10.1007/s40593-025-00481-x.

Nowak, G., Speed, O. and Vuk, J. (2023) 'Microlearning activities improve student comprehension of difficult concepts and performance in a biochemistry course', *Currents in Pharmacy Teaching and Learning*, 15(1), pp. 69–78. Available at: https://doi.org/10.1016/j.cptl.2023.02.010.

OECD (2019) *TALIS 2018 Results (Volume I): Teachers and School Leaders as Lifelong Learners*. OECD Publishing (TALIS). Available at: https://doi.org/10.1787/1d0bc92a-en.

OECD (2023) *PISA 2022 Results (Volume II): Learning During – and From – Disruption*. OECD Publishing (PISA). Available at: https://doi.org/10.1787/a97db61c-en.

Ormrod, J.E., Anderman, E.M. and Anderman, L.H. (2020) *Educational psychology: developing learners*. Tenth edition. Hoboken, NJ: Pearson.

Paas, F., Renkl, A. and Sweller, J. (2003) 'Cognitive Load Theory and Instructional Design: Recent Developments', *Educational Psychologist*, 38(1), pp. 1–4. Available at: https://doi.org/10.1207/S15326985EP3801_1.

Prasittichok, P. and Smithsarakarn, P.N. (2024) 'The Effects of Microlearning on EFL Students' English Speaking: A Systematic Review and Meta-Analysis', *International Journal of Learning, Teaching and Educational Research*, 23(4), pp. 525–546. Available at: https://doi.org/10.26803/ijlter.23.4.27.

Prestridge, S. (2012) 'The beliefs behind the teacher that influences their ICT practices', *Computers & Education*, 58(1), pp. 449–458. Available at: https://doi.org/10.1016/j.compedu.2011.08.028.

Rahbar, S., Zarifsanaiey, N. and Mehrabi, M. (2024) 'The effectiveness of social media-based microlearning in improving knowledge, self-efficacy, and self-care behaviors among adult patients with type 2 diabetes: an educational intervention', *BMC Endocrine Disorders*, 24(1), p. 99. Available at: https://doi.org/10.1186/s12902-024-01626-0.

Redecker, C. (2017) *European framework for the digital competence of educators: DigCompEdu.* LU: Publications Office. Available at: https://doi.org/10.2760/178382.

Richardson, M.X., Aytar, O., Hess-Wiktor, K. and Wamala-Andersson, S. (2023) 'Digital Microlearning for Training and Competency Development of Older Adult Care Personnel: Mixed Methods Intervention Study to Assess Needs, Effectiveness, and Areas of Application', *JMIR Medical Education*, 9, p. e45177. Available at: https://doi.org/10.2196/45177.

Rof, A., Bikfalvi, A. and Marques, P. (2024) 'Exploring learner satisfaction and the effectiveness of microlearning in higher education', *The Internet and Higher Education*, 62, p. 100952. Available at: https://doi.org/10.1016/j.iheduc.2024.100952.

Román-Sánchez, D., De-La-Fuente-Rodríguez, J.M., Paramio, A., Paramio-Cuevas, J.C., Lepiani-Díaz, I. and López-Millan, M. (2023) 'Evaluating satisfaction with teaching innovation, its relationship to academic performance and the application of a video-based microlearning', *Nursing Open*, 10(9), pp. 6067–6077. Available at: https://doi.org/10.1002/nop2.1828.

Salavera, C., Usán, P. and Jarie, L. (2017) 'Emotional intelligence and social skills on self-efficacy in Secondary Education students. Are there gender differences?', *Journal of Adolescence*, 60(1), pp. 39–46. Available at: https://doi.org/10.1016/j.adolescence.2017.07.009.

Schmitz, M.-L., Antonietti, C., Cattaneo, A., Gonon, P. and Petko, D. (2022) 'When barriers are not an issue: Tracing the relationship between hindering factors and technology use in secondary schools across Europe', *Computers & Education*, 179, p. 104411. Available at: https://doi.org/10.1016/j.compedu.2021.104411.

Shulman, L. (1987) 'Knowledge and Teaching:Foundations of the New Reform', *Harvard Educational Review*, 57(1), pp. 1–23. Available at: https://doi.org/10.17763/haer.57.1.j463w79r56455411.

Sozmen, E.Y. (2022) 'Perspective on pros and cons of microlearning in health education', *Essays in Biochemistry*. Edited by L.V. Mello and H. Watson, 66(1), pp. 39–44. Available at: https://doi.org/10.1042/EBC20210047.

Symes, W., Lazarides, R. and Hußner, I. (2023) 'The development of student teachers' teacher self-efficacy before and during the COVID-19 pandemic', *Teaching and Teacher Education*, 122, p. 103941. Available at: https://doi.org/10.1016/j.tate.2022.103941.

Szota, M., Rogowska, A.M., Kwaśnicka, A. and Chilicka-Hebel, K. (2024) 'The Indirect Effect of Future Anxiety on the Relationship between Self-Efficacy and Depression in a Convenience Sample of Adults: Revisiting Social Cognitive Theory', *Journal of Clinical Medicine*, 13(16), p. 4897. Available at: https://doi.org/10.3390/jcm13164897.

Tejedor, S., Cervi, L., Pérez-Escoda, A. and Jumbo, F.T. (2020) 'Digital Literacy and Higher Education during COVID-19 Lockdown: Spain, Italy, and Ecuador', *Publications*, 8(4), p. 48. Available at: https://doi.org/10.3390/publications8040048.

Tondeur, J., Van Braak, J., Ertmer, P.A. and Ottenbreit-Leftwich, A. (2017) 'Understanding the relationship between teachers' pedagogical beliefs and technology use in education: a systematic review of qualitative evidence', *Educational Technology Research and Development*, 65(3), pp. 555–575. Available at: https://doi.org/10.1007/s11423-016-9481-2.

Vanassche, E. and Kelchtermans, G. (2014) 'Teacher educators' professionalism in practice: Positioning theory and personal interpretative framework', *Teaching and Teacher Education*, 44, pp. 117–127. Available at: https://doi.org/10.1016/j.tate.2014.08.006.

Waldia, N., Sonawane, S., Mali, M. and Jadhav, V. (2023) 'Microlearning strategies for teacher professional development in the era of fourth industrial revolution in India', *Journal of e-Learning and Knowledge Society*, pp. 74-81 Pages. Available at: https://doi.org/10.20368/1971-8829/1135866.

Wang, Q., Sun, F., Wang, X. and Gao, Y. (2023) 'Exploring Undergraduate Students' Digital Multitasking in Class: An Empirical Study in China', *Sustainability*, 15(13), p. 10184. Available at: https://doi.org/10.3390/su151310184.

Webb-Williams, J. (2014) 'Gender differences in school childrens self-efficacy beliefs: Students and teachers perspectives', *Educational Research and Reviews*, 9(3), pp. 75–82. Available at: https://doi.org/10.5897/ERR2013.1653.

Weigold, A. and Weigold, I.K. (2021) 'Measuring confidence engaging in computer activities at different skill levels: Development and validation of the Brief Inventory of Technology Self-Efficacy (BITS)', *Computers & Education*, 169, p. 104210. Available at: https://doi.org/10.1016/j.compedu.2021.104210.

Zancajo, A., Verger, A. and Bolea, P. (2022) 'Digitalization and beyond: the effects of Covid-19 on post-pandemic educational policy and delivery in Europe', *Policy and Society*, 41(1), pp. 111–128. Available at: https://doi.org/10.1093/polsoc/puab016.

Zarshenas, L., Mehrabi, M., Karamdar, L., Keshavarzi, M.H. and Keshtkaran, Z. (2022) 'The effect of micro-learning on learning and self-efficacy of nursing students: an interventional study', *BMC Medical Education*, 22(1), p. 664. Available at: https://doi.org/10.1186/s12909-022-03726-8.

Zeng, Y., Wang, Y. and Li, S. (2022) 'The relationship between teachers' information technology integration self-efficacy and TPACK: A meta-analysis', *Frontiers in Psychology*, 13, p. 1091017. Available at: https://doi.org/10.3389/fpsyg.2022.1091017.

Zhang, Y., Yu, M., He, Y., Liu, F. and Shen, J. (2025) 'Jaw relation recording and transferring: Using digital technology to augment its instruction in predoctoral prosthodontics', *Journal of Dental Education*, 89(2), pp. 186–198. Available at: https://doi.org/10.1002/jdd.13712.

From Self-Perception to Reality in Digital Competence: Findings from the DigComp 2.2 Framework

Andrea Luna [1,2], Mauro Ocaña[1,3], Ana María Ortíz-Colón[2] and Javier Rodríguez-Moreno[2]

[1]Departamento de Ciencias Humanas y Sociales, Universidad de las Fuerzas Armadas ESPE, Sangolquí, 171103, Ecuador

[2]Universidad de Jaén, España

[3]Facultad de Posgrado, Universidad Técnica de Manabí, Portoviejo 130105, Ecuador

apluna@espe.edu.ec (Corresponding Author)
mhocana@espe.edu.ec
aortiz@ujaen.es
jrmoreno@ujaen.es

https://doi.org/10.34190/ejel.24.3.4611

Abstract: The rapid expansion of digital technologies in higher education has revealed persistent discrepancies between self-perception and actual performance in digital skills. Self-assessment of digital competence is notoriously unreliable, often masking critical skill gaps by conflating confidence with actual competence. This study, involving 629 military university students, introduces an alternative methodology designed to overcome the limitations of simple self-perception measures. Results from the DigComp 2.2 framework were used to validate two complementary constructs: (1) the Response Consistency Index (RCI), a promising proxy for metacognitive self-awareness, and (2) the Digital Readiness Index (DRI), a second-order latent model built using Confirmatory Factor Analysis within a Structural Equation Modelling framework that integrates the RCI with perceived competences and objective performance. The findings confirm significant discrepancies between perception and competence, identifying profiles consistent with over- and underconfidence (also known as the Dunning–Kruger effect). Additionally, Problem Solving and Information and Data Literacy are identified as the most deficient domains—a weakness observed even in the highest-performing students. The evidence from this study strengthens the idea that digital readiness is not simply the possession of skills, but rather the intersection of competence and accurate self-awareness of that competence. In the era of artificial intelligence, this emphasises the need for precise educational diagnosis and pedagogical intervention to overcome the illusion of digital competence generated by mere self-reports. By providing instructors and instructional designers with diagnostic tools that go beyond declared competence, this approach contributes directly to the advancement of evidence-based, adaptive e-learning practice. The proposed framework supports the development of learning analytics dashboards and pedagogical interventions that respond not only to what students know, but to how accurately they perceive what they know. These validated instruments—the RCI and the DRI—can directly inform the design of adaptive e-learning interventions and personalised digital training pathways in higher education contexts.

Keywords: Digital competence, DigComp 2.2, Response Consistency Index (RCI), Digital Readiness Index (DRI), Dunning-Kruger effect, Metacognitive calibration

1. Introduction

The European Digital Competence Framework for Citizens (DigComp) is a popular initiative of the European Commission aimed at defining and assessing essential digital competences in the current technological landscape. Since its inception in 2013 and subsequent updates culminating in version 2.2 (2022), DigComp has become a widely adopted reference framework across multiple contexts. Digital competence is defined as the knowledge, skills, and experience required to understand, evaluate, implement, and promote digital technologies (Chen, Ebrahimi and Cheng, 2025). As such, DigComp has served as a foundational standard for the development and evaluation of digital competence within and beyond Europe (Balyk et al., 2019; Bartolomé et al., 2025; Cabezas-González, Casillas-Martín and García-Valcárcel, 2024). Despite the emergence of alternative approaches—such as using learning analytics for competence assessment (Guallichico et al., 2023) or developing new evaluation tools (Tarazona Gil, Lloret Català and Martínez-Usarralde, 2025)—DigComp 2.2 remains one of the most widely adopted and relevant frameworks.

1.1 Advantages of the DigComp 2.2 Framework

DigComp 2.2 offers a robust, adaptable framework for fostering digital competence across diverse sectors and populations. One of its key strengths lies in its comprehensive and flexible design, which supports the development of digital skills among citizens across a broad range of contexts—including both personal and

Cite this article: Luna, A. et al. 2026. "From Self-Perception to Reality in Digital Competence: Findings from the DigComp 2.2 Framework", *Electronic Journal of e-Learning*, 24(3), pp 137-153, https://doi.org/10.34190/ejel.24.3.4611

professional spheres (Al Khateeb, 2017). Its structure enables implementation in multiple domains, such as education, healthcare, and the architecture, engineering, and construction (AEC) industry (AlDhaen, 2025; Chen, Ebrahimi and Cheng, 2025).

In the educational domain, DigCompEdu—a specialised adaptation of the DigComp framework—has received positive evaluations for its emphasis on pedagogical application. Unlike other frameworks that primarily focus on technological proficiency, DigCompEdu integrates pedagogical dimensions, aiding in the identification of training needs and enabling the design of personalised professional development paths (Cabero-Almenara et al., 2021; Gabbi and Ancillotti, 2024; Gutiérrez-Castillo et al., 2023).

A core principle of DigComp is its promotion of lifelong learning and continuous self-development. It underscores the importance of not only updating one's own digital skills but also supporting others in the acquisition of these competences (Biggins, Holley and Zezulkova, 2017). The inclusion of updated dimensions such as data literacy, collaborative work, and enhanced cybersecurity in versions 2.1 and 2.2 reflects an awareness of increasingly complex digital environments and growing concerns related to privacy, safety, and digital well-being (Calero-Plaza, González-García and Fernández-Piqueras, 2025).

Furthermore, DigComp has provided a validated foundation for the development of reliable assessment instruments. Tools such as the "ECODIES" questionnaire for students and the "DigCompEdu Check-In" for educators demonstrate strong psychometric properties and practical utility (Betín de la Hoz et al., 2023; Casillas Martín, Cabezas González and García Peñalvo, 2020; Ergül and Taşar, 2023). These instruments play a critical role in diagnosing levels of digital competence and informing data-driven educational policy and planning (Casillas Martín, Cabezas González and García Peñalvo, 2020).

1.2 Disadvantages and Limitations of the DigComp 2.2 Framework

Despite its many benefits, DigComp 2.2 and its associated instruments have several limitations that warrant critical attention. One of the most prominent issues is the discrepancy between individuals' self-perception of digital competence and their actual performance (González-Mujico, 2024). While tools such as the DigCompEdu Check-In have demonstrated validity and reliability in measuring self-perceived competence (Cabero-Almenara et al., 2021), self-assessments often diverge from observable behaviours and real-world application (González-Mujico, 2024).

Overestimation of digital competence is a recurring concern. Many individuals tend to rate their abilities higher than they are in practice (Martzoukou et al., 2020; Marzan, Stamate and Spiru 2024, Momdjian, Manegre and Gutiérrez-Colón, 2025; Vuorikari, Pokropek and Muñoz, 2025). Empirical studies indicate that approximately 30% of participants overrate their skills, while up to 68% may perform ineffective or superficial actions intended to demonstrate competence (Muszyński et al., 2023). Conversely, underestimation—though less frequent—also occurs, particularly among older adults or highly experienced educators who may underestimate their digital fluency due to lack of confidence or changing standards (Liesa-Orus, Blasco and Arce-Romeral, 2023; Marzan, Stamate and Spiru, 2024; Schwarz et al., 2024).

These distortions contribute to what has been termed an "explanatory gap"—a mismatch between reported competence and actual capability—that cannot be adequately captured through self-report methods alone (Bartolomé et al., 2025; González-Mujico, 2024). This gap is further compounded by the Dunning–Kruger effect, a well-documented cognitive bias in which individuals with lower competence tend to significantly overestimate their abilities (Dunning, 2011). In such cases, low-performing individuals may believe they are above average, thereby reducing their motivation to seek training or support (Golz, Hahn and Zwakhalen, 2024; Schlösser et al., 2013).

Moreover, self-reporting is inherently prone to response biases, particularly social desirability bias, which can compromise data accuracy (Amin et al., 2025). Participants may respond in ways they believe are expected or socially acceptable rather than providing honest assessments of their actual skills, leading to inflated scores and potentially flawed conclusions (Montenegro-Rueda and Fernández-Batanero, 2024; Vuorikari, Pokropek and Muñoz, 2025).

Memory inaccuracies and recall errors further undermine the reliability of self-assessed data (Benaoui and Kassimi, 2021). Correlations between self-reports and objective cognitive performance measures have been found to vary widely—from as low as $r = 0.2$ to as high as $r = 0.8$—suggesting that while some self-assessments may be indicative of actual competence, others offer only a vague approximation (Muszyński et al., 2023).

1.3 Additional Research Gaps

A review of the current literature reveals other critical gaps that must be addressed to achieve a more complete understanding and more effective implementation of digital competence:

Validation beyond self-assessment: Much of the existing research relies heavily on self-perception as a measure of digital competence. However, this approach presents well-documented limitations, particularly in aligning perceived versus actual skill levels (González-Mujico, 2024). There is a clear need for more empirical studies that validate DigComp through direct assessment methods, such as structured classroom observation or task-based performance analysis. Comparing self-perceived competence with objectively measured performance can offer deeper insights into digital proficiency and help refine theoretical constructs in the field (González-Mujico, 2024).

Specific studies by population groups: While DigComp is designed as a flexible, overarching framework, more targeted studies are needed to address the unique contexts of specific user groups. For instance, older adults (Calero-Plaza, González-García and Fernández-Piqueras, 2025), nurses (Buhagiar, 2023; Golz, Hahn and Zwakhalen, 2024), trainee teachers (Lázaro-Cantabrana, Usart-Rodríguez and Gisbert-Cervera, 2019; Grande-de-Prado et al., 2020; Jogezai, Koroleva and Baloch, 2023), and healthcare professionals (AlDhaen, 2025; Karvouniari et al., 2024) face distinctive digital demands and challenges. Focused research could facilitate the development of adapted tools and strategies that better serve these communities.

Longitudinal and organisational studies: Another gap concerns the lack of longitudinal studies and organisational case studies that track digital transformation over time. Quantitative research that identifies both enablers and barriers to digital competence acquisition remains scarce (Chen, Ebrahimi and Cheng, 2025). Furthermore, investigations into how institutional culture and leadership affect digital adoption could offer valuable insights, particularly when studied within high-performing or digitally mature organisations. The same is true of linguistic and cultural contexts (Díaz-Burgos et al., 2024).

Based on the identified gaps, this study examines the discrepancy between self-perceived and actual digital competence by validating the Response Consistency Index (RCI) as a proxy for metacognitive self-awareness. The RCI is integrated into a composite Digital Readiness Index (DRI), which is applied to a cohort of military students to assess the reliability of self-assessment and its implications for targeted digital skills development.

1.4 Research Questions

Based on the identified gaps in the literature, this study addresses the following research questions:

RQ1: To what extent do self-assessed digital competence scores align with objective performance outcomes among military university students, and to what extent can the Response Consistency Index (RCI) serve as a reliable proxy for metacognitive self-awareness?

RQ2: Which specific DigComp 2.2 competence domains exhibit the greatest discrepancy between self-perception and actual competence, and are these discrepancies consistent across different performance levels?

RQ3: Can the Digital Readiness Index (DRI), a composite construct integrating perceived competence, response consistency, and objective performance, provide a more accurate and psychometrically robust assessment of digital readiness than traditional self-report measures alone?

2. Materials and Methods

This study employed a quantitative descriptive-analytical design. Ethical approval for this study was granted by the Joint Command of the Armed Forces in Ecuador as part of an approved institutional research project.

The study population comprised 629 military university students (mean age = 20.1 years). Participation was voluntary, and all respondents were enrolled in military academic programmes at the undergraduate level.

To assess digital competence, the DigComp 2.2 digital skills test was administered (Vuorikari, Kluzer and Punie, 2022). The instrument was developed based on the 21 competence areas outlined in the European Digital Competence Framework for Citizens, grouped into five core domains: information and data literacy, communication and collaboration, digital content creation, safety and problem solving. The test items were structured using a 4-point Likert scale, with response options ranging from: “Never” (1), “Sometimes” (2), “Almost always” (3), “Always” (4).

Each survey item was derived directly from the competence descriptors and proficiency-level examples published in the official DigComp 2.2 framework (Vuorikari, Kluzer and Punie, 2022). The instrument was adapted to the military higher education context through a pilot study with 45 students from the same institution, after which minor linguistic adjustments were made to improve clarity without altering the underlying competence constructs. Face and content validity were reviewed by a panel of three experts in digital education and one expert in military training, confirming adequate alignment between the items and the original DigComp descriptors.

These responses were used for statistical analysis and enabled the calculation of an RCI for each participant. The RCI evaluates the internal consistency of responses within each DigComp domain, calculated based on the intra-domain standard deviation. An RCI value approaching 1.0 indicates high consistency, whereas lower values (closer to 0) suggest potential contradictions, random answering patterns, or inflated self-assessments. Cases with RCI < 0.5 or RCI = 1.0 (indicative of extreme response behaviour) were flagged for further analysis. Based on these data, a DRI was developed which integrates both the self-reported competence level and the RCI, thereby weighting participants' reported skill levels by the internal consistency of their responses.

Following the diagnostic phase, a performance-based test was administered to contrast self-perception with reality. This instrument consisted of 3 tasks per item aligned with the examples of knowledge, skills, and attitudes specified in DigComp 2.2. This test served as a benchmark for actual digital competence, enabling comparison between self-perceived ability and demonstrated performance. Scoring was binary (1 = Correct, 0 = Incorrect) to minimize subjective interpretation. This objective score (C_obj) was used to calculate the Discrepancy Index (DI) as explained in section 2.1.3 Phase 3.

The performance-based test comprised 63 tasks (three per each of the 21 DigComp 2.2 competence areas) and was designed following the examples of knowledge, skills, and attitudes provided in the official DigComp 2.2 documentation (Vuorikari, Kluzer and Punie, 2022). Tasks included, for example, identifying a phishing email (Safety), selecting the most efficient search operator (Information Literacy), and troubleshooting a file-format compatibility issue (Problem Solving). Content validity was established through expert review by three digital education specialists, who rated each task for alignment with its target competence descriptor using a relevance scale. Inter-rater scoring reliability was assessed on a random subsample of 60 responses, yielding a Cohen's kappa of $\kappa = 0.91$, indicating near-perfect agreement. The binary scoring protocol was adopted to eliminate subjective interpretation and to maximise comparability with the Likert-based self-assessment instrument.

2.1 Procedure

Following data collection and preprocessing, internal consistency of the DigComp 2.2 instrument was assessed using Cronbach's Alpha (α). The results indicated that 17 out of 21 competence domains achieved an alpha value of $\alpha \geq 0.70$, suggesting an acceptable level of reliability for the majority of the instrument. However, certain domains exhibited lower internal consistency and warranted further scrutiny. Specifically, Domain 1.1 (Navigation and Search) yielded an alpha of $\alpha = 0.53$, indicating poor reliability. Additionally, Domain 1.3 (Data Management) and Domain 4.3 (Health Protection) recorded values of $\alpha = 0.69$ and $\alpha = 0.63$ respectively, falling into the range of questionable consistency. While previous studies have reported significantly higher internal reliability using DigComp-based instruments (Zhao et al., 2021, $\alpha = 0.978$), contextual and sample-specific factors may explain these discrepancies. In response to these initial findings, the data analysis was structured into four sequential phases to ensure methodological rigour and address potential reliability concerns.

The lower reliability observed in Domains 1.1, 1.3, and 4.3 may reflect the heterogeneous nature of the skills assessed within these areas and the limited number of items per domain, which is a known constraint in short-scale instruments (Taber, 2018). To mitigate potential bias, sensitivity analyses were conducted excluding these domains; results remained substantively unchanged, confirming that the overall pattern of findings is robust. Nevertheless, these reliability limitations are acknowledged as a constraint of the current instrument and are considered when interpreting domain-level results for Navigation and Search, Data Management, and Health Protection.

To facilitate the reader's comprehension of the overall analytical workflow, Figure 1 provides a schematic overview of the four-phase procedure. Each phase builds upon the outputs of the preceding one: Phase 1 computes the RCI from the self-assessment instrument; Phase 2 examines the relationship between RCI and domain scores through correlation and regression analyses; Phase 3 introduces the performance-based test to calculate the Discrepancy Index (DI) and detect the confidence–competence paradox; and Phase 4 integrates all prior metrics into the DRI through CFA/SEM modelling. This sequential design ensures that each analytical layer is grounded in validated prior results before contributing to the composite readiness model.

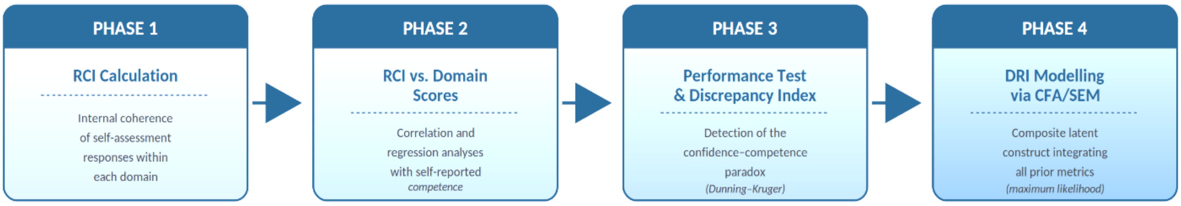


Figure 1: Schematic overview of the four-phase analytical procedure

Before detailing each phase, it is important to clarify the conceptual relationships among the three derived indicators. The RCI captures the internal coherence of self-assessment responses within each domain, serving as a reliability qualifier of self-reported data. The DI quantifies the direction and magnitude of the gap between self-perception and objective performance, identifying over- or underconfidence. The DRI is the integrative construct that synthesises both: it weights perceived competence by its consistency (RCI) and adjusts for systematic perception biases (DI), producing a single composite score that reflects not merely what students believe they can do, but how reliably and accurately they hold those beliefs. In this architecture, the RCI and DI are diagnostic inputs, while the DRI is the latent outcome.

2.1.1 Phase 1: Response Consistency Index (RCI)

In the first phase of analysis, an RCI was calculated for each participant to assess the internal coherence of responses within individual DigComp 2.2 domains. The RCI serves as a proxy indicator of metacognitive self-awareness, reflecting the degree to which respondents provide consistent answers to semantically similar or conceptually related items. To operationalise the RCI, inverse and redundant item pairs were identified within each domain. For example, one item might assess "confidence in using digital tools," while its inverse measures hesitancy or avoidance of similar tools. These pairings allowed for the detection of internal response contradictions that may suggest random guessing, misinterpretation, or inflated self-assessment.

The interpretation of response consistency as a proxy for metacognitive self-awareness is grounded in the theoretical framework of metacognitive monitoring (Flavell, 1979; Kruger and Dunning, 1999). Metacognitive calibration refers to the accuracy with which individuals judge their own cognitive performance. When a respondent provides internally consistent answers across semantically related items, this signals a coherent self-model of competence, which constitutes a behavioral indicator of effective metacognitive monitoring. Conversely, inconsistent responses to conceptually equivalent items suggest either a fragmented self-representation or a lack of reflective engagement with the assessment. The RCI thus operationalises the degree of internal coherence in self-evaluation, which prior research has identified as a behavioural correlate of metacognitive accuracy (Schlösser et al., 2013).

The RCI was computed as the mean absolute deviation between participants' actual responses and the expected pattern of internally consistent answers. Scores were then normalised on a continuous scale from 0 to 1, where:

RCI ≈ 1.0 indicates high response consistency, suggesting thoughtful and reliable self-assessment,

RCI < 0.5 reflects low consistency, potentially due to confusion, disengagement, or response bias,

RCI > 0.9 was used to identify participants with highly stable responses, who may serve as internal benchmarks or models of competence.

This index was not only used as a filtering criterion in data quality assurance but also integrated into the DRI in subsequent phases to weight self-assessment scores by response reliability.

2.1.2 Phase 2: Comparison of RCI indices with domain scores

In the second phase, RCI values were compared with domain scores derived from the DigComp 2.2 self-assessment instrument. Domain scores were calculated by computing the arithmetic mean of participant responses for each of the five competence areas. This enabled an aggregated measure of self-perceived proficiency within each domain. Although geometric aggregation is a commonly adopted option in the construction of composite indices—precisely because it strongly penalises imbalances among dimensions by following a non-compensatory logic (i.e., weaknesses in one domain cannot be fully counteracted by strengths in others)—this study intentionally chooses the arithmetic mean (AM) as the aggregation method. This choice is grounded in two key arguments that reinforce its theoretical and methodological appropriateness.

First, the DigComp 2.2 framework treats digital competences as clearly differentiated domains that are, nevertheless, positively correlated. That is, progress in one domain tends to facilitate or reinforce progress in others, thereby supporting a compensatory logic: strengths in certain domains can, to some extent, compensate for weaknesses in others without disregarding the importance of balanced development across all domains.

Second, the use of the AM maintains strong coherence with the structural equation modelling (SEM) approach employed in this study. In our proposed SEM framework, domain scores are treated as continuous indicators of a higher-order latent construct representing overall digital competence. Furthermore, this practice aligns with established psychometric standards for the aggregation of items or domains assessed via Likert scales in the field of competences (Vuorikari, Kluzer and Punie, 2022), thereby avoiding potential distortions that non-linear methods, such as geometric aggregation, might introduce in contexts where indicators reflect a common latent dimension.

To examine the relationship between response consistency and self-assessed competence, Pearson correlation analyses were conducted. Additionally, to control for the potential influence of demographic variables—including age, gender, and educational level—multiple linear regression models were applied. This allowed for a more nuanced interpretation of the association between RCI and self-reported domain scores, accounting for individual differences in participant profiles.

This phase aimed to determine whether higher response consistency (as measured by RCI) correlated with higher self-assessed competence, and whether this relationship held when controlling for relevant sociodemographic factors.

2.1.3 *Phase 3: Detection of the confidence versus competence paradox*

The third phase focused on identifying the confidence–competence paradox, a phenomenon closely associated with the Dunning–Kruger effect—a cognitive bias whereby individuals with lower ability levels tend to overestimate their competence, while more proficient individuals may underestimate their own skills (Kruger and Dunning, 1999).

This paradox was detected by measuring the discrepancy between self-perceived competence (subjective scores on confidence items) and objective competence (scores on knowledge and practical skills items). A Discrepancy Index (DI) was calculated for each participant to quantify the gap between self-perceived and actual competence levels:

$$DI_i = \sum (C_{sub} - C_{obj}) / \max(C_{sub}) - \min(C_{obj})$$

Where C_sub is subjective confidence and C_obj is objective competence. Positive DI values indicate overconfidence (paradox present), while negative values suggest underconfidence.

The DI served as a diagnostic metric to categorise individuals as: overestimators (high self-assessment, low performance), underestimators (low self-assessment, high performance) or accurate estimators (alignment between self-perception and actual competence). The DI enabled a detailed analysis of metacognitive accuracy across the sample and contributed to the overall construction of the DRI in the subsequent phase.

2.1.4 *Phase 4: Modelling the Digital Readiness Index (DRI)*

In the final phase, the DRI was conceptualised and modelled as a second-order composite construct that integrates the findings from the previous phases. To develop the DRI, a Confirmatory Factor Analysis (CFA) was conducted within a Structural Equation Modelling (SEM) framework using IBM SPSS AMOS software. CFA was employed to verify the underlying factor structure, specifically how the observed variables (i.e., test items) load onto latent constructs representing the five core DigComp 2.2 domains.

The DRI was operationalised as a second-order latent factor, derived from the five first-order latent variables corresponding to the five main areas of DigComp 2.2: Information and Data Literacy, Communication and Collaboration, Digital Content Creation, Safety, Problem Solving.

To account for response inconsistencies and cognitive biases observed in earlier phases, the model incorporated two weighting mechanisms: the RCI, which quantified the internal reliability of each participant's self-assessment using intra-domain standard deviation and, where applicable, serial correlation coefficients; and the Discrepancy Index (DI), which measured the divergence between self-reported competence and performance-based outcomes, capturing overestimation or underestimation effects consistent with the Dunning–Kruger bias. These indices were integrated into the model through adjusted factor loadings, ensuring that each domain's contribution to the DRI reflected not only perceived skill levels but also the internal consistency and metacognitive accuracy of each respondent's self-assessment. This approach allowed the DRI to more accurately represent digital readiness while correcting for known distortions in self-report data.

Specifically, the integration of RCI and DI into the SEM model followed a two-step approach. First, each participant's raw domain score was multiplied by their domain-specific RCI value, producing a consistency-adjusted competence score. This multiplicative weighting ensures that high self-reported scores accompanied by low response consistency are penalised proportionally. Second, the DI was incorporated as an observed covariate in the structural model, allowing the CFA to partial out the variance attributable to systematic overestimation or underestimation. The DRI was then estimated as the second-order latent factor from the five consistency-adjusted domain scores, with the DI informing the error structure. All model parameters—including factor loadings, residual variances, and second-order path coefficients—were estimated simultaneously via maximum likelihood in AMOS.

The model's fit was evaluated using widely accepted psychometric and SEM fit indices (Hu and Bentler, 1999):

Comparative Fit Index (CFI): Values above 0.95 were required to indicate excellent model fit relative to a null model.

Root Mean Square Error of Approximation (RMSEA): A value below 0.06, with 90% confidence intervals, denoted acceptable fit and accounted for model complexity.

Standardized Root Mean Square Residual (SRMR): Values under 0.08 were interpreted as evidence of limited misspecification and strong absolute model fit.

All fit statistics met or exceeded the recommended thresholds, with the final model yielding (CFI > 0.95, RMSEA < 0.05, SRMR < 0.06), indicating a robust fit between the hypothesized model and the empirical data. Furthermore, standardized factor loadings were all greater than 0.70, supporting convergent validity, while composite reliability (CR) values exceeded 0.80 across all domains. Discriminant validity was confirmed using the Fornell–Larcker criterion, verifying the distinctiveness of each latent construct within the model. These procedures contribute to the broader body of research on digital competency frameworks by integrating metacognitive self-awareness and psychometric robustness into a unified readiness model.

3. Results

Once the data has been processed using the DigComp 2.2 framework metrics and the RCI has been calculated, the results obtained are presented below. These are organized by competence domains, performance levels, and consistency patterns, allowing for a comprehensive analysis of the participants' digital skills.

3.1 Response Consistency Index (RCI) Results

Figure 2 presents the distribution of the Response Consistency Index (RCI) across the 629 participants. The following observations can be drawn from the histogram:

1. The horizontal axis (X-axis) shows the RCI, where 0 = maximum inconsistency and 1 = perfect consistency.
2. Vertical axis (Y axis) shows how many individuals fell into each RCI range. Higher bars mean more people in that consistency range. The peak of the histogram is around RCI = 0.70 to 0.85, suggesting that most respondents were fairly consistent, though not perfect.
3. This means that, while many people were fairly consistent, a few achieved a very high RCI (close to 1), but not the majority. An RCI of 1.0 indicates that the respondent gave exactly the same answer (e.g., all "Always" or all "Almost always") to every question within each domain (Information Literacy, Security, etc.).

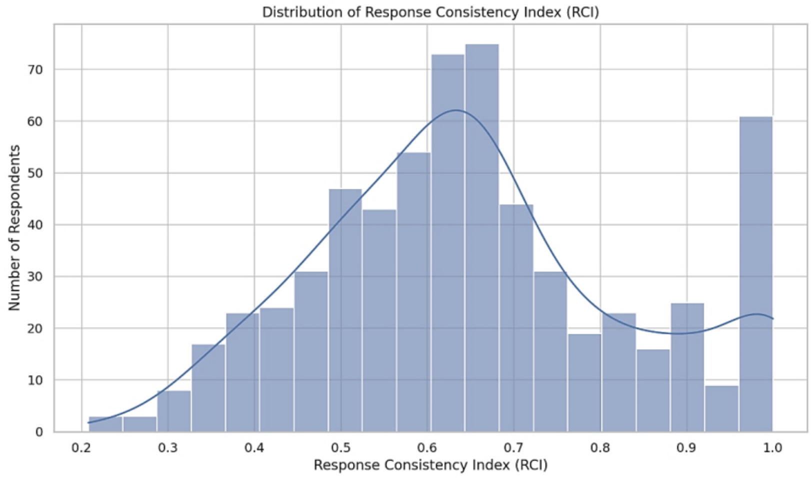


Figure 2: Distribution of Response Consistency Index (RCI)

As Fig. 2 shows, the RCI values are distributed fairly symmetrically, but slightly skewed to the right (with a tail toward 1.0). Potential scenarios and implications of extreme RCI values are discussed in the Discussion section**.**

3.2 RCI Scores Versus Domain Scores

The scatter plots shown in Figure 3 visualize the relationship between the RCI and the average score for each of the five digital competence domains. Each point represents one respondent. Analysis of these scatter plots reveals the following trends:

Information and data literacy: Fig 3 (A) shows a slight upward trend in RCI scores in relation to information and data literacy. Greater consistency tends to align with higher literacy scores: those who are more reflective about their information-seeking habits (e.g., verifying sources) tend to be more consistent. Some respondents with low RCI also scored high, suggesting possible overestimation or selective understanding.

Communication and collaboration: Fig 3 (B) shows a stronger positive correlation between the RCI Score and Communication and Collaboration. This means that digital collaboration skills (email etiquette, sharing, participation in online groups) may be more intuitive or practiced consistently, and therefore more stable across all self-assessments. This allows us to build on this stable foundation when designing collaborative digital training.

Digital content creation: Fig 3 (C) shows a slight positive correlation, with greater dispersion of the RCI Score vs. Digital Content Creation. While some very consistent respondents scored well, others did not: content creation may be more subjective or misunderstood (e.g., adapting images). This indicates that practical tasks should be considered to validate these skills rather than relying solely on self-reporting.

Safety: Fig 3 (D) shows a slight correlation with a wider dispersion of RCI vs. Safety Score. This indicates that cybersecurity practices vary widely in perception, even consistent users may misjudge the depth of their knowledge. For this, scenario-based tests can be used for real-life cybersecurity actions (e.g., phishing identification).

Problem solving: Fig 3 (E) shows the strongest positive trend of RCI vs. Problem Solving. Consistency in responses is largely associated with better problem-solving skills, supporting the idea that reflective and strategic thinkers tend to be methodical in all domains. This implies that these students could benefit from advanced digital autonomy tools (such as automation and data modelling). So far, we have calculated the average scores per competence domain for each respondent, the overall average score (across the five domains), and added each student's RCI to contextualize.

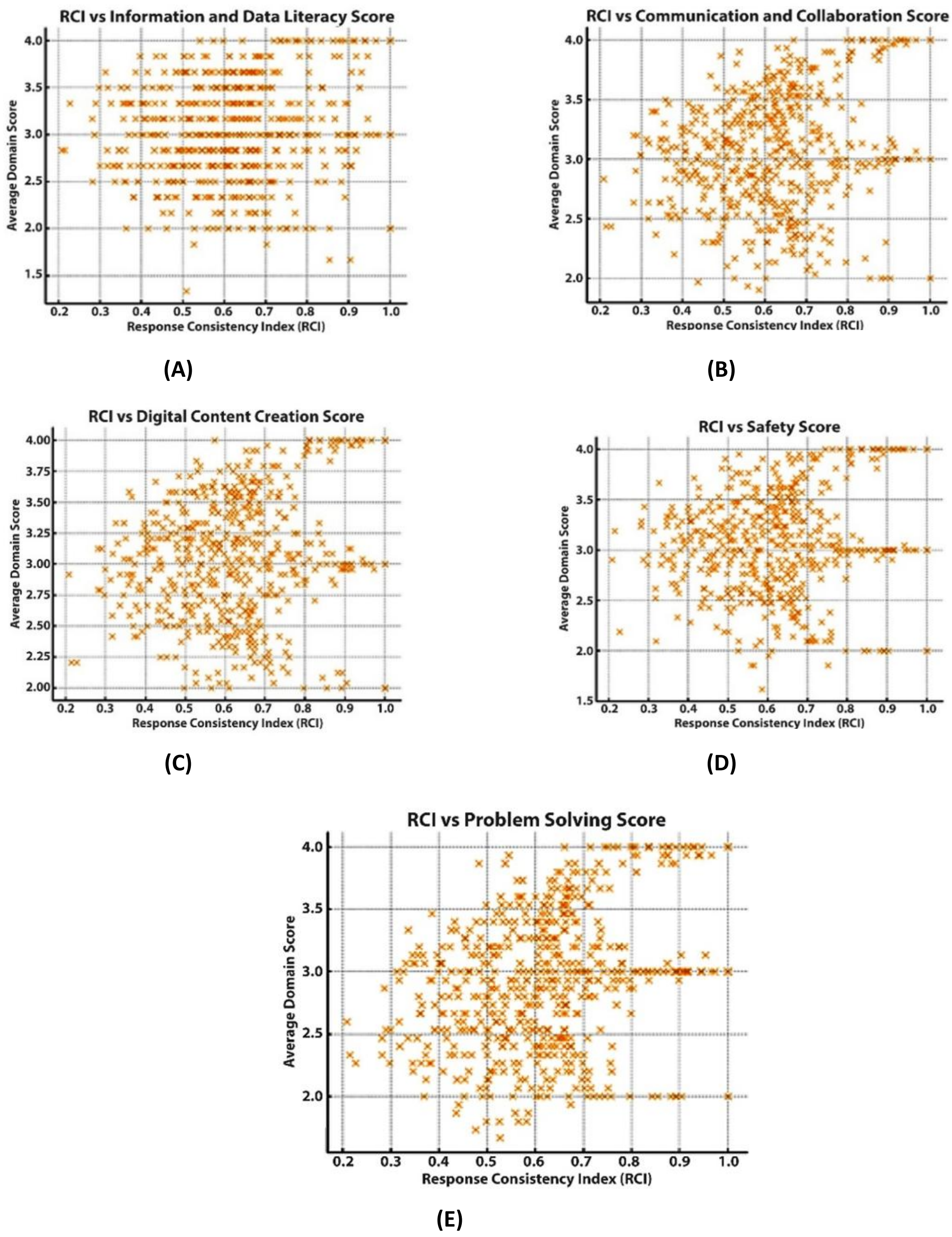


Figure 3: Scatter plots: RCI vs. DigComp 2.2. domain scores

3.3 Detailed Information by Group

Based on the results obtained, students were grouped into three performance levels: low, moderate, and high. For each group, the digital competence domain with the lowest score was identified in order to establish specific areas for improvement. The following table summarises these weaknesses, incorporating the average score obtained, as well as a pedagogical interpretation that also considers the RCI.

Below is a summary of the specific weaknesses identified by performance group, highlighting the weakest domain, its average score, and an interpretation aimed at guiding pedagogical decision-making (see Table 1).

Table 1: Specific weaknesses by performance group and consistency analysis

Weaknesses Breakdown Performance group	Weaker domain	Average score	Interpretation
Poor performance	Problem Solving	2.16	Lack of ability to identify or solve digital tasks using appropriate tools. They are probably unsure about how to customise digital tools (e.g., project management or collaboration platforms) or solve technical problems. In addition, they show moderate consistency (RCI = 0.67), which reinforces the need for intensive pedagogical support.
Moderate performance	Problem Solving	2.88	Although slightly better, it remains a significant weakness. It indicates the need for digital applications in scenarios. Its RCI (0.61) reveals a slight inconsistency, possibly related to unstable self-perception or lack of sustained commitment.
High performance	Information and data literacy	3.71	Even the best performers have relatively lower scores here, possibly struggling with source verification or advanced search techniques. Although their relative weakness lies in information literacy, they excel in digital security and collaboration (areas with higher scores).

The results allow us to draw up specific recommendations for each performance level, based on the weaknesses detected: it is evident that all groups need support in the development of problem-solving skills, although this support must be adjusted to the specific level of competence and consistency that each group presents.

Even high-achieving students need advanced training in information and data literacy, particularly in source verification, critical analysis of digital content, and identification of misinformation.

3.4 Detection of the Confidence-versus-Competence Paradox

This analysis identifies students whose self-assessment consistency (confidence) does not align with their actual digital competence (performance scores). To do this end, real-world digital tasks were used to contrast with self-perception. Two atypical profiles were identified, as described below:

3.4.1 Overly confident participants

Those who obtained an RCI ≥ 0.90, i.e., high consistency (strong self-confidence). However, their performance score is not high (low or moderate actual scores). For instance, some students identified got RCI = 1.0 because they always responded identically ("Always"). But their overall score was 2.0, which places them as low-performing students. This possible overestimation of their abilities may be due to misunderstanding of the scale, intention to appear competent, or ignorance of actual standards.

3.4.2 Uncertain participants

Those who obtained an RCI < 0.60, indicating an inconsistent self-assessment. Despite their low consistency, they showed high actual performance (true digital competence), with overall scores between 3.65 and 3.72, while their RCI was reported between 0.52 and 0.57, indicating that responses varied widely within the domains. These students may doubt their abilities or reflect more critically. Inconsistent responses, despite high performance, may reflect epistemic humility or self-aware recognition of areas for growth.

3.5 Modelling the Digital Readiness Index (DRI)

A DRI was constructed, a comprehensive and normalized score (0 to 1) that reflects the following from respondents:

- Competence in five digital domains, and

- Response consistency (RCI) as a measure of confidence or reliability.

Table 2 shows the weighting logic with the assigned factor loadings, composite reliability (CR), and average variance extracted (AVE).

Table 2: Weighting of domains for calculating the Digital Readiness Index (DRI)

Latent variable (Domain)	Indicator	Std. Factor Loading (λ)	CR	AVE
Information and Data Literacy	Item 1.1	0.82	**0.89**	**0.68**
	Item 1.2	0.79		
	Item 1.3	0.85		
	Item 1.4	0.81		
Communication and Collaboration	Item 2.1	0.85	**0.91**	**0.72**
	Item 2.2	0.88		
	Item 2.3	0.83		
	Item 2.4	0.84		
Digital Content Creation	Item 3.1	0.76	**0.86**	**0.61**
	Item 3.2	0.79		
	Item 3.3	0.81		
	Item 3.4	0.75		
Safety	Item 4.1	0.81	**0.87**	**0.63**
	Item 4.2	0.78		
	Item 4.3	0.82		
	Item 4.4	0.77		
Problem Solving	Item 5.1	0.90	**0.92**	**0.79**
	Item 5.2	0.93		
	Item 5.3	0.91		
	Item 5.4	0.89		

As shown in Table 2, the highest standardised factor loadings were observed in the Problem Solving domain ($\lambda = 0.89–0.93$), reflecting its strong contribution to the overall Digital Readiness construct. The RCI was integrated as a weighting mechanism to ensure that the DRI captures not only declared competence but also internal consistency and metacognitive self-awareness. The second-order SEM model confirmed that all five domains contribute meaningfully to the global Digital Readiness Index.

Based on this index, the top 10% of students were examined according to the Digital Readiness Index (DRI ≥ 0.90), and their characteristics and average scores were extracted to determine the behavioural patterns of top performers. Table 3 summarises the profile of this top-performing group, reporting their average domain scores, overall score, RCI, and DRI.

Table 3: Profile of the top 10% according to the Digital Readiness Index (DRI ≥ 0.90)

Summary metric of the profile of the best results	Top 10% average	Interpretation
Information and Data Literacy	3.86	Almost always check the sources and keywords.
Communication and Collaboration	3.98	Skilled communicators and digital collaborators.
Digital Content Creation	3.99	Expert in adaptation, design, and editing of digital media.
Safety	3.98	Strong awareness of cybersecurity and privacy.

Problem Solving	3.97	High capacity to customize tools to solve digital problems.
Overall score	3.95	Steady progress in all areas.
RCI (Response Consistency Index)	0.93	Highly consistent, thoughtful, and confident.
Digital Readiness Index	0.97	Practically perfect: model digital citizens.

The characterisation of the top 10% based on the DRI reveals a group of students who are not only technically competent, but also consistent and reflective in their self-assessments. These model digital citizens are a key reference point for institutional initiatives, such as mentoring programmes, peer training, or leadership in digital environments. Identifying them allows us not only to recognise high performance, but also to leverage their potential as multipliers of good practices in digital skills.

4. Discussion

Although there is general agreement on the importance of digital skills, e.g. from the early years of education (Albán Bedoya and Ocaña-Garzón, 2022; Ocaña, Almeida and Albán, 2022; Quinga et al., 2022), for teaching and learning (Bazurto Palma and Ocaña-Garzón, 2023; Ocaña et al., 2023) or professional development, doubts remain about the accuracy of DigComp 2.2.

In light of this, and in response to the research questions posed in Section 1.1, the objective of this study was twofold: first, to validate a method for assessing self-perception of digital competences (RQ1) and, second, to diagnose actual competence gaps in military university students (RQ2). The integration of these findings into the DRI construct (RQ3) addresses the overarching aim of this research. Findings confirm that simple self-perception is an insufficient and often misleading indicator of digital competence.

Our main contribution is the validation of the Response Consistency Index (RCI) as a proxy for metacognitive self-awareness and its integration into a Digital Readiness Index (DRI). This discussion is articulated along three axes: (1) the interpretation of the RCI as a qualifier of self-perception and its link to the Dunning-Kruger effect; (2) the analysis of critical competence gaps (Problem Solving and Information and Data Literacy); and (3) the value of the DRI as an alternative construct.

Theoretically, the study draws on two complementary frameworks. First, it is anchored in the metacognitive monitoring literature (Flavell, 1979), which posits that accurate self-assessment depends on higher-order cognitive processes that evaluate one's own performance. Second, it engages directly with the Dunning–Kruger model (Kruger and Dunning, 1999), extending its application from general cognitive tasks to domain-specific digital competences. The principal contribution of this study is therefore twofold: it provides an empirical operationalisation of metacognitive calibration within the DigComp 2.2 framework (via the RCI) and demonstrates that integrating calibration metrics into composite readiness indices (via the DRI) yields a more valid assessment than self-report alone.

4.1 Beyond Self-Perception: The RCI and the (In)Expert Paradox

The discrepancy between self-perceived and actual digital competence is a persistent and well-documented problem in the literature. Multiple studies, ranging from accounting students (McCourt Larres, Ballantine and Whittington, 2003; Ballantine, McCourt Larres and Oyelere, 2007) to older adults (Barboutidis and Stiakakis, 2023), demonstrate a systematic tendency to overestimate one's own abilities. The problem is exacerbated because self-assessment tools, often based on frameworks such as DigComp, can be vague and heavily biased by participants' personal beliefs (Siiman et al., 2016).

Our results not only confirm this discrepancy but deepen it. We identified profiles consistent with the Dunning-Kruger effect (Kruger and Dunning, 1999): a low-performing group with high consistency (overconfidence) and a high-performing group with low consistency (underconfidence). This duality aligns with previous research that, while confirming general overestimation, also points out that students with higher abilities tend to be more accurate in their self-assessments (McCourt Larres, Ballantine and Whittington, 2003).

Here, we interpret the RCI values as an indicator of metacognitive realism. The finding that 9.7% of the sample achieved a perfect RCI (1.0) is particularly revealing. Far from representing mastery, this group contained individuals who answered "Always" to everything, masking actual low competence. This suggests that "perfect consistency" in self-assessment may, paradoxically, be a warning sign of a lack of critical reflection.

In contrast, the RCI's weaker correlation with domains such as Safety and Digital Content Creation is equally significant. We argue that this is not a failure of the index, but rather an accurate capture of reality. A student may believe their security practices are adequate (e.g. password creation) without being aware of more complex threats. As Barboutidis and Stiakakis (2023) point out, discrepancies between self-assessment and objective assessments often emerge in specific DigComp domains.

These findings directly address RQ1, confirming that the RCI effectively captures metacognitive calibration and reveals systematic patterns of over- and underconfidence consistent with the Dunning–Kruger effect.

4.2 Critical Gaps: Problem-Solving and Information and Data Literacy

Diagnosing weaknesses by domain offers the most direct pedagogical implications. The fact that Problem Solving is the weakest domain for the low- and moderate-performing groups is a troubling finding. This competence represents the ability to apply digital knowledge strategically, and its absence indicates that students lack the ability to transfer skills to solve real-world problems.

However, the most critical finding is the relative weakness in Information and Data Literacy (RCI=0.69) even in the high-performing group. Our top-performing students, although technically competent, still show difficulty assessing credibility and verifying sources.

This finding is not isolated; recent studies in different disciplines, such as nursing (Martzoukou et al., 2024) and teacher training (Alonso García et al., 2024), also report insufficient levels of competence in information and data literacy. This refutes the myth of the "digital native" and aligns with research showing that technological fluency does not equate to critical literacy (cf. Hargittai, 2010). Indeed, mere frequency of internet use does not guarantee proficiency; research based on performance tests suggests that extensive use may lead to incorrect usage patterns and is inversely related to digital competence (Barboutidis and Stiakakis, 2023).

In response to RQ2, these results identify Problem Solving and Information and Data Literacy as the competence domains with the most pronounced discrepancies between self-assessed and actual competence, a pattern that persists even among top-performing students.

4.3 The DRI as an Alternative Model of "Digital Readiness"

Existing digital competence indices often rely solely on summing self-perception scores. The DRI proposes a methodological advance by integrating competence (DigComp) and self-awareness (RCI) into a single construct validated using CFA/SEM.

The DRI conceptualizes "digital readiness" not only as the sum of skills, but as the intersection of competence and accurate self-awareness of that competence.

The profile of the top 10% students validated by the DRI corroborates this model. These "model digital citizens" are not only exceptional in their skills (average 3.95), but also exceptionally realistic in their self-assessment (RCI 0.93). They do not suffer from overconfidence, a finding that contrasts with the widespread overestimation reported in the literature (e.g., Ballantine, McCourt Larres and Oyelere, 2007) and reinforces the idea that genuine digital mastery is, in essence, a metacognitive exercise.

More broadly, identifying training needs based solely on self-assessments can be misleading. Self-perception data may not reflect a real training need (Liesa-Orus, Blasco and Arce-Romeral, 2023), which can lead to poor decision-making, a limited understanding of actual competences, risks in research and evaluation, or difficulties in application design (Montenegro-Rueda and Fernández-Batanero, 2024; Palacios-Rodríguez et al., 2023; Silva, Usart and Lázaro-Cantabrana, 2019).

Collectively, these findings address RQ3, demonstrating that the DRI—by integrating competence, consistency, and metacognitive accuracy—constitutes a more valid and informative measure of digital readiness than self-perception scores alone.

5. Conclusions

This study empirically demonstrates that, in the era of digital literacy, how a student perceives their skills is as important as what skills they believe they possess. Blind reliance on mere self-assessment is, at best, incomplete and, at worst, misleading, as it creates an "illusion of competence" that masks the most critical skill gaps.

The fundational contribution of this research is the provision of a validated method for decoupling confidence from competence. The RCI is not a simple quality control measure; it stands as a robust proxy for an individual's metacognitive calibration. Applying this lens, our findings expose an educational paradox: students' apparent digital fluency coexists with profound strategic deficiencies, notably in Problem Solving and, critically, in Information and Data Literacy, even among the highest-achieving cohorts.

The DRI is, therefore, not just another index, but an integrated construct that conceptually redefines "digital readiness." We propose that genuine readiness lies not in the sum of technical skills, but in the integration of competence and accurate self-awareness of that competence.

Despite the robustness of the model, we acknowledge several limitations. First, the sample was drawn from higher-education military institutions, which may limit generalisability. Second, although the RCI and DI assess self-perception, the study still relies on self-report, a method whose reliability has been questioned.

Regarding the generalisability of these findings, it should be noted that the proposed methodology—particularly the RCI and DRI constructs—is not inherently limited to military populations. While the current sample was drawn from a specific institutional context, the underlying analytical framework is transferable to other higher education settings, professional training programmes, and diverse demographic groups. Future research should validate these constructs across diverse cultural, linguistic, and disciplinary contexts to test their broader applicability. Cross-national studies using the DigComp 2.2 framework would be particularly valuable in this regard.

A future line of research is the triangulation of these findings with performance-based tasks, an approach already used successfully to identify usage patterns and compare self-reported performance with objective performance. Furthermore, the DRI and the RCI should not only as diagnostic tools, but also as tool for pedagogical intervention. Future studies should explore whether pedagogical interventions focused on metacognition can improve RCI and whether this improvement leads to an improvement in actual digital competence.

For educational policy and curriculum design, the message is unequivocal: it is no longer enough to teach digital tools; it is imperative to teach students to think critically about how they use those tools. The future of digital literacy lies not only in closing the skills gap, but in closing the much more dangerous gap of self-awareness.

Acknowledgements

We would like to thank the Joint Command of the Armed Forces of Ecuador and Club Semillero de Investigación ESPE.

AI Statement: Generative artificial intelligence tools were used strictly to improve the language quality, readability, and syntax of this manuscript.

Ethics Statement: This study was conducted as part of an institutional research project approved and authorized by the Joint Command of the Armed Forces of Ecuador, which granted the corresponding permissions for its execution. All participants took part voluntarily and provided their informed consent, and the data were processed in accordance with confidentiality agreements.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Funding: No funding was received for this research**.**

Data Availability Statement: The datasets generated and analyzed during the current study are not publicly available due to confidentiality agreements but are available from the corresponding author on reasonable request.

Authors' contributions: Andrea Luna conceptualized the study, conducted the investigation, curated the data, and prepared the original draft of the manuscript. Mauro Ocaña designed the methodology, performed the formal analysis, and contributed to writing—review and editing. Ana María Ortíz Colón supervised the project, administered the research activities, and contributed to the final manuscript revision. Javier Rodríguez Moreno was responsible for validation, resources, and visualization. All authors read and approved the submitted version and agree to be accountable for the content of the work.

References

Al Khateeb, A.A.M. (2017) "Measuring Digital Competence and ICT Literacy: An Exploratory Study of In-Service English Language Teachers in the Context of Saudi Arabia," International Education Studies, 10(12), pp. 38–51. Available at: https://doi.org/10.5539/ies.v10n12p38.

Albán Bedoya, I. and Ocaña-Garzón, M. (2022) “Educational Programming as a Strategy for the Development of Logical-Mathematical Thinking,” in M. Botto-Tobar et al. (eds.) Emerging Research in Intelligent Systems. Cham: Springer International Publishing, pp. 309–323. Available at: https://doi.org/10.1007/978-3-030-96043-8_24.

AlDhaen, F.S. (2025) “AI-Powered Transformation of Healthcare: Enhancing Patient Safety Through AI Interventions with the Mediating Role of Operational Efficiency and Moderating Role of Digital Competence—Insights from the Gulf Cooperation Council Region,” Healthcare, 13(6). Available at: https://doi.org/10.3390/healthcare13060614.

Alonso García, S. et al. (2024) "Analysis of self-perceived digital competences in future educators: A study at the University of Granada," JOTSE: Journal of Technology and Science Education, 14(1), pp. 4–15. Available at: https://doi.org/10.3926/jotse.2521.

Amin, S.M. et al. (2025) “Nursing education in the digital era: the role of digital competence in enhancing academic motivation and lifelong learning among nursing students,” BMC Nursing, 24(1), p. 571. Available at: https://doi.org/10.1186/s12912-025-03199-2.

Ballantine, J.A., McCourt Larres, P. and Oyelere, P. (2007) “Computer usage and the validity of self-assessed computer competence among first-year business students,” Computers & Education, 49(4), pp. 976–990. Available at: https://doi.org/10.1016/j.compedu.2005.12.001.

Balyk, N. et al. (2019) “Design of approaches to the development of teacher’s digital competencies in the process of their lifelong learning” Available at: https://core.ac.uk/download/pdf/222804591.pdf.

Barboutidis, G. and Stiakakis, E. (2023) “Identifying the Factors to Enhance Digital Competence of Students at Vocational Training Institutes,” Technology, Knowledge and Learning, 28(2), pp. 613–650. Available at: https://doi.org/10.1007/s10758-023-09641-1.

Bartolomé, J. et al. (2025) “Using Eye-Tracking Data to Examine Response Processes in Digital Competence Assessment for Validation Purposes,” Applied Sciences, 15(3). Available at: https://doi.org/10.3390/app15031215.

Bazurto Palma, C.J. and Ocaña Garzón, M. (2023) “The use of a virtual environment to improve students’ Listening skills: a learning analytics approach,” Ciencia Latina Revista Científica Multidisciplinar, 7(2), pp. 2937–2953. Available at: https://doi.org/10.37811/cl_rcm.v7i2.5537.

Benaoui, A. and Kassimi, M. (2021) “Using machine learning to examine preservice teachers’ perceptions of their digital competence,” E3S Web of Conferences, 297. Available at: https://doi.org/10.1051/e3sconf/202129701067.

Betín de la Hoz, A.B. et al. (2023) “Statistical Validation of the ‘ECODIES’ Questionnaire to Measure the Digital Competence of Colombian High School Students in the Subject of Mathematics,” Mathematics, 11(1). Available at: https://doi.org/10.3390/math11010033.

Biggins, D., Holley, D. and Zezulkova, M. (2017) “Digital competence and capability frameworks in higher education: importance of life-long learning, self-development and well-being,” EAI Endorsed Transactions on e-Learning, 4(13). Available at: https://doi.org/10.4108/eai.20-6-2017.152742.

Buhagiar, M.R. (2023) “Self-perception of digital competence by nurse educators during the COVID-19 pandemic”. Master of Education (M.Ed.) thesis. University of Johannesburg. Available at: https://ujcontent.uj.ac.za/esploro/outputs/graduate/Self-perception-of-digital-competence-by-nurse/9931607907691

Cabero-Almenara, J. et al. (2021) “The Teaching Digital Competence of Health Sciences Teachers. A Study at Andalusian Universities (Spain),” International Journal of Environmental Research and Public Health, 18(5). Available at: https://doi.org/10.3390/ijerph18052552.

Cabezas-González, M., Casillas-Martín, S. and García-Valcárcel, A. (2024) “Influence of the Use of Video Games, Social Media, and Smartphones on the Development of Digital Competence with Regard to Safety,” Computers in the Schools, pp. 1–20. Available at: https://doi.org/10.1080/07380569.2024.2409688.

Calero-Plaza, J., González-García, R.J. and Fernández-Piqueras, R. (2025) "Validation of a Scale to Measure Digital Competence in the Elderly Population," International Journal of Instruction, 18(1), pp. 77–94. Available at: https://www.e-iji.net/dosyalar/iji_2025_1_5.pdf.

Casillas Martín, S., Cabezas González, M. and García Peñalvo, F.J. (2020) “Digital competence of early childhood education teachers: attitude, knowledge and use of ICT,” European Journal of Teacher Education, 43(2), pp. 210–223. Available at: https://doi.org/10.1080/02619768.2019.1681393.

Chen, S., Ebrahimi, O. and Cheng, C. (2025) “New Perspective on Digital Well-Being by Distinguishing Digital Competency from Dependency: Network Approach,” Journal of Medical Internet Research. Available at: https://www.jmir.org/2025/1/e70483/.

Díaz-Burgos, A. et al. (2024) “Psychological and Educational Factors of Digital Competence Optimization Interventions Pre- and Post-COVID-19 Lockdown: A Systematic Review,” Sustainability, 16(1). Available at: https://doi.org/10.3390/su16010051.

Dunning, D. (2011) “Chapter five - The Dunning–Kruger Effect: On Being Ignorant of One’s Own Ignorance,” in J.M. Olson and M.P. Zanna (eds.) Advances in Experimental Social Psychology. Academic Press, pp. 247–296. Available at: https://doi.org/10.1016/B978-0-12-385522-0.00005-6.

Ergül, D. and Taşar, M. (2023) “Development and validation of the teachers’ digital competence scale (TDiCoS),” Journal of Learning and Teaching in Digital Age. Available at: https://dergipark.org.tr/en/pub/joltida/article/1204358.

Flavell, J.H. (1979) "Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry," American Psychologist, 34(10), pp. 906–911. Available at: https://doi.org/10.1037/0003-066X.34.10.906

Gabbi, E. and Ancillotti, I. (2024) “Towards an extended framework for digital competence of educators. The validation process through experts’ review,” Italian Journal of Educational Technology, 32. Available at: https://doi.org/10.17471/2499-4324/1338.

Golz, C., Hahn, S. and Zwakhalen, S. (2024) “Psychometric validation of the digital competence questionnaire for nurses,” SAGE Open Nursing. Available at: https://doi.org/10.1177/23779608241272641.

González-Mujico, F. (2024) “Measuring student and educator digital competence beyond self-assessment: Developing and validating two rubric-based frameworks,” Education and Information Technologies. Available at: https://doi.org/10.1007/s10639-023-12363-7.

Grande-de-Prado, M. et al. (2020) “Digital Competence and Gender: Teachers in Training. A Case Study,” Future Internet, 12(11). Available at: https://doi.org/10.3390/fi12110204.

Guallichico, G. et al. (2023) "Assessing Digital Competence Through Teacher Training in Early Education Teachers," in M. Botto-Tobar et al. (eds.) Applied Technologies. Cham: Springer Nature Switzerland, pp. 55–68. Available at: https://doi.org/10.1007/978-3-031-24978-5_6.

Gutiérrez-Castillo, J.J. et al. (2023) “Development of Digital Teaching Competence: Pilot Experience and Validation through Expert Judgment,” Education Sciences, 13(1), p. 52. Available at: https://doi.org/10.3390/educsci13010052.

Hargittai, E. (2010) “Digital Na(t)ives? Variation in Internet Skills and Uses among Members of the ‘Net Generation,’” Sociological Inquiry, 80(1), pp. 92–113. Available at: https://doi.org/10.1111/j.1475-682X.2009.00317.x.

Hu, L.-t. and Bentler, P.M. (1999) “Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives”. Structural Equation Modeling, 6(1), pp.1–55. Available at: https://doi.org/10.1080/10705519909540118.

Jogezai, N., Koroleva, D. and Baloch, F. (2023) “Teachers’ digital competence in the post COVID-19 era: The effects of digital nativeness, and digital leadership capital,” Contemporary Educational Technology. Available at: https://doi.org/10.30935/cedtech/13620.

Karvouniari, A. et al. (2024) “Translation and Validation of Digital Competence Indicators in Greek for Health Professionals: A Cross-Sectional Study,” Healthcare, 12(14). Available at: https://doi.org/10.3390/healthcare12141370.

Kruger, J. and 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), pp. 1121–1134. Available at: https://doi.org/10.1037/0022-3514.77.6.1121.

Lázaro-Cantabrana, J.L., Usart-Rodríguez, M. and Gisbert-Cervera, M. (2019) “Assessing Teacher Digital Competence: the Construction of an Instrument for Measuring the Knowledge of Pre-Service Teachers,” Journal of New Approaches in Educational Research, 8(1), pp. 73–78. Available at: https://doi.org/10.7821/naer.2019.1.370.

Liesa-Orus, M., Blasco, R.L. and Arce-Romeral, L. (2023) “Digital competence in university lecturers: A meta-analysis of teaching challenges,” Education Sciences, 13(1). Available at: https://doi.org/10.3390/educsci13050508.

Martzoukou, K. et al. (2020) “A study of higher education students’ self-perceived digital competences for learning and everyday life online participation,” Journal of Documentation. Available at: https://doi.org/10.1108/JD-03-2020-0041.

Martzoukou, K. et al. (2024) “A cross-sectional study of discipline-based self-perceived digital literacy competencies of nursing students,” Journal of Advanced Nursing, 80(2), pp. 656–672. Available at: https://doi.org/10.1111/jan.15801.

Marzan, M., Stamate, A. and Spiru, L. (2024) “Digital Competence Among Older Adults: Is Self-Assessment Reliable?,” Cureus, 16(1). Available at: https://doi.org/10.7759/cureus.73543.

Mccourt Larres, P., Ballantine, J. and Whittington, M. (2003) “Evaluating the validity of self-assessment: measuring computer literacy among entry-level undergraduates within accounting degree programmes at two UK universities,” Accounting Education, 12(2), pp. 97–112. Available at: https://doi.org/10.1080/0963928032000091729.

Momdjian, L., Manegre, M. and Gutiérrez-Colón, M. (2025) “Bridging the digital competence gap: A comparative study of preservice and in-service teachers in Lebanon using the DigCompEdu Framework,” Technology , Knowledge and Learning, 30. Available at: https://doi.org/10.1007/s10758-024-09794-7.

Montenegro-Rueda, M. and Fernández-Batanero, J.M. (2024) “Adaptation and validation of an instrument for assessing the digital competence of special education teachers,” European Journal of Special Needs Education, 39(3), pp. 367–382. Available at: https://doi.org/10.1080/08856257.2023.2216573.

Muszyński, M. et al. (2023) “Can Overclaiming Technique Improve Self-Assessment Tools for Digital Competence? The Case of DigCompSat,” Social Science Computer Review, 41(6), pp. 2318–2341. Available at: https://doi.org/10.1177/08944393221117269.

Ocaña, M. et al. (2023) "Aplicaciones móviles en el desarrollo del lenguaje: Un enfoque comparativo entre padres y educadores," Revista Alpha & Omega, 1(1), p. 10. Available at: https://doi.org/10.24133/ALPHAOMEGA.VOL01.01.2023.ART01.

Ocaña, M., Almeida, E. and Albán, S. (2022) "How Did Children Learn in an Online Course During Lockdown? A Piagetian Approximation," in M. Botto-Tobar et al. (eds.) Emerging Research in Intelligent Systems. Cham: Springer International Publishing, pp. 259–272. Available at: https://doi.org/10.1007/978-3-030-96046-9_20.

Palacios-Rodríguez, A. et al. (2023) “Teacher Digital Competence in the education levels of Compulsory Education according to DigCompEdu: The impact of demographic predictors on its development,” Interaction Design and Architecture (s) Journal-IxD&A, 57, pp. 115–132. Available at: https://doi.org/10.55612/s-5002-057-007.
Quinga, Y. et al. (2022) “Virtual Activities to Strengthen Basic Math Skills in Children,” in M. Botto-Tobar et al. (eds.) Emerging Research in Intelligent Systems. CIT 2021, Springer, Cham: Springer International Publishing (Lecture Notes in Networks and Systems), pp. 173–185. Available at: https://doi.org/10.1007/978-3-030-96046-9_13.
Siiman, L.A. et al. (2016) “An Instrument for Measuring Students’ Perceived Digital Competence According to the DIGCOMP Framework,” in P. Zaphiris and A. Ioannou (eds.) Learning and Collaboration Technologies. Cham: Springer International Publishing, pp. 233–244. Available at: https://doi.org/10.1007/978-3-319-39483-1_22.
Schlösser, T. et al. (2013) “How unaware are the unskilled? Empirical tests of the ‘signal extraction’ counterexplanation for the Dunning–Kruger effect in self-evaluation of performance,” Journal of Economic Psychology, 39, pp. 85–100. Available at: https://doi.org/10.1016/j.joep.2013.07.004.
Schwarz, S. et al. (2024) “Digital Competence Scale (DCS),” Nordic Journal of Digital Literacy, 19(3), pp. 126–143. Available at: https://doi.org/10.18261/njdl.19.3.2.
Silva, J., Usart, M. and Lázaro-Cantabrana, J. (2019) “Teacher’s Digital Competence among Final Year Pedagogy Students in Chile and Uruguay.,” Comunicar: Media Education Research Journal, 67(21). Available at: https://eric.ed.gov/?id=EJ1229187.
Taber, K.S. (2018) "The Use of Cronbach's Alpha When Developing and Reporting Research Instruments in Science Education," Research in Science Education, 48(6), pp. 1273–1296. Available at: https://doi.org/10.1007/s11165-016-9602-2.
Tarazona Gil, C., Lloret-Català, C. and Martínez-Usarralde, M.-J. (2025) “Validación de un Toolkit de Aprendizaje-Servicio digital para la Educación Superior,” Profesorado, Revista de Currículum y Formación del Profesorado, 29(2), pp. 1–23. Available at: https://doi.org/10.30827/profesorado.v29i2.32534.
Vuorikari, R., Kluzer, S. and Punie, Y. (2022) “DigComp 2.2: The Digital Competence Framework for Citizens — With new examples of knowledge, skills and attitudes”. Luxembourg: Publications Office of the European Union. Available at: https://doi.org/10.2760/115376.
Vuorikari, R., Pokropek, A. and Muñoz, J. (2025) “Enhancing digital skills assessment: introducing compact tools for measuring digital competence,” Technology, Knowledge and Learning. Available at: https://doi.org/10.1007/s10758-025-09825-x.
Zhao, Y. et al. (2021) “Digital Competence in Higher Education: Students’ Perception and Personal Factors,” Sustainability, 13(21). Available at: https://doi.org/10.3390/su132112184.