ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DOTTORATO DI RICERCA IN

Psychology

Ciclo XXXVI

Settore Concorsuale: 11/E1 - PSICOLOGIA GENERALE, PSICOBIOLOGIA, PSICOMETRIA

Settore Scientifico Disciplinare: M-PSI/03 - PSICOMETRIA

Exploring the Psychometric Dimensions: Theoretical and Practical Applications of Artificial Intelligence and Gamification in Education, Learning, and Neuropsychological Assessment

Presentata da: *Matteo Orsoni*

Coordinatore Dottorato

Prof. Elisabetta Crocetti

Supervisore

Prof. Mariagrazia Benassi

Co-Supervisore

Prof. Elvis Mazzoni

Esame finale anno 2024

DOTTORATO DI RICERCA IN

Psychology

Ciclo

XXXVI

Settore Concorsuale: 11/E1 - PSICOLOGIA GENERALE, PSICOBIOLOGIA, PSICOMETRIA

Settore Scientifico Disciplinare: M-PSI/03 - PSICOMETRIA

Exploring the Psychometric Dimensions: Theoretical and Practical Applications of Artificial Intelligence and Gamification in Education, Learning, and Neuropsychological Assessment

Presentata da: *Matteo Orsoni*

Coordinatore Dottorato

Prof. Elisabetta Crocetti

Supervisore

Prof. Mariagrazia Benassi

Co-Supervisore

Prof. Elvis Mazzoni

Esame finale anno 2024

Table of Content

Acknowledgements

I would like to express my sincere gratitude to all those who have contributed to the completion of this doctoral thesis. Without their support, guidance, and encouragement, this endeavor would not have been possible. First and foremost, I am deeply grateful to my supervisor, Prof. Mariagrazia Benassi, and my co-supervisor, Prof. Elvis Mazzoni, for their invaluable mentorship, insightful feedback, and unwavering support throughout the entire research process. Their expertise and dedication have been instrumental in shaping this work. I extend my heartfelt appreciation to the members of my doctoral committee, for their constructive criticism, scholarly advice, and time invested in reviewing and evaluating this thesis. I am grateful to University of Bologna for its financial support, which made this research possible. I am grateful for the invaluable contributions of my colleagues and collaborators, both within the University of Bologna, the entire SPEV group, and beyond, who have significantly enriched my research journey. Special recognition goes to Dr. Milos Kravcik, Dr. Nghia Duong-Trung, Prof. Martin Grüttmüller, Mr. Alexander Pögelt, and Prof. Marco Scutari. Their insightful ideas, stimulating discussions, and fruitful collaborations have not only enhanced the quality of my work but also made the research process immensely rewarding. I am immensely grateful to Caterina for her support and for providing me with the strength to persevere, even during the most challenging moments. Her constant encouragement has been a source of reassurance and motivation, for which I am truly thankful. I am deeply thankful to my family for their unconditional love, encouragement, and patience throughout this demanding undertaking. Their belief in me has been a constant source of strength and motivation. Finally, I want to extend my heartfelt gratitude to all my friends, whether we've shared a long history or crossed paths during this journey. It's challenging to mention everyone individually here, but I sincerely thank each of you for your understanding, encouragement, and moments of joy and relief throughout these demanding years. This thesis is a testament to the collective effort and support of all those

mentioned above, as well as many others who have contributed in various ways. Thank you all for being part of this significant chapter in my academic and personal growth.

Abstract

In recent decades, the intersection of artificial intelligence (AI) and gamification has reshaped learning, education, and neuropsychological assessment. Tracing its roots to mid-20th-century pioneers like Alan Turing, this convergence reflects historical progress. As personal computing emerged in the late 20th century, AI principles integrated into education alongside the rising interest in gamification. Titled "Exploring Psychometric Dimensions: Theoretical and Practical Applications of Artificial Intelligence and Gamification in Education, Learning, and Neuropsychological Assessment," this dissertation aims to delve into the theoretical foundations and real-world applications of incorporating artificial intelligence and gamification in the realms of education, learning, and neuropsychological assessment with a psychometric perspective. The integration of AI and gamification goes beyond traditional pedagogical methods, promising personalized educational experiences. AI's data analysis, machine learning, and adaptive algorithms complement gamification's game design elements, engaging learners and enhancing motivation. In neuropsychological assessment, this fusion offers an innovative framework for evaluating cognitive functions with the potential for more accurate evaluations and enhanced participant engagement. Structured into five main sections, the dissertation begins with a clear introduction, outlining the two primary themes: artificial intelligence and gamification. Part I, "Theoretical Contributions," explores how AI influences human development and essential skills in education and examines gamification's knowledge and constraints. Part II, "Applications on Neuropsychological Assessment and Psychometrics", delves into the integration of AI and gamification in cognitive assessment. Part III, "Applications on Training", focuses on AI's role in training and recommender systems in learning. The dissertation concludes with a general discussion summarizing findings and suggesting future directions. Through critical examination and interdisciplinary exploration,

this research contributes to understanding the transformative impact of AI and gamification on learning, education, and neuropsychological assessment practices.

Introduction

In recent decades, the intersection of artificial intelligence (AI) and gamification has become a powerful force, transforming the landscape of learning, education, and neuropsychological assessment. To understand the origins of this convergence, we trace its roots back to the mid-20th century when early AI pioneers like Alan Turing laid the theoretical groundwork for what we now observe as a dynamic interplay between AI and gamification. As technology advanced and personal computing emerged in the late 20th century, the stage was set for the integration of AI principles into various fields, including education. This period saw a surge in interest and research exploring the potential of AI in educational settings, while simultaneously, the concept of gamification gained momentum, drawing inspiration from historical roots in game-based learning and behavioral psychology.

The confluence of these historical trajectories in AI and gamification has culminated in a dynamic and transformative force within the domains of learning, education, and neuropsychological assessment. This convergence reflects not only decades-long scientific and technological progress but also presents unprecedented opportunities to redefine educational practices and cognitive assessments. "Exploring the Psychometric Dimensions: Theoretical and Practical Applications of Artificial Intelligence and Gamification in Education, Learning, and Neuropsychological Assessment" aims to explore both the theoretical underpinnings and real-world applications of integrating artificial intelligence and gamification within the domains of education, learning, and neuropsychological assessment, all viewed through a psychometric lens.

The integration of AI and gamification in learning and education marks a paradigm shift, transcending traditional pedagogical methods. AI, with its capacity for data analysis, machine learning, and adaptive algorithms, promises personalized and interactive educational experiences. Simultaneously, gamification introduces elements of game design to engage learners and enhance motivation, creating a synergistic approach that captivates the modern

learner. In the domain of neuropsychological assessment, the fusion of AI and gamification presents an innovative framework for evaluating cognitive functions and neurological disorders. This marriage of technology and assessment strategies holds potential for more accurate and efficient evaluations while introducing a dynamic element that may enhance participant engagement and compliance.

This goal of this dissertation is to delve into the profound implications of applying AI and gamification in learning, education, and neuropsychological assessment. By critically examining the current landscape, identifying challenges, and proposing novel methodologies; through an interdisciplinary lens, this research seeks to contribute to the ongoing dialogue surrounding the transformative impact of these technologies on human cognition, education, and neuropsychological assessment practices, with a specific focus on their psychometric dimensions.

This dissertation is organized into five main sections. The first provides a general introduction, outlining the two primary themes explored throughout the work: artificial intelligence and gamification (Section 1.1). The intention is to provide clarity and align the reader with the concepts and content discussed in the subsequent sections.

Following the introduction, Part I, titled "Theoretical Contributions," presents two theoretical works. These contributions delve into theoretical perspectives and offer new research reflections intended to guide researchers, teachers, educators, and students in contemplating the integration of new technologies such as artificial intelligence and gamification in educational contexts. The first theoretical work (Section 2.1) explores how AI can influence human behavioral development and the acquisition of essential skills and competences, including Creativity, Critical Thinking, Problem Solving, and Computational Thinking, within an educational framework. The second work (Section 2.2) delves into the existing knowledge and

constraints surrounding the use of gamification in the realm of learning and education, adopting a methodological standpoint to elucidate these aspects. Proceeding further, Part II, entitled "Applications on Neuropsychological Assessment and Psychometrics" and consisting of Sections 3.1, 3.2, and 3.3, explores the incorporation of AI and gamification in cognitive assessment and in the validation of questionnaire psychometrics. Sections 3.1 and 3.2 emphasize the application of AI in revealing cognitive structures inherent in students' profiles through diverse AI methodologies. In contrast, Section 3.4 introduces an innovative approach to address questionnaire psychometric validation, incorporating Information Theory, Machine Learning, and Bayesian Networks.

Transitioning to Part III, titled "Applications on Training," the focus shifts to the utilization of AI in training and recommender systems in learning (Section 4.1). This section employs a combined approach, integrating Item Response Theory and Reinforcement Learning to dynamically address mathematical item presentation, with a primary emphasis on achieving the learner's objectives. Finally, Section 5.1 concludes with a general discussion summarizing findings and providing insights into future directions.

1.1 Artificial Intelligence, Machine Learning, and Deep Learning

1.1.1 A brief history of Artificial Intelligence

The history of Artificial Intelligence (AI) unfolds as a captivating journey characterized by innovation, significant advancements, and the continual pursuit of emulating human intelligence within machines.

Interestingly, its origins can be traced back to the early 1940s, notably in the short story "*Runaround*" penned by the American Science Fiction writer Isaac Asimov (1942). This story introduced the famous Three Laws of Robotics, serving as inspiration for the subsequent generation of computer scientists, mathematicians, and cognitive scientists in the field of robotics. Nevertheless, it is more widely recognized that Alan Turing, the English mathematician, played a pivotal role in elevating Artificial Intelligence from the realm of science fiction to a tangible reality (Haenlein & Kaplan, 2019). Famous for the code braking machine "*The Bombe*", in 1950 published a groundbreaking article titled "Computing Machinery and Intelligence", wherein he posed the fundamental question: "Can machines think?"(Turing, 1950). Turing in this seminal work outlined the principles for developing intelligent machines and providing a method to assess their intelligence. This assessment, known as the Turing Test, remains a significant benchmark for determining the intelligence of an artificial system. According to the test, if a human interacting with both another human and a machine cannot differentiate between the two, the machine is deemed intelligent.

Only six years later, in 1956, the word Artificial Intelligence finally get officially coined. This was made from two eminent scientists Marvin Minsky, a cognitive scientist, and John McCarthy, a computer scientist, organized the *Dartmouth Summer Research Project on Artificial Intelligence* (DSRPAI) at Dartmouth College in New Hampshire (McCarthy et al., 2006). This approximately eight-week-long workshop, served as the starting point for the "AI spring" period and brought together individuals who would later be recognized as the founding

fathers of AI. The initial decades witnessed optimism and ambitious goals, often fueled by the belief that machines could mimic human cognitive abilities. Early AI projects focused on symbolic reasoning and problem-solving, leading to the development of expert systems. However, the field faced challenges, and an "AI winter" occurred when expectations exceeded technological capabilities (Haenlein & Kaplan, 2019). The "AI winter" persisted for several years; however, in the late 20th century, a resurgence of AI occurred due to various factors. Notably, this period witnessed substantial enhancements in computer power capabilities. Breakthroughs in hardware, including the development of faster processors and more efficient storage systems, empowered AI systems to handle vast amounts of data at increasingly faster speeds. This surge in computational power facilitated the implementation of more complex algorithms and models, thereby enhancing the effectiveness of AI applications. Additionally, the increased availability of larger datasets for training AI algorithms, particularly in Machine Learning, proved beneficial, allowing systems to learn and adapt from data with greater accuracy. Finally, breakthroughs in cognitive science, which is the study of how the mind processes information, provided valuable insights into human intelligence and paved the way for more sophisticated AI approaches. Researchers began incorporating principles from cognitive science to enhance the design and functionality of AI systems (Fan et al., 2020). Discoveries about neural connections in the human brain, revealed through microscopes, served as inspiration for the development of artificial neural networks (Hebb, 1950). Electronic detectors unveiled the brain's convolution property and multilayer structure, inspiring the creation of convolutional neural networks and deep learning (Krizhevsky et al., 2012; LeCun et al., 1989). The identification of the attention mechanism, achieved through positron emission tomography (PET) imaging, inspired the design of attention modules (James, 2007). Functional magnetic resonance imaging (fMRI) results, revealing insights into working memory, provided inspiration for the memory module in machine learning models, contributing to the evolution

of long short-term memory (LSTM) (Hochreiter & Schmidhuber, 1997). Observations of changes in the spine during learning, conducted with two-photon imaging systems, influenced the development of the elastic weight consolidation (EWC) model for continual learning (Kirkpatrick et al., 2017). The paragraph titled "Exploring the Intersection between Neuroscience and Artificial Intelligence" will delve into the influence of neuroscience on the progress of AI solutions. In the following section, I will introduce key terminology essential for providing readers with the foundation to comprehend subsequent paragraphs.

1.1.2 Decoding AI Terminology: Unraveling the Distinctions between Artificial Intelligence, Machine Learning, and Deep Learning

In recent years, there has been a tendency to use the terms Artificial Intelligence (AI), Machine Learning (ML), and Deep Learning (DL) interchangeably. However, it is crucial to clarify the hierarchical structure inherent in these terms. The purpose of this section is to elucidate this structure, and the most effective approach is to begin by providing individual definitions. In addition, the Figure 1-1, serves to graphically elucidate this structure. The term "Artificial Intelligence" is closely associated with human intelligence. As explained in the introduction by Russell and Norvig (2021), AI refers to the creation of artificial agents or machines capable of performing tasks typically associated with humans. These tasks encompass learning, perception, reasoning, and specific activities such as playing chess, driving cars, creating poetry, or making diagnoses. Conversely, the term "Machine Learning" refers to algorithms capable of discerning patterns from data without explicit human guidance. Within this dissertation, the employed methods encompass algorithms and techniques for the purpose of learning representation or extracting information from the data. Machine Learning (ML) can be broadly classified into various types depending on the learning approaches and tasks undertaken. The most pertinent categories include Supervised, Unsupervised, and

Reinforcement Learning. Furthermore, although Semi-Supervised and Self-Supervised algorithms can be considered to broaden the spectrum of ML methodologies, it's important to note that these two approaches will not be addressed in this dissertation. Several widely used machine learning algorithms are found in the current literature, including but not limited to Classification and Regression Trees (CART) (Breiman, 1984), Random Forest (RF) (Breiman, 2001), Support Vector Machines (SVM) (Corinna Cortes & Vapnik, 1995), Bayesian Networks (BNs) (Koller & Friedman, 2009), Artificial Neural Networks (ANN), K-means (MacQueen, 1967), Fuzzy C-means (Nayak et al., 2015a), Gaussian Mixture Models (GMM) (D. Reynolds, 2009), and various others. Advancing in the hierarchy, Deep Learning can be characterized as a Machine Learning approach that utilizes Artificial Neural Networks in various forms as its foundational components. DL comprises algorithms designed to learn at multiple levels, each corresponding to distinct levels of abstraction (Chen, 2015). This approach, inspired by the structure of the brain, has found applications in various fields including artificial intelligence, image processing, robotics, and automation (Polson & Sokolov, 2018). Deep learning techniques, which can be supervised or unsupervised, are particularly effective in discovering abstract features in data and have shown state-of-the-art performance in areas such as object perception, speech recognition, and natural language processing (Firdaus & Dixit, 2018).

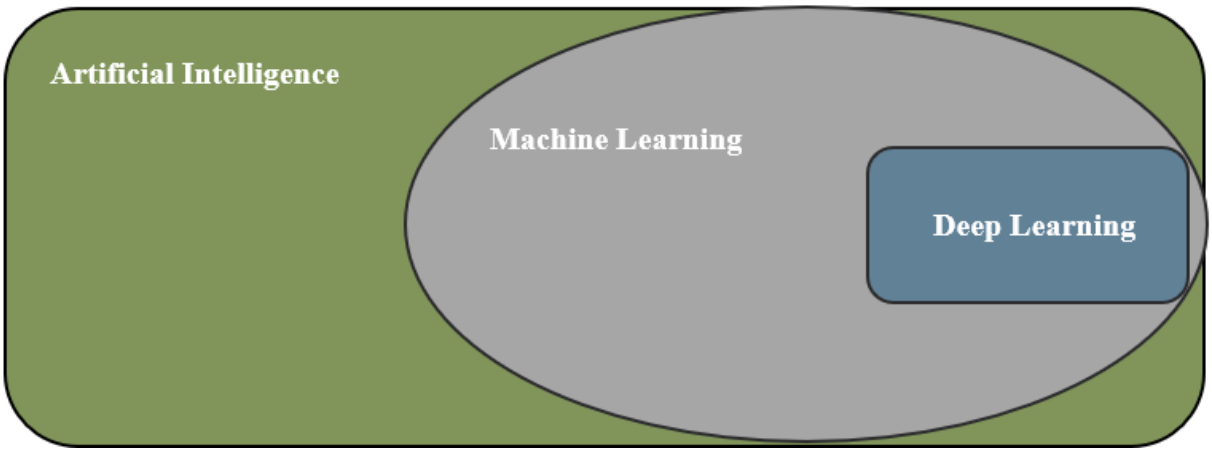

Figure 1-1-1 The hierarchical structure of Artificial Intelligence

1.1.3 Supervised Machine Learning

The most common branch of Machine Learning algorithms (Lecun et al., 2015), are the supervised ones (SML). The term "supervised" denotes that the learning process is guided by explicit instructions. During this learning phase, the training data acts as an informative channel, highlighting optimal associations between input data and their corresponding output labels. Through this instructional process, the algorithm discovers and assimilates patterns (data mining), relationships, and decision-making criteria, thereby cultivating the ability to apply this acquired knowledge to novel, unseen instances (Alloghani et al., 2020).

To deepen the comprehension of Supervised Machine Learning (SML) operations, let's delve into an illustrative example using psychological data. Imagine a scenario where the goal is to create a system capable of discerning between students with and without dyslexia. In this case, an algorithm processes cognitive neuropsychological data as input. Through this process, the algorithm identifies associations between the data and the classification of students as dyslexic or non-dyslexic, leveraging diagnoses provided by psychologists. The training of the machine learning model involves fine-tuning the internal parameters, or weights, guided by an objective function. This function quantifies the error between the model's output scores and the desired

pattern of scores derived from psychologist diagnoses. The iterative nature of this process refines the model's ability to classify students by adjusting the weight vector in the opposite direction to the calculated gradient vector.

Establishing solid foundations is crucial when developing effective machine learning solutions. In the upcoming section, I will succinctly delineate these best practices. The goal is to provide readers unfamiliar with the building blocks of machine learning implementation a clearer understanding, ultimately preparing them for Part II of this dissertation.

The overall goal of supervised ML implementation is to create a model able to classify or predict an output variable by discovering patterns or rules among the input variables. The objective is to provide a solution as accurate and reliable as possible. The **bias-variance tradeoff** is a fundamental principle in machine learning that necessitates careful consideration. This concept revolves around the challenge of finding an optimal equilibrium between model complexity and generalization. Essentially, it highlights the delicate balance between two types of errors a model can make. Bias represents the error stemming from simplifying a real-world problem with a basic model, resulting in underfitting and diminished generalizability. On the other hand, variance is the error arising from employing a complex model overly sensitive to fluctuations in the training data, leading to overfitting and similarly diminished generalizability. Achieving the right tradeoff involves adjusting the model's complexity to minimize both bias and variance, ultimately enhancing predictive performance on new, unseen data. Striking this balance is crucial for the development of models that exhibit robust generalization capabilities.

The initial phase of developing a reliable supervised machine learning solution begins with the problem understanding and the data at hand. The **pre-processing** phase, is dedicated to laying the foundation for model training, emphasizing the importance of ensuring the input data is of high quality and thoroughly understood. It is recommended to apply different steps to reach the goal. Initially, conducting an exploratory analysis of the data should provide insights into the

distribution of features and identify the presence of outliers or missing data. When these aspects are identified, it becomes necessary for the machine learning expert to address them strategically, aiming to minimize bias in the model solution. Furthermore, it is advisable, in most machine learning solutions, to standardize or normalize features. This ensures that the model receives features on a consistent scale. The choice of standardization or normalization method depends on factors such as the distribution of features and the type of machine learning application being implemented. Additional preprocessing techniques that enhance machine learning solutions include Feature selection and Feature engineering. Feature selection involves choosing the most informative features (independent variables) from a set of all available features. This step becomes crucial, particularly in datasets where the number of features (P) is significantly larger than the number of subjects (N). On the other hand, Feature engineering involves creating new features, often based on domain knowledge and preliminary data analysis. Essentially, feature engineering aims to derive novel input features from existing ones, intending to improve the overall performance of machine learning models (Orrù et al., 2020). The preprocessing step concludes with the division of the dataset into training-validation and test sets. In machine learning practices, these three datasets serve distinct purposes: the training data are where the ML algorithm learns the rules and relationships within the data. Typically, the dataset is split with 70% of the data allocated to the training dataset, 10% to the validation set, and the remaining 20% to the test set. It's important to note that these percentages are not fixed and can be adjusted based on the specific characteristics of the problem at hand. The validation and test sets serve a common purpose: assessing the model's performance on data it has not seen during training. These datasets play a crucial role in ensuring that the model can generalize well to new, unseen instances, providing valuable insights into its overall predictive capabilities. After completing the pre-processing phase, the **model training** phase commences, focusing on the creation and training of the model. The initial step in this phase is model

selection. As previously mentioned, there are numerous machine learning solutions available today, and choosing the appropriate one necessitates a deep understanding of the problem at hand, the nature of the data, and the exploration of different algorithms to identify the most effective option. Furthermore, during the model training process, it is crucial to carefully monitor for signs of overfitting or underfitting and make necessary adjustments to optimize model performance. A recommended practice to mitigate overfitting, ensuring greater robustness and generalization, involves the implementation of cross-validation and regularization techniques. Finally, every algorithm comes with a set of parameters that can be adjusted to enhance or optimize model performance. This optimization process is known as fine-tuning hyperparameters. To effectively explore the hyperparameter space, is possible to employ efficient tuning techniques like grid search or random search.

Following the training phase, the **model evaluation** phase starts. In this step is essential to assess the performance of the best model on unseen data. The selection of appropriate evaluation metrics varies depending on the nature of the problem—whether it is a classification or regression task and is primarily dictated by the specific characteristics of the problem being addressed. Finally, to gain a thorough understanding of the model predictions is a fundamental step. This is called model interpretability and utilize methods like feature importance analysis, Shapley Additive Explanations (SHAP) (Lundberg & Lee, 2017), or Local Interpretable Model-Agnostic Explanations (LIME) (Ribeiro et al., 2016) methods to uncover insights into the decision-making process of the model.

This provides a concise overview of the fundamental components associated with supervised machine learning implementation. For further exploration, interested readers can delve into valuable resources listed in the following references. (Gareth James Trevor Hastie, Robert Tibshirani, 2013; Lundberg & Lee, 2017; Orrù et al., 2020; Raschka & Mirjalili, 2019; Ribeiro et al., 2016).

1.1.4 Unsupervised Machine Learning

Unsupervised models, differently to the supervised ones, are constructed using unlabeled examples and involve grouping these examples based on their similarities (Orrù et al., 2020).

In unsupervised machine learning, two primary techniques are prominent: clustering and dimensionality reduction. Despite appearing similar initially, these techniques serve distinct purposes. Clustering is geared towards grouping similar data points, facilitating the discovery of patterns or structures within the data. On the other hand, dimensionality reduction is aimed at simplifying the dataset by reducing its feature space. This technique is employed to manage and visualize data more effectively, as well as to enhance the performance of machine learning models (Pavithra & Parvathi, 2017). Indeed, in some situation can be seen as Feature engineering technique.

Some classic and influential techniques for both linear and nonlinear dimensionality reduction include: Principal Component Analysis (PCA) (Kurita, 2019) and t-Distributed Stochastic Neighbor Embedding (t-SNE) (Maaten & Hinton, 2008). These are used to transform high-dimensional data into a lower-dimensional space. On the other hand, some popular clustering algorithms include, but are not limited to: k-means (MacQueen, 1967), Fuzzy C-means (Nayak et al., 2015a), and Gaussian Mixture Models (D. Reynolds, 2009). In section 3.2 and 3.3 I will move further on these topics by using different dimensionality reduction techniques jointly with Clustering techniques to discover internal data structures.

1.1.5 Reinforcement Learning

Reinforcement Learning (RL) stands out as the third and distinct machine learning paradigm, differing from the more familiar Supervised and Unsupervised approaches. In Reinforcement Learning an agent learns to make decisions by interacting with an environment over the time in order to maximize a certain reward (Sutton & Barto, 1998). To grasp the dynamics and

fundamental concepts within RL, as well as the interaction between the agent and the environment, it is essential to introduce certain notations. Furthermore, Figure 1-2 provides a visual representation to enhance the comprehension of the entire process in the classical RL framework.

Firstly, we can say that there is an agent (e.g. robot, human or software) that gives a **state** S_t by the environment at the time t. To answer to the state the agent can choose an **action** A_t at the same time according to an internal **policy** or strategy $\pi(S_t)$. This chosen action propels the agent back into the environment, which, in turn, responds by providing a new state, advancing in time to S_{t+1}, and simultaneously furnishing the agent with a scalar **reward** R_{t+1}.

A single transition is formed by this sequence, and an agent's engagement with its environment involves multiple such transitions. When examining these transitions, there is a clear assumption that the future is independent to the past, given the present. In simpler terms, the subsequent state depends solely on the current state, action, and environmental properties, without being influenced by any previous states or actions. This assumption is recognized as the Markov assumption, making the entire process a Markov Decision Process (MDP).

Then in the classical RL framework, the general goal of the agent is to maximize the reward obtained over several transitions, by discovering an optimal policy that, when employed to choose actions, yields an optimal reward over an extended period also called **expected return**. The predominant approach to achieving this maximization goal is to optimize the discounted reward. This involves incorporating a discount rate, denoted as γ^k, at each time step k during the calculation of the expected return. The discount rate ranges from 0 to 1 ($0 \leq \gamma \leq 1$), and it dictates the current value of future rewards. Specifically, a reward received k time steps in the future is valued at γ^{k-1} times its immediate worth (Sutton & Barto, 1998). As γ approaches to 1, then the return objective gives increasing emphasis to future rewards.

The expected return is then strictly related to another fundamental concept in RL, that is the concept of **value functions**. These functions determine “how good” (Sutton & Barto, 1998) is for the agent to be in a given state. In other words, this notion of "how good" represents the anticipation of future rewards that the agent can expect, according to the actions. The value functions serve as a crucial component for policy learning. Formally, the policy can be perceived as the mapping from states to the probabilities of choosing each available action.

Many RL algorithms employ a value function to assign utility to states and actions. Specifically, the value function for a state S_t represents the anticipated reward the agent expects to receive when initiating from that state and consistently following a particular policy indefinitely.

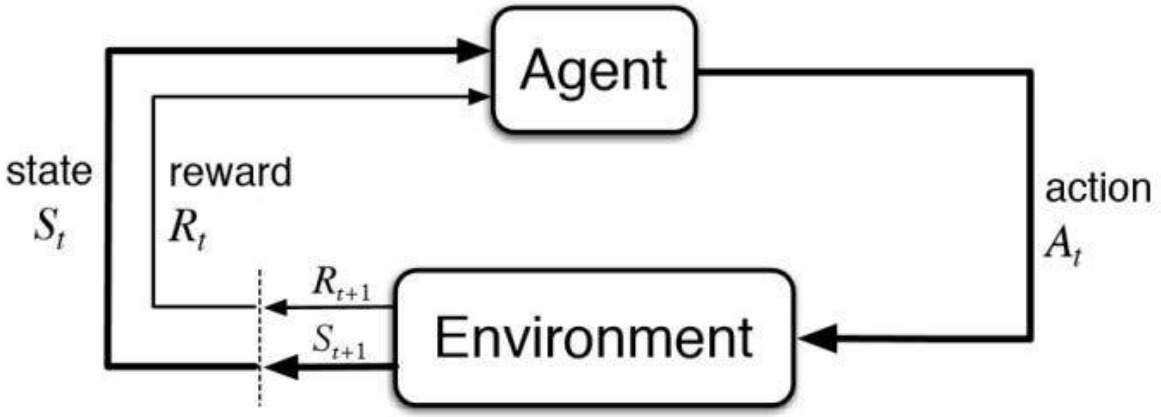

Figure 1-1-2 The classical RL framework. According to Sutton and Barto (1998)

Then the value function of a state s under a policy π denoted as $v_\pi(s)$ can be represented as MDP in the following way:

$$v_\pi(s) = E\left[\sum_{k=t}^{\infty} \gamma^{k-t} R(S_k, \pi(S_k))\right] \quad (1)$$

Where the equation calculates the expected return $E[\cdot]$ by summing up the discounted rewards over an infinite time horizon, considering the rewards obtained at each time step when following policy π, and is called the state-value function.

Another type of value function is expressed through the assignment to a state-action pair. In this *action-value function*, the estimation involves anticipating the reward expected when a specific policy is pursued after taking a particular action in that state. The action-value function can be described as:

$$\boldsymbol{q_{\pi}(s,a) = E\left[R_{k+1} + \sum_{k=t+1}^{\infty} \gamma^{k-t} R(S_k, \pi(S_k))\right]} \quad (2)$$

where it calculates the expected return by summing up the discounted rewards over an infinite time horizon, starting from the immediate reward obtained after acting a in state s. It considers the rewards obtained at each subsequent time step when following policy π from the new states.

In general, RL algorithms based on value functions typically focus on optimizing these estimates rather than directly optimizing the policy. After learning the optimal value function, selecting the highest value actions at each state defines an optimal policy. This process, known as value iteration is employed in various contemporary reinforcement learning algorithms (Subramanian et al., 2022). To derive the optimal policy, it is essential to articulate the connection between the value of a state and the values of its successor states, considering the expected cumulative reward. The Bellman Expectation Equation for the state-value and action-value functions compute this relationship. Where:

$$V_{\pi}(s) = \sum_{a} \pi(a \mid s)(R(s,a) + \gamma \sum_{s'} P(s' \mid s,a) V_{\pi}(s')) \quad (3)$$

Is the Bellman Expectation Equation for the state-value function, $P(s' \mid s, a)$ is the transition probability to state s' from state s after taking the action a.

The following equation:

$$Q_\pi(s,a) = \mathrm{R(s.a)} + \gamma \sum_{s'} P(s' \mid s,a) \sum_{a'} \pi(a' \mid s') Q_\pi(s',a')$$

(4)

Is the Bellman Expectation Equation for the action-value function.

To achieve the optimal policy, we then need to consider the Bellman Optimality Equation. For the state-value function is computed as:

$$V^*(\mathrm{s}) = max_a \left(\mathrm{R(s,a)} + \gamma \sum_{s'} P(s' \mid s,a) V^*(\mathrm{s}')\right)$$

(5)

In general, the Bellman Expectation Equation for $V_\pi(s)$involves calculating the expected vale of the next state's value $V_\pi(\mathrm{s}')$ under the policy π. On the other hand, the Bellman Optimality Equation for $V^*(\mathrm{s})$ entails finding the maximum value of the next state's value $V^*(\mathrm{s}')$ over all possible actions. The connection between these equations is established when considering the optimal policy π_* where $V^*(\mathrm{s}) = V_{\pi*}(\mathrm{s})$. In this specific scenario, the Bellman Expectation Equation effectively transforms into the Bellman Optimality Equation for state-value function.

In a similar way for the action-value function is:

$$Q^*(s,a) = \mathrm{R(s,a)} + \gamma \sum_{s'} P(s' \mid s,a) max_{a'} Q^*(\mathrm{s}', \mathrm{a}')$$

(6)

And like state-value function, the Bellman Expectation Equation for $Q_\pi(s,a)$ involves calculating the expected vale of the next state's value $Q_\pi(\mathrm{s}', \mathrm{a}')$ under the policy π. On the other hand, the Bellman Optimality Equation for $Q^*(\mathrm{s}, \mathrm{a})$ entails finding the maximum value of the

next state's value $Q^*(s', a')$ over all possible actions. The connection between these equations is established when considering the optimal policy π_* where $Q^*(s, a) = Q_{\pi*}(s, a)$. In this specific scenario, the Bellman Expectation Equation effectively transforms into the Bellman Optimality Equation for action-value function.

Over the years, various RL algorithms have been developed, each exhibiting unique characteristics. In the following, I've outlined some key elements that differentiate these algorithms:

- Model-free vs Model-based:

  In Model-free RL, the algorithms directly focus on learning a policy or value function, bypassing the construction of an explicit model of the environment. Differently, the Model-based RL algorithms build an explicit model of the environment and use it for decision-making. An example of Model-free RL algorithm is the Q-learning (C. J. C. H. Watkins & Dayan, 1992). In Q-Learning, an agent learns to make decisions without building an explicit model of the environment. The agent interacts with the environment, observes states, takes actions, receives rewards, and updates a Q-table (or Q-function) based on the experienced rewards. The Q-table represents the expected cumulative rewards for each state-action pair. The agent uses this table to make decisions by selecting actions that lead to the highest Q-values (Sutton & Barto, 1998). An example of Model-based RL algorithm is the Monte Carlo Tree Search (MCTS) (Świechowski et al., 2023). Monte Carlo Tree Search is a model-based RL algorithm that simulates future trajectories to build an explicit model of the environment. It combines exploration and exploitation strategies to traverse the state-action space efficiently. MCTS maintains a tree structure where nodes represent states, actions, and their associated statistics. The algorithm iteratively expands the tree, simulates trajectories, and updates the model to guide future decisions (Sutton & Barto, 1998).

- Value-based vs Policy-based vs Actor-Critic:

  The Value-based RL algorithm focuses its attention on estimating and optimizing value functions. An example of Value-based RL is the Q-learning algorithm (C. J. C. H. Watkins & Dayan, 1992).

  Differently, the Policy-based RL algorithm takes a direct approach by focusing on learning the policy itself. Instead of estimating the value functions, policy-based algorithms seek to optimize the policy directly to maximize the expected cumulative reward. This category includes algorithms like REINFORCE (R. J. Williams, 1992), and Proximal Policy Optimization (PPO) (Schulman, Wolski, Dhariwal, Radford, & Openai, 2017).

  Actor-Critic methods are a class of reinforcement learning algorithms that blend elements of both policy-based and value-based approaches. The Actor is responsible for learning and updating the policy, while the Critic is responsible for estimating the value function. The Actor and Critic interact closely during the learning process. The actor selects actions based on its policy, and the critic evaluates the chosen actions, providing feedback to both the actor and itself (Sutton & Barto, 1998). Examples of Actor-Critic methods are: Asynchronous Advantage Actor-Critic (A3C) (Mnih et al., 2016), DDPG Deep Deterministic Policy Gradient (DDPG) (Lillicrap et al., 2016).

- On-policy vs Off-policy:

  In On-Policy algorithms, the learning agent interacts with the environment and collects experiences (state-action pairs) based on its current policy. It updates its policy based on these experiences and continues to explore and learn from the ongoing interactions (Sutton & Barto, 1998). An example of On-policy algorithm is SARSA (Rummery & Niranjan, 1994).

Off-Policy algorithms, on the other hand, allows the learning agent to gather experiences from a set of data generated by a different policy. The agent doesn't necessarily follow the policy used to generate the experiences but can learn from and evaluate different policies, providing more flexibility in the learning process. An example of Off-policy algorithm is Q-learning (C. J. C. H. Watkins & Dayan, 1992).

- Online RL vs Batch RL

  Another key element is the way in which the agent interacts with the learning environment, particularly in terms of the data it uses for training. In Batch RL the learning agent collects a fixed dataset of experiences from the environment before any learning takes place. The agent uses the entire batch of data to update its policy or value functions in a batch mode (Levine et al., 2020).

  In Online RL the learning agent interacts with the environment in real-time, collecting experiences one at a time. After each interaction, the agent updates its policy or value functions based on the most recent experience. The learning process is incremental, with the agent continuously updating its knowledge after each interaction. This allows online RL to adapt to changes in the environment more dynamically. However, this flexibility can pose a risk, particularly because it makes the system more susceptible to the influence of noisy or misleading data (Ball et al., 2023; François-Lavet et al., 2018).

Moreover, most of the more advanced algorithms in Reinforcement Learning involves the integration of deep neural networks into the learning process formulating a subfield of algorithms, some already presented, called Deep Reinforcement Learning. The main difference of Deep RL over RL is the representation and approximation of the value functions and policies (François-Lavet et al., 2018). I won't delve deeply into the details of the RL framework, but my aim is to provide a brief introduction to one of the most crucial and recent frameworks in the field of AI. The interested reader can found relevant resources here (François-Lavet et al., 2018;

Levine et al., 2020; Sutton & Barto, 1998; Świechowski et al., 2023). This brief overview aims to offer readers a foundational understanding of RL, facilitating comprehension of the study presented in Part III of this dissertation. In the upcoming section, I will elucidate the profound interconnection between Neuroscience, Psychology, Cognitive Science, and Artificial Intelligence. I will illustrate how progress in the former disciplines influences advancements in the latter, and conversely, how developments in Artificial Intelligence contribute to progress in Neuroscience, Psychology, and Cognitive Science.

1.1.6 Exploring the Intersection between Neuroscience and Artificial Intelligence

The historical journey of AI has unveiled the positive synergy between cognitive science or neuroscience and the revival of AI from the late 20th century to the present day. As we explored the origins of AI, it became evident that many pioneering researchers in this emerging field had backgrounds in cognitive science or neuroscience.

Additionally, breakthroughs in cognitive science played a crucial role in enabling the development of some of the most essential AI algorithms that continue to be utilized today. The intersection of neuroscience and artificial intelligence (AI) is marked by a reciprocal exchange of insights and methodologies. In the pursuit of AI, researchers strive to investigate theories and construct computer systems capable of executing tasks that emulate human or biological intelligence, encompassing functions like perception, recognition, decision-making, and control (Russell & Norving, 2021). In parallel, neuroscience, aims to scrutinize the structures, functionalities, and operational mechanisms inherent in biological brains, including how information is processed, decisions are made, and interactions with the environment occur (G. A. Miller, 2003). Given these definitions, the close association between AI and neuroscience is apparent.

1.1.6.1 The brain networks and the Artificial Neural Networks

Understanding how brain networks function served as the inspiration for the earliest form of Artificial Intelligence. In the early 20th century, the advent of microscopy enabled researchers to observe the connections between neurons in the neural system. Motivated by these neural connections, computer scientists developed the artificial neural network (ANN), marking one of the earliest and most successful models in AI history.

In 1949, Hebbian learning was introduced (Hebb, 2005). This learning approach was directly influenced by the dynamics of biological neural systems. It operates on the principle that a synapse between two neurons strengthens when the neurons on either side of the synapse (input and output) exhibit highly correlated outputs. The Hebbian learning algorithm increases the connection weight between two neurons if their outputs are highly correlated. A modification of Hebbian learning evolved with the introduction of the Perceptron (Rosenblatt, 1958). Proposed by Frank Rosenblatt in 1958, the Perceptron serves as the fundamental unit for processing information in a neural network. It is a single-layer artificial neural network with a multidimensional input, laying the groundwork for the development of multilayer networks. In more detail, the Perceptron takes multiple binary inputs, each associated with a specific weight, and generates a single binary output. Its operation involves summing the weighted inputs and applying an activation function to the result. If the computed sum exceeds a predetermined threshold, the Perceptron outputs 1; otherwise, it outputs 0. Although a single Perceptron has limitations in solving complex problems, it serves as the foundation for more advanced neural network architectures like multilayer perceptrons. These architectures can handle intricate tasks through layered connections and non-linear activation functions.

1.1.6.2 The vision and the Convolution Neural Networks

The insights gained from Hubel and Wiesel's work in 1959 (Hubel & Wiesel, 1959), elucidating how humans process images in the visual system, paved the way for the development of Convolutional Neural Networks (ConvNets) in the late 1980s. Their initial investigations involved single-cell recordings from the mammalian visual cortex, revealing the filtering and pooling of visual inputs in simple and complex cells in the V1 area.

Hubel and Wiesel's research showcased that the brain's visual processing system performs convolutional operations and possesses a multilayered structure. This finding suggested that biological systems employ successive layers with nonlinear computations to convert raw visual inputs into a progressively complex set of features. This process ensures that the vision system remains invariant to transformations, such as pose and scale, during the recognition task (Fan et al., 2020). The architecture of ConvNets, with its convolutional and pooling layers, mirrors the lateral geniculate nucleus (LGN)–V1–V2–V4–inferotemporal (IT) hierarchy in the visual cortex ventral pathway (Felleman & Van Essen, 1991; Lecun et al., 2015). Different studies have shown how the ConvNets can serve as a reliable approximation of how human visual cortex works. In 2014, Cadieu and colleagues demonstrated that when ConvNet models and monkeys are presented with the same image, the activations of high-level units in the ConvNet account for half of the variance observed in random sets of 160 neurons within the monkey's inferotemporal cortex (Cadieu et al., 2014).

1.1.6.3 The tuning process of the brain and the backpropagation algorithm

The process by which artificial networks update weights is a pivotal area of research in artificial intelligence. Currently, the widely employed method for this task is the back-propagation algorithm, introduced by Rumelhart, Hinton, and colleagues in 1986 (Rumelhart, Hinton, et al., 1986). Interestingly, it's worth noting that neuroscientists and cognitive scientists were the initial proponents of this idea, rather than computer scientists or machine learning researchers.

(Rumelhart, McClelland, et al., 1986; Fan, Fang, Wua, Guo, & Dai, 2020). The back-propagation algorithm draws inspiration from the microstructures within neural systems, where the biological brain's neural system undergoes a gradual tuning process through learning. This procedure aims to minimize errors and maximize the reward of the output (Fan, Fang, Wua, Guo, & Dai, 2020).

1.1.6.4 The attention and the attention module

A distinct aspect of artificial intelligence, known as the attention module, has been integrated into AI practices. This module draws inspiration from the psychological concept of attention, emphasizing that intelligent agents should selectively concentrate on relevant information rather than processing all available data. This approach aims to enhance the cognitive process (James, 2007).

The introduction of the attention module was influenced by advancements in medical imaging techniques, such as PET or fMRI studies in the late 20th century, which extensively explored the attention mechanism in the brain (Scolari et al., 2015). Insights gained from studying the biological brain paved the way for AI researchers to incorporate attention modules into artificial neural networks. These modules were implemented either temporally (Bahdanau et al., 2014) or spatially (Reed et al., 2015) resulting in improved performance in deep neural networks for natural language processing and computer vision, respectively. By integrating an attention module, the network gains the ability to selectively focus on significant objects or words while disregarding irrelevant ones. This selective attention enhances the efficiency of both the training and inferential processes, surpassing the capabilities of conventional deep networks.

1.1.6.5 The Working Memory and the Long Short-Term Memory

The human capacity to retain and manipulate information in memory is a crucial ability, as identified by Baddeley and Hitch in the early 1970s, termed as Working Memory (Baddeley & Hitch, 1974). Since the 1990s, researchers have utilized PET and fMRI to investigate working memory in biological brains, pinpointing the prefrontal cortex as a key component (Goldman-Rakic, 1991; Jonides et al., 1993; G. McCarthy et al., 1996).

Building on insights from brain science, AI researchers have sought to integrate a memory module into machine learning models. One prominent method is Long Short-Term Memory (LSTM) (Hochreiter & Schmidhuber, 1997), foundational for tasks like natural language processing, video understanding, and time-series analysis.

Recent studies have demonstrated that models equipped with a working memory module can excel in complex reasoning and inference tasks, such as determining the shortest path between specific points and deducing missing links in randomly generated graphs (Graves et al., 2016). Leveraging previous knowledge, these models can also engage in one-shot learning, requiring only a few labeled samples to grasp a new concept (Santoro et al., 2016).

1.1.6.6 The decision-making process. Learning from the agent-environment interaction: the Reinforcement Learning

Reinforcement Learning (RL) emerged from the convergence of concepts in artificial intelligence, neuroscience, and cognitive science. Several principles from behaviorism have been translated into computational RL algorithms. As presented in section 1.2.5, RL serves as a versatile decision-making framework applicable to various scenarios whenever an artificial agent is faced with multiple action choices (Matsuo et al., 2022).

The concept of reinforcement learning is closely intertwined with established principles in behavioral psychology and neuroscience, such as Pavlovian classical conditioning (Rehman et al., 2023) and Thorndike’s law of effect (1898), also known as instrumental conditioning. A

key distinction between these two types of algorithms lies in the nature of their outcomes: classical conditioning outcomes are independent of the subject's actions, whereas instrumental conditioning outcomes are contingent upon the subject's actions, and where a Stimulus (S) – Response (R) association, called *habits*, is learned. These associations are called habits due the fact once learned are autonomous from the outcome. In more detail, in response to a situation S, the animal initiates a corresponding action R. If the outcome is favorable, the association between S and R is reinforced; conversely, if the outcome is unfavorable, the association weakens. This process enhances the likelihood of the animal exhibiting advantageous responses in similar situations. This process is akin to how reinforcement learning (RL) solutions operate, illustrating how artificial agents can effectively address instrumental conditioning challenges (Maia, 2009). The agents optimize their responses in various situations, seeking to maximize positive rewards while minimizing exposure to negative ones.

While the connection between instrumental conditioning and reinforcement learning is evident, the link with classical conditioning is less clear. The proficiency to optimize actions for maximum rewards and minimal punishments requires the ability to anticipate future outcomes. As a result, reinforcement-learning systems frequently incorporate this predictive capability. A particularly effective method for predicting future reinforcements is through temporal differences, which provides a comprehensive explanation for behavioral patterns, and neural findings on classical conditioning (Maia, 2009). In instrumental conditioning, over the S-R responses, the animals learn Action (A) – Outcome (O), and S – A – O responses. These associations are called *goal-directed.* As suggested by Dickinson (1985), habits remain unaffected by changes in outcome manipulation value, whereas goal-directed actions promptly respond to revaluation procedures (Adams, 1982). This difference shed light on the difference between model-free and model-based algorithm presented in section 1.2.5 (Maia, 2009).

One popular model-free approach, which has parallels in brain science, is the Temporal Difference Learning (TD). TD learning focused on learning the state-value function $V_\pi(s)$ using the TD error δ.

$$\delta_t = r_t + \gamma V(s_{t+1}) - V(s_t) \quad (7)$$

that describes the difference between a real transition and the expectation. Then, the value function is updated as:

$$V(s_t) = V(s_t) + \alpha\delta_t \quad (8)$$

Where α is the learning rate.

TD learning algorithms aim to minimize TD errors to enhance the accuracy of Q values, mirroring behaviors observed in neuroscience. A corresponding reward prediction error (RPE) has been identified in the activities of dopamine neurons across various brain regions, such as the ventral tegmental area (VTA), midbrain, dorsolateral prefrontal cortex (PFC), anterior cingulate cortex (ACC), and striatum (Oyama et al., 2010; Sul et al., 2010).

The outcome of the TD error is an update of the value function, essentially derived from the reward prediction that serves as the basis for generating the reward prediction error (RPE) (C. Fan et al., 2023). According to Cai and colleagues (2011), it has been discovered that the brain employs encoding for both state-value functions and action-value functions. Specifically, neural signals associated with state-value functions are identified in the ventral striatum, anterior cingulate cortex (ACC), and amygdala. These signals play evaluative roles for all available choices. Meanwhile, signals linked to action-value functions are stored and updated at the synapses between cortical axons and striatal spiny dendrites. They prove valuable in selecting a particular action, particularly before and during a motor response. Furthermore,

neural signals representing the selected action values corresponding to post-decision state values have been identified in the orbitofrontal cortex, medial frontal cortex, dorsolateral prefrontal cortex (dlPFC), and striatum (Cai et al., 2011; C. Fan et al., 2023).

Revisiting the differentiation between habits and goal-oriented actions, the literature suggests a neural distinction between these two systems. While habits appear to be modulated by the dorsolateral striatum, goal-directed actions seem to depend on the dorsomedial striatum and prefrontal cortex (Maia, 2009). The associations between stimuli and responses (S-R) that are modulated by the dorsolateral striatum can be linked to a RL architecture, specifically the Actor-Critic algorithm. In this setup, the critic's role is to compute the values of states, denoted as $V(s_t)$, and subsequently use them to calculate the prediction error (see Equation 7). Then, an area that can subserve the role of critic should establish projections to and receive projections from the dopaminergic system. This connectivity is crucial since values are utilized in calculating prediction errors, which, in turn, are employed to update these values (Maia, 2009). In a rat-based electrophysiological study, it was observed that neurons in the ventral striatum primarily represent predicted rewards as opposed to actions. Conversely, neurons in the dorsal striatum are found to represent actions rather than predicted rewards. This observation supports the notion that the ventral striatum bears similarity to the critic network, while the dorsal striatum shares resemblance with the actor network (C. Fan et al., 2023).

According to the neurobiology of the brain, recent literature (C. Fan et al., 2023), suggested a review of the most advanced topics and algorithms in RL and their connection to the brain. The authors aimed to present a bottom-up perspective, starting from micro-neural activity, and progressing to macro-brain structures and cognitive functions.

From micro-neural brain activity, the authors suggested three different RL algorithms. The Distributional RL (DRL), the Stigmergy RL (SRL), and the Successor representation RL (SR-RL).

The primary distinction between traditional RL methods and **Distributional RL** lies in how they compute long-term rewards. Traditional RL methods calculate a single average value for long-term rewards, whereas DRL focuses on modeling a distribution of expected returns. This algorithm works by updating a random variable $Z(s,a)$, where its expectation corresponds to the action value $Q(s,a)$, denoted as $Q(s,a) = E[Z(s,a)]$.

$$Z_{(s_t,a_t)} = r_t + \gamma Z_{(s_{t+1},a_{t+1})} \quad (9)$$

DRL was initially introduced in computer science before being explored for its implications in neural mechanisms. Notably, DRL has demonstrated biological plausibility in dopaminergic and cortical processes. In dopamine neuron responses, recorded from the mouse VTA area, DRL excels in predicting Reward Prediction Error (RPE) turning points and future rewards (Dabney et al., 2020). Additionally, the prefrontal cortex (PFC), particularly the ACC, emerges as a robust candidate for DRL (Muller et al., 2024).

Stigmergy RL, is grounded in the concept of stigmergy, first introduced by Grassé (1959). This concept effectively resolves the "coordination paradox," addressing how insects, despite their limited intelligence and absence of apparent communication, can successfully collaborate on complex tasks such as building a nest (Heylighen, 2016). According to Heylighen (2016), stigmergy involves four essential components: medium, trace, condition, and action, forming a feedback loop between agents and their environment. The medium serves as an information aggregator, facilitating multi-agent collaboration. The trace is a digital pheromone left by agents

in the medium, signaling environmental changes caused by their actions. These traces can superimpose, diffuse, and decay over time.

Briefly, in SRL, as agents engage with the environment, they leave a trace, referred to as a digital pheromone, in the medium. This trace comprises instructional records, encompassing information such as value, time, and location. The medium, essentially a digital pheromone map, conveys the distribution of this information, guiding subsequent action selections by providing conditions. The digital pheromones from different agents can linearly superpose, mutually diffuse, and gradually decay over time. Consequently, the medium undergoes constant updates through mutual communications among agents in the respective area. As the reader will see in the next paragraph, the neuroscientific evidence of SRL, lies in the synaptic connection and in the role of astrocytes. Indeed, recent experimental evidence resembling the importance of astrocytes in the regulation of synaptic transmission. Owing to the enrichment of various receptors, astrocytes can be involved in many neural modulations, and the interaction between synapses is mainly reconciled by the propagation of calcium ion within astrocytes (C. Fan et al., 2023). These evidences have been reinforced by a work of Xu and colleagues (2018). They found that the stigmergy mechanism share numerous similarities with the interactive activities among synapses in the brain.

The third RL algorithm of this section, focused on **Successor Representation RL (SR-RL)**. This opens a third option for RL to learn value functions. Instead of considering model-free or model-based approach, the fundamental concept of SR-RL is to create a "predictive map" of the environment, encapsulating the long-range predictive relationships between different states of the environment (C. Fan et al., 2023). The neural contribution of SR is evident in several studies. The predominant hypothesis is based on the function of hippocampal neurons (Gershman, 2018). To elaborate, if we consider a group of neurons encoding spatial functions for each state in the brain, the resulting population code closely resembles the classical place

fields found in the hippocampus. Furthermore, the resemblance between the hippocampus and SR has been identified through functional magnetic resonance imaging (Momennejad et al., 2017).

From macro-neural brain activity, Fan and colleagues (2023) proposed four different RL algorithms that can have connection with brain science: The Hierarchical RL (HRL), the Meta RL (MRL), the Prefrontal RL (PRL), and lastly the PFC–BG interaction-inspired RL. In general, prefrontal cortex (PFC) and basal ganglia (BG) are two key structures associated with Reinforcement Learning (RL). Regarding the PFC, neurophysiological experiments have indicated that the medial prefrontal cortex (mPFC) plays a role in regulating RL parameters, including the learning rate and exploration rate (Domenech et al., 2020). Additionally, representations in the entorhinal and ventromedial prefrontal cortex (vmPFC) extend beyond specific RL tasks, contributing to a broader understanding and generalization of RL problem frameworks (Baram et al., 2021).

Hierarchical RL (HRL) fundamentally integrates temporally abstract actions, allowing for the organization of a series of interconnected low-level actions into hierarchical subgoals. This integration notably enhances the scalability and learning efficiency of the system (M. M. Botvinick, 2012). These temporally abstract actions are termed "options," and decision policies are developed and optimized over these options instead of individual actions. Each option maintains an option-specific prediction error known as pseudo reward prediction error (PPE), and an option concludes when a specific subgoal is attained. Readers interested in understanding how to mathematically estimate options and the option-value function can refer to (Bacon et al., 2016). HRL offers a robust computational model for comprehending abstract action representations, suggesting the presence of a cognitive hierarchy within the Prefrontal Cortex (PFC). The neural mechanisms responsible for hierarchically organized behavior are believed to be associated with the dorsolateral prefrontal cortex (dlPFC). Neuropsychological

studies have indicated a positive correlation between Pseudoreward Prediction Errors (PPEs) and the activation of the Anterior Cingulate Cortex (ACC) (Chiang & Wallis, 2018). These findings provide support for the notion that the PFC plays a crucial role in encoding subgoals and PPEs, facilitating Hierarchical RL in the brain.

Meta Reinforcement Learning (MRL) has emerged as a promising strategy to address the high sample complexity of Reinforcement Learning (RL) algorithms. It focuses on training agents to acquire transferable knowledge that can generalize to new tasks by leveraging their prior experiences (C. Fan et al., 2023). In MRL the intelligent agent utilizes a shared framework across various tasks during meta-training. This approach facilitates quick adaptation to new tasks during meta-testing, even with a limited number of experiences. MRL received great attention also in neuroscientific community. A work of Tsutsui and colleagues (2016), suggested a relevant role of PFC in MRL. The authors found that the PFC represents both the expected values of actions and states but also encodes the history of actions taken and their corresponding rewards.

Prefrontal RL. This type of RL system stems from the existing evidence, partly presented, indicating that the system is influenced by two prediction error signals. The evidence suggests that the brain employs two distinct systems to guide action selection. The first is a reflexive, model-free RL represented by a Reward Prediction Error (RPE), reporting the disparity between actual and expected rewards. The second is a deliberative, model-based RL employing a State Prediction Error (SPE) to learn and enhance the understanding of the environment's structure.

The concept of Prefrontal RL involves making inferences about the reliability of model-based and model-free systems based on the relative magnitude of the SPE and RPE. By estimating reliability signals, it becomes possible to determine the probability of a model-based RL (PMB), providing insights into the dominance between model-based and model-free systems (S. W.

Lee et al., 2014). Concerning the regions associated with the value signals of the two Reinforcement Learning systems, it is observed that the action value of the model-based system is linked to activity in the orbital area, medial Prefrontal Cortex (mPFC), and certain portions of the Anterior Cingulate Cortex (ACC). Conversely, the action value of the model-free system is associated with activity in the dorsomedial Prefrontal Cortex (dmPFC), dorsolateral Prefrontal Cortex (dlPFC), and the supplementary motor area (Piray et al., 2016).

Inspired to the work of O'Reilly and Frank (2006), where the decision-making process in the brain involves interactions between the prefrontal cortex (PFC) and basal ganglia (BG), in particular the PFC is believed to store contextual reward information in working memory, exerting a top-down influence on the action selection process in the BG, a PFC–BG **Interaction-Inspired Reinforcement Learning (PB-RL)** has been proposed. In this kind of algorithm, the dopamine reward is utilized to evaluate actions in the BG and update working memory in the PFC (F. Zhao et al., 2018).

The last RL category proposed by Fan and colleagues (2023) focused on RL algorithms related to Cognitive Functions. The **Attentional RL** takes inspiration from the cortico – BG – thalamocortical loop. According to Yamakawa (2020), the striatum receives prediction signals from the neocortex. The basal ganglia generate an attention signal, acting as a gate that releases the suppression of the thalamic relay cell through the globus pallidus and mediates the prediction signal. This algorithm would mimic the relation between attention and learning. The brain's attention mechanism enables individuals to concentrate on task-relevant dimensions, leading to improved performance, accelerated learning, and simplified generalization. Neuroscience research highlights a reciprocal relationship between attention and learning. Attention narrows learning to relevant environmental aspects and influences value calculation and updating, while attentional filters adapt dynamically based on ongoing decision outcomes (Leong et al., 2017).

The **Episodic RL** is the last algorithm presented in this brief overview. This takes inspiration from the Episodic memory. Episodic memory, facilitated by the hippocampus and related medial temporal lobe structures, plays a crucial role in providing comprehensive and temporally extended information about interdependent actions and rewards from individual experiences (Greenberg & Verfaellie, 2010). This capability allows organisms to approximate value functions in complex state spaces, learn efficiently with limited data, and establish long-term connections between action and reward functions. These fundamental capabilities contribute to the efficient and precise decision-making observed in humans (Fan et al., 2023). Briefly, in ERL, episodic memories are utilized to create estimates of state- and action-value functions through a nonparametric approximation. The most straightforward implementation of ERL involves storing historical trajectories (Gershman & Daw, 2017).

Other research proposes the integration of neuroscience principles into the implementation of Reinforcement Learning (RL) algorithms. Additionally, some RL algorithms introduce novel perspectives on how the brain operates during the decision-making process. For more details, interested readers can refer to (M. Botvinick et al., 2020; Matsuo et al., 2022; Subramanian et al., 2022; Sutton & Barto, 1998).

1.1.6.7 Where the most advanced AI models are in the brain: the Plausibility of Transformers into the astrocyte's biology.

Regarding the reciprocal influence of decision-making processes guided by discoveries in RL and vice versa, noteworthy advancements have recently emerged in AI, particularly within the realm of well-known Large Language Models (LLMs).

In the paper titled "Attention is all you need," Vaswani and colleagues (2017) introduced a neural architecture that underlies numerous recent innovations in AI, such as Generative Pretrained Transformer-3 (GPT-3) (Brown et al., 2020), and Chat Generative Pretrained

Transformer (ChatGPT) (OpenAI, 2022), called Transformer. The preliminary idea of Transoformers architecture was to overcome the limitations of Recurrent Neural Networks (RNNs) (Vaswani et al., 2017). Unlike RNN, which handle inputs sequentially, Transformers have immediate access to all past inputs. This is facilitated by their self-attention mechanism. Through this, Transformers can learn long-range dependencies between words in a sentence without the need to recurrently maintain a hidden state over extended time intervals (Kozachkov et al., 2023). However, as we have already observed, while a straightforward biological interpretation is conceivable for recurrent and convolutional neural networks, it is not the case for Transformers. This discrepancy is primarily due to the self-attention mechanism employed by Transformers, specifically in how the self-attention matrix is computed. Briefly, the computation involves distinct steps: 1) calculating all pairwise dot products between "tokens" (e.g., words in a sentence, patches in an image, etc.), 2) exponentiating these dot product terms, and 3) normalizing the rows of this matrix to sum to one. Importantly, these operations exhibit a fundamental nonlocality in both temporal and spatial dimensions, rendering their interpretation in biological terms challenging (Kozachkov et al., 2023). Nevertheless, a recent study of Kozachkov and colleagues (2023) proposes that a biological interpretation of the Transformer architecture may be discovered within the realm of astrocyte biology. The astrocytes are the most well-studied type of glial cell, and in the work of Halassa and colleagues (2007) has been estimated how a single astrocyte cell forms connections with thousands to millions of nearby synapses. Throughout much of the brain, neurons and astrocytes exist in close proximity. In regions such as the hippocampus, nearly 60% of axon-dendrite synapses involve the envelopment of astrocyte cell membranes, known as processes (Semyanov & Verkhratsky, 2021). In the cerebellum, this proportion is even higher. This common three-way arrangement, involving the presynaptic axon, postsynaptic dendrite, and astrocyte processes, is referred to as the tripartite synapse (Perea et al., 2009). Furthermore, astrocyte have receptors

that align with the neurotransmitters released at the synaptic sites they envelop. For instance, astrocytes in the basal ganglia respond to dopamine, while those in the cortex are responsive to glutamate (Verkhratsky & Butt, 2007). Additionally, the brain features extensive communication among astrocytes themselves. It has been reported how astrocytes establish expansive networks with one another (Halassa et al., 2007), and communicate via calcium waves (Kuga et al., 2011). On the basis of these neurobiological findings, Kozachkov and colleagues (2023) constructed a computational neuron-astrocyte model with the aim of offering a computational and normative explanation of how communication between astrocytes and neurons supports brain function. Moreover, from a more intriguing perspective related to our topic, their goal was to provide a biologically plausible account of how Transformers could potentially be implemented in the brain. The evidence they presented implies the biological feasibility of this idea. However, as the authors noted, further progress in astrocyte biology is necessary to gain a more comprehensive understanding of the biological mechanism underlying the Transformer architecture.

1.1.7 The AI and gamification in Neuropsychological Assessment

To explore the integration of AI in neuropsychological assessment, we should begin by identifying potential areas where AI can be applied. Casaletto and Heaton (2017) identify three core objectives of clinical neuropsychological assessment: 1) detecting neurological dysfunction and providing guidance for differential diagnosis, 2) characterizing changes in cognitive strengths and weaknesses over time, and 3) offering guidance for recommendations related to everyday life and treatment planning. These objectives present promising opportunities for implementing AI into neuropsychological assessment practices, not only to aid in the diagnosis of cognitive disorders and neurological conditions but also to predict

cognitive decline, assess the impact of interventions, recommend suitable interventions, and contribute to early detection and personalized treatment plans.

An examples on how AI can be integrated into neuropsychological practice, can be found in a work of Langer and colleagues (2022). The authors suggest implementing an AI solution to overcome the current scoring limitations, and then to enhance the rating score of the Rey-Osterrieth complex figure (ROCF). ROCF is a test that assesses nonverbal visuo-spatial memory capacity across diverse age groups, from childhood to old age (Shin et al., 2006). Indeed, a notable limitation of the ROCF quantitative scoring system is the subjective nature of labeling portions of the figure as "accurate" or "inaccurate," which may vary among clinicians. Additionally, scoring could be influenced by factors such as motivation, tiredness, or inadvertent biases during clinician-patient interactions. Consequently, an automated system that provides reliable, objective, robust, and standardized scoring while saving clinicians' time is not only economically advantageous but also crucial for achieving more precise scoring and subsequent diagnoses (Langer et al., 2022). The authors proceeded to employ a multi-head convolutional neural network to address the problem. Their investigation revealed that the AI system surpassed clinicians in score attribution, establishing it as a more reliable tool for the task.

Other applications of AI in neuropsychological assessment, are related to cognitive screening and diagnosis. AI algorithms can assist in the early detection and diagnosis of cognitive disorders by analyzing patterns and anomalies in cognitive performance data. Current research delves into the possibilities of utilizing AI in neuropsychological diagnosis, with a specific focus on categorizing neurological and psychiatric disorders through the analysis of MRI data (Zhang et al., 2021). Reinforcing this exploration is the suggestion of a medical AI agent designed for neuropsychiatric diagnoses. This agent incorporates sensors to gather physiological parameters and patient responses (Rao et al., 2020). Additionally, attention is

drawn to the application of machine learning techniques on structural MRI data for diagnosing depression (Takamiya et al., 2019). AI and ML applications have also found various applications in the developmental population, including the detection of Dyslexia (Giri et al., 2020; Kaisar, 2020), Dyscalculia (Subramanyam et al., 2019), and Neurodevelopmental disorders (Uddin et al., 2019). In conclusion, the expanding applications of AI in neuropsychological assessment mark a significant step forward in enhancing cognitive screening and diagnosis. AI algorithms play a crucial role in the early detection of cognitive disorders, meticulously analyzing patterns and anomalies in cognitive performance data. Ongoing research is actively exploring the potential of AI in the nuanced field of neuropsychological diagnosis. As AI and ML applications evolve, their role in neuropsychological assessment holds promise for advancing diagnostic accuracy and contributing to more effective interventions in cognitive health.

The term "gamification," introduced by Deterding and colleagues in 2011 (Deterding et al., 2011), originates from the digital media industry in 2008. They argue that "gamification" is founded on the growing significance of video games, especially their premise that video games are primarily designed for entertainment, aiming to motivate users to engage with them for extended periods. As per Deterding and colleagues (2011), gamification is defined as "the use of game design elements in non-game contexts." Following this definition, incorporating game elements into products unrelated to the gaming context is believed to enhance their enjoyment and engagement. Other definitions have been presented in subsequent years, all building on the ideas proposed by Deterding and colleagues (2011). Kapp (2012) defines gamification as the use of “game-based mechanics, aesthetics and game thinking to engage people, motivate action, promote learning, and solve problems”. To Hamari and colleagues (2014) “Gamification” is “the phenomenon of creating gameful experiences”, whereas for Werbach (2014) gamification is “the process of making activities more game-like”.

Gamification, nowadays has a wide range of applications, particularly in management, where it has been used in finance, corporate governance, risk management, and human resource management (Wanick & Bui, 2019). It is also increasingly being used in various other fields, including business, banking, education, and medicine (Figol et al., 2021). The effectiveness of gamification in these areas is often analyzed based on psychological and social motivations, as well as through the use of game mechanics and playability metrics (Aparicio et al., 2012). Concerning gamification in neuropsychological assessment, it has been proposed as a way to increase participant engagement and motivation in cognitive tasks, potentially improving data quality (Khaleghi et al., 2021). A thorough examination of the applications and effectiveness of gamified cognitive assessment and training was conducted by Lumsden and colleagues (2016) through a systematic review. The review encompassed 33 pertinent studies that investigated 31 gamified cognitive tasks applied across various disorders and cognitive domains. The findings indicate that gamification has been employed to address primary working memory and general executive functions. However, the impact on task performance was varied, showing mixed effects (Lumsden et al., 2016). In another application, a study assessed cognitive control by employing a mobile game with gamification elements and compared the outcomes to those obtained through traditional assessments. The findings revealed that gamification techniques can enhance learning and cognitive function effectively through consistent engagement and exercise (Gkintoni et al., 2021). However, the integration of gamification into cognitive training and assessment is not without its limitations. Firstly, there is a shortage of comprehensive studies that delve into the design challenges and potential drawbacks associated with using gamification for cognitive assessment and training. Moreover, is the possibility of fostering dependency or overload on game elements, risking the compromise of the learner's intrinsic motivation, autonomy, or creativity (Lumsden et al., 2016). The validation of gamified tasks presents challenges due to varying standards, making

it challenging to disentangle the impact of gamification from the intervention in training games (Gkintoni et al., 2021). To address these issues, it is crucial to employ gamification judiciously and strategically, adopting a balanced approach that incorporates other forms of motivation while refraining from making it the sole focus of the learning process.

1.1.8 The AI and gamification in Learning, and Education

The present era is marked by numerous changes and challenges, largely driven by the rapid growth of Information Communication Technologies (ICT). In this dynamic environment, individuals are compelled to acquire competencies and knowledge that differ significantly from those required just two decades ago. The key to navigating these profound changes lies in education and continuous learning. Considering this, the European Commission has recently emphasized the importance of identifying the skills that contemporary students must develop. These skills have been consolidated under the term "21st-century skills". These skills aim to transition the future generation of EU citizens from a knowledge-centric culture to one centered on competence (Benvenuti et al., 2023; Mazzoni et al., 2022). Yet, to cultivate these skills, education must also progress. A recent study by Benvenuti and colleagues (2023) outlines how emerging technologies such as artificial intelligence and robotics can play a pivotal role in steering this evolutionary process. While this cultural shift is still in its early stages, noteworthy advancements have already been integrated into the education and learning landscape, driven by increasingly sophisticated AI solutions.

A recent review of AI in education and learning, Zhang and Aslan (2021), highlighted at least six domain or AI applications as: chatbot, expert systems, intelligent tutoring systems or adaptive learning systems, and educational data mining, personalized learning systems or environments, and visualizations and virtual learning environments (VLE).

Okonkwo and Ade-Ibijola (2021) highlighted their recognition in chatbot as a valuable technology in enhancing learning within educational settings by fostering a more personalized learning experience. Generally, chatbots are interactive agents designed for engaging conversations and providing prompt responses to users (Smutny & Schreiberova, 2020). The main areas of application focused on teaching and learning activities in education, as well as research and development implementation, administration, assessment, and advisory services (Okonkwo & Ade-Ibijola, 2021). The benefits of integrating chatbots in education and learning encompass streamlined content integration, rapid access to information, heightened motivation and engagement, support for multiple users, and immediate assistance. These chatbots facilitate the collection and storage of diverse information in a centralized unit (Information unit), allowing authorized users quick and easy access. Furthermore, they encourage personalized learning, provide instant user support, and permit simultaneous access to the same information by multiple users. However, the integration of chatbots in education encounters challenges related to ethical considerations, evaluation processes, user attitudes, supervision, and maintenance issues. These challenges have the potential to impact the adoption and utilization of chatbots in educational settings, potentially influencing users' perceptions and restricting the applications of chatbot systems. Finally, the review emphasizes that future research should focus on technological advancements, ethical principles, and usability testing as crucial aspects for a comprehensive understanding of chatbot usage in education and learning.

In the domain of Adaptive Learning Systems and Intelligent Tutoring Systems, AI emerges as a driving force in the evolution of educational platforms (Zhang & Aslan, 2021). Adaptive systems play a crucial role by tailoring educational content intricately to meet the specific needs, learning styles, and progress of individual students. This results in a highly personalized and engaging learning experience where the content is finely tuned to align with each student's unique requirements. Beyond enhancing overall comprehension, this adaptive approach fosters

a dynamic and responsive educational environment attuned to the diverse learning preferences of students. One such example is the Intelligent Tutoring System (ITS), a support system for learning that includes visual lesson guides and interactive exercises. These systems offer immediate feedback and personalization for various users. The main objective of ITS is to support the teaching process while reducing the need for human teacher intervention. This not only saves labor costs but also contributes to improved educational outcomes (Trung et al., 2023).

AI is utilized in Educational Data Mining (EDM). In this field, AI techniques are instrumental in scrutinizing extensive datasets originating from educational environments. Specifically, EDM involves the application of Data Mining methods to analyze student information, educational records, exam results, class participation, and the frequency of students' asking questions (Yağcı, 2022). This data-centric approach involves extracting valuable insights to inform various aspects of the educational landscape. For instance, in a review of EDM application Trung and colleagues (2023) presented a series of applications. One of the current trends in EDM is the use of Machine Learning and Deep Learning techniques to predict student characteristics. This involves utilizing academic, demographic, and psychological data, such as learning styles, cognitive profiling, and introversion/extroversion, to estimate an overall value associated with the student. This value is closely linked to academic ability, achievement, learning style, and emotional states during studying. The significance of this value lies in its potential use by teachers, particularly for providing advice to students and paying closer attention to those facing challenges, thereby enabling more precise and tailored teaching strategies.

Another crucial task within EDM involves the detection of undesirable student behavior. This task focuses on identifying anomalies in student learning progress, including indicators such as

dropout, distraction, and academic failure. The goal is to promptly alert educators, enabling them to take timely actions and address potential challenges (Trung et al., 2023).

EDM has also found application within the framework of Social Network Analysis (SNA). A Social Network (SN) refers to a social structure comprising individuals, groups, or organizations and their interactions. Social Network Analysis (SNA) involves conceptualizing individuals or groups as "nodes" and representing their relationships with "edges." This process entails analyzing the patterns created by these objects to assess their impact on the individuals, known as actors. A social network encompasses a group of actors with unique characteristics and various types of relationships, falling into two categories: directional and nondirectional (Trung et al., 2023). In the context of education, the primary focus of SNA is on students, learners, and their relationships within the group, particularly in the context of collaborative teamwork.

Additional approaches employed in EDM include student modeling, which involves creating a profile for each student that captures traits and characteristics. The goal is to develop personalized systems or recommendations that teachers can use for more tailored learning experiences. Another related concept, but focused on groups, is to offer personalized content to students who share similar characteristics. This approach aims to create customized content for each group, thereby enhancing the quality of teaching and learning (Trung et al., 2023).

Other notable applications of EDM for both teachers and students include the development of recommendation systems, automated grading systems, plagiarism detection systems, classroom monitoring, attendance checking systems, feedback mechanisms to support educators, and the creation of course concept maps. Additionally, EDM plays a role in decision support systems. To a more in depth discussion of these applications the interested reader can see (Trung et al., 2023).

In general, AI plays a pivotal role in education, providing solutions to contemporary challenges and elevating the overall learning experience (Ahmad et al., 2021; Harry, 2023). Its contributions include enabling personalized learning, efficient assessment, and data-driven decision-making, ultimately leading to improved student outcomes (Harry, 2023). Nevertheless, there are challenges such as privacy concerns and the possibility of bias that require careful consideration (Harry, 2023). Despite these obstacles, the potential impact of AI in learning and education is substantial, presenting an opportunity to revolutionize the sector and enhance learning outcomes (Mijwil et al., 2022; Panigrahi, 2020). Finally, the intersection of AI with Learning, and Education, holds the potential to revolutionize traditional approaches, making education more accessible, personalized, and effective.

I won't delve into the topic of gamification in learning and education here. However, in Section 2.2, I will conduct a comprehensive examination of the impact of gamification on education and learning (Orsoni et al., 2023). This investigation will specifically address methodological issues and controversies surrounding gamification. Additionally, I will explore potential moderators identified in recent reviews, systematic reviews, and meta-analyses, shedding light on the intricate connection between gamification, learning, and education. These moderators encompass aspects such as study design, theory foundations, personalization, motivation and engagement, game elements, game design, and learning outcomes.

In the upcoming section, I will present two theoretical contributions. The first pertains to the role of AI in learning and education (Section 2.1), while the second focuses on gamification (Section 2.2). In Section 2.1, the aim is to highlight new research reflections and perspectives that could assist researchers, teachers, educators, and students in contemplating the integration of emerging technologies, such as artificial intelligence and robot tutors. This exploration delves into how these technologies can impact human behavioral development and the acquisition of skills and competences, specifically creativity, critical thinking, problem-solving,

and computational thinking, within an educational context. The analysis suggests a perspective on the effectiveness of creativity, critical thinking, and problem-solving in promoting computational thinking. Furthermore, it explores how Artificial Intelligence (AI) could serve as a valuable tool for teachers in fostering creativity, critical thinking, and problem-solving in schools and educational environments.

Section 2.2 will feature a study examining the role of gamification in learning and education from a methodological perspective. Despite researchers' attempts to assess the impact of gamification in educational settings, several methodological challenges persist. The scarcity of studies with robust methodological rigor diminishes the reliability of results. To this end, the goal of the study is to pinpointed key concepts elucidating methodological issues in the application of gamification in learning and education, leveraging controversies identified in existing literature. The ultimate objective was to establish a checklist protocol facilitating the development of more rigorous studies within the gamified-learning framework. The checklist proposes potential moderators that elucidate the relationship between gamification, learning, and education, as identified by recent reviews, systematic reviews, and meta-analyses. These moderators encompass study design, theoretical foundations, personalization, motivation and engagement, game elements, game design, and learning outcomes.

Part I - Theoretical contributions

2.1

AI in human behavioral development. A perspective on new skills and competences acquisition for the educational context.

Benvenuti, M., Cangelosi, A., Weinberger, A., Mazzoni, E., Benassi, M., Barbaresi, M., & Orsoni, M. (2023). Artificial intelligence and human behavioral development: A perspective on new skills and competences acquisition for the educational context. *Computers in Human Behavior*, 148, 107903. https://doi.org/10.1016/j.chb.2023.107903

2.1 AI in human behavioral development. A perspective on new skills and competences acquisition for the educational context.

2.1.1 Abstract

Despite the significant emphasis placed on incorporating 21st century skills into the educational framework, particularly at the primary level, recent scholarly works indicate considerable variation in the implementation of these skills across different countries and regions, suggesting a demand for further research specifically focusing on primary education. The indications of the Digicomp framework and 21st-century skills in Europe have outlined the key competences for lifelong learning needed for all citizens, including teachers and students. In this perspective, Education plays a fundamental role in ensuring that citizens acquire the required skills. The objective in the common European framework is clear: to initiate a transition from the culture of knowledge to the culture of competence. Nowadays, technological advancement allows the researchers to create and combine different frameworks with the perspective of an even more tailored, and engaged education, some examples derived from the implementation of Virtual Reality (VR) and Augmented Reality (AR), in the combination of Gamification and AI, or the development of Intelligent Tutoring Systems (ITS) to foster and create an even more personalized learning and teaching. Following these premises, in this paper, we want to point out new research reflections and perspectives that could help researchers, teachers, educators (and consequently students) to reflect on the introduction of new technologies (e.g., artificial intelligence, robot tutors) and on how these can affect on human behavioral development and on the acquisition of new skills and competences (Specifically: Creativity, Critical Thinking, Problem Solving, and Computational Thinking) for the educational context. The analysis carried on, suggests a perspective on how creativity, critical thinking, and problem-solving can be effective in promoting computational thinking, and how Artificial Intelligence (AI) could be

an aid instrument to teachers in the fostering of creativity, critical thinking, and problem-solving in schools and educational contexts.

2.1.2 Introduction

Information and Communication Technologies (ICTs) play a relevant role in how European societies perceive, discuss, and approach global challenges, including the COVID-19 pandemic, political destabilization, and climate change. Emerging technologies could be key to understanding and overcoming such challenges but are simultaneously perceived as threats to how we live together in a different social context (European Commission, s.d.).

Artificial intelligence (AI), for example, has accelerated the development of medical breakthroughs, but the threats to humanity are well known if AI is left unchecked, for example AI used in educational or vocational training, that may determine the access to education and professional course of someone's life (e.g., scoring of exams). In this regard, EU proposed a regulation on AI (https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence), and the regulatory framework on artificial intelligence (https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai). The proposed AI regulations are a first step in the direction to a trustworthy AI. While most AI systems pose limited to no risk and can contribute to solving many societal challenges, certain AI systems have to treat in a more cautious way to avoid undesirable outcomes. Implementing AI algorithms in the field of learning require to the developers consider various factors, ranging from the sensitivity of the data utilized for training the algorithms to the reliability and trustworthiness of these algorithms. In line with this trajectory, a novel and burgeoning field of research known as Explainable Artificial Intelligence (XAI) has emerged. The primary aim of researchers in this field is to furnish comprehensive explanations and interpretations for the decision-making processes employed by AI systems (Gohel et al., 2021). Nevertheless, it is essential to note that examining how these AI systems function in real-world contexts and assessing their alignment with the intended purposes under expert supervision is another crucial perspective that merits significant attention by researchers and practitioners in the field (Orsoni et al., 2023). These

challenges continue yet learning and working online has sustained societies during a pandemic, overcoming time and space limits and barriers. Artificial Intelligence systems will continue to have a tremendous impact on how we address major challenges, as well as how we live our daily lives and learn, changing our behavior (Gillath et al., 2021). Thus, schools need to provide an appropriate education in a ubiquitously digitalized world and within an accordingly complex and changing career landscape. Some research has highlighted that the worker of the future (student of today) is expected to develop critical thinking, problem-solving, communication, and teamwork since these qualities have significant impacts on the development of innovation (and the use of AI systems) (Chen et al., 2020; Göksel & Bozkurt, 2019). Hence, current, and future generation of workers need to be prepared for the functional use of emerging technologies (i.e., a use that sustains personal and social development, but also the development of knowledge and skills), preventing the risks of the dysfunctional one (i.e., a use that doesn't sustain human development and could also determine problems in many aspects of human life (https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai). Using and reflecting on AI in schools, often subsumed as "digitalization of education", is neither systematically addressed in the European educational context, nor is it subject of standardized let alone technology-enhanced, automatized assessment, which would provide instant feedback to stakeholders such as (head) teachers, parents, school boards, and policymakers.

An important reflection follows from the ethical point of view about what behaviors can or cannot develop such a system (e.g., schools), especially now in which those behaviors have an impact on individuals (Langer & Landers, 2021). For these reasons, the European Union has proposed guidelines and ethics to guide the interaction between humans and the AI system. The goal is to ensure that people develop trust towards this technology and can use it feeling safe, including in school contexts. To all that has been said so far, we must add the robot side which, currently, is the ideal match for artificial intelligence. Considering the perspective of

Developmental Robotics, if we want to create artificial intelligence systems that expand following the same dynamics and phases of human development, it is necessary to equip them with a body side that allows them to build knowledge based on environmental interaction. Without the physical-social environment with which to interact, it would be impossible to hypothesize that artificial intelligence could follow, in its development, the dynamics of the human one. For example, an interesting aspect to consider is the use of AI systems, applied to robotics, to create robots sustaining human development in knowledge and skills. Interacting with humans in different periods of development, AI robots could adapt their interactive behavior to act in the human zone of proximal development. This Vygotskian concept defines humans' development potential when they operate with more experienced partners than alone. Studies building on socio-cognitive conflict (Benvenuti & Mazzoni, 2020; Mazzoni & Benvenuti, 2015) have also highlighted the importance of interactions and, particularly, the relevance of sharing different points of view and negotiating them to join more advanced solutions in complex tasks. These studies, together with those conducted in the field of divergent thinking, social creativity, and networked flow dynamics, advanced a perspective of robot/AI systems that evolve in a way that could sustain human cognitive development, improving the human knowledge and skills in the same way, or in a better way, than a human partner could do.

In the workplace, robots can prevent humans from many heavy and tiring activities, safeguarding their physical and mental health. There are currently many experiences with promising results in which robots are used for the education of children, but there is a lack of a shared perspective and plan on what skills should be developed in the school environment to cope with and use AI in educational contexts. This perspective and reflective paper brings together different views and concepts of developmental and educational psychology (starting from a literature review) but also explores more technical fields to offer a perspective on the

lines of research that could be taken to offer tools to teachers and students, to prepare them for the challenges of the future (and for the future labor market).

2.1.3 State of the art

In response to the pandemic emergency, Information and Communication Technologies (ICT) have highlighted their potential in many fields, particularly in educational contexts. On the one hand, ICT- enabled distance learning and classes were carried on without interruption; on the other, the isolation of pupils, particularly of adolescents, was undoubtedly a negative influence on the ICT-enhanced educational context. The lack of social interactions and motivation leads to feelings of loneliness and dejection. Additionally, it strongly limited the ability to learn in a social context. This indicates a clear need to exploit novel technologies to promote a way of learning that is grounded in interactions and sociality.

From a piagetian constructivist perspective (Ackermann, 2001; Piaget, 1962), the process of understanding the world is the result of the relationship established between a thinking and acting subject and the object of his own experience. In addition, Papert (1980, 1993) underlined the importance of technological artifacts in learning, not as supporting this process but as in simulating reality. From Papert's point of view, knowledge cannot simply be transmitted as it is from one person to another, but each subject reconstructs information in a personal and original way. According to this, the use of technological devices (e.g., computers, tablets, and robots) represents an effective method for building knowledge, allowing students to apply theoretical knowledge to practice. Even more, the use of a physical artifact (e.g., a robot tutor) determines an effective learning process as it makes students reflect on the knowledge they possess and how to apply it to the reality on which they are acting (Mubin et al., 2013). In his researches and works Papert highlights how the use of robotics kits, far from transmitting computer skills, generates curiosity and stimulates creativity and motivation to learn, allowing

one to build and enter in touch with powerful new ideas (Papert, 1980, 1993). Moreover, following the idea that learning is an active process based on experience and that social interactions can facilitate it, learners might make understanding more effective by working together. This means that technological innovation in education should be able to expand teachers/learners' opportunities for collaborative interaction and let them explore new strategies for teaching/learning (Braun et al., 2020). Moreover, schools need to provide an appropriate education in a ubiquitously digitalized world within complex and changing training needs and career landscapes.

The actual digital transformation is deeply changing most human sectors and the importance of transversal knowledge, skills, and competencies training is growing both in the labor market and as essential abilities for participating in European society. It has been highlighted that the citizens of the future are expected to develop critical thinking, problem-solving, communication, and teamwork since these qualities have a significant impact on the development of innovation (Fadel & Trilling, 2009). Communication, cooperation, and problem-solving are, almost by definition, the future skills demanded. Together with ICT literacy, content creation abilities and safety constitute the so-called 21st-century skills (Ferrari et al., 2012). Novel technological tools are key for the construction of 21st-century skills, but how can they develop uniformly for all students in educational contexts? This can be better understood within the Activity Theory approach applied to an education system (Batiibwe, 2019; Engeström, 2014; P. Zhang & Bai, 2005), in which emerging technologies mediate the relationship between the actors and the knowledge construction. A strong tenet of Activity Theory is that cognitive development and learning happen first at the social level, thanks to dynamics such as interaction, points of view sharing, socio-cognitive conflict dynamics, and negotiation, and then, it is interiorized by individuals (Fig. 2-1-1) (see Fig. 2-2-2). Contradictions (e.g., the use of digital technologies and distance education during the pandemic

situation) are the motor of change, in as much as needs go beyond the solutions adopted to date and promote the so-called "learning by expanding" (Engeström, 2015) based on Piagetian processes of assimilation and accommodation, to find a new balance in the system (e.g., schools). In this panorama, European Union addressed new strategy for high quality, inclusive, and future-oriented education, aiming to "contributing to the development of quality education by encouraging cooperation between Member States and, if necessary, by supporting and supplementing their actions (Treaty of the Functioning of the European Union Article 165)" (https://ec.europa.eu/commission/presscorner/detail/it/MEMO_17_1402). Despite this, the use of emerging technologies in schools and educational context, often subsumed as "digitalization of education", is not "equally addressed" in Europe, as deepen outline in PISA-OECD data (https://www.oecd.org/pisa/publications/). In this regard, building on promising approaches to learning analytics, progress in this area is bound to the definition of recommendations and methodological approaches that will guide teachers to develop didactic and educational activities based on technological tools (e.g., educational robotics, CT platforms, etc.) and support the schools' journey towards digital readiness.

For all these reasons, following the recommendation of the European Commission's Digital Education Action Plan (2021–2027) (https://education.ec.europa.eu/focus-topics/digital-education/action-plan), this paper supports the fostering of the development of a high-performing digital education ecosystem, and encourages teachers in promoting 21st-century skills through digitalization (e.g. the use of technologies and robot tutor) during their didactic activities, particularly proposing ideas that can favor the development of those skills that were particularly addressed as fundamental: Creativity, Critical Thinking, Problem Solving, and Computational Thinking. Two principal questions guide this perspective paper: a) how creativity, critical thinking, and problem solving can be effective in promoting computational

thinking, and b) how Artificial Intelligence can be an aid instrument to teachers to foster creativity, critical thinking, and problem-solving in schools and educational contexts.

2.1.4 Methods

To better understand how to start building shared tools to develop the skills described above, starting from the EU indications, we addressed a review of the existing literature.

This work was arranged using the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) protocol (Shamseer et al., 2015). We pursued a systematic literature search across three academic databases (PsycINFO, Scopus, and WOS) searching for keywords 'Artificial Intelligence AND Problem Solving OR Critical Thinking OR Creativity OR Computational Thinking AND (Education OR School OR Learning OR Teaching OR Classroom OR Education system). During the revision process, we filtered only articles, reviews, systematic reviews, and meta-analyses published in English in the last five years (2018–2023), we excluded papers published before 2018, books, chapters book, commentary, keynote presentations, panel discussions, dissertations, work-in-progress articles and works that were not conducted within the context of learning, and education.

The revision has been conducted by using Rayyan software (Ouzzani et al., 2016). Additional records have been found by using the software Connected Papers (Tarnavsky et al., 2020).

95 studies have been established as eligible for further investigation. 917 have been evaluated as duplicates, and then excluded in the next steps. 822 articles were excluded after title screening. The remaining 83 were processed for abstract and full text evaluation. After that, only 20 articles were considered relevant. Moreover, we carried out a bibliographic investigation from some recent meta-analysis and perspective articles by using the Software Connected Papers (Tarnavsky et al., 2020), then, we mainly focused on those articles found aiming to suggest educational model indication for developing Critical Thinking, Problem

Solving and Creativity using AI and Computational Thinking (Alam, 2022; Bocconi et al., 2022; Chassignol et al., 2018; van Laar et al., 2019).

The development of these skills is also an important issue of the Digital Education Action Plan 2021–2027 of the European Commission, where quality Computing Education is addressed as a key element under the priority "Enhancing digital skills and competencies for the digital transformation". Relevant to this work is digital competence, which concerns the responsible use of digital technologies for learning, at work, and participation in society. It consists of eight points: information and data literacy, communication and collaboration, digital content creation (including programming), safety (digital well-being and competencies related to cybersecurity), problem-solving and Critical Thinking. In this vision, the skills acquired in one domain could support competencies developed in another. This is the case with the skills related to Critical Thinking, and Creativity, which are embedded throughout the key competencies (Mubin et al., 2013). From this perspective, the necessity to introduce Computer Science (CS) practices, particularly Computational Thinking (CT), coding, and programming already in compulsory education has arisen. Nowadays Critical Thinking, Creativity, collaboration, communication, and CT are the core skills that must be learned by students (DigiComp Framework, s.d.). This would meet the needs of growing young people that could be creators and not just consumers of technology (Papert, 1993). However, activities that include AI systems that could help teachers and educators to develop didactical activity in schools and educational contexts is still not complete in the literature (and not uniform). In this regard, the following sections (Results of the review) will try to answer to the two review questions, giving a more extensive overview of different activities on how Creativity, Critical Thinking, Problem Solving, could be foster developed, and implemented with Computational Thinking and AI (with e.g., robot tutor) in educational and didactic teaching. First, the paper will discuss about Computational Thinking, Programming and Coding in Schools' Curriculum. Second, the

connection between Creativity, Computational Thinking and Programming and how to foster it by means of AI, will be analyzed. Finally, will be the turn of Critical Thinking, Problem solving, and their connection to Computational Thinking and Programming, and how to foster them by using AI.

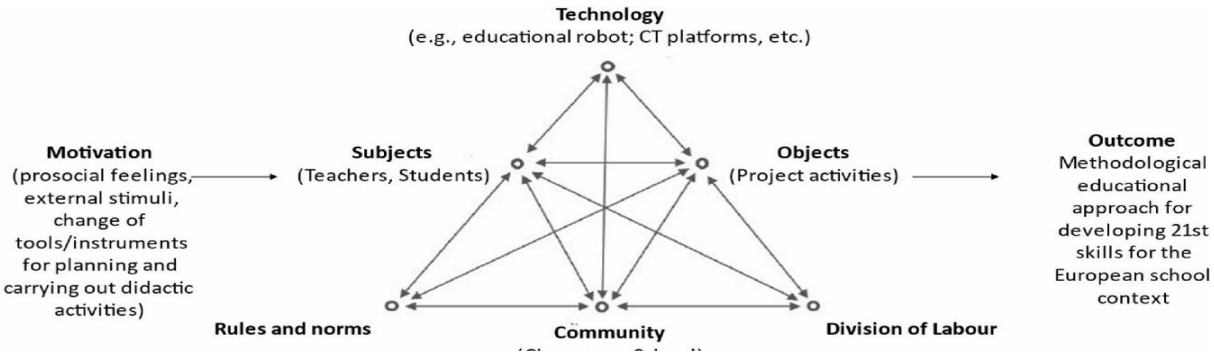

Figure 2-1-1 Activity theory applied to educational context.

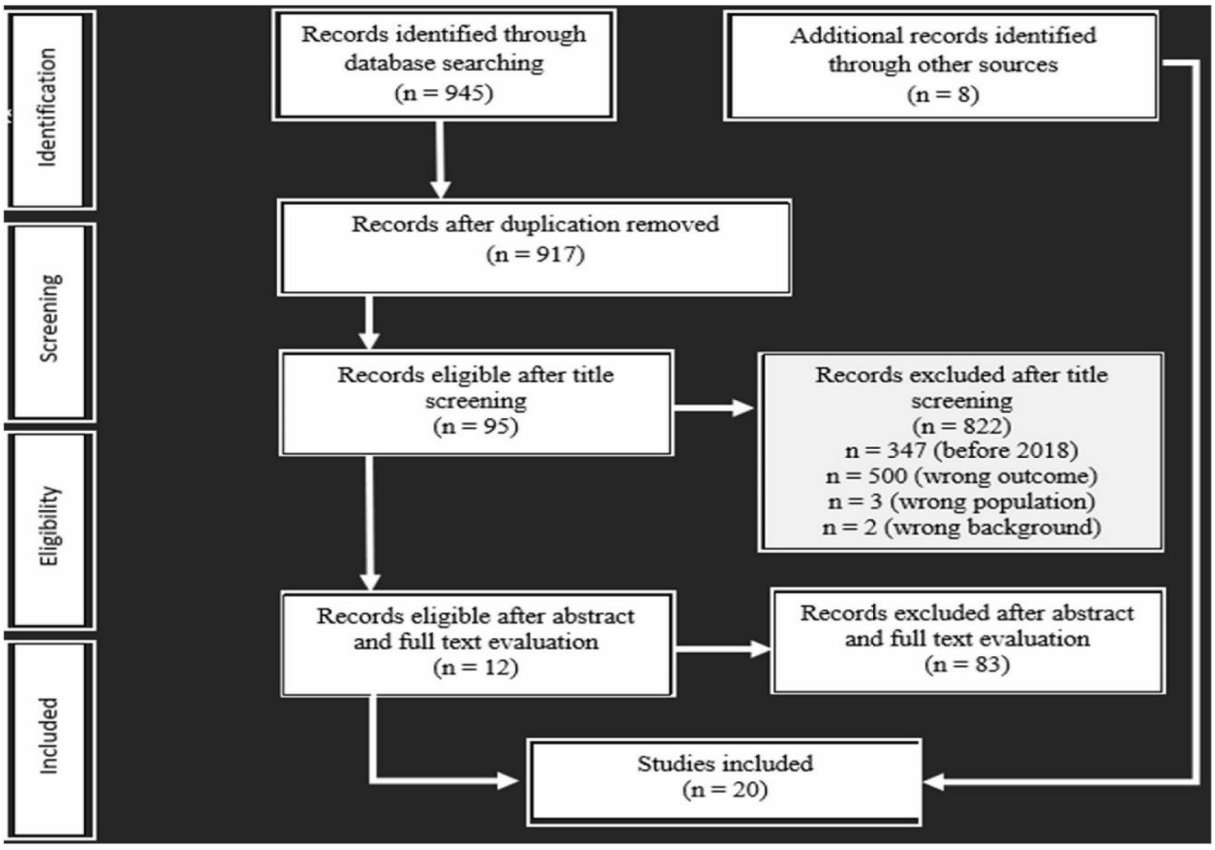

Figure 2-1-2 Summarizes the PRISMA flowchart of the present study process.

2.1.5 Results of the review

2.1.5.1 Computational thinking, programming, coding in schools' curriculum

Computational Thinking (CT) has been defined in several works by Wing (2006, 2010, 2017) and nowadays Wing's definition is considered the reference point in the discussion on CT. To Wing, "Computational thinking is the thought processes involved in formulating a problem and expressing its solution(s) in such a way that a computer – human or machine – can effectively carry out" (Wing, 2017). Then, CT is a set of concepts and skills involving abstraction, algorithmic thinking, automation, decomposition, debugging, and generalization (Bocconi et al., 2016, 2022). These skills are suitable in compulsory education, allowing students to move beyond operable and technical skills, creating problem solvers than just beneficiaries of the technology, developing creativity, and problem-solving capabilities (Yadav et al., 2014). Moreover, CT allows approaching problem-solving in a manner that results in solutions that can be reusable in different contexts (Shute et al., 2017).

One of the constituents of CT is programming (Bocconi et al., 2016, 2022). We could define programming as the activity of analyzing a problem, designing a solution, and implementing it. This has been indicated by DigComp, the European framework for digital competencies, as one of the constituents for EU citizens (Ferrari et al., 2013). Differently, coding is the step of implementing solutions in a particular programming language. According to Bers and colleagues (2019) , "coding is a playground", a new literacy for the 21st century, and a new language for children. Through coding, children can learn to code via fun, play, and creativity (Bers et al., 2019). Literature suggests programming as an efficient framework for fostering CT skills (Angeli & Giannakos, 2020; L. Sun et al., 2022).

In recent years, computational thinking and programming/coding are a reality of compulsory education in different EU countries (Bers et al., 2019; Bocconi et al., 2016, 2022). The United

Kingdom, in 2013, has incorporated computer science in the early years of its school curriculum (U.K. Department for Education, 2013). Moreover, an interesting report promoted and funded by the Nordic@BETT2018 Steering Group (Bocconi et al., 2018), e.g., shows that in Finland, Sweden, Denmark, and Norway, CT and programming are already included in the primary and secondary schools' curricula (but not all over Europe, specifically in the south), sometimes as transversal competencies and within existing subject matter (e.g., in Finland and Sweden) or as a new (elective) subject (e.g., in Denmark and Norway). A deep reading of the report highlights the relevance of two key transversal competencies to foster computational thinking and programming in compulsory schools: critical thinking and creativity (Bocconi et al., 2018).

2.1.5.2 Creativity and its connection to computational thinking and programming and how to foster it by using AI.

Creativity consists of a core skill for promoting personal growth (Papadakis & Kalogiannakis, 2018) and is embedded throughout the key competencies for lifelong learning (Mazzoni et al., 2022). However, there is still debate about what is creativity. Over the years, researchers developed different conceptualizations and definitions of this term, even if it is possible to find a certain consensus in the simplest definition of creativity. Kaufman and Glaveanu (2019), refer to creativity as something both new and task appropriate. In addition, it is possible to focus on three mental operations that underlie creativity (Antonietti & Molteni, 2014). The first one is related to broadening the mental field, linked to the subject's ability to conceive unique and different ideas e.g. divergent thinking concept (Guilford, 1950), to generate solutions of which at least one survives the judgment (Johnson-Laird, 1998), or the subject's capability of holding a mental wealth of information able to enhance the probability to find elements related to each other for creating something new. In the second mental operation, creativity allows connecting usually conceived antithetical and distant mental fields (Rothenberg, 1979). Lastly, about the third mental operation, a creative act is present when there is a reorganization of the mental

field. Only in recent years, creativity has been embraced as a relevant element in computer science for its importance in supplying motivation and interest in the field, but also in improving performance and knowledge acquisition (Israel-Fishelson & Hershkovitz, 2022). Although the literature suggests a bidirectional link between creativity, computer science, and CT, in this work we mainly focus on how creativity can influence CT. Moreover, we present a possible perspective in which AI has been implemented to improve the creativity of participants.

Israel-Fishelson and Hershkovitz (2022), highlighted as creativity may facilitate the resolution of algorithmic problems, the development of computational products, and new knowledge. Liu and Lu (2002) found how standardized creativity tests allow for prediction creativity in solving programming problems among undergraduate students. Similar results have been found in the work of Perez-Poch and colleagues (2016). The authors found a significant positive correlation between the levels of creativity and programming skills among engineering students. In detail, a high level of creativity predicted achieving excellence in programming. These results have been corroborated by Hodges and colleagues (2013). The idea behind this work was to improve CT by fostering creative thinking. Creative thinking is personalized thinking leading to creative results (Hodges et al., 2013). They found that the implementation of creative thinking exercises in CS courses improved computational knowledge and skills (Miller et al., 2013).

In general, these pieces of evidence suggest a relevant role of creativity and give value to its integration into compulsory education that would foster CT and programming skills. Thus, it is important to explore whether and how Artificial Intelligence can be a valuable tool in fostering human creativity, and consequently CT. The idea of AI helping humans to achieve better creative performances is undoubtedly fascinating. The branch of Computer Science that deals with this aspect is called computational creativity. Wingström and colleagues (2022), observed how nowadays computational creativity focuses on two lines of research. The first explores the

capabilities of AI algorithms to recreate human-level creativity while the second is merging the creativity of humans and AI in a reciprocal course. Concerning co-creativity, Maher and colleagues (2013), suggested three roles of computers: 1) as supporters of the human creative process by giving tools and procedures; 2) as enhancers of human creative ability by providing knowledge and promising creative cognition; 3) as generators, by offering to the user, creative elements to interpret, evaluate and integrate as creative products.

Unfortunately, the approach to co-creativity is young and most of the co-creative AI is in the arts domains (Wingström et al., 2022). According to this line, one recent work (Rong and colleagues (2022), explored how fine art training based on Virtual Reality and Artificial Intelligence can enhance the creativity and concentration of middle school students. The study was done by comparing the students' creativity, distraction, and anxiety levels before and after AI and VR course training. The results showed significant improvement in creativity levels (assessed with the "Creative Thinking Test for Middle School Students"), and significantly reduced distraction and anxiety levels. The authors claim that the training proposed can adequately improve students' creativity and concentration, and at the same time, reduce students' test anxiety.

Another work by (Liapis and colleagues (2016), presented a computational approach by using mixed-initiative tools aiming to support and foster human creativity by improving lateral thinking with educational activities. In this work, four mixed-initiative tools or games were presented. The goal of fostering lateral thinking was carried out by the computer supported by AI that proactively contributes to the design process by creating suggestions for the human user to consider. In this perspective, human and computers do affect each other; the action done by the computer reformulate the human's mental associations, but also the action taken by the human constrains the search space of the algorithm, enabling it to focus on specific possible solutions to a problem (Liapis et al., 2016). Authors suggested in their results how this co-

creative approach was able to foster human creativity by improving the lateral thinking of humans. Unfortunately, the work considers only qualitative and observational data limiting the generalizability of results.

2.1.5.3 Critical thinking, problem solving, their connection to computational thinking and programming, and how to foster them by using AI.

The other transversal competence considered is critical thinking. Critical thinking is merely the ability to think critically and is a key to individual civic engagement and economic success (Willingham, 2019). As for creativity, there is no general definition for critical thinking, but researchers highlighted some agreement about the characteristics inherent to it, like analysis and synthesis, making judgments, decision-making, drawing warranted conclusions, and generalizations (Buckley, 2012). According to Fagin and colleagues (2006), three are the key parts of critical thinking: clarity (the ability to understand the information received), accuracy (the ability to investigate the distance between the information and factual reality), and relevance (the ability to evaluate if the information received is pertinent). The suggestion is that critical thinking might be considered a prerequisite to problem-solving (Buckley, 2012). Even if the literature did not deeply explore yet the relationship between critical thinking and CT, an interesting work of Buckley (2012), pointed to a connection between these two forms of thinking. The idea formulated, focused on perceiving a problem as an obstacle. The author claimed that to overcome the obstacle was possible to apply a linear problem-solving strategy or a 3-D problem-solving model. Both models consider critical thinking as a non-algorithmic higher order of thinking that directly affects knowledge acquisition. Then, critical thinking becomes a prerequisite for knowledge acquisition. By using critical thinking, the subject becomes aware of the problem and then the information is extrapolated and critically analyzed. Starting from the relevant knowledge extracted in this way, it is possible to apply CT and then solve the problem.

How Artificial Intelligence could be a valuable tool in fostering human critical thinking, and problem-solving? Critical thinking and problem-solving, are the key element in the decision-making process and all these three elements are interconnected to achieve the best solution given a problem (Özgenel, 2018). Starting from this point, a recent line of research focused on the use of AI and metacognition in the learning process to enhance students' problem-solving capabilities. Metacognition is the ability to think about one's cognition (Cortese, 2022). According to Molin and colleagues (2020), students with a higher level of metacognitive skills are mainly prone to self-regulated learning, which is an approach linked to learning where students set their goals, and track, regulate and control actions, cognition, and motivation to achieve these goals. Confidence is the measure by which metacognition is measured in the field of psychology and neuroscience (Cortese, 2022). According to Cortese (2022) is possible to bring together the aspect of confidence with the mathematical formalism of Reinforcement Learning that fits well with the question of how to explain learning and how confidence can affect learning and vice-versa. The focus on metacognition as an element to enhance students' problem-solving and decision-making capabilities, and how AI can be beneficial for this purpose, has been investigated in a recent work by Callaway and colleagues (2022). The objective of this work was to improve the planning strategies of students facing different problems. By adopting the Mouselab-MDP paradigm (Callaway et al., 2017), the authors developed an intelligent cognitive tutor that employs metacognitive feedback to teach planning. The idea of metacognitive feedback is to give people feedback on how instead of what they decide to do. The authors based on the theory of metacognitive reinforcement learning developed a system able to discover the optimal cognitive strategies and accelerate metacognitive learning in people by suggesting optimal feedback signals. The presented approach was validated by the authors in six different experiments. The results showed how practicing with this system allowed people to be more effective than traditional methods. In

more detail, the group that used the metacognitive feedback showed significantly better results than the other groups (feedback related to action and no feedback). In addition, by applying this method the authors found how improvements were also transferred in new situations, and retained over time (Callaway et al., 2022).

2.1.6 Conclusion

To be a citizen of the 21st century requires one to master different skills and competencies to be an effective worker, for personal realization and development. The school and the teachers are the key elements to educating students in this transformation process where computational thinking, critical thinking, problem-solving, creativity, and the remaining skills are taking a leading role in this even more digitalized world. With this paper we try to propose a perspective on how creativity, critical thinking, and problem-solving can be effective in promoting computational thinking, and how Artificial Intelligence can be an aid instrument to teachers in the fostering of creativity, critical thinking, and problem-solving in schools and educational contexts.

Literature suggests how AI is used in education with different applications like chatbots, intelligent tutoring, automated grading systems, and recommended systems, but its application in the field is still limited compared to others, like medicine and business (Celik et al., 2022).

This aspect is also reflected in this study, where very few articles have been considered eligible for the aim of the article itself. One possible reason has been presented in the work of Celik and colleagues (2022), where there was evidence of the resistance of decision-makers such as teachers, educators, and traditional textbook publishers to the use of AI, but also the knowledge of stakeholders, including students, about AI plays a relevant role in its application.

According to this line, Marrone and colleagues (2022) investigated how students perceived AI in fostering creativity in the school context. They found four key factors describing the

relationship between AI and creativity: social, affective, technological, and learning. Concerning the affective one, the authors observed an effective response in students based on their degree of familiarity with AI; students who were more familiar with AI concepts or applications reported being more comfortable in using AI technologies compared to the students who were not.

2.1.6.1 Future directions

Considering the precious aspects, to be able to implement new technologies, as a driving force for change in teaching activities, it is necessary to keep in mind that the school is a cornerstone for promoting the skills of the 21st century. Even if the dissemination of these technologies and activities in school curricula in Europe is not uniform, it is necessary to continue to disseminate (also through scientific research in this field) the dissemination of techniques that teachers and educators can use. In this regard, our invitation is to follow the indications of the Digital Education Action Plan (2021–2027), however enhancing collaboration between schools (e.g., using eTwinning (https://school-education.ec.europa.eu/en/etwinning)) throughout Europe in order to reduce the existing gap in the development of skills that currently exists between north and south (Bocconi et al., 2022). Additionally, evidence has suggested few studies have implemented AI as a method to help students and individuals foster creativity and problem solving (e.g. Alam, 2022; Callaway et al., 2022; Chen et al., 2020). This depends a lot on finding resources and on the skills that teachers have in being able to use such technological tools. In this regard, it would be necessary to promote lifelong learning, with a view to a lifelong learning program (https://lllplatform.eu/) also for teachers (https://school-education.ec.europa.eu/en/about/etwinning-future-teachers). Always taking advantage of the networks of connections existing throughout Europe. Teaching and awareness of what AI can and cannot do as a tool is a key step in making it more familiar in the educational context. A tool you can rely on.

2.1.6.2 Limits

One of the most important limits of this paper is that it doesn't consider the ethical aspects (considering also GDPR's data protection) of the use of AI in many fields. Indeed, one of the most relevant paper's aims are primarily focused on the functional use of AI in fostering the so- called soft skills or life skills, without forgetting the dysfunctional or critical effects of its use (although not central). Thus, future studies, more focused on ethical effects of the use of AI to develop and foster soft skills, should deepen the critical aspects related, e.g., to data protection, data collection, and awareness to interact with non-human agents.

2.2

Learning landscape in gamification: the need for a methodological protocol in research applications.

Orsoni, M., Dubé, A., Prandi, C., Giovagnoli, S., Benassi, M., Mazzoni, E., & Benvenuti, M. (2023). Learning Landscape in Gamification: The Need for a Methodological Protocol in Research Applications. Perspectives on Psychological Science, 0(0). https://doi.org/10.1177/17456916231202489

2.2 Learning landscape in gamification: the need for a methodological protocol in research applications.

2.2.1 Abstract

In education, the term “gamification” refers to of the use of game-design elements and gaming experiences in the learning processes to enhance learners’ motivation and engagement. Despite researchers’ efforts to evaluate the impact of gamification in educational settings, several methodological drawbacks are still present. Indeed, the number of studies with high methodological rigor is reduced and, consequently, so are the reliability of results. In this work, we identified the key concepts explaining the methodological issues in the use of gamification in learning and education, and we exploited the controverses identified in the extant literature. Our final goal was to set up a checklist protocol that will facilitate the design of more rigorous studies in the gamified-learning framework. The checklist suggests potential moderators explaining the link between gamification, learning, and education identified by recent reviews, systematic reviews, and meta-analyses: study design, theory foundations, personalization, motivation and engagement, game elements, game design, and learning outcomes.

2.2.2 Introduction

Educational games were the second most studied educational technology of the last decade: The amount of articles on educational games grew 255%, and the amount on gamification grew an astounding 2,687% (Dubé & Wen, 2021). This research spans a vast range of fields and is not specific to any one educational context. According to Landers (2014), both serious games and gamification have as purposes the improvement of learning outcomes, but the processes involved to achieve such gains are quite different. In the serious games field, games are designed to affect learning directly. In other words, the instructional content and activities within the serious game are tantamount to learning activities (Landers, 2014). In gamification, game elements are designed to influence learning indirectly by acting on learner behaviors or attitudes (e.g., participants' engagement and motivation), which improves learning as a result (Landers, 2014). In this work, we focus on gamification without focusing on serious games. Deterding and colleagues (2011) defined gamification as "the use of game design elements in non-game contexts." Following this definition, the game elements could affect the context experience by increasing the motivation and by augmenting the engagement. Likewise, Kapp (2012), Hamari and colleagues (2014), and Werbach (2014) defined the term gamification as "the process of making activities more game-like." According to Dichev and Dicheva (2017), the specific use of gamification in education refers to the inclusion of gaming elements in the design of learning processes. Indeed, as reported in Zainuddin and colleagues' (2020) review, including 46 empirical studies, three were the most relevant positive applications of gamified learning: learning achievement, motivation and engagement, and interaction and social connection. Despite the excitement for the positive outcomes in the application of gamified elements in learning and educational contexts, most of the works have tended to have inconclusive results (Bai et al., 2020; Huang et al., 2020; Sailer & Homner, 2020). This point is expanded on below. It is possible to recognize at least two types of constraints concerning

the use of gamification in learning and education studies: methodological and specific constraints. Methodological constraints include the aspects related to methodological issues that have been emphasized in literature over time, whereas specific constraints pertain to the key aspects discussed in the literature on gamification. According to methodological aspects, former literature has stressed a lack of understanding of which education level should be incorporated for optimal benefits (De Sousa Borges et al., 2014), varying impacts on student engagement depending on intrinsic or extrinsic motivation (Faiella & Ricciardi, 2015; J. Xu et al., 2021), insufficient empirical data and lack of comparative- and longitudinal-study designs, underdeveloped theoretical foundations and conceptual ambiguity (Seaborn & Fels, 2015), small sample sizes, a lack of experimental design, an absence of explicit motivation measurements, and a lack of using validated psychometric instruments (Antonaci et al., 2019; Ortiz et al., 2016; Sailer & Homner, 2020). Moreover, many studies lacked an experimental design that included both control and experimental groups (Alomari et al., 2019; Ofosu-Ampong, 2020; Ortiz et al., 2016). Indeed, research on gamification is limited and often lacks controlled experimental designs; few studies have examined the effects of individual gamification elements in a controlled manner (Bozkurt & Durak, 2018; Hung, 2017; Majuri et al., 2018). Dichev and Dicheva (2017) highlighted that studies generally focus on game performance as a measure of the effect of gamification without measuring educational outcomes. Usually, the focus is only on short-term outcomes, simplifying the phenomenon and failing to take into account contextual factors and individual differences, with limited exploration of game-design practices and ethical considerations related to long-term impacts and personal data (Rapp et al., 2019; Zainuddin et al., 2020). Metwally and colleagues (2021), Nair and Mathew (2021), Behl and colleagues (2022), Nadi-Ravandi and Batooli (2022), and Saleem and colleagues (2022) identified several challenges that need to be addressed. These included a lack of understanding of gamification techniques and instructional theories, a debate

about the use of point-badge-leaderboard (PBL) elements, potentially negative effects on intrinsic motivation, and unclear impacts on learning or knowledge levels. In addition, the authors highlighted that the lack of methodologically correct experimental designs, the lack of solid theoretical basis in many studies, and multiple technological difficulties could hinder the effective implementation of gamification in educational activities. Regarding specific constraints, we found that personalization has been considered by different studies (Aljabali & Ahmad, 2019; Denden et al., 2022). Aljabali and Ahmad (2019) noted that there is a lack of understanding of how to design game mechanics that promote desired outcomes and cater to individual learner characteristics. Most studies treat gamification as a generic construct and fail to investigate the impact of personalized gamification on learning outcomes (Denden et al., 2022). In addition, there is a tendency to adopt a one size-fits-all approach, and the literature is fragmented, including insufficient descriptive statistics for meta-analysis (Oliveira et al., 2022). Other limitations concern the game design in gamification environments. In general, it is suggested that there is a need for more personalization and integration of motivational and instructional design in gamification. Facey-Shaw and colleagues (2017) emphasized difficulty in comparing the effectiveness of badge designs because of their variety. Lack of formal design support and frameworks for many gamification experiences makes it difficult to apply procedures and features of case studies in different contexts (Laine & Lindberg, 2020; Mora et al., 2017). A very recent work of Khaldi and colleagues (2023) noted that on 39 articles investigated, a significant portion of applied gamification research is not rooted in theoretical frameworks and does not employ them in the design of gamified learning systems. Although some experimental studies have endeavored to adapt psychological and educational theories from the literature as gamification approaches, the resulting systems lack clarity. In general, despite the lack of a comprehensive theory of gamification in education, many theories from social, cognitive, and educational psychology are used to identify how gamification enhances

motivation, engagement, and learning. The most widely adopted theory is self-determination theory (SDT; Ryan & Deci, 2000), and flow theory (Csikszentmihalyi, 1990) is also relevant for active engagement and learning. The only one specifically developed for gamified learning is Landers's (Landers, 2014) theory of gamified learning. Other theories from developmental and educational psychology and social psychology can also be relevant, but some gamification research lacks a theoretical framework (Krath et al., 2021). Finally, some studies indicated as a critical aspect the limited number of respondents involved in studies, inconsistent findings on the effect of gamification on academic achievement, and different effect sizes found in previous meta-analyses, suggesting that the effectiveness may depend on external and internal factors such as gamification designs, pedagogical contexts, learners' frustration, and distraction (Dikmen, 2021; Ortiz-Rojas et al., 2017; Q. Zhang & Yu, 2022). In this study, we aim to synthesize existing literature on gamification in learning and education and propose a checklist protocol based on recent evidence to facilitate design, production of more rigorous studies, ability to have more reliable results, and enhancement of the quality evaluation of gamification studies in education. This is in response to the recent need for a validated checklist to assess the quality of future research in gamification, as suggested by Metwally and colleagues (2021). The proposed checklist protocol is intended to focus on the most recent evidence and aligns with current needs in the field. This work has been structured into distinct sections, each of which covers a specific aspect of gamification. Beginning with the method, we then provide a comprehensive discussion on its efficacy in the context of learning. This discussion encompasses an analysis of the core elements that have been extracted from the qualitative analysis of the reviews, systematic reviews, and meta-analyses included in this work. Finally, these aspects are used to develop an informative checklist protocol that may serve as a useful resource for researchers and practitioners.

2.2.3 Method

This work was arranged using the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) protocol in its latest version (Page et al., 2021). We pursued a systematic literature search across four academic databases (ACM Digital Library, PubMed, WOS, and Scopus) searching for keywords "Gamification AND learning" and filtering for "Review," "Systematic Review," "Meta Analysis," "Literature Review," and "Systematic Literature Review"; year range was between 2011 and 2023. The decision to choose this range of years was made to ensure the inclusion of works related to gamification between Deterding et al.'s (2011) definition and the present day. In addition, Caponetto et al. (2014) and Ortiz and colleagues (2016) discovered through literature review that the term "gamification in education" did not appear in article titles until 2011. The inclusion criteria were that the articles must be written in English. Articles written in languages other than English were excluded. In addition, single papers, keynote presentations, panel discussions, dissertations, work-in-progress articles, and papers that focused on serious games, game-based learning, revisions, systematic revisions, or meta-analyses that were not conducted within the context of learning, education, or school were also excluded. The revision was conducted using Rayyan software (Ouzzani et al., 2016). Furthermore, to explore additional findings, a bibliographic investigation was conducted using recent meta-analyses as sources, specifically those authored by Bai et al. (2020), Sailer and Homner (2020) and Huang et al. (2020). The software Connected Papers (Tarnavsky et al., 2020) was employed for this purpose. In addition, the software Elicit (Ought, 2023), an artificial-intelligence-based tool using large language models such as GPT-3, was used to perform a literature review. The query posed was, "How effective is gamification at promoting learning?"; we included a filter for reviews, systematic reviews, and meta-analyses only. The flow diagram in Figure 1 summarizes the PRISMA process. The inspection through the databases resulted in a total of 1,257 works eligible for further investigation. One hundred

eighty articles were evaluated as duplicates and then excluded in the next steps. A total of 953 were screened for relevance. Nine hundred two articles were excluded after title screening. The remaining 51 were processed for abstract and full-text evaluation. After that, only 28 were considered relevant. The ground for exclusion is presented in the PRISMA flow diagram (Fig. 2-2-1). According to the other two methods used for identification of studies, 122 were processed, and 44 were selected to be eligible for the revision. A total number of three articles were not retrieved. A total of 72 articles were considered for the present work, and according to the findings, a checklist was developed. A table of findings of the articles is included in the Supplementary Materials (https://osf.io/6kbn2/?view_only=19213e2c0ccd4c93a41f1055298310b5). In the subsequent sections, we outline the development process of the checklist. The initial phase consisted of a qualitative analysis that included descriptive statistics regarding the included articles. Specifically, the frequencies of reviews, systematic reviews, and meta-analyses were recorded for the period between 2011 and 2023. In addition, the core elements or focal points of the articles were identified, and their distribution over the years was analyzed. Subsequently, on the basis of the identified core elements, we developed subsections to present the findings, limitations, and key elements that researchers should consider when developing a study on gamification in education and learning and then to create the checklist.

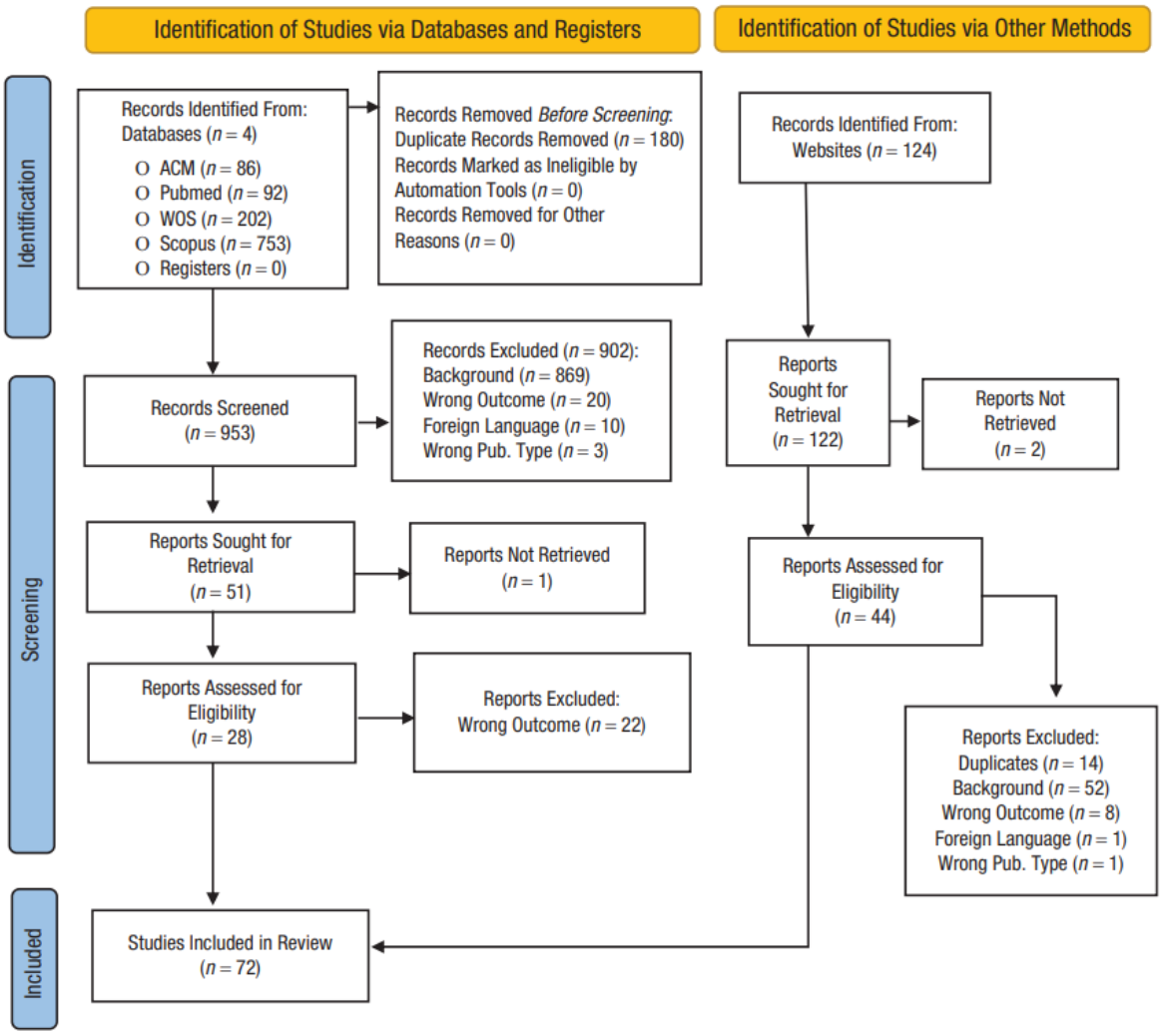

Figure 2-2-1 Preferred Reporting Items for Systematic Reviews and Meta-Analyses flow diagram of present study process.

2.2.3.1 Exploring Frequencies and Core Elements: Descriptive Analysis of Included Articles

Out of the total 72 articles included in this study, 36 were reviews (50%), four were critical reviews, 22 were systematic reviews (30.5%), 11 were meta-analyses (15%), one was a systematic deductive analysis (1.5%), one was a systematic mapping review (1.5%), and one was a systematic metareview (1.5%). Furthermore, the articles under investigation were categorized into core elements or focal points; some dealt with specific aspects of gamification in education and learning, whereas others were a combination of single elements. These core elements include game elements, game design, general aspects of gamification in learning and

education, learning outcomes, motivation, personalization (tailored gamification, adaptive gamification), theory, game elements and learning outcomes, game elements and motivation, and motivation and engagement. A concise summary of the fundamental components can be provided by categorizing them into seven fields, which were also employed for the inspection and checklist development. These fields encompass the following: the broad aspects of gamification (comprising the study design), theoretical foundations, personalization, motivation and engagement, game elements, game design, and learning outcomes. To better illustrate the distribution of these core elements over the years, a stacked chart (Fig. 2-2-2) was developed to highlight trends and tendencies. The year 2021 had the highest number of publications in the field of gamification in learning and education, with a total of 14 articles identified. Moreover, it is noteworthy that several publications dealing with the general aspects of gamification in learning and education have remained constant over the years. However, systematic works that account for the aspects of personalization have received greater attention in recent years, with a more pronounced focus since 2018. Likewise, systematic works covering learning outcomes have gained increased attention, with a rising trend since 2017. According to the type of publication and the core elements, the meta-analyses focused on learning outcomes (n = 8), general aspects of gamification in learning (n = 1), motivation (n = 1), and game elements and learning outcomes (n = 1). Most of the systematic reviews focused on investigating general aspects of gamification (n = 13), followed by personalization (n = 3), learning outcomes (n = 2), game elements (n = 2), game design (n = 1), theory (n = 1), motivation (n = 1), and game elements and intrinsic motivation (n = 1). Of the reviews, most of the studies focused on the general aspects of gamification (n = 14), followed by game elements (n = 7), game design (n = 4), personalization (n = 4), motivation (n = 1), motivation and engagement (n = 1), and game elements and motivation (n = 1). All the critical reviews focused on general aspects.

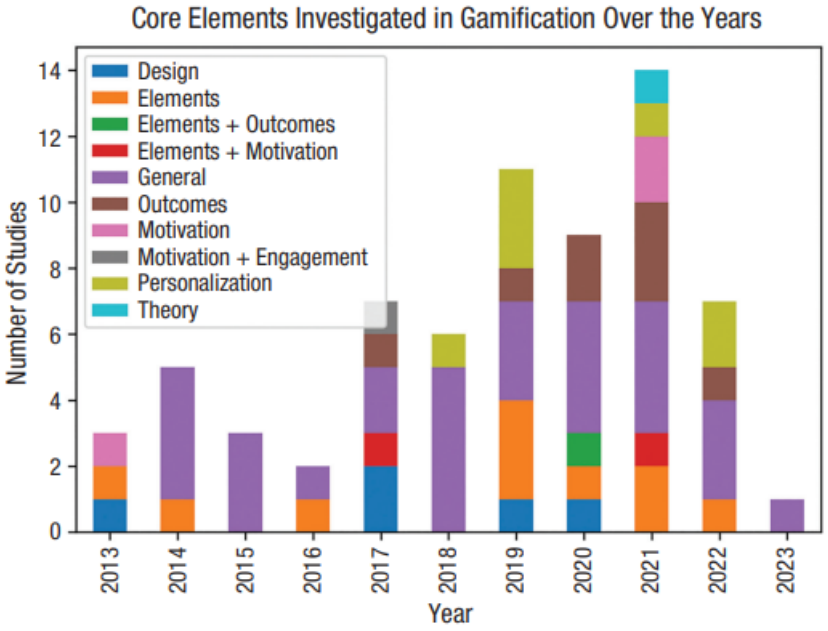

Figure 2-2-2 Distribution of the articles from 2013 to 2023 according to the core elements investigated.

.

2.2.4 Gamification of Learning: What We Found and What Should Be Addressed

2.2.4.1 Investigating the general aspects of gamification

In this section, we categorize and refer to the literature that covers the general aspects of gamification. These studies broadly investigated the effects or applications of gamification in learning and education without any specific focus on its core aspects. Nah and colleagues (2013) identified five principles that guide gamification in education. First, games should have multiple layers of goals to ensure goal orientation. Second, recognition of players' achievements enhances their motivation and engagement. Third, positive reinforcement through points or virtual currency can promote learning, whereas negative feedback can offer corrective information. Fourth, competition sustains engagement and focus on the learning task. Finally, a fun component or orientation is crucial for motivating and engaging learners in educational games. According to Wilson and colleagues (2015), a gamified system has three core elements: a user, a nongame task, and a set of game-design elements that motivates the

user to execute the task. We present the evidence about the effects of gamification over the years, the methodological concerns, and other aspects such as educational level, educational courses, and duration of interventions.

2.2.4.2 Effects of gamification in learning and education

Gamification has been studied with a focus on enhancing student engagement, motivation, and learning outcomes. Although studies have shown the positive effects of gamification, mixed results have been reported, depending on the implementation context. Motivation, engagement, self-efficacy, and flow/cognitive absorption are the most significant constructs in gamification research. In addition, gamification has been shown to improve learning achievement, social connection, creativity, and self-directed study. However, the effectiveness of gamification in promoting learning and participation is still debated in literature given that weaker statistical differences have been observed between gamified and nongamified environments. In more detail, De Sousa Borges et al. (2014) found that previous research on gamification in education has focused mainly on evaluating student engagement, whereas Caponetto et al. (2014) reported \that gamification techniques have also been used to develop attitudes and behaviors, such as collaboration, creativity, and selfdirected study. Several studies (Faiella & Ricciardi, 2015; Gerber, 2014; Hamari et al., 2014; Sanmugam et al., 2015; Surendeleg et al., 2014) have provided empirical evidence supporting the effectiveness of games in enhancing learning, engagement, and motivation. However, Seaborn and Fels (2015), Ortiz et al. (2016), and Dichev and Dicheva (2017) noted that the effectiveness of gamification varies depending on the implementation context, resulting in mixed results. According to Inocencio (2018), motivation, engagement, self-efficacy, and flow/cognitive absorption are the most significant constructs in gamification research because they have consistent theoretical frameworks and reliable scales. Although satisfaction and attitude are commonly used, their effectiveness is not as strong. Majuri et al. (2018) found a generally positive effect of gamification, although there

is also a substantial amount of research with mixed or null results. Indriasari et al. (2020) described gamification as having positive effects on student engagement, and Kalogiannakis et al. (2021) identified motivation and engagement, learning achievements, and social interaction as the most affected learning outcomes. Zainuddin et al. (2020) found that the positive themes that emerged from gamification studies included learning achievement, motivation and engagement, and interaction and social connection. Manzano-León et al. (2021) added positive effects on student academic performance at different educational levels, especially in university education, in which academic achievement is emphasized. Similar results were highlighted by Metwally et al. (2021), who found that gamification can enhance motivation and engagement in education, particularly through extrinsic rewards, such as achievement and progression, and improve various aspects of children's learning, including cognition, skills, socialemotional abilities, and attitudes. Nair and Mathew (2021) corroborated the notable positive effect on gamification on learning outcomes, learner motivation, and engagement. In most studies, gamification was found to have a significant impact on learning; 47 studies exhibited statistically significant outcomes in the dependent variable. These positive outcomes were substantiated by Saxena and Mishra (2021) and Devendren and Nasri (2022) in classroom settings. In a recent work, Nadi-Ravandi and Batooli (2022) tried to summarize evidence from a sociometric, content, and co-occurrence perspective for studies between 2000 and 2021. Authors reported how the application of gamification in education is still challenging because of inconclusive or contradictory results. In gamified education, motivation, learning, and engagement are the most important concepts. Benefits include increased learner competition, practical skills, and perceived learning. Increased participation can improve learning skills and academic achievement. Although educational interventions have been effective in promoting learning, motivation, and participation, most studies did not definitively

establish the effect of gamification, and weaker statistical differences between gamified and nongamified environments were observed.

2.2.4.3 Methodological issues and concerns

Over the years of research on gamification in learning and education, several concerns about methodological issues have been substantiated. These issues include small sample sizes, lack of validated measurements, unclear reporting, and absence of control groups. A need for more validation research to test innovative gamification techniques and methods is even more clear, as are established guidelines on how to effectively implement gamification in education. Despite efforts to develop more engaging and effective gamified systems, there is a lack of methodological rigor, and a common language is needed for research. However, with the rapidly evolving field, there is also a need for higher quality studies that include two groups with pretest and posttest measures.

De Sousa Borges et al. (2014), Devers and Gurung (2014), Faiella and Ricciardi (2015), and Ortiz et al. (2016) identified a need for more validation research to test innovative gamification techniques and methods. The authors reported studies often had methodological limitations, including small sample sizes, lack of validated psychometric measurements, absence of control groups, unclear reporting of results, short experiment time frames, and no multilevel measurement models. Dichev and Dicheva (2017) found inconclusive results in most of the studies investigated, largely because of methodological inadequacies. They suggested a lack of established guidelines on how to effectively implement gamification in education and an inadequacy about the existing high-quality evidence on its long-term benefits. Bozkurt and Durak (2018) reported that nearly half of the articles lack theoretical or conceptual frameworks. In addition, more recent articles have also raised concerns regarding methodological limitations in gamification research. Koivisto and Hamari (2019), Rapp et al. (2019), and Metwally et al. (2021) highlighted a lack of theoretical and methodological rigor (lacks control groups, clear

reporting, adequate sample sizes, and experimental time frames) despite efforts to develop more effective and engaging gamified systems. They emphasized the need for a common language, the use of a broader range of theories, and the use of rigorous scientific-validity-methods constructs in gamification research in learning and education. Nadi-Ravandi and Batooli (2022) suggested that those drawback elements (lacks well-controlled empirical studies or high-quality studies considering, e.g., two groups with pretest and posttest measures) and scarcity of methodological rigor are typical of areas of research still in development, which gamification in learning and education actually is. Unfortunately, this aspect results in a very low number of eligible studies to develop quantitative analysis compared with the overall published; the presence of inconsistent, contradictory results; and a focus on elements for which the effect is not reflected in the quantitative analysis.

2.2.4.4 Educational levels

Regarding the educational levels fostered by gamified research and applications, most of the research has been conducted in higher education, and very limited studies have focused on primary education. Empirical studies on gamification have predominantly been carried out in university settings with a primary focus on adult participants. Consequently, there is a lack of research on the use of gamification in K–12 education (Dichev & Dicheva, 2017; Metwally et al., 2021; So & Seo, 2018), which highlights the need for further exploration in this area.

De Sousa Borges et al. (2014) and Caponetto et al. (2014) noted that most of the gamification research in education has centered on higher education, mainly in the university setting, with few studies conducted in elementary education. Of 51 articles investigated, Dichev and Dicheva (2017) found that 44 were centered on the university level and that only seven were centered on K–12 education. Among the K–12 studies, three involved elementary school students, two focused on middle school students, and two examined high school students. Similar results were corroborated by Ortiz Rojas et al. (2017). Of 23 articles investigated, most research focused on

higher education (n = 19), followed by high school (n = 2) and middle school (n = 2). So and Seo (2018) identified significant research gaps in educational game research in Asian K–12 schools. Kocakoyun and Ozdamli (2018) and Zainuddin et al. (2020) found that most studies have concentrated on adult participants or higher education students. A slightly different finding was found by Huang et al. (2020). Most of the research on gamification in formal education has been carried out with undergraduate students (k = 13, n = 1,724), followed by K–12 students (k = 10, n = 920). Kalogiannakis et al. (2021) and Metwally et al. (2021) concluded that the focus on gamification research in K–12 education is limited, as suggested in the previous years, and confirmed that most studies involved students from higher or secondary education. Likewise, a meta-analysis by Dikmen (2021) between 2010 and 2020 in Turkey revealed that the studies analyzed were conducted across middle school, high school, and university levels. However, no studies on the impact of gamification on academic achievement were found in primary schools.

2.2.4.5 Educational courses

Since the beginning of gamification research on learning and education, computer science (CS) and information technology (IT), engineering, and management have been primary contributors. However, the recent literature suggested there is emerging interest from the fields of arts and humanities, environmental science, and psychology (Saxena & Mishra, 2021). Dichev and Dicheva (2017) examined 32 academic studies organized into six categories: CS/IT, math, multimedia/communication, medicine/biology/psychology, languages, and others. Science, technology, engineering, and mathematics (STEM) domains comprised most studies, accounting for 19 out of 23 studies noted by Ortiz Rojas et al. (2017), and computing had the largest share (39%) of fields involved according to Limantara et al. (2019). Business, science, medical, and accounting fields each constituted 9% of studies, and remaining studies spanned various fields, including art, humanities, mathematics, language, and education. Indriasari et al.

(2020) found that STEM domains are frequently reported areas for gamified peer-review activities, Metwally et al. (2021) identified CS and IT as the most commonly studied subjects in gamification research, and Saxena and Mishra (2021) proposed emerging interest in gamification from the fields of arts and humanities, environmental science, and psychology.

2.2.4.6 Duration of interventions

Regarding the duration of gamification interventions, we found that the literature has not extensively addressed this aspect. However, some meta-analyses have used duration as a moderator to evaluate the impact of gamification on learning outcomes (Kim & Castelli, 2021; Yıldırım & Şen, 2019). As noted by Zainuddin et al. (2020), most studies have been conducted within a few weeks or months. Even if Ortiz Rojas et al. (2017) identified a clear tendency among researchers to avoid a novelty effect by conducting longer interventions, Saputro et al. (2017), Saxena and Mishra (2021), and Alsawaier (2018) stated that longitudinal studies are necessary to assess the actual impact of gamification in motivation, engagement, and learning outcomes.

2.2.4.7 Focusing on Personalization

In recent years, there has been a growing interest in personalization in gamification for learning and education, which is in response to often inconsistent and conflicting research results in the field. Studies depicted in this section have shown that gamification effectiveness depends on individual characteristics, such as demographic variables, expectations, learning style, behavior, and skill/knowledge. Personalized gamification has the potential to improve the learning experience by recognizing and catering to the diverse needs of learners to enhance motivation and performance. Moreover, understanding different learning styles is essential for designing and delivering personalized interventions that yield optimal outcomes. However, the effectiveness of gamification personalization in improving students' learning outcomes remains

inconclusive. It is essential to cater to individual learners' needs to ensure gamified learning success, and educational designers need to acquire an empirical understanding of outcomes, learning objectives, and content to enhance the effectiveness of gamification in education. Although gamification has been shown to have a positive impact on education, we note that negative effects may arise because of individual differences and behaviors (Denden et al., 2022; Saleem et al., 2022).

In one the first works on this aspect, Hamari et al. (2014) hinted at customizing gamified learning to accommodate individual differences among students, and Sanmugam et al. (2015) proposed using Bartle's player-motivation types to assist in recognizing and addressing different student skills and personalities, which helps identify user types for the system. According to Dichev and Dicheva (2017), Ortiz et al. (2016), Hung (2017), Caporarello et al. (2019), and in more recent years, Denden et al. (2022), Bennani et al. (2021), and Oliveira et al. (2022), the effectiveness of gamification was found to depend on individual characteristics and needs, such as demographic variables, gender, personality traits, learning types, gaming frequency, player types, individual study design, expectations, and culture. To address the unique needs of individual learners, gamification requires customization of game elements. Oliveira et al. (2022) analyzed 21 studies from various countries to assess the impact of personalized gamification on learning outcomes. Most studies focused only on gamer types for personalization and ignored other important factors, such as culture and gender. Regarding the methods implemented in literature to personalize the gamified environment, Aljabali and Ahmad (2019) and Rozi et al. (2019) noted the FelderSilverman learning-styles model and Kolb's learningstyle model as the most used learning-style models. Bartle's player type has been used to identify different player types, whereas the Five Factor Model has been extensively used to examine personality traits. Hallifax et al. (2019) reported that there are two main types of adaptive gamification systems in education: static and dynamic. Static systems adapt game

elements using a learner profile, whereas dynamic systems adapt according to learner activity. Moreover, in recent years, other areas that investigate the impact of personalization in gamification interventions have been included, such as ontology, artificial intelligence, and intelligent tutoring systems (Aljabali & Ahmad, 2019). The findings revealed that the personalized mode had higher engagement levels and learning outcomes compared with the nonpersonalized mode, improved users' satisfaction (Behl et al., 2022), and improved learning motivation and achievement in elementary students (Aljabali & Ahmad, 2019). However, Oliveira et al. (2022) noted that although tailored systems were more effective in certain situations, nontailored systems were more effective in others, highlighting the importance of adapting gaming features to increase learner engagement. Moreover, customization to the learner's proficiency level has been seen to prevent frustration and monotony (Saxena & Mishra, 2021). Aljabali and Ahmad (2019) found extroverted and introverted individuals perceived the playfulness of leaderboards differently. However, the effectiveness of gamification personalization in improving students' learning outcomes remains unclear (Oliveira et al., 2022).

2.2.4.8 Focusing on Motivation and Engagement

Enhancing motivation and engagement are two of the most important objectives of gamification in learning and education. According to Brooks et al. (2012), motivation guides behavior and decision-making, whereas engagement is a dynamic force associated with various actions and tasks (Frydenberg et al., 2005). Existing research indicates that it is important to evaluate the motivation levels and intrinsic motivation of learners. In addition, Limantara et al. (2019) proposed the "model of student participation," which considers how students were enrolled in the study and their underlying motivation for participating in the gamified study. In general, the findings depicted in this section suggest that the impact of gamification on motivation remains inconclusive and that incorporating game elements in learning environments can significantly

enhance student engagement. In a recent meta-analysis, Mamekova et al. (2021) suggested that gamification can enhance motivation to learn, but for only about one third of students. Nevertheless, as suggested by Ortiz Rojas et al. (2017), there is a need for assessing motivation explicitly in future studies on gamification in learning and education. One direction could be the motivation evaluation by using psychometrically validated measures such as the Intrinsic Motivation Inventory, as suggested by Seaborn and Fels (2015), or other validated measures identified in Touré-Tillery and Fishbach (2014). In one of the first works on this aspect, Glover (2013); but also later, Koivisto and Hamari (2019), and Mohammed and Ozdamli (2021) suggested that the careful implementation of gamification can motivate learners to complete activities and promote desirable behavior while discouraging undesirable behavior. The crucial factor to consider when evaluating the usefulness of gamification for a group of learners is their level of intrinsic motivation (Sanmugam et al., 2015). If their intrinsic motivation is already high, the addition of extrinsic motivation through rewards could have a counterproductive effect, making gamification unsuitable in such a scenario. In Xu et al. (2021), out of 58 studies reviewed, 35 studies (59.32%) found that gamification improves motivation, whereas three studies (5.08%) found that gamification did not improve motivation. For the remaining 20 studies (33.9%), results were either inconclusive or not relevant to the research question of the study. Furthermore, seven out of 10 studies found that gamification improves intrinsic motivation. Similar results were found in a meta-analysis by Mamekova et al. (2021). In the study, the authors included seven articles between 2011 and 2021. They suggested that gamification in education can enhance students' motivation to learn, but for only about one third of the students. In addition, the effectiveness of gamification might vary depending on whether the game type is appropriate for the learning content. Sailer and Homner (2020), in their meta-analysis, found a significant, small effect of gamification on motivational (Hedges's $g = 0.36$, $SE = 0.09$, 95% confidence interval [CI] = [0.18, 0.54], $p < .01$,) learning outcomes

with an additional significant and substantial amount of heterogeneity (I^2 = 75.13%; Shamseer et al., 2015). Concerning the moderators, gamification interventions lasting half a year or less showed significantly larger effects on motivational learning outcomes (g = 0.59, 95% CI = [0.39, 0.59], k = 6, n = 932) than intervention lasting 1 day or less (g = 0.19, 95% CI = [–0.07, 0.45], k = 9, n = 1,145). Furthermore, the effects in higher education settings (g = 0.52, 95% CI = [0.33, 0.71], k = 7, n = 1,025) and work-related education settings (g = 0.72, 95% CI = [0.25, 1.19], k = 2, n = 53) were significantly larger than those found either in informal training or school settings. This is in line with previous research in which contextual factors were found to differentially affect the experience of gamification in each situation (e.g., demographic and personality factors), the associations attached to the task or activity in general, and the temporal and spatial context (Majuri et al., 2018). Moreover, effects differed between experimental and quasi-experimental studies; the latter showed a significant medium-sized effect compared with the nonsignificant effects of the former. However, this pattern changes if only studies with high methodological rigor are included (i.e., experimental designs or quasi-experimental designs with pretest and posttest measures; Sailer & Homner, 2020). Zhang and Yu (2022) found gamification has varying effects on different types of motivation. Overall, across 10 studies, gamification showed a moderate effect on motivation (Cohen's d = 0.77, 95% CI = [0.53, 1.01], p < .001, k = 10, I^2 = 77.4%). Regarding the intrinsic motivation, the authors found a positive effect of gamification (d = 0.64, 95% CI = [0.37, 0.91], p < .001, k = 5, I 2 = 66.9%), observed also in extrinsic motivation (d = 0.92, 95% CI = [0.50, 1.34], p < .001, k = 5, I 2 = 84.4%). Regarding student engagement, Alsawaier (2018) noted that incorporating game elements and designing gamified courses with appropriately challenging tasks into learning environments may significantly enhance student engagement, but the impact on motivation remains inconclusive. However, no longitudinal study investigated the most effective game components that promote intrinsic motivation.

2.2.4.9 Focusing on Game Elements

Game elements in educational activities have been shown to promote a sense of enjoyment, challenge, and achievement among students. Game elements such as leaderboards, levels/milestones, challenges/quests, badges, immediate feedback, social-engagement loops, teams/social dynamics, and visual 3D/sound can enhance engagement, motivation, and involvement in learning. Although PBLs remain the most common game-design elements used to enhance motivation, other elements, such as collaborative work, virtual maps, and skill trees, have also been proposed. The effects of game elements on motivation, attitudes, and performance vary depending on gender and personality, and each game element should be carefully selected according to rigorous research. Studies have found an overall slightly positive effect of gamification on student-learning outcomes, with leaderboard, badges/ awards, and points/experience employed in most studies. However, studies not using leaderboards resulted in a higher statistically significant effect size than those studies that did use it. In one of the first works on gamification and game elements, Nah et al. (2013) noted that gamification offers various elements, such as leaderboards, levels/ milestones, onboarding, challenges/quests, badges, immediate feedback, social-engagement loops, teams/ social dynamics, rules, marketplace/economies, visual 3D/sound, avatars, customization, narrative context, and role-play, that can improve user engagement in learning. These components can provide a sense of achievement, reward, personal control, social interaction, and meaning to learning tasks while also simulating real or fantasy worlds and teaching abstract concepts or subjects. In addition, storytelling through narrative context can induce psychological responses and drive fulfillment of goals, ultimately enhancing user motivation, focus, and involvement in learning. Ortiz and colleagues (2016), Saputro et al. (2017), Alomari et al. (2019), Antonaci et al. (2019), Ofosu Ampong (2020), Huang et al. (2020), and Xu et al. (2021) indicated that gamification research commonly combines elements such as PBLs and challenges, levels, and avatars. The PBL triad

has been found to maintain student engagement and motivation, create a sense of competition, and improve learner performance (Alomari et al., 2019; Antonaci et al., 2019). However, Dichev and Dicheva (2017) stated that relying solely on the use of PBLs may not be sufficient to address the relevant motivational factors.

Antonaci et al. (2019) found that the effects of badges may differ according to gender and personality. Badges can be used to establish clear goals or encourage social comparison. Leaderboards have been found to positively affect attitudes toward gamification, learning performance, engagement, enjoyment, and goal commitment, especially in challenging tasks. Like badges, the effects of leaderboards vary depending on personality and can facilitate social comparison, which positively influences performance by providing information on user points and stimulating competition among users. Leaderboards were found to increase positive competition and motivation in 32 studies, but some students felt less motivated because of the added competition (Xu et al., 2021). Bernik et al. (2022) highlighted the use of a leaderboard and top-scoring student list along with continuous feedback, virtual meetings, and a socially oriented system for effective gamification. Points, scores, and rankings in gamification have been found to have positive effects on motivation, engagement, performance, and emotional states. Limantara et al. (2019) suggested points as the most motivating game elements for assignments. However, these effects may vary depending on gender and personality, but the use of points can foster social comparison and encourage users to undertake challenging tasks. The PBL triad was found to create extrinsically motivating conditions to encourage intrinsic motivation (Xu et al., 2021). Rewards, progress bars, feedback, and avatars are also considered effective in promoting motivation and engagement in learning. Saputro et al. (2017) noted that intrinsic motivation can be increased through a sense of autonomy, competence, relatedness, and purpose, which can be fostered through various game-design elements, such as collaborative work, virtual maps, and skill trees. Howard-Jones and Jay (2016) focused on the

role of reward in educational games from a cognitive neuroscientific perspective. They stated interventions using uncertain rewards can be effective but have limited evidence. However, understanding how rewards affect memory formation can aid in implementing gamification in education. Huang et al. (2020) investigated 30 studies trying to give some indications about the effects of game elements on learning outcomes. An overall slightly positive effect was found ($g = 0.464$, 95% CI = [0.244, 0.684], $p < .001$), with a substantial amount of heterogeneity ($I^2 = 88.21\%$). The authors found that studies not using leaderboards resulted in a higher statistically significant effect size ($g = 0.771$, 95% CI = [0.286, 1.256], $k = 8$, $n = 724$) than those studies that did use it ($g = 0.358$, 95% CI = [0.107, 0.608], $k = 23$, $n = 2,359$), and similar effect size has been found between using or not using badges/awards and points/experience design elements. A significant medium effect size was found in the use of responsive feedback ($g = 0.509$, 95% CI = [0.185, 0.833], $k = 19$, $n = 2,148$). The presence of timed activities showed a small effect size ($g = 0.236$ 95% CI = [–0.199, 0.670], $k = 6$, $n = 710$) not statistically significant compared with the absence ($g = 0.529$, 95% CI = [0.268, 0.790], $k = 24$, $n = 2,373$). Moreover, the presence of a collaboration design element showed a significant effect size ($g = 0.609$, 95% CI = [0.222, 0.997], $k = 9$, $n = 719$), whereas the absence of competition showed a major effect size ($g = 0.590$, $k = 9$, $n = 665$) compared with the presence ($g = 0.412$, $k = 21$, $n = 2,318$). Finally, the highest effect size was observed in the use of quests/missions/modules ($g = 0.649$, 95% CI = [0.279, 1.02], $k = 12$, $n = 1,142$). In addition, a significant medium effect size was found in undergraduate students ($g = 0.638$, 95% CI = [0.378, 0.898], $k = 13$, $n = 1,724$), whereas the K–12 students ($g = 0.306$, 95% CI = [–0.156, 0.767], $k = 10$, $n = 920$) showed a nonsignificant one. Cavalcanti et al. (2021) focused on investigating the effects of feedback on students' performance activities. They found that 65% of the articles concluded that feedback had a positive impact on students' performance, and 41.26% used feedback to support self-regulation. In addition, Willert (2021) focused on feedback in gamified education.

It found six different types of feedback can be implemented in educational games. Formative feedback assesses the quality of students' responses and can enhance their competence, whereas summative feedback summarizes the students' achievement status or end of a course unit and can influence future decisions. Immediate feedback is provided virtually during tests or given soon enough after submission to affect the student's next task. Self-regulation feedback supports students in monitoring and adjusting their actions toward learning goals. Scaffolding provides support to students in their learning process and can be gradually faded out as competence increases. Social or peer feedback is when feedback is given to tasks and assignments from one student to another. Investigating 50 articles, Willert (2021) found that feedback types are distributed as follows: 31 are formative/process, 19 are summative/corrective, 17 are immediate/rapid, 12 are self-regulation, nine are social/peer, and four are about feedback through scaffolding. In general, the purpose of implementing feedback is to enhance student engagement, give a better sense of progression and goal orientation, help students in their work or progress, improve the correctness of submitted assignments, increase student motivation, enhance perceived competence, empower students, and add enjoyment and fun to the learning process. The feedback implementation allowed an overall satisfaction with the new course or system, better engagement, higher rates of submitted assignments, increased student motivation, better self-pacing of learning, qualitative improvements in code, higher student satisfaction, and better onboarding for inexperienced participants. However, some focal points have been highlighted. Alomari et al. (2019) suggested how each game element should be carefully selected using rigorous research. Ofosu-Ampong (2020) emphasized the role of having a clear experimental approach, without which it is difficult to determine which game elements are most effective for a specific activity and group of learners. Finally, Saleem et al. (2022) suggested that the effectiveness of gamification in education remains a contentious issue

because incorporating gamification elements has not resulted in significant improvements in students' group cohesion, talent, motivation, and intrinsic drive.

2.2.4.10 Including Game Design

Game design considers the underlying design principles that make games engaging. Effective gamification requires a deep understanding of game-design principles and how they can be applied to learning objectives. The literature suggests that effective gamification in learning requires a deep understanding of game design principles, such as providing students with freedom to fail, offering frequent feedback, designing progression, and using storytelling. Badge-system design is critical and should consider functions, structure, and design. Successful game design requires defining clear objectives, considering feasibility, and understanding stakeholders. Psychological factors, such as fun, motivation, and social interaction, are also important. Finally, game designers should engage diverse players by providing challenges at adjustable difficulty levels, allowing sufficient time to solve challenges, promoting creativity and self-expression, and employing social play, storytelling, and fantasy. One of the first works on the application of game design principles to gamification in learning and education was carried out by Stott and Neustaedter (2013). The authors noted that these principles can create a more enjoyable and effective learning experience for students. These include providing students with the freedom to fail and experiment without fear of irreversible damage, offering rapid and frequent feedback, designing progression in the form of scaffolded instruction or levels, and using storytelling to contextualize learning elements. Several years later, Facey-Shaw et al. (2017) and Park and Kim (2019) presented works focused on the design of a badge system. According to the authors, badge system design is a critical element of the process of motivating, recognizing, and showcasing formal and informal learning using digital badges. It has been suggested that badge systems have three core dimensions, including the functions or purpose of badges, the structure of badge systems, and the design and interaction with badges.

Park and Kim suggested focusing on three conditions when developing badge design: distinguishing between physical and conceptual learning activities, distinguishing between individual and interaction induced learning, and reviewing the time and effort required for earning badges. The study proposed eight badge types for three badge-design conditions with a statistically significant difference between them ($\chi 2$ = 1,117.7; $p < .001$). The authors indicated that badges are useful tools for promoting self-directed learning and providing various benefits, such as flexible learning environments, goal setting, progress tracking, and planning. Moreover, badges have been shown to positively affect critical thinking, teamwork, leadership, and other skills and knowledge. However, they may not always be effective in instilling interest. Indeed, learners were generally comfortable displaying badges within a social learning environment but less comfortable sharing badges with external audiences (Facey-Shaw et al., 2017). Other than badge design, Mora et al. (2017) noted 10 relevant ingredients for successful game design. These include self-representations, three-dimensional environments, narrative, feedback, reputations, ranks and levels, marketplaces and economies, competition under rules, teams, communication, and time pressure. In addition, others highlighted in their work include engagement cycle, end game, rules, storytelling, the importance of defining clear objectives, considering feasibility and investment, understanding stakeholders in the design process, and psychological factors, such as fun, motivation, social interaction, and desired behaviors. Finally, Laine and Lindberg (2020) identified 56 motivators that contribute to motivated engagement in educational games, which were grouped into 14 classes based on their similarities. The authors suggested game designers engage diverse players by providing challenges at adjustable difficulty levels, favoring simple challenges, and allowing sufficient time to solve challenges. Players should have the ability to make choices and use input mechanisms suitable for them. Game designers should foster creativity and self-expression, promote exploration, ensure fairness, and set clear and achievable goals. The game should be

relevant to the player's context and involve game resources to increase engagement. In addition, social play, storytelling, and fantasy should be employed. Finally, the authors claimed that many of the motivators were initially intrinsic, but game mechanics supporting these motivators can produce different motivational results depending on the context of use.

2.2.4.11 Focusing on Learning Outcomes (Cognitive, Behavioral, and Affective)

Gamification is a technique that has been used to improve learning outcomes (cognitive, behavioral, and affective) in various educational settings. However, research has yielded mixed results. This section depicts some of the most relevant findings. Ortiz Rojas et al. (2017) investigated the effects of gamification on learning performance. They found that although only nine studies demonstrated a positive impact, it is crucial to examine why the remaining 14 studies showed negative or mixed results. The authors stated that various factors, including mediating variables, the choice of measurement instrument, sample size, and study duration, could have influenced the outcomes. Sailer and Homner (2020) investigated the effects of gamification on cognitive and behavioral learning outcomes. The results yielded a significant, small effect of gamification on cognitive ($g = 0.49$, $SE = 0.10$, 95% CI = [0.30, 0.69], $p < .01$) and behavioral ($g = 0.25$, $SE = 0.11$, 95% CI = [0.04, 0.46], $p < .05$) outcomes, with an additional significant and substantial amount of heterogeneity ($I^2 = 72.21\%$, I 2 = 63.80% respectively; Shamseer et al., 2015). From the moderator analysis, results indicate that the inclusion of game fiction ($g = 0.41$, 95% CI = [0.31, 0.51], $k = 3$, $n = 254$) and social interaction, specifically the competition-collaboration combination ($g = 0.70$, 95% CI = [0.41, 0.99], $k = 3$, $n = 135$), were particularly effective at fostering behavioral learning outcomes. However, by considering only studies with high methodological rigor, only cognitive learning outcomes showed a small effect of gamification ($g = 0.42$, $SE = 0.14$, 95% CI = [0.14, 0.68], $p < .01$). But this is a much smaller number of studies ($k = 9$) and total sample ($N = 686$) than the more inclusive analysis that contained studies with a lower methodological rigor ($k = 22$). In addition,

no moderators were found to significantly moderate the effects of gamification on cognitive learning outcomes in this more conservative analysis. In addition, Kim and Castelli (2021) investigated with a meta-analysis the effects of gamification on behavioral change in education, assessed through test score or participation level. From 18 eligible studies, authors found a moderate significant effect size (d = 0.48, 95% CI = [0.33, 0.62]), higher in participation level (d = 0.60, 95% CI = [0.40, 0.77], n = 15,322) than test score (d = 0.30, 95% CI = [0.03, 0.18], n = 3,059). These results are in line with those found by Sailer and Homner (2020). In this context, the gamification appeared to be effective both for adult (d = 0.95, 95% CI = [0.70, 1.12], n = 12,455) and K–12 (d = 0.92, 95% CI = [0.29, 1.55], n = 146) interventions, although not for college students (d = 0.15, 95% CI = [–0.04, 0.35], n = 5,780). Concerning the intervention length, those with less than 1 hr (d = 1.57, 95% CI = [1.25, 1.90], n = 492) was the most effective than 2 to 16 weeks (d = 0.39, 95% CI = [0.21, 0.57], n = 12,282) and 1 to 2 years (d = −0.20, 95% CI = [–0.47, 0.09], n = 18,381) in behavioral change. Ritzhaupt and colleagues (2021), through a meta-analysis, investigated the impact of gamification in formal education settings on affective and behavioral outcomes. Authors included 19 studies with affective outcomes and 13 with behavioral outcomes. In this work, the label “affective outcome” is analogous to “motivational outcome” in the work of Sailer and Homner (2020). Regarding the affective outcomes, a significant medium effect size was found similar to the work of Sailer and Homner (g = 0.574, 95% CI = [0.384, 0.764], p < .001). In addition, a high amount of heterogeneity (I^2 = 73.51%) was found. In accordance with the rest of literature, Ritzhaupt and colleagues (2021) found that leaderboards, badges/ awards, and points/experiences were the most frequently observed design elements also for affective outcomes. Studies with leaderboards resulted in a notable effect on affective outcomes (g = 0.643, 95% CI = [0.420, 0.866], k = 13, n = 1,560) compared with those without (g = 0.397, 95% CI = [0.071, 0.772], k = 6, n = 414). This result suggests that competition in educational settings has the highest effect

size on affective outcomes; no other statistically significant differences were discovered between the presence and absence of specific game elements. However, it is also true that the other game elements were rarely observed in the studies, suggesting future lines of research. Considering behavioral outcomes, Ritzhaupt and colleagues (2021) found a significant medium effect size ($g = 0.740$, 95% CI = [0.465, 1.014], $p < .001$) with a high amount of heterogeneity ($I^2 = 83.26\%$). This is in line with what was found by Sailer and Homner (2020) and Kim and Castelli (2021). PBLs were the most frequently used game-design elements, but no statistically significant differences resulted with and without each of these. However, other nonfrequently used game elements showed more interesting results. The presence of nonlinear navigation resulted in a statistically significant difference on behavioral outcomes compared to its absence ($g = 1.362$, 95% CI = [0.903, 1.822], $k = 1$, $n = 133$). However, this result is based on one study only ($n = 133$). The absence of adaptivity/personalization ($g = 0.806$, 95% CI = [0.515, 1.096], $k = 12$, $n = 1{,}498$) and narrative/storytelling ($g = 0.791$, 95% CI = [0.482, 1.101], $k = 12$, $n = 1{,}397$) resulted in a statistically significant difference on behavioral outcomes compared with their presence. Bai and colleagues (2020) conducted a meta-analysis of 24 quantitative studies and a synthesis of 32 qualitative studies, all containing a control condition and meeting Medical Education Research Study Quality Instrument (MERSQI) standards for the field, to examine the impact of gamification on academic learning outcomes in K–12 education. Overall, they found a medium effect of gamification on learning ($g = 0.50$, 95% CI = [0.28, 0.72], $p < .001$), with substantial heterogeneity ($I^2 = 88.2\%$; Shamseer et al., 2015). To account for the large variance in effect sizes, the moderator analysis included (a) the type and number of game elements used, (b) the quality/ level of the control group, (c) intervention characteristics (e.g., sample size, subject, duration, flipped classroom or not, integration of gamification into instructional activities or not, use of tangible rewards), and (d) participant characteristics (student level, geographic region). Results indicated that effect size significantly increased with

sample size, decreased with interventions greater than 1 month, was greatest in classrooms from Western Asia (i.e., majority of published works), and did not differ within any other set of moderators. Bai and colleagues' (2020) qualitative synthesis highlighted four reasons students liked gamification: (a) fosters enthusiasm, (b) provides performance feedback, (c) gives a sense of recognition, and (d) promotes goal setting. In addition, they identified two reasons students disliked gamification: not adding additional utility and causing anxiety or jealousy because of social comparisons/competition. Critically, the large variability in effect size was not explained by the number (one vs. six) or choice of game elements used. This is likely caused by too few studies meeting the standards for inclusion; 42 were screened out for lacking a control group (n = 13), not meeting the criteria for a gamified course (n = 8), and providing insufficient data (n = 21). Clearly, more gamification studies need to meet inclusion standards to facilitate cross-study comparisons and a better understanding of which game elements matter. Another meta-analysis, by Fadhli et al. (2020), focused on the effects of gamification in different learning outcomes (cognitive, skills, attitude, language, health, and social-emotional abilities). The difference between pretest and posttest measures express a positive impact of gamification in fostering learning outcomes in 6- to 10-year-old children (d = 1.01, 95% CI = [0.98, 1.05], k = 6), with I 2 = 0.53. However, the findings cannot be considered conclusive because of the limited number of studies included. Yıldırım and Şen (2019), who conducted a meta-analysis to evaluate the effectiveness of gamification in students' achievements by using the educational course as moderator, found that gamification's effect on student achievement did not show significant differences in both technology-based and nontechnology-based courses: technology-based courses (N = 15; g = 0.482, 95% CI = [–0.007, 0.970], p = .053) and nontechnology-based course (N = 30; g = 0.588, 95% CI = [0.346, 0.829], p < .001). However, given the effect sizes, it seems that nontechnology-based courses have a greater advantage in using gamification for academic achievement compared with technology-based courses.

Another meta-analytic work on the effect of gamification in university students' academic achievement was conducted by Dikmen (2021) in the Turkic population. Dikmen (2021) incorporated 52 primary studies and discovered a favorable association between gamification and academic achievement ($d = 0.862$, 95% CI = [0.68, 1.04], $p < .001$, $k = 52$), with a large amount of heterogeneity (mode: $Q = 266.417$, $p < .001$). Different moderators were investigated (educational level, educational course, class size, and publication years). The analyses showed a nonsignificant moderator effect of educational level in terms of the effect of gamification on academic achievement. Educational course was found as a positive moderator; the largest effect size was observed in the science course ($d = 0.993$), and the smallest was observed in the mathematics course ($d = 0.416$). The class size and the publication years were not considered positive moderators in this study. To sum up, the present study corroborated the beneficial impact of gamification on students' academic accomplishment. This finding was supported by Zhang and Yu (2022), who demonstrated that gamification enhances learning performance ($d = 0.85$, 95% CI = [0.32, 1.37], $p < .001$ $k = 6$). However, some meta-analyses have reported different effect sizes, possibly because of cultural differences in gamification of learning and the grouping of courses in previous studies. Bai et al. (2020) discussed the importance of educational levels in moderating the impact of gamification on academic achievement. However, the current study's findings suggest that could be not true. This indicates that gamification can be effective across all levels of students and is not limited to a specific age group. According to educational level, the findings appear inconsistent with those of previous meta-analyses. Indeed, Yıldırım and Şen (2019) grouped the courses as technology-based and nontechnology-based. Furthermore, the limited inclusion of subject disciplines in previous studies may have contributed to these differences. Finally, a systematic review of Nurtanto et al. (2021) corroborated the previous findings on the positive effects of gamification in affective, behavioral, and cognitive outcomes. Concerning the affective domain, the authors found that

gamification increases enthusiasm, motivation, and other emotional responses. Regarding the cognitive outcomes, gamification has been found to have a positive impact on student retention, Other positive benefits of gamification are related to behavior change, with improvements in teamwork, communication skills, social skills, digital literacy, critical thinking, and digital literacy.

2.2.5 A Checklist for Research in Gamification

Gamification in learning and education is a complex system that involves various aspects, including user characteristics, learning outcomes, system implementation, and the development of elements within the system. In the previous sections, we discussed the most important characteristics identified in the literature, revealing mixed sentiments regarding the results, key considerations, and methodological constraints. The purpose of this article is to create a checklist (see Appendix), as suggested by Metwally et al. (2021), that can guide researchers and developers in conducting high-quality research on gamification in learning and education. This checklist considers the most critical elements, moderators, and mediators that may affect gamification success; methodological considerations; and essential elements that should not be excluded in the study design. In addition, a gamified learning environment is a high-cost process involving different professionals. Having a starting point could result in a reduction of the production costs. To this end, we turn to findings from the 72 studies investigated. The checklist was created using key constructs identified according to the present systematic review. Our analysis concentrated on seven primary aspects: study design, theoretical foundations, personalization, motivation and engagement, game elements, game design, and learning outcomes. In total, the checklist comprises 24 items. Some of them have been structured to have a 4-point quantitative Likert scales ranging from −1 to 3, with the idea of directing the researcher in the implementation of the more or less effective elements compared with what is known today contextually to the period in which we wrote this work. The value −1 of the scale

reflects which elements or study design have been seen as having a negative impact on learning or in the methodological rigor. The value 0 reflects elements having a neutral impact on learning or in the methodological rigor, 1 reflects a low positive impact, 2 reflects a medium positive impact, and 3 reflects a high positive impact. These values, if related to a learning aspect investigated in the meta-analyses, are mirrored to the effect size discovered, according to Cohen (1992). Furthermore, we have incorporated a point-based system into our approach, which serves as an indicator of the evidence reviewed in the preceding sections. This addition facilitates researchers and practitioners in assessing the quality of their work. Before initiating their research or designing a study, researchers can complete the checklist, which highlights the critical elements. This step will enable them to evaluate the quality of their work using the point-based system, which assigns a score ranging from 0 to 20. A higher score indicates the inclusion of aspects that improve the research's quality and methodological rigor, whereas a lower score suggests that the study may have overlooked critical factors that are necessary for a rigorous evaluation of gamification's impact on learning and education. The first set of criteria (Items 1–6) concerns study design. Given calls to develop more methodologically rigorous experiments, we encourage researchers to set up experimental or quasi-experimental studies, employ pre-post assessments, and include control groups (Bai et al., 2020; Huang et al., 2020; Kalogiannakis et al., 2021; Sailer & Homner, 2020; Seaborn & Fels, 2015). In addition, researchers should consider, as control measures, the type of activities the control group has carried out (e.g., passive, active, or no activities) and whether the control groups is equivalent to the experimental group at pretest (e.g., in previous knowledge). Other important considerations involve accounting for education level and course type as covariates, especially when analyzing multiple levels and types of courses. In addition, there has been a focus on the potential for conducting longitudinal studies to assess the long-term effects of gamification and using a sizable sample to enable more sophisticated analyses, such as mixed or multilevel

models. Item 7 concerns the theory behind gamification in learning and education. In detail, Khaldi et al. (2023) highlighted how many studies lack a foundation in theoretical frameworks and do not incorporate them in the development of gamified learning systems. Contextualizing the study results based on a reference theory of gamification in learning and education could enhance the quality and the clarity of the study itself. The next set of criteria (Items 10–16) concerns personalization aspects. As emphasized earlier, exploring individual behaviors and characteristics has become an essential element in determining which gamification systems are best suited for specific individuals. Gender, personality traits, learning types, gaming frequency, player types, individual study design, expectations, and culture have been identified as crucial factors in this regard. Synthesizing evidence from studies that consider one or more of these personalized elements as moderators can help create more high-quality research on gamification in learning and education. Focusing on player types, researchers studying gaming personality hold that different player types exist (i.e., different characteristics and preferences for specific game elements) and can affect players' perceptions of gamification design elements (Santos et al., 2021). Gaming personality may account for how interactive and engaging gamification may be for some students (Tu et al., 2015). The Bartle test of gamer psychology (Bartle, 1996) or its successor (González Mariño et al., 2018) are frequently applied by researchers to understand and categorize online game players into four gaming personalities according to their their gaming preference. In recent years, other models have been proposed. The BrainHex model (Nacke et al., 2011) is based on players' neurobiological characteristics. It consists of seven player types, called "archetypes," that typify a particular player experience. Marczewski (2015) proposed the gamification user types hexad, a model specific to gamification, based on SDT (Ryan & Deci, 2000), in which six user types are motivated by different combinations of intrinsic or extrinsic motivational factors. At this point, the question of which of these gamer personality models best explains gamification in education arises (and

is still up for debate), and future work is needed to understand whether gamer personality is a critical element in gamification. Item 17 concerns the motivation and engagement outcomes. We categorized these outcomes separately because existing literature has emphasized the significance of assessing motivation using psychometrically validated measures both before and after implementing the gamified intervention (Ortiz-Rojas et al., 2017; Seaborn & Fels, 2015). Thus, a study that evaluates motivation both before and after implementing a gamified intervention has a higher methodological rigor than one that does not. The next set of criteria (Items 18–21) concerns the effects of game elements (Dichev & Dicheva, 2017; Dicheva & Dichev, 2015; Seaborn & Fels, 2015). As reported previously, conducting a gamification study requires the selection of specific game elements by the researchers (Dichev & Dicheva, 2017). Some of the literature indicates that the use of PBLs is most likely to produce learning outcomes (Limantara et al., 2019; Zainuddin et al., 2020). However, previous reviews contain inconsistent and contradictory results and found that researchers do not always identify the game-design elements used in their study or systematically inspect their impact on learning outcomes (e.g., Alomari et al., 2019; Bai et al., 2020; Ofosu-Ampong, 2020; Seaborn & Fels, 2015). Thus, checklist Item 18 provides researchers with a comprehensive list of game-design elements for potential inclusion in their study. Furthermore, the effectiveness of any one game element or combination of elements may vary as a function of other factors (i.e., Items 5, 6, 8, 11–16), which should be addressed with specific and appropriate statistical analysis accordingly. Regarding feedback, in accordance with SDT (Ryan & Deci, 2000), literature suggests that affective feedback combined with gamification is linked to positive, intrinsically motivated behavior (Hassan et al., 2019). However, it is still not clear how different types of feedback are related to gamification environment, game-design elements, and participants' characteristics in improving learning outcomes (Hassan et al., 2019; Seaborn & Fels, 2015). For this reason, studies of gamification should indicate the presence and type of feedback in their design and

investigate it comprehensively (Item 19). In Item 22, we focused on game design. According to the literature, effective gamification in learning requires a deep understanding of game-design principles (Laine & Lindberg, 2020; Mora et al., 2017). Thus, incorporating game-design principles in research yields a high-quality research level. In the last criteria (Items 23, 24), we summarized evidence derived from the reviews, systematic reviews, and mostly meta-analyses divided for the learning outcomes. For the behavioral learning outcome, we summarized evidence from the meta-analyses of Sailer and Homner (2020), Kim and Castelli (2021), and Ritzhaupt and colleagues (2021). Overall, small to medium effect sizes were found. Interventions from less than 1 hr to 16 weeks were the most effective, eliciting a medium to high effect size. Moreover, investigating the effect on the target population, we found that adults and K–12 populations saw the most benefit, with a high effect size, whereas an inconsistent result was found in the undergraduate/ college population. Furthermore, the most relevant game-design elements were the presence of nonlinear navigation, game fiction, competition-collaboration, and active instructions. In addition, a negative effect of adaptivity/personalization and narrative/storytelling was found. Considering the motivational/affective learning outcomes, we summarized the evidence from the meta-analyses of Sailer and Homner (2020) and Ritzhaupt and colleagues (2021). Overall, a small to medium effect size was found. Interventions less than 6 months were the most effective, eliciting a medium effect size. Moreover, investigating the effect on the target population, we found that the higher/undergraduate students and K–12 populations were those with a medium to high effect size, whereas an inconsistent result was found in school settings. Furthermore, the most relevant game-design elements were the presence of leaderboards and competition-collaboration. Considering the cognitive learning outcomes, we summarized evidence from the meta-analyses of Sailer and Homner (2020). Overall, a medium effect size was found. Interventions from less than 1 day to less than 6 months were the most effective, eliciting a

medium effect size. Moreover, investigating the effect on the target population, we found that the higher/undergraduate students and the K–12 students were those with a medium effect size. No game-design element was found to significantly moderate the effect of gamification on cognitive outcomes. This result is consistent with those found in Vermeir and colleagues (2020). Finally, we considered effects of gamification on student learning outcomes by investigating the meta-analyses of Bai and colleagues (2020), Huang and colleagues (2020), Dikmen (2021), Yıldırım and Şen (2019), and Zhang and Yu (2022). Overall, a medium or high effect size was found. Interventions less than 1 week or between 1 and 3 months were the most effective, eliciting a medium to high effect size. A medium effect size was found for both technology-based courses and nontechnology-based courses. The class size and the publication years showed a neutral impact. The effect on specific target populations showed a medium effect size for undergraduate students. Considering the game-design elements, we found a medium effect size for responsive feedback, collaboration, quests/missions/modules, and the combination of PBLs.

2.2.6 Conclusion

Introducing gamification into learning and education is a multifaceted system that necessitates researchers and practitioners to consider various elements for a fruitful implementation. Despite the considerable amount of educational-gamification research conducted in the past decade alone (Dubé & Wen, 2022), much of this work has highlighted the need to enhance the quality and methodological rigor of research in this field. In Sailer and Homner's (2020) meta-analysis of 786 studies, 427 (54.3%) were excluded for research-design issues or for lacking a control group, and only 38 (4.83%) were considered sufficiently methodologically robust. Huang and colleagues (2020) excluded 379 articles because they lacked a control group, and only 30 studies were eligible for meta-analysis. The same result was found in Ruthzaupt and colleagues (2021), in which 379 studies were excluded because of the absence of a control group or other

methodological issues. In this study, we aimed to investigate the necessary elements and propose a checklist protocol for conducting high-quality and methodologically rigorous research in the field of gamification in learning and education. This is based on the evidence presented by a systematic review conducted between 2011 and 2023, which considered reviews, systematic reviews, and meta-analyses. The evidence showed how seven core elements have to be considered in the implementation process: study design, theoretical foundations, personalization, motivation and engagement, game elements, game design, and learning outcomes. The necessity of this tool is further reinforced by a recent study conducted by Metwally et al. (2021). The proposed checklist is expected to serve as an initial reference for researchers and developers to conduct studies that encompass the essential elements reported in the literature to design products that are of high quality and methodological rigor. Moreover, it is important to recognize that this study represents an initial step, given that the tool was developed based on the existing literature. A forthcoming investigation will focus on the validation of the current checklist.

2.2.7 Appendix

Enhancing Research Quality on Gamification in Learning and Education: A Checklist Protocol for Researchers and Practitioners	**Points**
Study design	
1) In which country you are developing the study? _____	
2) Which type of Experiment are you planning to condict? □ Experimental (3) □ Quasi-Experimental (2) □ Qualitative (-1) □ Other	 [2] [1] [0] [0]
3) Have you considered a Pre-Post study with two groups? □ Yes (3) □ No (-1)	 [1] [0]
4) Presence of Control Group: □ Yes (3)	 [1]

□ No (-1) The instruction of the control group is: □ Passive (e.g., listening to lectures, watching instructional videos, reading textbooks) □ Active (explicitly prompting the learners to engage in learning activities (e.g., assignments, exercises, laboratory experiments)) □ No instruction Comparisons between the groups at pre-test □ No statistical difference (equivalent groups) □ A statistical difference (non-equivalent groups) □ No comparison	[0] [1]
5) Educational level targeted: □ Elementary School Students □ Middle School Students □ High School Students □ Undergraduate Students □ Postgraduate Students Are you also considering the specific population targeted as covariate in the analyses? □ Yes □ No □ No, I used only one population at the same grade	 [1] [0] [1]
6) Educational course targeted: □ STEM field: Specify: _____ □ Non-STEM field: Specify: _____ Are you also considering the specific course targeted as covariate in the analyses? □ Yes □ No □ No, I used only one course in the study	 [1] [0] [1]
7) Have you planned a longitudinal study? □ Yes □ No	 [1] [0]
8) Have you considered a sufficiently large sample size considering the covariates to be included in the model? □ Yes □ No	 [1] [0]

Theory	
9) Have you considered contextualizing the results based on a reference theory of gamification in learning and education? □ Yes □ No Which? □ Motivational: □ Social-Determination Theory □ Intrinsic-Extrinsic Motivation □ Expectancy Theory □ Goal-setting Theory □ Flow Theory □ Learning: □ Classical Conditioning □ Operant Conditioning □ Theory of Gamified Instructional Design □ Social Theory □ Socio-Cognitive Conflict □ Social Learning Theory □ Other: _____ _____	[1] [0]
Personalization	
10) In which country you will plan the study? _____ _____	
11) Are you considering the gender as covariate in the analyses? □ Yes □ No □ No, I used only one population	[1] [0] [1]
12) Have you considered investigating the personality traits in the effect of gamificatio? □ Yes □ No Which instrument? □ Five Factor Model (FFM) □ Other	[1] [0]

_____ _____	
13) Have you considered investigating the learning types in the effect of gamification? □ Yes □ No Which instrument? □ Felder-Silverman Learning Styles Model (FSLSM) □ Kolb's learning style model	[1] [0]
14) Have you considered investigating the gaming frequency in the effect of gamification? □ Yes □ No	[1] [0]
15) Have you considered investigating the player types in the choice of game design elements? □ Yes □ No Which instrument? □ The Bartle Test of Gamer Psychology □ BrainHex Model □ Hexad □ Other _____ _____	[1] [0]
16) Have you considered to use a static or dynamic gamification system? □ Static gamification system □ Dynamic gamification system	
Motivation and Engagement	
17) To ensure high-quality research in the field, it is essential to conduct an evaluation of motivation using psychometrically validated measures. Have you made arrangements to assess motivation both before and after implementing the gamified intervention? □ Yes □ No Which instrument? □ Intrinsic Motivation Inventory (IMI) (Ryan, 1982) □ Other	[1] [0]

_____
_____

Game Design Elements

18) Game Design Elements (Components) select one or more:

□ Achievements	□ Content	□ Levels
□ Avatars	□ Unlocking	□ Points/Experiences
□ Badges/Awards	□ Gifting	□ Virtual Goods
□ Timed activities	□ Leaderboards	□ Narrative/ Story telling
□ Collections	□ Quest/Missions/Modules	□ Feedback
□ Teams (collaborative)	□ Competition	□ Game fiction
□ Adaptivity/ Personalization	□ Collaboration	Other: _____

□

19) Use of Feedback:

□ Yes [1]
□ No [0]

Which type of feedback, according to Willert (2021)

□ Formative
□ Summative
□ Immediate
□ Self-regulation
□ Scaffolding
□ Social or Peer
□ Other

_____
_____

In which circumstances?

_____
_____
_____

20) In your research, are you going to consider evaluating the impact of single or multiple game elements, and the interaction between them with specific statistical analyses?

□ Yes [1]
□ No [0]

Which element/s you will inspect?

1) _____
_____

2) _____
_____

3) _____ _____	
21) In your research, are you considering inspecting the interactions between different elements of gamification (e.g. game element/s, feedback/s, player's personality, learning types) with specific statistical analyses? □ Yes □ No Which? 1) _____ _____ 2) _____ _____ 3) _____ _____	[1] [0]
Game Design	
22) Are you considering game design principles in your study? □ Yes □ No	[1] [0]
Learning Outcomes	
23) What specific learning outcome could be enhanced using the gamified system? □ Cognitive □ Motivational/Affective □ Behavioral □ Students learning □ Creativity □ Other: _____	
24) What we know today, divided for learning outcome Behavioral: • A general small to medium effect of gamification (1; 2) Intervention length: o Intervention less than 1 hour (3) o Intervention between 2-16 weeks (2) o Intervention between 1-2 years (- 1) Population Target: o Adults (3) o K-12 students (3) Gamification elements:	

○ Non-linear navigation (3) ○ Game fiction (2) ○ Competition-Collaboration (2) ○ Adaptivity / Personalization (-1) ○ Narrative / Story telling (-1) Motivational/Affective: • A general small to medium effect of gamification (1; 2) Intervention length: ○ Intervention less than 6 months (2) ○ Intervention less than 1 day (-1) Population target: ○ Higher/undergraduate (2) ○ K-12 students (3) Gamification elements: ○ Leaderboards (2) ○ Competition-Collaboration (2) Cognitive: • A general medium effect of gamification (2) Intervention length: ○ Intervention less than 1 day (2) ○ Intervention less than 1 months (2) ○ Intervention less than 6 months (2) Population target: ○ School (3) ○ Higher/undergraduate (2) Gamification elements: ○ Inconsistent results (0) Students learning: • A general medium effect of gamification (2; 3) Intervention length: ○ Intervention less than 1 week (2) ○ Intervention between 1-3 months (3) Course target: ○ Technology based courses (2) ○ Non-technology based courses (2) Population target: ○ Undergraduate students (2) Class size (0) Publication years (0) Gamification elements:	

○ Responsive feedback (2) ○ Collaboration (2) ○ Quests/Missions/Modules (2) ○ Badges + Leaderboards + Points (2)	
Total points:	/20

2.2.8 Supplementary Materials

Summary of findings table

Author	Title	Year	Methods	Study design	Number of studies. (part.)	Core	Mediator and Moderators	Target population	Findings	Future research
Aljabali and Ahmad	A Review on Adopting Personalized Gamified Experience in the Learning Context	2018	Review	N/A	13	Personalization	• Learning style • Player type • Personality traits • Other (ontology, AI, ITS)	54% Undergraduate students 8% Postgraduate Students 15% Elementary students 23% N/A	Significant evidence toward student outcomes and performance using personalized gamified learning.	More studies in the field of gamified learning personalization
Alomari et al.	The role of gamification techniques in promoting student learning: a review and synthesis	2019	Review	PRISMA	40	Game elements	• Game elements	Higher Education	points (75%); badges (68 %), leaderboards (63%), levels (38%) are the most used game elements.	More studies on how gamification techniques affect the behavior of learners.
Alsawaier	The Effect of Gamification on Motivation and Engagement	2017	Review	N/A	N/A	Motivation and Engagement	• Learner motivation • Learner engagement • Game elements	N/A	Results suggest that incorporating game elements in the learning environment may significantly increase students' engagement, but there is inconclusive evidence regarding the impact on motivation. It is crucial to design	Future research should include conducting longitudinal studies to gain a comprehensive understanding of the impact of gamification on learners' engagement

									gamified courses with appropriately challenging tasks that match students' abilities to maintain engagement, as overly complex challenges may lead to disinterest, anxiety, and frustration. However, a longitudinal study is necessary to fully capture the long-term effects of gamification on learners' motivation and task engagement.	and motivation, rather than solely relying on full deployment of gamification features. Additionally, it is crucial to investigate the most effective components of game elements that foster intrinsic motivation. Finally, mixed-method design research is needed to understand learners' perceptions of gamification interventions holistically, and to elucidate the relationship between gamification, engagement, and motivation.
Bennani et al.	Adaptive gamification in E-learning: A literature review and future challenges	2021	Review	Qualitative	N/A	Personalization	• Students' profiles • Learning styles • Game elements • AI	N/A	Gamification's effectiveness varies across individuals due to specific contexts, and it is crucial to customize game elements to cater to the unique needs of each learner. Educating oneself on different learning styles helps design and deliver tailored interventions for optimal outcomes. Gamification can be a powerful tool to enhance the learning experience by recognizing and addressing the diverse needs of learners.	Explore the relationship between gamification mechanisms, dynamics, and user characteristics. An effective design should be proposed that adapts game design and gamification elements according to each learner's profile. The learner's implicit information should be detected and analyzed to tailor the gamification

										process based on the player type, preferences, interactions, motivation, and feedback. A mechanism for communication and recording learning achievements and experiences should be provided. Additionally, theories and algorithms of AI should be applied to develop a system that can learn from users' experiences and adapt to different learning situations
Bernik et al.	Computer Game Elements and its Impact on Higher Education	2022	Review	N/A	N/A	Game elements	• Game elements	Higher Education	The optimization of e-learning through gamification involves including a leaderboard and top list, providing continuous feedback, occasional virtual meetings, and social interaction. Bonuses should be made available to students in special circumstances with rewards aligned to the level of activity difficulty that is gradually increased over time	N/A
Caponetto et al.	Gamification and Education: A Literature Review	2014	Review	N/A	119	General	• Learner Engagement • Learner Motivation	University (43%) Upper secondary school (2%) Lower secondary school (4%)	Gamification techniques are being used to enhance motivation and engagement in learning tasks across different education levels and subject areas. While there is a strong prevalence of empirical	It is crucial to plan and design learning interventions carefully.

								Primary school (3%) Other (48%)	studies of gamification at the university level, they are also being adopted in various educational contexts to address transversal attitudes and behaviors such as collaboration, creativity, and self-guided study.	
Caporarello et al.	One Game Does not Fit All. Gamification and Learning: Overview and Future Directions	2019	Review	N/A	N/A	Personalization	• Learner Behavior • Learner Attitude • Learner achievement/performance	N/A	The design and educational effectiveness of gamification in education were studied, with the majority of studies providing positive or neutral results. No significant correlation was found between gamification and student performance in any of the case studies. However, the limited number of studies on the pre-experience moment limited the ability to identify significant trends. These findings highlight the importance of tailoring gamified systems to their prospective users' backgrounds and expectations.	
Devers and Gurung	A Critical Perspective on Gamification in Education	2014	Review	N/A	N/A	General	• Methodological considerations	N/A	To test the effectiveness of gamification in education, measure learning before and after introducing gamification. Statistical significance is an important factor to consider when evaluating the success of gamification, as the ultimate goal is to improve student learning outcomes.	N/A

Dichev and Dicheva	Gamifying education: what is known, what is believed and what remains uncertain: a critical review	2017	Critical Review	N/A	51	General	• Educational level • Subject course • Learning activities • Game elements	Higher education (86%) Elementary school (6%) Middle school (4%) High school (4%)	Gamification in education involves incorporating game design principles into the learning environment to increase learner motivation and engagement. However, there are no clear guidelines for how to do this effectively, and there is insufficient high-quality evidence to support the long-term benefits of gamification in education. The emphasis on points, badges, and leaderboards may not be enough to address relevant motivational factors, and understanding the target population is crucial for successful gamification.	More research is needed to understand how to gamify an activity based on the specifics of the educational context, and to explore the effect of game design elements across different learning contexts
Facey-Shaw et al.	Educational Functions and Design of Badge Systems: A Conceptual Literature Review	2017	Review	N/A	61	Game Design (Badge)	• Functions or Purpose of Badges • Structure of Badge Systems • Design and Interaction with Badges	N/A	The potential of badges for learning across educational levels has been confirmed, but the variety of badge designs and functions limits the ability to compare their effectiveness. Instructors and designers face the challenge of maximizing learning benefits and minimizing negative effects. The use of leaderboards to track learner progress was popular, but results were mixed, with some learners finding them demotivating. Learners were generally comfortable with displaying badges within social learning environments but less comfortable with sharing them externally.	Future research should focus on optimizing badge system design by utilizing badge values and other parameters and analyzing incentives for user contribution. Additionally, future research should investigate the impact of specific structural features of badges on learning to enhance their effectiveness. These findings highlight the need for further research to improve the design and

										implementation of badge systems for enhanced learning outcomes.
Faiella and Ricciardi	Gamification and Learning: a review of issues and research	2015	Critical Review	N/A	N/A	General	• Learner motivation • Learner engagement • Learning Outcomes	N/A	The potential of gamification to improve learning experiences and outcomes has not been fully established experimentally, and there is a need for customization of gamified learning to consider how different students are affected. However, the empirical studies on gamification suffer from methodological limitations such as small sample sizes, no well-validated psychometric measurements, and a lack of clarity in research reports.	Further research is needed to address these limitations and provide a clearer understanding of how to use gaming elements in the educational process.
Fui-Hoon Nah et al.	Gamification of Education: A Review of Literature	2014	Review	N/A	N/A	Game Elements	• Game elements	N/A	These findings discuss different gamification elements that can be used to enhance learner motivation and engagement in educational settings. The elements include points, levels, badges, leaderboards, prizes and rewards, progress bars, storyline, and feedback. While each element has its benefits, the effectiveness of these elements can vary depending on the context and the individual learners. The immediate and frequent feedback was found to be particularly helpful in engaging learners and enhancing their learning effectiveness.	Future research should focus on the impact of gamification in education by using a design science approach and scientific methodologies such as experiments and surveys to evaluate their effectiveness. There is a need to conduct systematic evaluations to establish a clear understanding of the impact of gamification in education.

Fui-Hoon Nah et al.	Gamification of Education Using Computer Games	2013	Review	N/A	N/A	General (focus on Game elements)	• Game elements	N/A	The main principles of gamification are goal orientation, achievement, reinforcement, competition, and fun orientation. These principles can enhance learners' motivation and engagement when incorporated into educational games. Positive reinforcement through points or virtual currency can promote learning, while negative feedback can offer corrective information. Competition can increase engagement, and having a fun component is crucial for effective gamification.	Evaluate the impact of different system design elements on learner engagement and learning achievement
Gerber	Problems and Possibilities of Gamifying Learning: A Conceptual Review	2014	Review	N/A	N/A	General	• Game elements	N/A	Gamification offers a crucial concept and potential benefit in its ability to engage and motivate individuals. Additionally, gamified experiences hold the potential to tap into collective intelligence, enhancing their effectiveness.	To use gamification successfully in education, educators must move beyond basic game mechanics and understand what makes games truly immersive and successful. This requires a deeper understanding of play theorists, learning principles in video games, and sound research in human psychology. Gamification should not be used as a quick fix for a broken system, but rather should be approached in a way

										that is tailored specifically to education. To do this, educators should invest time in playing games and becoming more cognizant of their own learning experiences.
Glover	Play as you learn: gamification as a technique for motivating learners	2013	Review	N/A	N/A	Motivation	• Leraner motivation • Learner engagement • Game elements	N/A	Gamification is a useful approach to enhance engagement in learning, but it should not be considered as the only solution. It can increase motivation and encourage positive behavior if implemented thoughtfully, using appropriate game elements. However, it is important to consider the level of intrinsic motivation among learners, as excessive extrinsic motivation can demotivate them. Additionally, rewards should be both attainable and desirable but not too common, allowing learners to feel a sense of pride and accomplishment upon receiving them.	N/A
Hallifax et al	Adaptive gamification in education: A literature review of current trends and developments	2019	Review	N/A	20	Personalization	• Adaptive gamification • Game elements	N/A	The research paper investigates the current types of adaptive gamification in education and their impact on learners. The two main categories of adaptation are static and dynamic, with static adaptation based on learner profiles and dynamic adaptation based on learner performance. The	Future research should focus on developing richer learner models, exploring dynamic adaptation methods, conducting longer and more structured studies, and standardizing metrics for measuring

									majority of systems use this information to select appropriate game elements, with only a few adapting the game elements themselves. Short studies tend to show positive impacts on learner motivation and performance, while longer studies show more mixed results. The impact of adaptive gamification is typically measured in terms of learner motivation and performance.	learner performance and motivation.
Hamari et al.	Does Gamification Work? — A Literature Review of Empirical Studies on Gamification	2014	Review	N/A	24	General	• Game elements • Psychological outcome • Behavioral outcome • Type of studies • Context	N/A	The majority of studies reviewed in the paper reported positive effects of gamification, but only in part of the relationships between gamification elements and outcomes. Confounding factors were found. However, the review also revealed that there are methodological limitations in many studies that need to be addressed in future research. These limitations include small sample sizes, lack of validated psychometric measurements, lack of control groups, lack of clarity in reporting results, and short experiment timeframes. Additionally, no single study used multi-level measurement models including all motivational affordances, psychological outcomes, and behavioral outcomes.	Future research on gamification should examine the impact of the context of the gamified system through experimental conditions and employ a more robust methodological approach to refine the research. It is also important to ensure that future studies use comparable methods, as many of the existing studies relied on qualitative methods.

Howard-Jones and Jay	Reward, learning and games	2016	Review	N/A	N/A	Game element (reward)	• Game elements	N/A	The authors examined the significance of rewards in education, particularly in educational games, from a cognitive neuroscientific perspective. They highlighted the importance of rewards and motivation in education to influence behavior and achieve long-term goals. They suggested that understanding how rewards impact memory formation can help in implementing gamification in education, and that interventions with uncertain rewards can be effective but lack adequate evidence.	N/A
Hung	A Critique and Defense of Gamification	2017	Critical Review	N/A	N/A	General	• Game elements	Higher Education	The article discusses the criticisms that gamification has faced as well as its potential as a tool for improving instructional design. It argues that the effectiveness of gamification depends on how it is designed and implemented, and that it is not inherently good or bad. In summary, gamification can be a valuable tool in education if used thoughtfully and meaningfully	Research on gamification should be expanded to involve instructors and researchers from diverse disciplines to examine its impact on a wider range of students. Additionally, future studies should investigate whether specific gamification designs are more effective in certain disciplines compared to others.
Kocakoyun and Ozdamli	A Review of Research on	2018	Review	PRISMA	313	General	• N/A	N/A	The results of this study indicate that quantitative research methods	Future research could focus on examining

	Gamification Approach in Education								are more commonly used in gamification studies. Most of these studies have been conducted with adults, which is expected since gaming is more prevalent among adults. Mobile environments are the most commonly used platforms for gamification research. The use of motivational theories is also frequent in gamification research. Additionally, the most commonly used game components include goal-setting, rewards, and progress tracking. Finally, the focus of gamification research has been primarily on mobile learning.	game designs that are appropriate for the gamification approach. More studies using achievement tests as a quantitative measure could also be conducted. Researchers could use this study as a guide to integrate different game components into educational environments, create independent learning areas, and explore different motivational theories. Additionally, there is a need for further research on the effectiveness of gamification in different learning areas and environments.
Koivisto and Hamari	The rise of motivational information systems: A review of gamification research	2019	Review	N/A	273	General	• Learner motivation • Game elements • Learner behavior • Methodological consideration	N/A	The empirical research on gamification has mainly focused on education, learning, health, and exercise, with a particular emphasis on individualistic motivations such as self-care and self-management. Current research is largely centered on implementing the core elements of gamification such as points, badges, and leaderboards, with a focus on the positive impact of	Future gamification research should focus on exploring cooperative and collective gamification approaches, diversifying the use of gameful affordances, and widening the thematic perspective of the domains being investigated. It should

									technology on human motivation and behavior. Gamification attempts to support people's goals and tasks through motivational information systems. Feedback provided by gamification can be cognitive, affective, and social. Most cited definitions of gamification describe it as a process that linearly affects psychological states, experiences, and behavior. However, most empirical research is conducted without control groups, and study designs often lack control between various affordances implemented in the studied systems. Sample sizes and experimental timeframes have also been limited in gamification research, and reporting of methods, data, analysis, and results in research papers is often unclear.	also explore the potential negative effects of gamification and how to mitigate them, as well as pay attention to the pre-determinants/requirements of gamification success and the role of the user in the effectiveness and adoption of gamification. Moreover, future research should incorporate the context of gamification deployment and the different types of feedback, while acknowledging the dynamic, cyclical nature of gamification. Additionally, future gamification research should aim for consistency in measurement instruments and research models, employ controlled experimental research methods, and increase sample sizes and time spans of studies. Finally, clear and comprehensive reporting of research should be prioritized in

										future gamification research
Laine and Lindberg	Designing Engaging Games for Education: A Systematic Literature Review on Game Motivators and Design Principles	2020	Review	Kitchenham and Charters (2004)	41	Game Design	• Challenge • Control • Creativity • Exploration • Fairness • Feedback • Goals • Learning • Profile and Ownership • Relevance and Relatedness • Resources and Economy • Social Play • Storytelling and Fantasy	N/A	The findings suggest that game designers should aim to engage diverse players by providing adjustable difficulty levels, simple challenges, and sufficient time to solve challenges. Feedback and goals are important motivators, and repeatable challenges containing learning content are particularly beneficial. Game designers should also consider player control and input modalities suitable for the target players and contexts. Fostering creativity and self-expression in players, providing means for players to contribute to game content, and promoting exploration through freedom and experimentation are important. Appropriate feedback, clear and achievable goals, and avoiding one-size-fits-all gameplay experiences are also important. Finally, the study highlights principles for two categories of game design: Learning and Profile/Ownership, as well as the importance of relevance and relatedness, social play, and storytelling and fantasy.	One potential avenue of research is to examine the effectiveness of gamification in practice by investigating how well the game is received by its intended users. Another area of research could focus on the game development process itself, including the elicitation of requirements, design, implementation, and testing.
Limantara et al.	The elements of gamification learning in higher	2019	Review	N/A	30	Game elements	• Game elements	Undergraduate students	Researchers have used various participation patterns, such as compulsory, voluntary, random,	Future research can explore various aspects of learning

	education: a systematic literature review						• Model of student participation • Educational course		and rewarded, to explore the implications of gamification in learning. The majority of studies found that gamification positively impacted the learning process, including increasing student learning, motivation, and engagement in the classroom	gamification. Analyzing gamification tools and software and their effects on student motivation and scores, examining how different types of courses influence the effects of gamification, and incorporating database sources to improve research results. Additionally, future researchers may apply gamification elements in the information systems field and compare their impact with previous research.
Majuri et al.	Gamification of education and learning: A review of empirical literature	2018	Review	N/A	128	General	• Game elements • Psychological outcome • Behavioral outcome	N/A	The findings of the analysis suggest that gamification in education aligns with the general research on gamification in terms of implemented features and psychological outcomes. The focus on behavioral outcomes in education is mainly on quantifiable educational outcomes such as course and assignment grades. The majority of the studies report positive results, but there is also a significant amount of research with null or mixed results.	In future research, contextual factors such as demographic, personality, and associations attached to the task or activity, as well as different learning styles, should be taken into account to better understand the varying results of gamification. Study designs should be improved to isolate the effects of specific elements in educational

										settings, with more controlled designs and multiple sources of data. The scope of affordances implemented in education should be expanded, focusing on more social and immersive elements, and inducing social interaction with gamification solutions.
Metwally et al.	Revealing the hotspots of educational gamification: An umbrella review	2021	Review	PRISMA	46	General	• Game elements • Educational course • Learning outcomes • Game design	N/A	The findings suggest that gamification in education can improve motivation and engagement, with emphasis on extrinsic factors like reward, achievement, and progression over social and immersion-based factors. There is a tendency to repeat research designs and underuse certain methodologies. Standard design models and innovative approaches are both important. While systematic reviews with quantitative methods are prevalent, there is a shortage of other types of reviews. Promising review studies have used rigorous scientific validity evaluation methods and analyzing sets of methods. Researchers are encouraged to develop a validated checklist to assess the quality of gamification research.	Future research should focus on the development and maintenance of game mechanics and student satisfaction in gamification. Additionally, gamification should be implemented in emerging technologies, and adaptive and personalized gamification in education requires further exploration. Theoretical frameworks need to be developed to connect educational gamification practices with learning theories, and storylines should be integrated to immerse

										students in learning tasks. Design and user interfaces of gamified applications need to be enhanced, and universal design principles should be expanded. Future research should consider consistency in research measurement tools, sample sizes, and study duration, and incorporate precise quantitative data and measurements. The impact of each affordance on learning outcomes and student interests should be studied individually, and gamification studies should be expanded to primary and secondary education. Personalized gamification experiences and their effects on learning outcomes and student interests, player taxonomies for personalization, and the relationship between higher education learners' types and gamified experiences should be explored.

Ofosu-Ampong	The Shift to Gamification in Education: A Review on Dominant Issues	2020	Review	Kirriemuir and McFarlane (2004).	32	Game elements	• Cognitive • Attitude • Performance • Motivation • Engagement • Interpersonal relationship	N/A	The literature review showed that effective game mechanics like virtual goods, trophies, and redeemable points can improve engagement and performance, while poor game features or mechanics result in failed educational goals. The use of penalties and award systems was found to improve student participation and attendance. The most prevalent gamification elements were awards, points, badges, levels, and quests, with points, leaderboards, and badges being the dominant ones. However, there is a need for a systematic experimental approach to identify the most effective game elements or configurations for promoting engagement and supporting learning. Overall, some elements of games were found to be more effective than others in motivating learners.	Future research is needed to understand the specific game elements that are preferred by learners in different contexts and educational institutions, as well as to identify design principles governing gamification. Longitudinal studies and systematic experimental approaches can help map out the effectiveness of game element configurations in supporting learners. A more complex model that includes moderating and mediating variables should also be developed. Additionally, gamified courses should focus on students' progression and provide quick feedback to encourage engagement.
Ortiz et al.	Gamification in higher education and stem: a systematic review of literature	2016	Review	N/A	30	General (focus on STEM)	• Game elements • Sample size • Educational level • Data collection methods	Higher Education	Gamification studies often use a combination of elements including badges, points, and leaderboards, as well as challenges, levels, and avatars. However, only badges have been widely studied in	Future research should focus on increasing technological support to involve other STEM areas in empirical studies of gamification.

									isolation. Points, challenges, quests, and leaderboards are less frequently studied as unique elements. The majority of gamification studies are conducted in computer science courses, with limited presence in other STEM fields. Researchers report a mix of positive and negative results from gamification interventions, with few studies using validated psychometric measurements to assess personality, flow, motivation, and goal orientation. The effectiveness of gamification interventions depends on individual study designs and many variables are not considered when designing gamification studies, such as motivation, player types, and personality. Improved research designs generally result in more positive and mixed results.	To improve the consistency of results, future studies should consider more variables such as motivation, player types, and personality in their designs. A promising approach for measuring motivation is the intrinsic motivation inventory developed by Ryan (1982), which measures different dimensions of motivation. Overall, more comprehensive research designs are needed to fully understand the potential benefits and limitations of gamification in education
Park and Kim	A Badge Design Framework for a Gamified Learning Environment: Cases Analysis and Literature Review for Badge Design	2019	Review	N/A	943 (badge cases)	Game design (Badge)	• Learning activity • Individual or interactive learning • Time and Effort	N/A	the study recommends eight badge types for three badge design conditions and finds a significant difference in chi-square (1117.7, P<.001). The use of badges can promote self-directed learning by improving learning sustainability, motivation, and goal-setting. Badges also offer benefits such as flexibility in learning, visualization of completed goals, and planning for future activities. The study found that badges had a positive	N/A

									impact on critical thinking, teamwork, leadership, and unrecognized abilities or knowledge/skills.	
Rapp et al	Strengthening gamification studies: Current trends and future opportunities of gamification research	2019	Critical Review of Gaimfication	N/A	N/A	General	• Theories • Methodological rigor	N/A	In summary, the gamification research space lacks theoretical and methodological rigor, but there is a positive trend towards more effective and engaging gamified systems. However, there is a tendency to use a limited number of theories and constructs in gamification designs, and there is a need for a common language for research. Tangible user interfaces and wearable technologies are emerging as new approaches in game research, with a focus on physicality and interaction opportunities.	There is a call for a comprehensive study of design to improve game design.
Saleem et al	Gamifcation Applications in E learning: A Literature Review	2022	Review	N/A	N/A	General	• Leraner motivation • Learner engagement • Game elements	N/A	In summary, gamification in education remains a controversial topic as the use of gamification elements has failed to improve students' sense of group and has not substantially enhanced their talents, desire for achievement, and inner inspiration. The study suggests that it is crucial to find a way to meet each player's needs to ensure the success of gamified learning. The leading cause of why learning by gamified applications has been unsuccessful is the use of game elements, instructional	N/A

									design, and technical problems. Therefore, educational designers need to gain an empirical understanding of outcomes, learning goals, and content when assessing individual play selection to improve the effectiveness of gamification in education.	
Sanmugam et al	Gamification as an Educational Technology Tool in Engaging and Motivating Students; an Analyses Review	2015	Review	N/A	N/A	Motivation, Engagement and Personalization	• Learner motivation • Learner engagement • Game elements • Learner behavior	N/A	In summary, the study suggests that gamification elements have the potential to increase student motivation and engagement. Bartle's player motivation types can be used to identify and cater to different student skills and personalities, helping to identify the type of users of the system. Games can impact the cognitive, emotional, and social aspects of players, making them suitable for gamification in education. To ensure the success of meaningful gamification, it is crucial to prioritize the needs of users over the needs of an organization. Focusing solely on game mechanisms can create a false scenario in achieving a goal. Therefore, when considering whether gamification can benefit a group of students, it is essential to identify their levels of motivation.	Future studies should focus in identify the level of motivation before applying gamification.
Saputro et al	A review of intrinsic motivation	2017	Review	N/A	36	Game elements and Motivation	• Game elements • Learner motivation	N/A	In summary, while gamification has shown to have a positive effect on motivation, some researchers	Future research should focus on exploring various approaches and

	elements in gamified online learning								have obtained different results, and issues raised vary depending on factors such as age, gender, personal traits, and gaming personalities. The PBL Triad (points, badges, and leaderboards) is the most commonly used game design elements in gamification. Researchers continue to modify and propose new game design elements to enhance students' motivation and engagement in online-based learning. These elements include level, unlock level, meaningful choice, progress bar, skill tree, AvatarWorld, narrative, leaderboards, onboarding, quests, mission, lives, badges, performance graphs, XP, grades, level, dashboards, collaborative work, competition, social status, quests, storyline, avatar, teammates, and virtual maps. These elements aim to give students a sense of autonomy, competence, relatedness, and purpose in gamified online learning.	methods to validate the interaction of gamification elements systematically and scientifically, rather than just examining its overall effect. Gamification design should emphasize the relationship between game dynamics, gamification contexts, gaming personalities, personality traits, gender, situational conditions, and activity characteristics to provide a strong foundation for designers. Future studies should consider the number of participants involved and the duration of experiments to assess the long-term effects of gamification. There is a need for standardization in the field of gamification to assess the success rate of studies accurately, allowing for comparison of results and proper meta-analysis. Future research should determine game elements that can

										enhance intrinsic motivation and explore the long-term effects of gamification utilization, particularly on students' intrinsic motivation.
Stott and Neustaedter	Analysis of Gamification in Education	2013	Review	N/A	N/A	Game design	• Game dynamics	N/A	Game dynamics such as Freedom to Fail, Rapid Feedback, Progression, and Storytelling are effective elements that educators should prioritize in their gamification efforts. These game dynamics are based on established pedagogical practices and have been demonstrated to be effective.	N/A
Surendeleg et al.	The Role of Gamification in Education – A Literature Review	2014	Review	N/A	N/A	General	• Game elements • Learner engagement • Learner motivation	N/A	The use of game elements in learning can enhance student motivation and engagement. Feedback, leaderboards, points, and levels are commonly used game elements in gamified applications. Empirical evidence shows that games can effectively enhance learning and increase engagement levels, leading to a positive attitude towards learning and increased productivity.	Future research should focus on investigating the impact of gamification on the development of lifelong skills in adult learners.
Willert	A Systematic Literature Review of Gameful Feedback in Computer Science Education	2021	Review	N/A	50	Game elements (feedback)	• Feedback • Game elements	N/A	There are different types of feedback and learning support methods. The first discusses how assessments of student responses, known as formative feedback, can	As a suggestion for future research, it is recommended to explore the use of contextualized game-elements as a way of presenting feedback to

									shape and improve student competence. The second defines summative feedback as a summary of a student's achievement or course unit status, which may not have an immediate impact on learning but can affect future decisions. The third describes immediate feedback as feedback given virtually at the time of a test, or fast enough to impact the student for the next task. The fourth talks about self-regulation feedback and its role in enhancing student self-regulation by supporting monitoring and adjusting of learning goals and actions. The fifth explains scaffolding, a support structure used to aid student learning, which can be gradually faded out as student competence increases. The final discusses social/peer feedback, which involves feedback on tasks and assignments given by one student to another. The sentences suggest that certain approaches have a positive impact on students' learning behavior. Feedback and game-elements have a positive influence on student results and motivation, particularly in creating a goal orientation. The use of game-elements contributes to increased motivation by creating	students when creating tools or environments to support their learning process.

									a goal-oriented environment for students.	
Wilson et al	Gamifcation Challenges and a Case Study in Online Learning	2015	Review	N/A	N/A	General	• Learner engagement • Learner motivation • Game elements	N/A	Gamification can enhance online learning by promoting early engagement and effective teamwork through game-based mechanics. However, there is no guarantee that gamification will lead to success in an online course.	Future research on gamification should focus on carefully planning and designing the game mechanics and rewards that align with the beliefs and values of participants. By selecting metaphors and game characteristics that support how players feel about the tasks, and blending user interaction design strategies with game-based mechanics, a successful gamification experience can be achieved.
Antonaci et al.	The Effects of Gamification in Online Learning Environments: A Systematic Literature Review	2019	Systematic Review	PRISMA	61	Game elements on learning behaviors	• Game elements	N/A	Effects of badges/reward, learderboards, point/score/ranking on different learning behaviors (motivation, engagement, performance, attitude toward gamification, collaboration and social awareness)	Rigorous research on how to design proper gamification and to study the effects on human behavior. Different implementation of game elements than PBL. Identify game elements that can generate a sense of community and interdependence.

Behl et al.	Gamification and e-learning for young learners: A systematic literature review, bibliometric analysis, and future research agenda	2022	Systematic Review	PRISMA	32	Personalization	• Game elements • Personalization • Learner Style • Learner Engagement	Young Learners	Personalization in e-learning allows for users' needs to be fully met, which in turn increases their satisfaction.	More studies on game elements and tailored gamification
Borges et al.	A Systematic Mapping on Gamification Applied to Education	2014	Systematic Mapping Review	N/A	26	General	• Learning outcomes • Behavioral outcome • Learner engagement • Socialization	N/A	Most research on gamification in education has been focused on higher education, with only a small number of studies conducted in elementary education. The primary objective of the majority of studies is to evaluate student engagement through gamification. However, there is a lack of validation research to propose and test novel techniques and methods in well-designed experiments.	Involving teachers as end-users in further research is essential to gain valuable insights into the application and impact of gamification in learning environments, thereby improving its use in educational settings.
Bozkurt and Durak	A Systematic Review of Gamification Research: In Pursuit of Homo Ludens	2018	Systematic Review	Lexical Analysis, Keyword Analysis, Social Network Analysis, Citation Analysis	N/A	General	• Learner Motivation • Learner Engagement • Behavioral outcome	N/A	Conceptual/descriptive methodologies are the most commonly used approach in gamification research, but other types of methodologies are increasing. Education, teaching, and learning; engagement, motivation, and behavior change; and gamified designs are emerging patterns in gamification research. The main focus of gamification research revolves around issues such as engagement, motivation,	To enhance the gamification field, research on emerging phenomena should be explored using different methodologies, including qualitative, mixed, data mining and analytics, and practice-based methods. While conceptual/theoretical and quantitative research paradigms have been used,

									behavioral change, and user experience. It was found that almost half of the articles lacked theoretical or conceptual frameworks.	incorporating these additional methodologies can contribute to the field.
Cavalcanti et al.	Automatic feedback in online learning environments: A systematic literature review	2021	Systematic Review	StArt (State of the Art through Systematic Review)	63	Game element (Feedback)	• Learner performance	N/A	Most of the reviewed papers found that feedback had a positive impact on students' performance, but it's unclear if it was due to the tool or final feedback product, while some papers reported increased performance but with some dissatisfaction with the feedback message, and 41.26% of the articles used feedback to support self-regulation, which is aligned with good practices of feedback	Future research should align proposed feedback systems with educational research to improve the final results of the feedback process in terms of learning outcomes, learning processes, and student satisfaction. Educational research has identified factors that should be considered when creating feedback, such as increasing student awareness about the learning goal, progress, and subsequent goals required to achieve the overall goal. Additionally, future research should consider feedback as a dialogic process, which is not addressed in the papers reviewed.
Denden et al.	The role of learners’	2022	Systematic Meta-Review	PRISMA	40	Personalization	• Learner’s characteristic • Player types	N/A	Individual learner characteristics like gender, personality traits, gaming frequency, and player	Future research should consider the individual differences to design

	characteristics in educational gamification systems: a systematic meta-review of the literature						• Gaming frequency • Social type • Learning outcomes • Personality traits		types can influence their experiences with gamification systems. Although gamification has a positive impact on education, negative effects can occur due to individual differences and behaviors during computer-based learning.	gamification systems that cater to diverse learners' needs.
Devenderen and Nasri	Systematic Review: Students' Perceptions of the Use of Gamification	2022	Systematic Review	PRISMA	29	General	• Game elements • Learner motivation • Learner engagement	N/A	This systematic review found that gamification is effective in improving students' motivation, involvement, attitudes, and interests. However, there is still room for improvement in areas such as internet access, teaching methods, and digital support resources. The imbalance in selecting the correct gamification elements needs to be addressed, and the effectiveness of gamification needs to be studied based on learning theory.	Future research on gamification should focus on implementing it as a classroom assessment method to reduce students' anxiety and increase transparency in the assessment process. Additionally, using gamification elements in project-based learning (PBL) in primary schools can motivate students and improve their skills, both individually and in groups, while also helping weaker students. The reward system in gamification can increase student responsibility and interest in completing PBL assignments.

Indriasari et al.	Gamification of student peer review in education: A systematic literature review	2020	Systematic Review	N/A	39	General	• Educational course • Learner engagement	Higher Education	The study found that artifact assessment and creation are the most frequently gamified actions in the context of peer review models, and that quantity and quality of both artifacts and feedback are the most popular reward criteria. Science, Technology, Engineering, and Mathematics (STEM) are the most commonly reported areas for gamified peer review activities. Although existing literature reports positive effects of gamification on student engagement, the study suggests a narrow range of student actions that have been incentivized. Further research is needed to explore other actions that can be incentivized in peer review activities.	Future research can explore the potential of gamification to encourage student reflection on feedback received. These activities have been understudied in the context of gamification and can offer valuable directions for further investigation.
Inocencio	Using Gamification in Education: A Systematic Literature Review	2018	Systematic Review	N/A	95	General	• Theories • Learning outcomes • Learner motivation • Learner engagement	N/A	The study identified motivation, engagement, self-efficacy, and flow/cognitive absorption as the most important constructs in gamification research, with reliable scales and consistent theories. While satisfaction and attitude are commonly used, they are not as effective.	The study also suggests that future research should explore the effect of extrinsic rewards on experiential outcomes and investigate learning performance as a downstream effect of studying behavior. Future research should focus on exploring transfer and other high-order learning outcomes in a gamification

										context, as these areas have not been extensively studied.
Kalogiannakis et al	Gamification in Science Education. A Systematic Review of the Literature	2020	Systematic Review	PRISMA	24	General	• Game elements • Educational level • Science education	Primary School Secondary School Higher education	The study provided insights into the application of gamification in science education, including the popular content areas, educational levels, and current trends. Additionally, the study found that gamification has the potential to improve data collection through assessment tools, leading to more information about students' learning processes. Finally, the study identified the primary gaming elements used in science education through gamification. In science education, creating a competitive environment is a controversial topic often used to improve learning outcomes by combating negative emotions and experiences. The most affected learning outcomes were identified as motivation and engagement, learning achievements, and social interaction.	Future research should consider using multiple databases, such as JCR or Scopus, in addition to Google Scholar, to minimize bias. Furthermore, future researchers could broaden their scope by including other types of publications such as dissertations and conceptual papers to gain more extensive information and insight.
Khaldi et al	Gamification of e-learning in higher education: a systematic literature review	2023	Systematic Review	N/A	90	General	• Game elements • Theories	Higher Education	The study found that most applied gamification research is not based on theory and has not used gamification frameworks in the design of gamified learning systems. Although some studies attempted to adapt psychological and educational theories as	Future research should focus on the pedagogical aspect of learning systems and the task under gamification. The effectiveness of theory-driven versus data-driven gamification

									gamification approaches, the resulting systems are not clear, and there is no clear rationale for choosing specific game elements. As a result, it is concluded that these gamification approaches are not effective in helping designers and practitioners gamify their learning systems.	approaches needs to be compared, and a hybrid approach proposed to solve design issues. Efforts should also focus on building a holistic approach by considering personalization, gamified subject, educational context, learner culture, preferences, level, playing motivations, and experience with games. Finally, statistical analyses and comparative studies should be conducted to validate existing gamification approaches in the literature.
Krath et al	Revealing the theoretical basis of gamification: A systematic review and analysis of theory in research on gamification, serious games and game-based learning	2021	Systematic Review	ROSES	32	Theory	• N/A	N/A	Gamification can transparently illustrate goals and their relevance, allow users to set their own goals, provide direct feedback on their actions, reward users for their performance and communicate the relevance of their achievements, allow users to see their peer's performance, connect users to support each other and work towards a common goal, adapt tasks and complexity to the abilities and knowledge of the user, nudge users towards the	Future empirical research should focus on testing, challenging, and refining the principles of gamification that have been theoretically deduced. The aim is to gain a more concrete and precise understanding of the "how" and "why" of gamification.

									actions necessary for achieving the goals, allow users to choose between several different options to achieve a certain goal, and simplify content as gamification systems are usually easy to use.	
Manzano-León et al.	Between Level Up and Game Over: A Systematic Literature Review of Gamification in Education	2021	Systematic Review	PRISMA	14	General	• Game elements • Learner motivation • Learner engagement • Learner achievement/performance	N/A	The study found that gamification can have positive effects on student motivation, engagement, and academic performance at different educational levels, particularly in university education where there is a greater focus on increasing academic achievement.	Future research could focus on the relevance of player types and how to personalize educational gamification programs based on the individual characteristics of each student. It is also important to consider demographics such as age, gender, and previous experience with video games. Additionally, the suitability of gamification for all students should be explored based on the gamification elements used and the students' educational needs and interests. Finally, further research is needed to understand the potential of gamification for long-term retention of positive results.

Mohammed and Ozdamli	Motivational Effects of Gamification Apps in Education: A Systematic Literature Review	2021	Systematic Review	N/A	37	Motivation	• Game elements	N/A	Gamification apps are increasingly being used in schools, colleges, and higher learning institutions, and are yielding positive results. The game elements associated with gamified teaching methods are changing students' attitudes and behaviors towards learning activities, increasing their desire to learn, and inspiring them to participate more. Gamification apps have the ability to occupy and motivate individuals, making teaching and learning more effective. Both tutors and students using gamification systems have developed advanced curiosity and desire to learn, and the use of videos as lecture notes has allowed for collaboration, independent learning, and skill testing.	Future research should focus on investigating the motivational effects of gamification systems on tutors using either a quantitative or qualitative approach. Additionally, future studies should explore how gamification systems are implemented in organizations.
Mora et al.	Gamification: a systematic review of design frameworks	2017	Systematic Review	N/A	27	Game Design	• Fun • Learner Motivation • Socialization • Behavioral outcome • Personalization	Higher Education	The study identified ten ingredients relevant for successful game design, including self-representations, narrative, feedback, and social interaction. The gamification design process should consider principles, clear objectives, desired behaviors, profiling of players, and measurement metrics. Fun and motivation were also important factors, with most frameworks incorporating social interaction and storytelling. While some frameworks emphasized the	Future research could focus on the development and extension of a complete framework for personalization in higher education environments. This would involve considering principles and knowledge acquired from previous work and applying them to diverse case studies. Current literature

									importance of analytics and user experience, ethical issues were not widely considered. Further research is needed to investigate the impact of gamification on user experience.	focuses on ad hoc experiences rather than formal design processes, so a more comprehensive framework could be beneficial.
Nair	Learning through Play: Gamification of Learning A Systematic Review of Studies on Gamified Learning	2021	Systematic Review	N/A	64	General	• Attitude • Learner behavior • Learner motivation • Learner engagement • Perception • Reaction	School and Higher Education in 61% of articles	The results of the studies suggest that gamification has a significant positive impact on the learning process, learner engagement, and motivation. Gamification was found to increase learning outcomes and bring about desired changes in employee behavior. A majority of the studies evaluated found that gamification has a significant impact on learning, with 47 studies showing statistically significant results between gamification and changes in the dependent variable. Additionally, gamification was found to improve learner engagement and reaction to the training	More studies using true experimental design are needed to establish a causal relationship between gamification and learning outcomes. Additionally, there is a need to identify other environmental factors that could moderate the effectiveness of a gamified module, which would help practitioners design their interventions better.
Nurtanto et al.	A Review of Gamification Impact on Student Behavioral and Learning Outcomes	2021	Systematic Review	Kitchenham 2010	40	Learning Outcomes and Behavioral	• Affective • Behavioral • Cognitive • Student achievement/performance	N/A	The application of gamification in learning processes has a positive impact on the affective, cognitive, and behavioral domains of learning. There has been a significant increase in enthusiasm and internal motivation as key variables in the affective outcomes of gamification. In terms of	N/A

									cognitive outcomes, gamification has a positive impact on student retention. Additionally, gamification has been found to improve behavioral outcomes such as teamwork, communication skills, social skills, and digital literacy. Finally, gamification has been proven to increase student engagement, intrinsic motivation, extrinsic motivation, interest, enjoyment, satisfaction, and innovation in learning activities.	
Oliveira et al.	Tailored gamifcation in education: A literature review and future agenda	2022	Systematic Review	Kitchenham (2004)	19	Personalization	• Learning style • Player Type • Demographic factors • Psychological states • Game elements • Country	N/A	A total of 21 studies from different countries were analyzed to identify the effectiveness of personalized gamification in enhancing learning outcomes. The majority of the studies considered only gamer types for personalization and neglected other important human aspects such as culture and gender differences. The results showed that tailored systems can be better in some cases, but non-tailored systems were more effective in others, indicating the relevance of adapting gaming features to enhance learners' engagement. However, the effectiveness of gamification personalization in improving students' learning outcomes remains inconclusive.	The review highlights the importance of considering individual human characteristics and conducting comparative studies to identify whether tailored gamified educational environments are better than non-tailored ones in terms of improving learning outcomes. Future research should explore the potential of automation to improve the design process and consider conducting longitudinal studies to examine the long-term effects of personalized gamification on students' learning

										outcomes. Additionally, the development of frameworks to support gamification designers in tailoring the educational environments is essential.
Ortiz et al.	Gamification and learning performance: A systematic review of the literature	2017	Systematic Review	five-stage framework of Arkey & O'Malley (2005)	23	Learner achievement/ performance	• Educational level • Game elements • Educational course • Duration • Sample size • Learner motivation • Learner engagement • Country	Higher education (N = 3951) 82.5% High School (N = 129) 2.7% Middle School (N = 709) 14.8%	Higher education institutions are the most likely to adopt gamification as a way to address student motivation and engagement issues. However, studies rarely focus on specific gamification elements, making it difficult to understand their impact on learning outcomes. Computer science fields are the most likely to adopt gamification innovations. Researchers tend to avoid a novelty effect by involving subjects over a longer period. The number of study respondents is often limited. Motivation and engagement are the most studied additional variables, and gamification has a mediating or moderating effect on learning. While only 9 studies showed a positive impact, it is important to analyze the reasons for negative or mixed results, including mediating variables, choice of measurement instrument, sample size, and study length. Overall, a controlled study design that considers sample size,	Future research should focus on investigating the direct and indirect effects of gamification on learning performance. It is essential to empirically support these effects through further research. Additionally, it is crucial to study specific gamification elements in isolation to determine their distinct impact on learning outcomes. To increase the reliability and generalizability of results, future studies should include larger sample sizes and longer intervention periods to avoid novelty effects.

									variables, and study length shows a promising increase in learning performance with gamification.	
Rozi et al	A Systematic Literature Review on Adaptive Gamification: Components, Methods, and Frameworks	2019	Systematic Review	N/A	25	Personalization	• Learner profile • Learnng style • Learner behavior • Learner skills/knowledge	N/A	In summary, this study identified 11 types of methods used in adaptive gamification, including scoring, clustering, and rule induction algorithms. The proposed framework by Hassan et al. consists of three elements for an adaptive gamification experience based on the dimensions of learning, including an adaptive gamification engine, adaptive component, and gamification display. Personalized adaptive gamification has the potential to increase motivation and performance, and clear design frameworks tend to be more successful. The components used in adaptive gamification include player/learner profiles, learning style, behavior, and skill/knowledge. The most popular methods in adaptive gamification are scoring and the Felder-Silverman Learning Style Model	In summary, future research should focus on the personalization of gamification in learning, as different motivations can lead to different responses. It is essential to personalize the learning method for each employee based on their individual character to improve the effectiveness of the gamification approach.
Saxena and Mishra	Gamification and Gen Z in Higher Education: A Systematic Review of Literature	2021	Systematic Review	N/A	29	General	• Country • Educational course • Game elements • Duration • Learning outcomes	Higher Education	Results suggest that games have the potential to enhance learners' motivation and engagement, enriching their intellectual activities in a classroom setting. Game-based techniques can be personalized to the learner's skill	Future research should investigate the potential correlations between gamification and student performance and identify which gamification element

									level to prevent frustration and boredom, with community-based performance evaluation and feedback providing a balance of individual and community involvement.	contributes most significantly to improvement. There is a need for further study to determine if gamification through its elements can impact actual learning outcomes while maintaining student engagement through motivation. Researchers should also examine whether these findings are applicable and sustainable in other subjects and among students from diverse cultural and educational backgrounds. To identify the impact of individual gamification elements, researchers should design studies that include player types, learning preferences, and personality as mediating or moderating variables. To ensure high-quality research, larger sample sizes and better research instruments should be utilized. Longitudinal studies should be performed to assess the long-term impact of gamification.

Seaborn and Fels	Gamification in theory and action: A survey	2015	Systematic deductive analysis	Meta - Synthesis	31	General	• Game elements • Contextual factors	N/A	Results indicate a positive-leaning but mixed view of gamification's effectiveness, with context-specific implementations impacting participants differently. The effectiveness of gamification varied among individuals and was influenced by demographic variables and expectations. The current state of gamification research design employs mixed methods and single-study approaches, utilizing various measures and instruments to capture quantitative and qualitative data in one-off experiments. User-centered design methodology may help identify intrinsic motivators for a given user population. There may not be an ideal gamified system, but instead, gamified systems may need to be selectively designed or flexible and inclusive to accommodate individual users' needs and preferences.	Future research should explore the range of contexts and game elements used in gamification and address design issues, such as statistical analyses and isolating the gamification effect. Validated instruments, such as the Intrinsic Motivation Inventory, can be used to assess motivation in gamified systems. Additionally, research should aim to identify the most and least promising game elements in specific contexts for particular end-users. Findings from studies on intrinsic motivation can be extrapolated to the design and evaluation of extrinsically-motivating gamification elements. More empirical, mixed methods research that employs statistical analysis and reports effect sizes for standard elements, dynamics, and experiences is necessary. Comparative studies with controls are also needed to determine the unique

										impact of gamification in comparison to other approaches.
So and Seo	A Systematic Literature Review Of Game- Based Learning And Gamification Research In Asia	2018	Systematic Review	N/A	22	General	• Learning outcomes	K-12	The reviewed articles showed a positive impact of gamification on learning outcomes, but there are some research gaps in the educational game research in Asian K-12. These gaps include a lack of diversity in subject disciplines and game genres, reliance on media-comparison experiments, and concerns about sustainability and scalability. Future research is needed to address these gaps and to shift from a focus on content to context in educational game research.	Future research should aim to examine the impact of gamification through diverse research methods. Longitudinal studies are needed to examine the impact of game-based learning and gamification across timescales and multiple spaces. Researchers should provide lessons learned for scalability and sustainability, and also conduct research studies in resource-poor and under-developed countries.
Xu et al	Psychological interventions of virtual gamification within academic intrinsic motivation: A systematic review	2021	Systematic Review	PRISMA	105	Game elements and Intrinsic motivation	• Game elements • Learner motivation	N/A	The results indicate that gamification is linked to increased intrinsic motivation, which has been demonstrated through a variety of observational, self-report, data analysis, and systematic review measures. Gamification methods also reinforced extrinsic motivation by utilizing points, badges, and leaderboards to boost individual intrinsic motivation. The continued use of gamification methods contributed to the development of	N/A

									intrinsic motivation by fostering an internal drive to complete tasks and shifting the source of motivation from external to internal. In the majority of studies (n = 53), a combination of the three most popular gamification elements (points, badges, and leaderboards) were utilized, and they were identified as the most effective methods for increasing intrinsic motivation through gamification.	
Zainuddin et al	The impact of gamification on learning and instruction: A systematic review of empirical evidence	2020	Systematic Review	N/A	46	General	• Learner motivation • Methodological approach • Duration • Game elements • Learner achievement/ performance • Socialization • Personalization	N/A	Questionnaires were the most frequently used method in gamification research, followed by experimental tests, interviews, observations, and document analysis. Most intervention studies were conducted within a few weeks or months, with SDT and flow theory being the most commonly used psychological theories. Gamification studies mostly involved adult learners or higher education students, with points, leaderboards, and badges being the most frequently used game mechanics. Three positive themes emerged: learning achievement, motivation and engagement, and interaction and social connection. Gamification should align learning objectives with a student's intrinsic motivation and understanding of	Future research should focus on gamifying students' learning activities and exploring the effects of gamification across different course subjects. Theoretical foundations of gamification in education should be discussed and other theories explored beyond SDT, flow theory, and goal-setting theory. Longitudinal studies are needed to understand the long-term effects of gamification on learners. More investigation of gamification at the

									the content. Incentive systems can undermine intrinsic motivation. Instructional and motivational design theories to support gamification thresholds are scarce.	primary or secondary school level is also recommended, along with a critical examination of how gamified systems can be applied in low-tech information environments.
Bai et al.	Does gamification improve student learning outcome? Evidence from a meta-analysis and synthesis of qualitative data in educational contexts	2020	Meta-Analysis	PRISMA	24 (total N = 3202)	Learning outcomes	• Learner achievement/performance • Game elements • Educational courses • Educational level • Sample size • Duration of interventions • Country	From elementary to postgraduate students	This work studied the impact of gamification on academic learning outcomes in K-12 education. They found a medium effect of gamification on learning and identified several factors that moderate this effect. Their qualitative synthesis revealed four reasons students liked gamification and two reasons they disliked it. However, the large variability in effect size was not explained by the number or choice of game elements used, highlighting the need for more studies meeting inclusion standards to better understand which game elements matter.	Future research should explore the influence of user types or traits on their interest in gamification. Personalized gamified systems can be developed for individual users based on their user types or traits. Several studies have proposed different lists of gamification user types, and further research is needed to identify the most common traits or types found in the broader user

										population and how they can be best supported in a gamified system to design more personalized gamified systems catering to different users' preferences.
Dikmen	Does gamification affect academic achievement? A meta-analysis of studies conducted in Turkey	2021	Meta - Analysis	PRISMA	52	Learning achievement/performance	• Educational level • Subject course • Publication year • Class size	N/A	This study focused on the effect of gamification on academic achievement in Turkey. The results showed that gamification has a large positive effect on students' academic achievement, explaining 74% of the variance in academic achievement. The effect of gamification on academic achievement was found to be similar to previous meta-analyses conducted in Turkey but different from studies conducted in other countries, possibly due to cultural differences. Gamification was found to be effective in all student levels and across all publication years, but had varying effects on different subject disciplines. Furthermore, publication bias may influence the effect size of studies with larger samples.	Future research could investigate the effect of gamification on academic achievement across different courses and in various countries. Due to the limited statistical data in some studies, it is essential for future research to provide comprehensive statistical results to ensure their inclusion in meta-analyses. This would help to provide a more comprehensive understanding of the effects of gamification on academic achievement in different contexts.
Fadhli et al.	A Meta-Analysis of Selected Studies	2020	Meta-Analysis	PRISMA	6	Learning outcomes	• N/A	Children (6-10) years	The summary of the effects of six studies on gamification for	N/A

	on the Effectiveness of Gamification Method for Children								children aged 6-10 years shows that the standardized mean difference (SMD) is 1.01 with confidence intervals of 0.98-1.05. I^2 equal to .53 The gamification method has been found to have a significant and beneficial effect on children. The results of a meta-analysis indicate that the post-test design of gamification methods can effectively improve children's cognitive skills, attitudes, language, health, and social-emotional abilities. In general, the study demonstrates that gamification has a positive effect on children's learning outcomes, across all aspects of their learning.	
Huang et al.	The impact of gamifcation in educational settings on student learning outcomes: a meta-analysis	2020	Meta - Analysis	PRISMA	30 (total N = 3083)	Game elements and learning outcomes	• Attitudes • Behavior outcome • Game elements • Type of publication • Subject course • Educational level	N/A	The research paper found that while the effect size of gamification on educational outcomes was small to medium according to Cohen's criteria, other researchers have noted that effect sizes of .40 and greater are practically relevant. The deployment of gamification has potential beyond the pointification design found in most studies. Not all gamification design elements have the same effect on student learning outcomes, and leaderboards may undermine the intended goal of improving learning outcomes. Collaboration	The future research recommendations suggest considering gamification design elements beyond "pointification," such as collaboration and quests/missions/modules, and exploring alternative designs that mimic popular video games. The study highlights an alarming difference between implementations in undergraduate and K-12 contexts, which requires

									and quests/missions/modules are promising gamification design features with statistically significant effect sizes higher than the overall effect size. The effect size for undergraduates was statistically significant, but the effect size for K-12 students was not statistically significant and nearly half the size.	more thorough exploration in future research. The research should address which aspects of gamification to deploy with a target population and determine the combination of gamification design elements that have the most potential for facilitating learning outcomes.
Kim and Castelli	Effects of Gamification on Behavioral Change in Education: A Meta-Analysis	2020	Meta - Analysis	N/A	83	Learner behavior	• Duration • Educational level • Game elements	K-12	The authors conducted a meta-analysis to determine the effectiveness of gamification on behavioral change in education, using test scores and participation levels as measures. They found a moderate effect size of gamification on behavioral change, which was higher for participation levels than test scores. The results were consistent with those of a previous study. Gamification was effective for both adults and K-12 interventions, but not for college students. The most effective intervention length was less than one hour.	Future research could focus on the use of objective, measurable treatments such as online badges and leaderboards in K-12 educational settings. Additionally, there is a need to explore the effectiveness of using a variety of gamification types, such as progress bars, points, and avatars, in diverse educational programs.
Mamekova et al.	A Meta-Analysis on the Impact of Gamification over	2021	Meta-Analysis	N/A	7 (total N = 448)	Motivation	• Game elements • Learning content	Higher Education	The findings suggest that gamification in education can enhance students' motivation to learn, but only for about one-third	uture research could investigate the effectiveness of using alternative game design

	Students' Motivation								of the students. The effectiveness of gamification might vary depending on whether the game type is appropriate for the learning content. However, even when gamification did not increase motivation for learning, it could still be effective for some students. Therefore, it is important to consider the suitability of the learning content for gamification.	elements, such as quests, in educational gamification. This may help to overcome the issue of overused elements and enhance students' motivation and engagement in the learning process.
Nadi-Ravandi and Batooli	Gamifcation in education: A scientometric, content and co occurrence analysis of systematic review and meta analysis articles	2022	Meta – Analysis and Systematic Review	N/A	25	General	• Country • Length of study • Learners' Number • Educational Course • Educational level • Game elements • Theories	N/A	The most important concepts studied in the field of gamified education are motivation, learning, and engagement, and benefits observed in studies related to gamification in education include increased learner competition, practical skills, and perceived learning. Perceived learning was widely concluded as a positive effect of gamification learning, although in some interventions, no improvement was observed in final exam scores. Increasing the level of participation can develop learning skills and academic achievement. Educational interventions were effective in promoting learning, motivation, and participation of learners, but the definite effect of gamification was not mentioned in most studies, and weaker statistical differences between gamified and non-	Future research should focus on conducting systematic review studies and meta-analyses that consider the seven identified items. There is also a need for further research to gamify face-to-face classes, interventions in other disciplines and courses, and higher quality studies (two groups with pre-test and post-test) to determine the effect of gamification on variables.

									gamified environments were observed	
Ritzhaupt et al.	A meta analysis on the infuence of gamifcation in formal educational settings on affective and behavioral outcomes	2021	Meta-Analysis	PRISMA	32 (total N= 3570)	Learning outcomes	• Game elements • Learner affective • Learner behavior • Educational courses • Educational level	From elementary to postgraduate students	They conducted a meta-analysis to investigate the impact of gamification on affective and behavioral outcomes in formal education settings. They found that leaderboards, badges/awards, and points/experiences were the most frequently observed design elements, with leaderboards having the highest effect size on affective outcomes. Non-linear navigation, adaptivity/personalization, and narrative/storytelling showed interesting results for behavioral outcomes, but more research is needed. The study suggests that gamification has a significant impact on both affective and behavioral outcomes in education.	Future research is needed to provide conclusive and generalizable findings in the domain of gamification, as many gamification design elements were rarely observed in the studies. More empirical work on gamification is necessary to move beyond the mere pointification found in most studies, and explore the mediators for student learning outcomes, including affective, behavioral, and cognitive outcomes simultaneously. However, few primary studies reported outcomes in all three domains, which makes it impossible to conduct a meta-regression model based on the current literature.
Sailer and Hommer	The Gamification of Learning: a Meta-analysis	2019	Meta - Analysis	PRISMA	38 (total N= 4883)	Learning outcomes	• Learner motivation/ affect	From elementary to postgraduate students	Sailer and Homner (2020) conducted a meta-analysis on the effects of gamification on cognitive, behavioral, and	Future research should focus on conducting follow-up tests to determine the endurance

							• Learner behavior • Learner cognition • Game elements • Educational level • Educational courses • Game fiction • Social interaction • Duration of interventions		motivational learning outcomes. They found a significant, small effect of gamification on cognitive and behavioral outcomes, with game fiction and social interaction being particularly effective for behavioral outcomes. They also found a significant effect on motivational outcomes, with shorter interventions and higher education/work-related settings showing larger effects. However, when only studies with high methodological rigor were considered, only cognitive learning outcomes showed a small effect of gamification.	of the effects of gamification, exploring theoretical avenues to create an empirical framework, investigating psychological needs and high-quality learning activities fostered by gamification, and studying learners' experiences and perceptions of gamification, their actual activities in interventions, the role of skill level, the influence of initial motivation, the adaptiveness of gamified systems, and individual characteristics.
Yıldırım and Sen	The effects of gamification on students' academic achievement: a meta-analysis study	2019	Meta - Analysis	PRISMA	40 (N total = 3487)	Learner Achievement/ performance	• Educational course • Educational level • Learner motivation	Primary school Secondary school High school University	The results suggest that gamification has a moderate level positive effect on student achievement with a mean effect-size value of 0.557, determined through the use of a random-effects model due to the heterogeneity among effect-size values. The effect of gamification on student achievement did not differ significantly between technology-based and non-technology-based lessons, supporting the idea that	N/A

									gamification can have similar effects across different disciplines. Additionally, the effect of gamification design on student achievement did not differ significantly between school levels, indicating that gamification is suitable for use in primary school through university. However, it is worth noting that there was no statistically significant effect of gamification at the high-school level, and further experimental studies are needed to confirm the effect of gamification in this context. Overall, the evidence suggests that gamification design can significantly increase student achievement across almost all school levels and is suitable for use in all types of lessons, not just technology-based ones.	
Zhang and Yu	Meta-Analysis on Investigating and Comparing the Effects on Learning Achievement and Motivation for Gamification and Game-Based Learning	2022	Meta - Analysis	PRISMA	27	Learner achievement / performance Learner motivation	• Learner achievement/performance • Learner motivation	N/A	The study found that gamification had positive effects on learning achievement. Gamification had stable impacts on intrinsic motivation, although its effects on learning achievement were relatively unstable due to factors outside of the learning environment. Gamification had more stable effects on intrinsic motivation than on extrinsic motivation and contributed to developing highly internalized	Future research should focus on exploring the potential benefits and limitations of using gamification in various occupational training contexts.

									extrinsic motivation that could develop into intrinsic motivation. Overall, gamification was found to enhance motivation, learning, and problem-solving skills.	

Part II

Applications on Neuropsychological Assessment and Psychometrics

3.1

Preliminary evidence on machine learning approaches for clusterizing students’ cognitive profile.

Orsoni M, Giovagnoli S, Garofalo S, Magri S, Benvenuti M, Mazzoni E, Benassi M. Preliminary evidence on machine learning approaches for clusterizing students' cognitive profile. Heliyon. 2023 Mar 16;9(3):e14506. doi: 10.1016/j.heliyon.2023.e14506. PMID: 36967938; PMCID: PMC10031485.

3.1 Preliminary evidence on machine learning approaches for clusterizing students' cognitive profile.

3.1.1 Abstract

Assessing the cognitive abilities of students in academic contexts can provide valuable insights for teachers to identify their cognitive profile and create personalized teaching strategies. While numerous studies have demonstrated promising outcomes in clustering students based on their cognitive profiles, effective comparisons between various clustering methods are lacking in the current literature. In this study, we aim to compare the effectiveness of two clustering techniques to group students based on their cognitive abilities including general intelligence, attention, visual perception, working memory, and phonological awareness. 292 students aged 11–15 years, participated in the study. A two-level approach based on the joint use of Kohonen's Self-Organizing Map (SOMs) and k-means clustering algorithm was compared with an approach based on the k-means clustering algorithm only. The resulting profiles were then predicted via AdaBoost and ANN supervised algorithms. The results showed that the two-level approach provides the best solution for this problem while the ANN algorithm was the winner in the classification problem. These results laying the foundations for developing a useful instrument for predicting the students' cognitive profile.

3.1.2 Introduction

Specific cognitive functions are typically assessed to explain learning heterogeneity in students, particularly for those with atypical development (Alloway & Elsworth, 2012; Foley-Nicpon et al., 2012; Menghini et al., 2010). Indeed, even if students with atypical development are grouped in exact diagnostic group by means of specific diagnostic criteria, it is possible to find cognitive subgroups within each diagnosis aiming to personalize teaching interventions (Catts et al., 2012; Heim et al., 2008). In research on typical developmental populations, cluster analysis techniques have been used. This is a type of multivariate analysis that helps to classify subjects into groups that are internally highly homogeneous and externally highly heterogeneous. Depending on the variables used to cluster the subjects, different solutions have been proposed to describe the cognitive profiles of students. For instance, Yokota et al. (2015) used a k-means clustering technique that considered four factors (verbal comprehension, perceptual organization, freedom from distractibility, and processing speed) of the Wechsler Intelligence Scale for Children (Third Edition). The authors discovered the presence of six cognitive subtypes that differed in verbal comprehension, perceptual organization, processing speed, and distractibility. The interesting part is that the authors validated the clustering solution by confirming differences among clusters based on fMRI measures. The study results suggest that cognitive profiles can differentiate children with typical development, and these differences may be reflected in specific neural patterns. Recently, Poletti et al. (2018) conducted a cluster analysis study on ten core subtests of the Wechsler Intelligence Scale for Children–Fourth Edition (WISC-IV) and identified four subgroups of students with Specific Learning Difficulties (SLD). These subgroups differed in their performance on the WISC-IV subtests, particularly in the areas of verbal comprehension, coding, and executive functions. The authors also observed that while impairments in reading and mathematics were associated with low reasoning and executive functioning, difficulties in written expression were linked to low verbal

and coding abilities. Despite the relevant evidence obtained from different studies, results from cluster analysis could be sometime considered unsatisfactory in terms of rigor in methodology and replicability (Clatworthy et al., 2005). A major limitation of previous studies is that they lacked two crucial steps for clustering techniques, namely, comparison between different clustering solutions and validation of the selected cluster solution (Benassi et al., 2020; Kraus et al., 2011). In Yokota et al. (2015) only k-means clustering method was used, and the number of clusters was selected arbitrarily by progressively increasing them until a minimum of one cluster containing less than 10% of the sample appeared. Although this method has an advantage of including a parsimonious selection criterion, it may not result in the most meaningful clusters. In Poletti et al. (2018), the authors relied on visual inspection of the agglomeration coefficients and dendrogram figure to identify the best cluster solution. Although they acknowledged the possibility of using multiple methods, their approach focused solely on supporting a single solution, rather than comparing various methods using statistical indices. Additionally, the authors evaluated agreement between clustering solutions using Cohen's kappa and Intraclass correlation coefficient but did not assess the accuracy of the proposed solution. While the method of comparing two solutions is intriguing, it fails to determine which of the two is superior, thus hindering a meaningful comparison. To date, as far as we know, no study has explicitly aimed to compare various clustering techniques and assess the feasibility of implementing efficient and replicable methods for the topic at hand. Therefore, further research is needed to investigate and compare multiple clustering approaches, while also evaluating their effectiveness and reproducibility. In this study, we aim to cluster different cognitive profiles of secondary school students by using a two-level approach based on the joint use of an artificial neural network, the Kohonen's Self Organizing Maps (Kohonen, 1990) and the k-means clustering algorithm. The proposed approach would be beneficial in: 1. Allowing to compare different clustering approaches and select the best one

on statistical index supporting the choice; 2. Enhancing the clustering solution accuracy. Recent findings in non-psychological fields showed an improvement in the clustering solution by applying the Kohonen's Self Organizing Maps (SOMs) (Kohonen, 1990) before the k-means or hierarchical clustering implementation compared to the clustering methods only (Dong et al., 2015; Juntunen et al., 2013; Palamara et al., 2011b). In this study, we evaluated the performance of k-means clustering with and without a SOM-based pre-processing step. We selected this clustering algorithm based on previous research (Dong et al., 2015; Juntunen et al., 2013; Palamara et al., 2011b), which showed superior performance compared to other algorithms such as hierarchical clustering. Despite the advantages of this approach, its application in the field of psychology, particularly in cognitive profiling, is still limited. Our objective was to identify the optimal clustering solution by comparing two clustering methods, based on cognitive functions that previous research had identified as the most distinguishing between students with SLD and controls, while also supporting academic skills. Specifically, we focused on executive functions, language, and visual perception abilities (logical reasoning, visual attention, visual perception, verbal comprehension, and working memory), as reported in studies (Allan et al., 2014; Carlson et al., 2013; Fenwick et al., 2016; Johnson, 2014; Kudo et al., 2015; Pradeep Kumar Gupta & Dr. Vibha Sharma, 2017; C. R. Reynolds & Shaywitz, 2009; Stevens & Bavelier, 2012; Vock et al., 2011). Next, we developed and tested the validity of a machine learning (ML) model to determine whether cognitive profiles could be accurately predicted by the model. This final step can serve to verify the replicability of the selected clustering solution.

3.1.3 Material and methods

3.1.3.1 Participants

To recruit participants for the study, several schools in the Emilia-Romagna Region of Italy were invited to participate. Of those invited, four consented to the use of an online digital game

for cognitive assessment (see Cognitive Assessment section), and three of those also agreed to standardized battery tests. This resulted in a total sample of 292 secondary school students (104 females (36%), age range: 11–15 years) for cluster analysis. Of these, 99 (33.96%) were attending secondary school, while 193 (66.04%) were attending secondary high school. All participants were Italian, and 12 (4.11%) were bilingual. Form the total, a subgroup of 105 students was selected for full clinical data collection, with 71 (29 females (41%), age range: 11–14 years) attending secondary school and 34 (13 females (38%), age range: 14–15 years) attending secondary high school. This subsample was assessed by four psychologists using standardized battery tests to evaluate the cluster solution's ability to differentiate between typically and atypically developing subjects. Of these, 30 (28.6%) met the criteria for a specific learning difficulty (SLD), with 7 having dyslexia, 7 having dyscalculia, and 16 exhibiting multiple disorders.

3.1.3.2 Cognitive assessment

All students cognitive abilities have been assessed by an online digital game called PROFFILO developed for the assessment of the student's cognitive profile (Matteo Orsoni et al., 2021), and with standardized tests for logical reasoning (Raven's Progressive Matrices) (Raven, 1989b), working memory (WISC-IV Inverse SPAN) (Orsini et al., 2012), and Visual Attention (NEPSY-II Visuospatial Attention subtest) (Brooks et al., 2010). PROFFILO was administered in class to the students and lasted 20/25 min. It is composed of five different sub-tests (games), each developed for the assessment of a specific cognitive function (logical reasoning, visuospatial attention, motion perception, phonological awareness, and working memory). A previous study showed a good correlation between these games and standardized tests for the evaluation of the same clinical functions, with the only exception of the phonological awareness game that, for this reason, was excluded from the subsequent analyses (Matteo Orsoni et al.,

2021). The tests used showed good convergent validity with standardized tests (see Supplementary Materials).

3.1.3.3 Reading, spelling and math assessment

Students aged 11–14 were assessed for reading and spelling abilities by standardized Italian reading test, DDE-2 and for math abilities, by AC-MT (Cornoldi et al., 2017). Students aged above 14–15 years, were assessed by means of Advanced MT-3 battery test (Cornoldi et al., 2017) both for reading and math abilities. The student was considered as having SLD when her/his standardized reading or spelling or math score was below 2SD, while the general intelligence evaluated by Raven's Matrices (Raven, 1989b) was within the normative range (scoring above 25th centile). As documented by an interview with teachers and parents, all participants had no evidence of brain injury, socio-cultural detriment, or relevant behavioral problems. In Table 3-1-1, one-way ANOVAs have been carried out to inspect differences between the. SLD group, and no-SLD in reading and arithmetic clinical tests.

3.1.3.4 From raw data to clustering: preprocessing, Self-Organizing Maps and K-means

To improve algorithm performance, we used Min-Max normalization as a preprocessing procedure on the sample. This procedure allowed for unifying the feature's orders of magnitude (Walesiak & Dudek, 2020). It performs a linear transformation on the original data, mapping a value in a range between [0,1], and it is not dependent on the distribution of the variable (Suarez-Alvarez et al., 2012; Visalakshi & Thangavel, 2009). Additionally, before applying the k-means algorithm, we used Self-Organizing Maps (SOMs) for a subsequent clusterization. SOMs are competitive or unsupervised artificial neural networks that provide a topological representation of the input data (Kohonen, 1990). A thorough explanation can be found in the supplementary materials. The pseudo-code (Algorithm 1) explains the SOM implementation. The k-means method (MacQueen, 1967) was then applied to find the best clustering solution.

The optimal number of clusters was selected using the Elbow method (Thorndike, 1953). We implemented two k-means cluster algorithms, with 1000 as the maximum number of iterations allowed and 100 as the number of random sets chosen. The cluster's quality was evaluated by qualitatively inspecting the cases gathered in each cluster, by reviewing all clusters by hand to evaluate the meaning of the membership of each data to a given cluster. Following (Palamara et al., 2011b), the cluster accuracy index I_c was calculated. It refers to a single cluster and takes the form of Eq. (10), where a_v is the number of correctly assigned cases and n_c is the number of cases grouped in the cluster.

$$I_c = \frac{a_v}{n_c} \tag{10}$$

We calculated the accuracy index by examining the number of variables that fell within the centroid's membership boundaries, as shown in Table 3-1-2. We graphically represented these boundaries for each cluster and variable and considered a subject misclassified if it fell outside of them. If a subject was misclassified in more than two variables, the entire case was deemed misclassified by the algorithm.

Algorithm 1 This program search the best seed avoiding local minima in Self-Organizing Maps implementation by starting from normalized Data

Require: : Data
  function MIN-MAX NORMALIZATION(x) ▷ Calculate the Min-Max Normalization of the features of interest
    $(x - min(x))/(max(x) - min(x))$
  end function
  for $i \leftarrow 1, nrow(Data)$ **do**
    for $k \leftarrow 1, ncol(Data)$ **do**
      Min-Max Normalization(Data)
    end for
  end for
  seed = 0 ▷ Searching for the best seed, trying to avoid local minima
  samplesize = $nrow(Data)$
  grid.size = $samplesize^{(1/2.5)}$
  for $j \leftarrow 1, 200$ **do** ▷ Run 200 SOMs
    set.seed(j)
    som.grid = (grid.size, hexagonal topology)
    som.model = (matrix(data normalized), grid = som.grid, 300) ▷ 300 is the number of times the complete data set will be presented to the network
    seed[j] = mean(SSE distances of the som.model)
  end for
  min seed = min(seed) ▷ Searching for the seed with minimum SSE
  set.seed(min seed)
  som.model2 = (matrix(data normalized), grid = som.grid, 300)

Age group	Test	SLD	no-SLD	$F_{(1,103)}$	p	η^2
11-14			***Reading tests***			
	Word reading Speed	-2.863(0.349)	-0.267(0.226)	38.978	< 0.001	0.361
	Word reading Accuracy	-2.619(0.478)	0.040(0.310)	21.766	< 0.001	0.240
	Word reading Speed	-2.573(0.376)	-0.093(0.241)	30.876	< 0.001	0.306
	Word reading Accuracy	-0.524(0.252)	0.745(0.162)	17.956	< 0.001	0.204
			Math tests			
	ACMT1a	-0.614(0.214)	0.333(0.137)	13.918	< 0.001	0.166
	ACMT2a	-1.818(0.268)	-0.359(0.172)	21.048	< 0.001	0.231
	ACMT3a	-1.327(0.275)	-0.328(0.176)	9.353	0.003	0.118
	ACMT4a	-1.402(0.229)	0.378(0.147)	42.850	< 0.001	0.380
	ACMT1v	-1.657(0.271)	0.009(0.178)	26.394	< 0.001	0.280
	ACMT2v	-2.245(0.405)	-1.140(0.262)	5.243	0.025	0.071
14-15			***Reading tests***			
	Word reading Speed	-3.576(0.375)	-0.531(0.225)	48.568	< 0.001	0.603
	Word reading Accuracy	-1.016(0.318)	0.263(0.191)	11.911	0.002	0.271
	Word reading Speed	-1.924(0.315)	-0.044(0.189)	26.239	< 0.001	0.451
	Word reading Accuracy	-0.847(0.300)	0.125(0.180)	7.711	0.009	0.194
			Math tests			
	MT3a	-0.198(0.292)	0.767(0.175)	8.042	0.008	0.201
	MT3t	-1.355(0.368)	-0.154(0.221)	7.831	0.009	0.197
	MT3af	-0.693(0.302)	0.120(0.181)	5.344	0.027	0.143

Table 3-1-1 z score in all the reading and math tests in SLD and no-SLD group (Mean and SE are reported). One-way ANOVAs for comparing SLD and no-SLD groups in reading, spelling, and math tests.

3.1.3.5 From clustering to prediction: imbalance classification, AdaBoost and artificial neural networks

The cluster solution revealed the presence of various groups of different sizes, resulting in an imbalanced classification problem. Imbalance arises when one or more classes have significantly lower proportions in the training data than the other classes (Kuhn & Johnson, 2013). As a result, the impact of class imbalance on classification performance metrics is a significant concern (Luque et al., 2019). To address the imbalance problem, we utilized the Synthetic Minority Over-Sampling Technique (SMOTE) proposed by Chawla et al. (2002). This approach combines up-sampling and down-sampling techniques that are determined by the class. Three parameters guided the SMOTE algorithm, including the amount of up-sampling, the amount of down-sampling, and the number of neighbors used to create new cases. During the up-sampling, SMOTE generated new cases by randomly selecting a data point from the minority class(es) and determining its K-nearest neighbors (KNNs). The new synthetic data point was a random combination of the selected data point predictors and its neighbors. Additionally, the SMOTE algorithm down-sampled cases from the majority class via random sampling to achieve balance in the training set (Kuhn & Johnson, 2013) The number of neighbors used in the algorithm implementation was set to 3. After the SMOTE implementation, we got a training sample of 540 subjects, 60 for each class. This sample characterized by real and synthetic data was used for the subsequent analyses where we compare the performances of two supervised ML algorithms in the prediction of our clusters both in the imbalanced dataset (only real data) and in the balanced (real and synthetic data) ones. The choice of using ML algorithms based on their best predictive ability than linear models [38,39]. After applying the Synthetic Minority Over-Sampling Technique (SMOTE) to balance the training set, we used the Adaptive Boosting algorithm (AdaBoost) (Freund & Schapire, 1996) and a fully connected Artificial Neural Network (ANN) to predict the clusters emerged from the previous steps and test the replicability of the solution. We compared the

performances of these two supervised ML algorithms, with AdaBoost being selected for its good performance in imbalance classification problems (Luque et al., 2019; Sun et al., 2007). The Adaptive Boosting algorithm is an ensemble method that functions in a boosting network. Boosting is a technique that can significantly reduce the error of any weak learning algorithm to create classifiers that only need to be slightly better than random guessing (Freund & Schapire, 1996). The AdaBoost algorithm assigns weights to each sample based on its importance and places the most weight on those examples that are most frequently misclassified by the previous classifiers. This emphasis may cause the learner to produce an ensemble function that differs significantly from the single learning (Sun et al., 2007). We implemented the algorithm on both balanced and imbalanced training samples. To prevent model overfitting and identify the optimal parameters, we performed 5-fold cross-validation and hyperparameter search on both samples. The tuning process involved two parameters: the number of estimators and the learning rate. For the number of estimators, the search range was set from 10 to 700 in increments of 10. For the learning rate, the search range was set from 0.0001 to 1 in increments of 0.1. The second algorithm employed a fully connected artificial neural network (ANN) with two hidden layers. The first comprised 512 units, while the second contained 256 units. Rectified linear unit (ReLU) activation functions were used for the hidden layers, with the Softmax function employed for the output layer. The Adam optimizer was employed with the Sparse Categorical Crossentropy function utilized as the loss function. The Accuracy metric was implemented to evaluate the model. To prevent overfitting, the model was trained for 600 epochs with a dropout rate of 0.5. Furthermore, an early stopping callback was implemented. The algorithm was also tested on both the balanced and imbalanced training samples.

3.1.3.6 Evaluate the performances.

To evaluate the performance of the model, the accuracy metric is commonly used. However, in cases of class imbalance, accuracy may not be a suitable measure as the minority class has little impact on the overall accuracy compared to the majority class (Sun et al., 2007).

Cluster- ID	*Logical Reasoning*	*Visual Perception*	*Visuospatial Attention*	*Working Memory*	*N (%)*
vhLR-aAll	0.17	0.28	0.53	0.55	52 (17.808%)
aALL-lWM	0.55	0.43	0.40	0.25	12 (4.110%)
vhLR-lWM	0.16	0.33	0.65	0.22	37 (12.671%)
vlVP-lVA	0.24	0.83	0.29	0.61	10 (3.424%)
vhLR-vlVA	0.18	0.30	0.13	0.46	23 (7.877%)
vhLRWM	0.16	0.39	0.54	0.93	22 (7.534%)
ahAll	0.21	0.29	0.76	0.52	75 (25.685%)
vhLR-lVP	0.18	0.67	0.68	0.54	19 (6.507%)
vhALL	0.10	0.18	0.79	0.81	42 (14.384%)

Table 3-1-2 Centroids for each cluster and variable found after the implementation of k-means and SOM and clusters' numerosity for both the sample used in the k-means clustering.

To address this issue, several other metrics can be derived from the confusion matrix to assess model performance. The confusion matrix compares the true classes with the predicted classes obtained from the model and can be used to calculate various error parameters based on the counts of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) values. These values form the basis for computing the Precision (PRE) and Recall (REC) error metrics. Precision (Eq. (11)) is the ratio of the number of correct predictions of an event (class) to the total number of times the model predicts it.

$$PRE = \frac{TP}{TP+FP} \quad (11)$$

The lower is the value of False Positive, higher is Precision. Recall Eq (12) reflects the model's sensibility. It is the ratio of the correct predictions for a class of the total cases in which it occurs.

$$REC = \frac{TP}{TP+FN} \quad (12)$$

Usually, Precision and Recall are combined to obtain the F1 score. F1 Eq. (13) represents the harmonic mean between Precision and Recall:

$$F1 = \frac{2}{\frac{1}{R}+\frac{1}{P}} \quad (13)$$

In general, the harmonic mean of two numbers is closer to the smaller of the two. Therefore, having a high F1 score indicates that both the Recall and Precision are relatively high (Sun et al., 2007). In this study, the performance of each class was evaluated based on metrics such as TP, TN, FP, FN, PRE, REC, and F1 scores. Moreover, the Balanced Accuracy Score and Weighted F1 Score were also calculated.

3.1.4 Software and packages

The analyses were carried on by using JASP statistical software (2022), R v4.03 (R Core Team, 2020), and Python v3.8 (Van Rossun & Drake, 2009). On R, the 'kohonen' package was used for the Self-Organizing Maps implementation, and the k-means algorithm was carried out within the 'stats' base package. In Python, the 'scikit-learn' package (Pedregosa et al., 2012) was used for the implementation of the Adaptive Boosting algorithm and the TensorFlow for the implementation of the Artificial Neural Network (ANN) (Abadi et al., 2016).

3.1.5 Results

3.1.5.1 Self-organizing maps and K-means

To evaluate the reliability to use the SOM as pre-processing for a subsequent clusterization, we compared the variance explained between the k-means cluster after the SOM implementation and the clusterization held by the normalized data alone.

The SOM algorithm was executed 200 times in order to minimize the mean distance between the codebook vector and the real vector (represented by the mean sum of square error, or SSE) to the closest unit on the map. The resulting mean SSE was 0.009. To visualize the distances in the original space, we employed the U-matrix method (Ultsch & Herrmann, 2007), and the resulting plot can be found in Supplementary Materials Fig.B1. This method calculates the average distances between the prototype vector of each cell and the prototype vectors of its neighboring cells, which are then represented by different color shades ranging from blue to red. Blue shades correspond to the smallest average distances, while red shades represent the largest ones. To graphically display the properties of the variables, we created a property plot for each variable, which can be found in Supplementary Materials Fig.B.2. These plots allow for the visualization of the similarity of a particular object to all units on the map, as well as how these units are organized. Before the k-means cluster implementation, the Elbow method was implemented both in the unit of the SOM and in the normalized data. To determine the optimal number of clusters using the method described above, we conducted a k-means run for 45 steps, with a maximum of 1000 iterations allowed. As shown in Fig. 3-1-1, the results indicated that for the normalized data, the optimal number of clusters was 9, resulting in a within-cluster sum of squares of 15.15. However, when using the data that had been pre-processed by the SOM, we identified 9 clusters with a within-cluster sum of squares of 4.66. It is possible to evaluate how the cluster solution found implementing the SOM as pre-preprocessing improving the clusterization by reducing the WCSS compared to k-means alone. The k-means implemented in normalized data and SOM showed the ratio between the sum of squares (BSS) and the total sum of squares (TSS) as equal to 0.673 (67.3%) and 0.712 (71.2%),

respectively. This result outline how the hybrid approach (SOM + k-means) enhances the BSS/TSS ratio of 3.9%, highlighting how the clusterization preceded by the SOM is the one that best embodied the properties of internal cohesion and external separation explaining most of the variance. In addition, both solutions were compared by using the BIC criterion. The results showed the hybrid approach as associated with the lowest BIC value (170.4), as compared to the single solution with k-means (219.4).

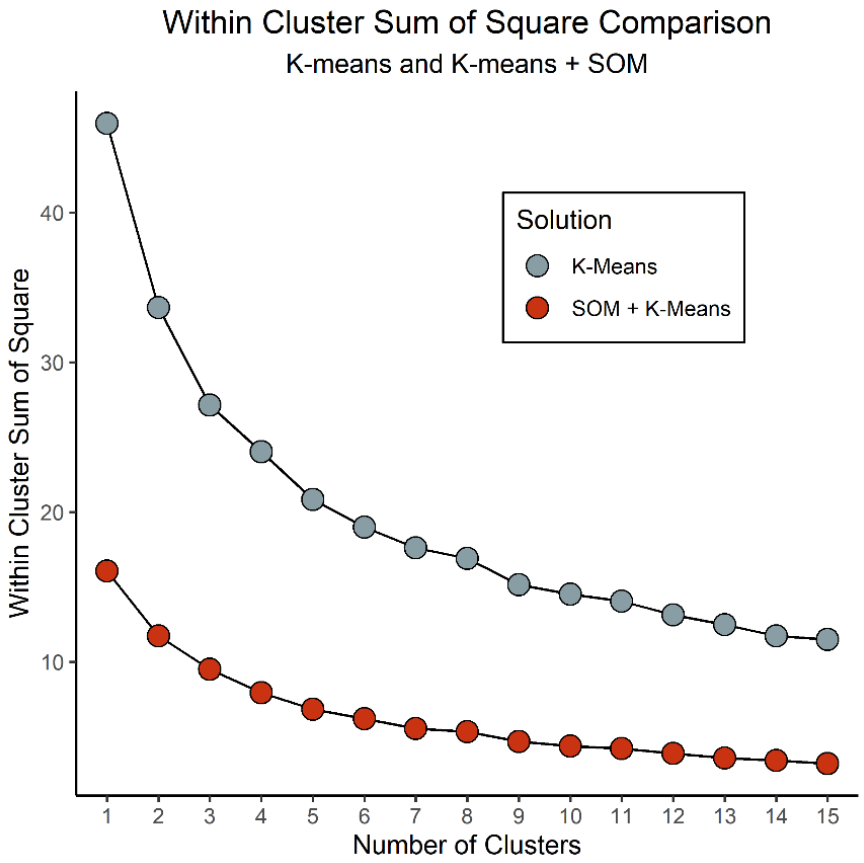

Figure 3-1-1 Within Cluster Sum of Square (WCSS) solution for the k-means and k-means + SOM solutions.

3.1.5.2 Clusters' description

Table 3-1-2 summarized the centroids' mean for each cluster and variable, along with their numerosity. To evaluate the performance, we proposed to consider different thresholds. In Visuospatial Attention, and Working Memory, subjects with values between 0 and 0.20 were classified as very low performance, 0.20-0.40 as low performance, 0.40-0.60 as average performance, 0.60-0.80 as high performance, and 0.80-1 as very high performance. In Logical Reasoning and Visual Perception, performance values are reversed. In other words, subjects with values between 0 and 0.20 were classified as very high performance, 0.20-0.40 as high performance, 0.40-0.60 as average performance, 0.60-0.80 as low performance, and 0.80-1 as very low performance. Moreover, these values allowed us to calculate the Cluster Accuracy Index as reported in Table 3. By inspecting the centroids found after clusterization, we highlighted the characteristics of each cluster. Cluster (1) Very high Logical Reasoning average All (vhLR-aAll) (n = 52) consists of subjects with very high performance in logical reasoning, high in visual perception, average in working memory, and visuospatial attention. Cluster (2) Average All, Low Working Memory (aALL-lWM) (n = 12) consists of subjects with average performance in logical reasoning, visual perception, and visuospatial attention but low in working memory. Cluster (3) Very High Logical Reasoning and Low Working Memory (vhLR-lWM) (n = 37) consists of subjects with very high performance in logical reasoning, high in visuospatial attention, and visual perception, and low performance in working memory. Cluster (4) Very Low Visual Perception and Low Visuospatial Attention (vlVP-lVA) (n = 10) consists of subjects with very low performance in visual perception, low visuospatial attention, and high in working memory and logical reasoning. Cluster (5) Very High Logical Reasoning and Very Low Visuospatial Attention (vhLR-vlVA) (n = 23) consist of subjects with very high performance in logical reasoning, average performance in working memory, high in visual perception, and very low in visuospatial attention. Cluster (6) Very high Logical Reasoning and Working Memory (vhLRWM) (n = 22) consists of subjects with very high performance in

logical reasoning and working memory, high in visual perception, and average in visuospatial attention.

Cluster-ID	*N*	*Accuracy %*
vhLR-aAll	52	100
aALL-lWM	12	91.7
vhLR-lWM	37	91.9
vlVP-lVA	10	80
vhLR-vlVA	23	91.3
vhLRWM	22	90.9
ahAll	75	88
vhLR-lVP	19	94.7
vhALL	42	90.5
Average weighted Accuracy		91.8

Table 3-1-3 Clusters and Accuracy Index overall and divided for each cluster.

Cluster (7) Average-High All (ahAll) (n = 75) It is the cluster with the highest representativeness and consists of subjects with high performance in all the tasks. Cluster (8) Very high Logical Reasoning, Low Visual Perception (vhLR-lVP) (n = 19) consists of subjects with very high performance in logical reasoning, low in visual perception, and high in visuospatial attention and working memory. Cluster (9) Very High All (vhALL) (n = 42) consists of subjects with a high or very high performance in all the tests. In particular, they present a very high performance in logical reasoning, high, visual perception, working memory, and high performance in visuospatial attention. The Cluster Accuracy Index (Ic), weighted for the cluster numerosity showed a 91.8% overall accuracy, as illustrated in Table 3-1-3. When inspecting the presence of SLD in 105 students, we found that 30 (28.6%) of them reached the criteria for an SLD (dyslexia or dyscalculia, or both). Furthermore, in Table 3-1-3 the presence of SLD in the cluster solution found has been summarized. Furthermore, according to the data presented in Table 3-1-4, we observed a higher frequency of students with SLD in certain clusters compared to those without SLD. Specifically, the clusters aALL-lWM and vhLR-lWM showed a prevalence of students with SLD that was 2.5 and 4.16 times greater, respectively,

than those no-SLD. Additionally, both clusters exhibited poor performance in the working memory task, which supports previous research identifying working memory as a cognitive risk factor for students with SLD (Geary & Hoard, 2001; McLean & Hitch, 1999; Moll et al., 2016). Table 3-1-2 displays the results of the cluster analysis, which grouped the original sample into nine different categories based on numerosity. This grouping was used to implement two classification algorithms: Adaptive Boosting and Artificial Neural Network (ANN) for both the original imbalanced sample and the balanced sample after the SMOTE process. A summary of the most relevant metrics can be found in Table 3-1-5.

Cluster-ID	*no SLD (%)*	*SLD (%)*
vhLR-aAll	16 (21.3)	7 (23.3)
aALL-lWM	3 (4)	3 (10)
vhLR-lWM	6 (8)	10 (33.3)
vlVP-lVA	2 (2)	0 (0)
vhLR-vlVA	6 (8)	1 (3)
vhLRWM	1 (1.3)	0 (0)
ahAll	21 (28)	7 (23.3)
vhLR-lVP	9 (12)	1 (3)
vhALL	11 (14.6)	1 (3)

Table 3-1-4 Distribution of SLD and no SLD students concerning the cluster solution found.

3.1.5.3 Adaptive Boosting

The results of the algorithm trained on the original sample of 233 students (80%) showed a learning rate of 0.5001 and the number of estimators of 550 as the best parameters during the training, reflecting on an accuracy score of 80.68% at the training set and 84.75% on the test set on a sample of 59 students (20%). The reliability of the classifier was compared with the imbalance of the training set. In our sample, Cluster (7) ahAll occupies 25.75% of the total frequency on the training set. Therefore, the algorithms' reliability was evaluated by considering that we can obtain the 25.75% of accuracy at the test set by predicting the majority

class without the help of any supervised classifier. In our situation, the global accuracy of 84.75% and the balanced accuracy score of 77.23% highlight the ability of this classifier to learn most of the rules for predicting starting with the feature variables under consideration. However, not all the classes have been predicted correctly. By inspecting the F1 score for each cluster in Table B1 (see Supplementary Materials) may be noted a good classification for clusters 1,5,6,7,8 and 9, whereas the other clusters showed a low F1 score meaning a worse classification rate. This resulted in a weighted F1 score of 0.846. The results of the Adaptive Boosting on the balanced sample of 540 students showed a learning rate of 0.8001 and the number of estimators of 690 as the best parameters during the training reflecting an accuracy score of 69.81% at the training set and 57.63% on the test set on a sample of 59 students (20%). This shed light on a balanced accuracy score of 59.83%. By inspecting the F1 score for each cluster in Table B2 (see Supplementary Materials) clusters 1, and 5, showed a good classification. Compared to the remaining clusters, all exhibited a low F1 score. This resulted in a weighted F1 score of 0.595.

3.1.5.4 Artificial neural networks (ANN)

After 140 epochs the model reached a global accuracy score of 94.42% at the training set and 89.83% at the test set. The balanced accuracy score of this model is 89.58%. By inspecting the score F1 for each cluster in Table B3 (see Supplementary Materials) all the clusters exhibit scores over 0.8. This resulted in a weighted F1 score of 0.899. The results after the implementation of the ANN on the imbalanced sample showed better results compared to the previous AdaBoost algorithm on the same sample. The ANN implemented in the balanced showed even better results. After 82 epochs the model reached a global accuracy score of 94.81% at the training set and 91.53% at the test set, resulting in a balanced accuracy score of 91.66%. By inspecting the score F1 for each cluster in Table B4 (see Supplementary Materials) no cluster exhibits scores below 0.7 while others over 0.8. This resulted in a weighted F1 score

of 0.916. These results displayed the ANN as the best algorithm for this type of problem both for the imbalanced and for the balanced sample.

Principal Metrics	*AdaBoost Imbalanced*	*AdaBoost Balanced*	*ANN Imbalanced*	*ANN Balanced*
Global Training Accuracy (%)	80.7	69.8	94.4	94.8
Global Testing Accuracy (%)	84.7	57.6	89.8	91.5
Balanced Accuracy (%)	77.2	59.8	89.6	91.7
Weighted F1 Score	0.85	0.59	0.90	0.92

Table 3-1-5 Metrics of the algorithms implemented both in the imbalanced and balanced sample. It is possible to observe how the ANN on the balanced dataset performs better than the others in all the metrics under evaluation.

3.1.6 Discussion

In this study, a novel clustering method was utilized for classifying cognitive abilities in secondary school students. This new approach involved preprocessing the k-means algorithm to achieve the most precise and reliable classification. The results revealed that the accuracy of classification and discrimination was at its peak when this method was applied. The study demonstrated that a hybrid clustering approach, which combined Kohonen's Self-Organizing Maps (SOMs) and k-means, enhanced the replicability of clustering among students with typical development. The efficiency of this profiling technique was confirmed by an ANN algorithm, suggesting that it is highly effective in profiling new users. Our findings confirm the results of prior research in various fields, as reported in (Dong et al., 2015; Juntunen et al., 2013; Palamara et al., 2011b). The self-organizing map (SOM) technique groups similar cases into map units during the initial clustering stage. This reduces the amount of data to be classified in subsequent clustering procedures and diminishes the amount of noise (Palamara et al., 2011b). As a result, applying the k-means algorithm to the map units divides the dataset into distinct partitions. Our study reveals that this approach provides a more accurate representation

of the clustering space, explaining a higher degree of variance than the single cluster method alone. Furthermore, the overall cluster accuracy is excellent, achieving 91.8%. The solution obtained indicated nine different groups having very high, high, average, low, or very low performance in the cognitive domains investigated. One group with difficulties in visual perception (vhLR-lVP), another group with impaired visuospatial attention (vhLR-vlVA), two groups with difficulties in working memory (aALL-lWM, vhLR-lWM), and one group with visuospatial and perceptual deficits (vlVP-lVA). This solution is partially in agreement with Yokota et al. (2014) study, indicating that perceptual organization and attention are important factors in clusterizing typically development children. Moreover, the proposed solution allowed us to distinguish between the distribution of the clusters in the SLD and no-SLD groups. In particular, by inspecting the aALL-lWM and vhLR-lWM clusters, were 2.5 and 4.16 times more for SLD students than no-SLD respectively. This corroborates previous findings where low working memory has been reported as a cognitive risk factor for developing dyslexia and dyscalculia (Geary & Hoard, 2001; McLean & Hitch, 1999; Moll et al., 2016). The application of this cluster approach puts a novelty in the psychological field given that ML is not extensively used in the analysis of psychological experiments as compared to other fields (e.g., genetics) (Orrù et al., 2020). The results showed the ANN algorithm carried out in the balanced sample as the best one for this problem. The average F1 score of 0.92 indicates a very good ability of the algorithm to learn the rules which hold the cluster differences by considering a wide range of cognitive variables. These results, although preliminary, reveal that this approach could be an efficient tool for clustering cognitive profiles. However, some limitations should be considered when interpreting the results of the present study. Above all, in the face of good results both for the cluster evaluation and prediction phases, the sample size consists of 292 students, thus reducing the generalizability of the results. Further studies will aim to increase the sample size, allowing a precise evaluation of the external validity of the clusters and,

arguably, the cluster prediction from ML algorithms. In addition, due to the nature of the tool used we have not been able to include the phonological awareness of the students inside the clustering procedure, but given its importance in learning, could be important its inclusion. This could point out more precisely students with educational special needs. Further studies would include a more reliable assessment of the phonological awareness inside PROFFILO and then include it in the clusterization procedure. Moreover, we would also focus on and compare other clustering algorithms (e.g. DBSCAN, spectral clustering, and gaussian mixture models) in the joint use with SOM.

3.1.7 Conclusion

The current study employs machine learning techniques to cluster the cognitive profiles of Italian secondary school students. The study measures cognitive abilities, including logical reasoning, visuospatial attention, motion perception, and working memory, using an online digital game called PROFFILO. The use of this clustering approach is a novelty in the psychological field, as machine learning is not widely used in psychological experiments as compared to other fields (Orrù et al., 2020). However, Orrù et al. (2020) enumerated the benefits of using ML in psychological research, including improved generalization, replication of results, and personalized predictions at a single subject level. The present study adds to this literature, suggesting that ML can be especially useful for clustering heterogeneous populations, as it improves classification accuracy and allows testing the replicability of results. Psychologists often need to explain the heterogeneity of clinical populations and find a way to group patients in order to settle down successful interventions. The presented results evidenced that the use of ML within a cognitive profiling test such as PROFFILO may have important practical implications for clinical practice. Indeed, having a clustering model that is validated as the most accurate as possible, and could be replicated in other samples, allows the clinician to implement personalized based models of intervention. Moreover, the model is advantageous

because it is expected to increase its validity and efficiency by adding cases and information. Furthermore, on the methodological point of view, this study is the first to compare the benefits and reliability of using both the Self-Organizing Maps algorithm and k-means for cognitive profiling, and to investigate the potential utility of supervised machine learning algorithms (specifically, AdaBoost and ANN) in predicting the cognitive profile of new users. The findings of this study demonstrate that applying a hybrid clustering approach, which involves multiple steps using Self Organizing Maps and k-means, can enhance the reliability of clustering when analyzing diverse measures, such as cognitive profiling. This approach provides a better understanding of how clusters are distributed in groups with and without specific learning difficulties (SLD). Overall, these results suggest that hybrid clustering techniques can be useful in the field of psychology to improve the dependability of clustering and the accuracy of solutions.

3.1.8 Supplementary Materials

3.1.8.1 Proffilo Description

PROFFILO is a digital assessment tool based on gamification and developed in the Unity platform. In PROFFILO, the student has to cope with five tasks interfacing with the demands of a robot.

Each task consists of a specific game involving a specific cognitive domain: phonological awareness, motion perception, visuospatial attention, verbal memory, and problem-solving.

In the Logic game, the robot asks the subject to identify the missing element that fulfills a pattern. The total number of correct answers attests to problem-solving strategies.

In the Vispa game, within 60 seconds, the robot asked the subject to find a target presented within a set of distractors. Four different levels of complexity are presented. The number of correct detections is a measure of visuospatial attention ability.

In the Motion perception game, the subject is asked to recognize the direction of moving stimuli obtained by white dots moving on a black background. This task allows to assess the subject‘s motion perception abilities.

NonWord Recognition game is a measure of phonological abilities and consists of two little CPU robots reporting two Italian phonologically similar words. Each robot could report a real word or a nonsense word; the subject has to recognize the robot reporting the real word.

In the Memory game, takes place the assessment of the working memory by a game in which the robot verbalizes numbers and the subject is instructed to report them backward by using the computer keyboard. The number of correct memorized numbers is reported.

PROFFILO was preliminarily validated on a sample of 81 students (32 Female, age range 11 and 14 years) attending Italian secondary school. To validate PROFFILO, each student was assessed with both PROFFILO games and the correspondent standard neurocognitive tests. In detail, Raven’s Progressive Matrices test was used to assess non-verbal intelligence (Raven,

1989a); NonWORD Repetition Test (PROMEA) was used to evaluate phonological ability (Vicari, 2007); Motion coherence test was used to measure motion perception abilities; NEPSY-II attention subtest was included to measure visuospatial attention (Brooks et al., 2010) and WISC-IV memory tests were used to evaluate verbal short-term memory (Orsini et al., 2012).

To evaluate the convergent validity of PROFFILO, the correlation between the scores obtained in each game of PROFFILO and the neuropsychological tests was evaluated by Pearson's correlation analysis. The results showed significant correlations between Logic and Raven ($r=-0.63$; $p<.05$), Memory and WISC-IV Span subtest ($r=0.49$; $p<.05$), Motion and motion coherence test ($r=-0.76$; $p<.05$) and Vispa and NEPSY test ($r=0.40$; $p<.05$), while NonWord Recognition game did not show a good correlation with NonWordRepetition test.

3.1.8.2 Self-Organizing Maps

A SOM carried out from data consists of neurons organized in a low (2 or 3) dimensional grid. Each neuron in the grid (map) is connected to the input vector through a d-dimensional connection weight vector $m = \{m1,...,md\}$ where d is the size of the input vector, x. In our situation $x = 4$. The connection weight vector is also named codebook vector (Drachen et al., 2009). A SOM aims to minimize the distance between the codebook vector and the real vector in an iterative way, according to a learning parameter η, calculating a Vector Quantization (VQ) error that produces an approximation to a continuous probability density function $p(x)$ of the vectorial input variable x using a finite number of codebook vectors m (Kohonen, 1990).

By the fact that SOM's performances are closely related to the initial input weights, 200 SOMs were developed, and the one that minimized the VQ error was then selected. A SOM of size 10x10 with a hexagonal toroidal topology was selected. We chose the hexagonal structure because it gives each unit more neighboring connections, allowing better interaction with the adjacent units. The algorithm was trained on 300 epochs to reach the minimum distance

between the codebook vector and the real vector. The learning process was carried out using the batch learning method. Due to the SOM's properties, this algorithm can be used for clustering data in an unsupervised way. Indeed, after the convergence, the units are organized in areas whose proximity in the grid space approximately reflects their proximity in the original space. Therefore, samples with a similar sequence of events are classified from prototype vectors that are close on the map (Palamara et al., 2011a).

3.1.8.3 Convergent Validity

The results showed statistically significant positive correlations between Logic and Raven (ρ = 0.468; p<.001), Memory and WISC-IV Inverse Span subtest (ρ = 0.356; p<.001), and Vispa and NEPSY test (ρ = 0.442; p<.001).

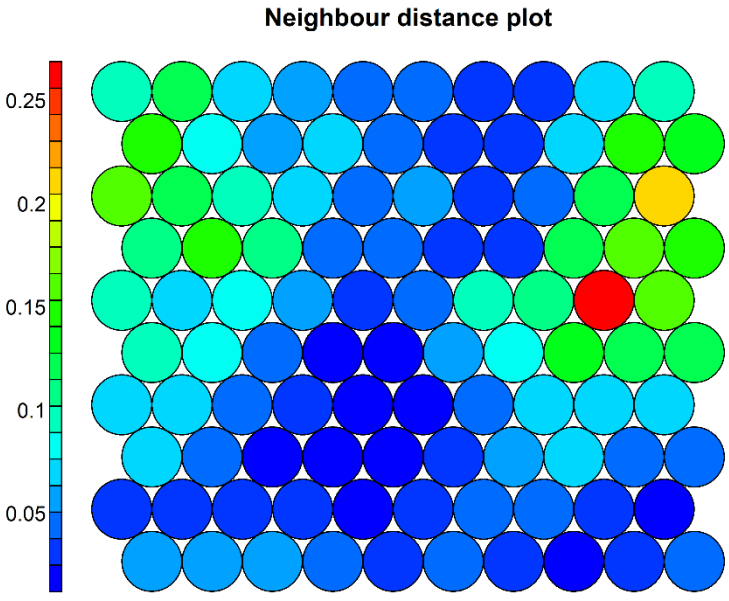

Figure B.1 The U-matrix found after the SOM training on the normalized data.

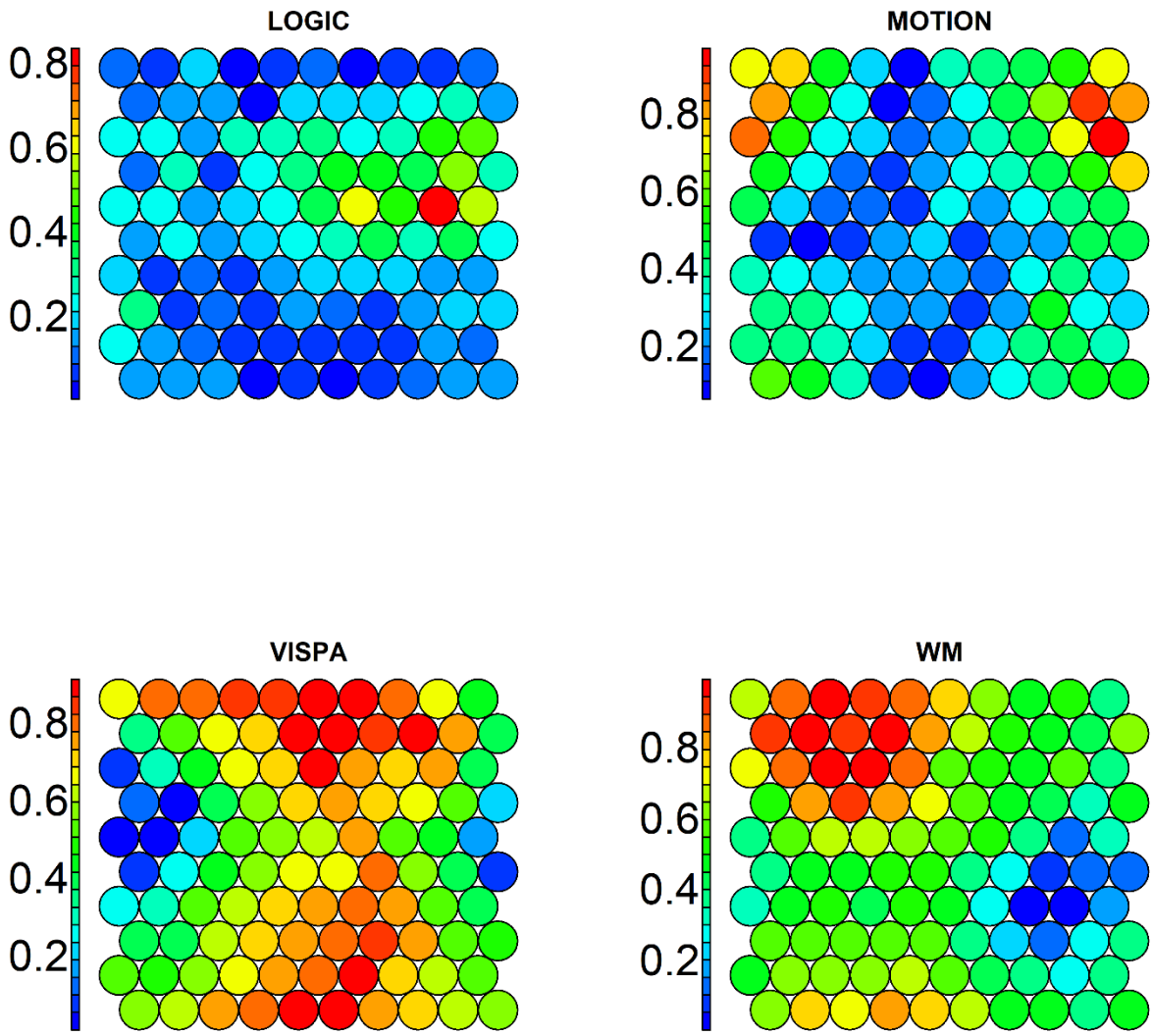

Figure B.2 Property plot of the four cognitive variables. The properties of each unit have been calculated and shown in color code from blue (lower values) to red (highest values). Usually, can be used to visualize the similarity of one particular object to all units in the map, to show the mean similarity of all units and the objects mapped to them.

Cluster-ID	*True Positive*	*False Positive*	*False Negative*	*True Negative*	*Accuracy by class*	*Precision*	*Recall*	*F1 Score*
vhLR-aAll	10	1	1	47	0.97	0.91	0.91	0.91
aALL-lWM	1	1	1	56	0.97	0.50	0.50	0.50
vhLR-lWM	5	2	2	50	0.93	0.71	0.71	0.71
vlVP-lVA	1	0	1	57	0.98	1.00	0.50	0.67
vhLR-vlVA	4	0	1	54	0.98	1.00	0.80	0.89
vhLRWM	3	0	1	55	0.98	1.00	0.75	0.86
ahAll	15	3	0	41	0.95	0.83	1.00	0.91
vhLR-lVP	4	2	0	53	0.97	0.66	1.00	0.80
vhALL	7	0	2	50	0.97	1.00	0.77	0.89

Table B.1 Evaluation Metrics in the imbalanced sample after the AdaBoost implementation.

Cluster-ID	*True Positive*	*False Positive*	*False Negative*	*True Negative*	*Accuracy by class*	*Precision*	*Recall*	*F1 Score*
vhLR-aAll	10	0	1	48	0.98	1.00	0.91	0.95
aALL-lWM	1	2	1	55	0.95	0.33	0.50	0.40
vhLR-lWM	5	16	2	36	0.69	0.24	0.31	0.36
vlVP-lVA	1	0	1	57	0.98	1.00	0.50	0.67
vhLR-vlVA	4	0	1	54	0.98	1.00	0.80	0.89
vhLRWM	2	1	2	54	0.95	0.67	0.50	0.57
ahAll	4	2	11	42	0.78	0.67	0.27	0.38
vhLR-lVP	3	3	1	52	0.93	0.50	0.75	0.60
vhALL	4	1	5	49	0.90	0.80	0.44	0.57

Table B.2 Evaluation Metrics in the balanced sample after the AdaBoost implementation.

Cluster-ID	*True Positive*	*False Positive*	*False Negative*	*True Negative*	*Accuracy by class*	*Precision*	*Recall*	*F1 Score*
vhLR-aAll	10	1	1	47	0.97	0.91	0.91	0.91
aALL-lWM	2	1	0	56	0.98	0.67	1.0	0.80
vhLR-lWM	5	0	2	52	0.97	1.00	0.71	0.83
vlVP-lVA	2	0	0	57	1.00	1.00	1.00	1.00
vhLR-vlVA	4	0	1	54	0.98	1.00	0.80	0.89
vhLRWM	4	0	0	55	1.00	1.00	1.00	1.00
ahAll	15	4	0	40	0.93	0.79	1.00	0.88
vhLR-lVP	3	0	1	55	0.98	1.00	0.75	0.86
vhALL	8	0	1	50	0.98	1.00	0.88	0.94

Table B.3 Evaluation Metrics in the imbalanced sample after the ANN implementation.

Cluster-ID	*True Positive*	*False Positive*	*False Negative*	*True Negative*	*Accuracy by class*	*Precision*	*Recall*	*F1 Score*
vhLR-aAll	10	1	1	47	0.97	0.91	0.91	0.91
aALL-lWM	2	1	0	56	0.98	0.67	1.00	0.80
vhLR-lWM	6	0	1	52	0.98	1.00	0.86	0.92
vlVP-lVA	2	1	0	56	0.98	0.67	1.00	0.80
vhLR-vlVA	4	0	1	54	0.98	1.00	0.80	0.89
vhLRWM	4	0	0	55	1.00	1.00	1.00	1.00
ahAll	14	1	1	43	0.97	0.93	0.93	0.93
vhLR-lVP	3	0	1	55	0.98	1.00	0.75	0.86
vhALL	9	1	0	49	0.98	0.90	1.00	0.95

Table B.4 Evaluation Metrics in the balanced sample after the ANN implementation.

3.2

Unlocking Cognitive Patterns: A Comparative Exploration of Linear and Deep Dimensionality Reduction Approaches in Clustering Students' Cognitive Profiles.

Orsoni, M., Giovagnoli, S., Garofalo, S., Mazzoni, N., Spinoso, M., & Benassi, M. (2024). Unlocking Cognitive Patterns: A Comparative Exploration of Linear and Deep Dimensionality Reduction Approaches in Clustering Students' Cognitive Profiles. (under review)

3.2 Unlocking Cognitive Patterns: A Comparative Exploration of Linear and Deep Dimensionality Reduction Approaches in Clustering Students' Cognitive Profiles.

3.2.1 Abstract

Cognitive profiling plays a crucial role in understanding learning dynamics, it contributes significantly to the development of students' metacognitive skills and awareness of the learning process, thereby facilitating the adoption of tailored learning experiences. Clustering, proves effective in cognitive profiling. However, the challenge of the "curse of dimensionality" introduces complexities that can impact the accuracy of cluster subject attribution. This paper investigates the evaluation of various cluster internal validation metrics and cluster stability using a dataset of 1626 participants comprising 54 items across six cognitive domains from the digital assessment tool, PROFFILO. We employ three clustering procedures—K-means, Gaussian Mixture Models, and Fuzzy-C Means—on raw data and apply linear (Principal Component Analysis) or non-linear (Variational Autoencoders), or a combination of PCA and VAE dimensionality reduction techniques. Results indicate that, for high-dimensional cognitive domains, a combination of PCA and VAE yields superior clustering quality. Conversely, in less high-dimensional domains, the VAE outperforms the PCA approach. In summary, the application of dimensionality reduction techniques demonstrates promising outcomes in student cognitive profiling, especially for data characterized by high dimensionality and heterogeneity. These findings have practical implications for advancing personalized learning experiences and enhancing our understanding of the intricate relationships within students' cognitive domains.

3.2.2 Introduction

Understanding the learner and pinpointing the specific characteristics that markedly contribute to the success of learning experiences is crucial (Altun, 2016). Cognitive profiles play a pivotal role in learning, showing a significant relationship with academic performance particularly in children (Altun, 2016; Nesayan et al., 2018; Webster, 2002). Nesayan et al. (2018), conducted research examining the relation between cognitive profiles and academic performance in 6 to 13 year old students. The findings revealed how children with enhanced cognitive functions, especially in tasks related to executive functions, demonstrated higher academic performance. This result, was consistent with findings reported in the existing literature (Becker et al., 2014; Dulaney et al., 2015; García-Madruga et al., 2014; Vandenbroucke et al., 2017). Then, a comprehensive representation of cognitive profiles, including cognitive style, learning style, and personality, is instrumental in cultivating metacognitive skills, heightening student awareness of their learning process (Webster, 2002), and facilitating the adoption of tailored learning experiences (Altun, 2016). Clustering constitutes a pivotal domain within unsupervised pattern recognition, and can be a relevant approach in finding suitable student's cognitive profile. Its fundamental purpose is to partition a collection of unlabeled samples into distinct subsets according to a predefined objective function. The primary aim is to minimize inter-partition similarity while concurrently enhancing intra-partition similarity (Jayanth Krishnan & Mitra, 2022). Nevertheless, a widely recognized challenge within the machine learning community, known as the "curse of dimensionality", wherein datasets exhibit exceptionally high dimensions, can exert a substantial impact on clustering methodologies (Altman & Krzywinski, 2018). Increasing dimensions introduces challenges such as higher error rates and exponential running times. While it theoretically implies more information, the practical reality involves increased noise and redundancy due to covariation among features (Dessureault & Massicotte, 2022). One way to address this problem is the use of dimensionality

reduction techniques. These can significantly improve the quality and computational efficiency of the clustering process (Anowar et al., 2021; Gotoh, 2004). Principal Component Analysis (PCA) stands out as a widely employed dimensionality reduction technique across diverse domains (Salem & Hussein, 2019; B. Zhao et al., 2021). Functioning as a linear method, PCA facilitates the linear transformation of input data, generating new independent variables known as principal components. The primary objective is to capture the utmost variance inherent in the dataset through this transformation (Anowar et al., 2021). A more contemporary approach, stemming from the realm of deep learning, utilizes autoencoders (AE) and variational autoencoders (VAE) to achieve the objective of identifying a lower-dimensional latent space for input data through the application of nonlinear functions (E. Lin et al., 2020; Yan et al., 2023). Yan et al. (2023), found that PCA and t-distributed stochastic neighbor embedding (t-SNE) are less effective than VAE in the task of learning latent representations of cells particularly with data exhibiting high heterogeneity. In a different study, Nguyen and colleagues (2021) incorporated Principal Component Analysis (PCA) and Clustering-based Autoencoder (CAE). PCA was utilized to establish a novel data representation space, with the aim of augmenting CAE's ability to capture latent and crucial features within the data. The proposed method demonstrated superior performance across various benchmark datasets in comparison to alternative approaches. The present study aims to investigate linear and nonlinear dimensionality reduction techniques and their influence on several clustering validation metrics commonly employed for assessing cluster quality. The proposed methodology commenced with the implementation of the widely recognized linear Principal Component Analysis (PCA), progressed to the nonlinear approximation of the latent space through the application of Variational Autoencoders (VAE), and ultimately examined the dual application of PCA followed by VAE. Subsequently, upon identifying the clustering solution, we proceeded to implement a Bayesian Network derived from the data. This network serves the dual purpose of

delineating the conditional relationships among items and clusters, as well as furnishing a model of the cognitive profile-associated environment to which students belong. The general overview of this work is structured as follows: In Section 2, we establish the foundation by presenting background information and a comprehensive overview of our proposed approach. Building upon this foundation, Section 3.2.3.1 introduces the data acquisition instrument. Subsequent sections then delve into the complexities of clustering. Section 3.2.3.2 places emphasis on cluster tendencies measures, shedding light on observed patterns. This is succeeded by a thorough exploration of cluster validation techniques in Section 3.2.3.3 and a detailed examination of various clustering methods in Section 3.2.3.4. The focus then transitions to dimensionality reduction techniques, with Section 3.2.3.5 elaborating on Principal Component Analysis (PCA), followed by a discussion on Variational Autoencoders (VAE) in Section 3.2.3.6. Section 3.2.6 elucidates the discovered insights, providing a nuanced understanding. Shifting the focus to the clustering-item structure of the test from a Bayesian Network perspective, Section 3.2.6.3 is presented. Culminating the discussion, Section 3.2.7 briefly explores the implications of our findings. Finally, Section 3.2.8 consolidates the key takeaways, leaving readers with our concluding remarks.

3.2.3 Method

3.2.3.1 Participants and Instrument

All procedures adhered to the ethical standards established by national committees on human experimentation and were in accordance with the Helsinki Declaration of 1975, as revised in 2008. Approval for the study was obtained from the University of Bologna Bioethics Committee. Both parents and youths provided written and online informed consent to participate in the study. Due to our data anonymization policy, demographic characteristics of the sample are only available for a subset of students (n = 292) as detailed in a prior study by

Orsoni et al. (2023). A total of 1626 participants within the age range of 6 to 16 were considered eligible for subsequent analyses. The cognitive abilities of all students were evaluated using an online digital game called PROFFILO, specifically designed for assessing students' cognitive profiles (Orsoni et al., 2023). The instrument convergent validity and specifications were already included in other published papers (Orsoni et al., 2021; 2023). Briefly, the PROFFILO assessment comprised six distinct sub-tests (games), each tailored to evaluate a specific cognitive function, namely logical reasoning, visuospatial attention, motion perception, phonological awareness, verbal comprehension and working memory, and lasted 20-25 minutes per participant. A total of 54 items have been administered to all the participants divided as follow:

- Logical Reasoning: 15 items. The test consists of a series of visual pattern matrices, each with one missing part. The task for the test-taker is to identify the missing piece from multiple choices. The data is binary, representing 0 for incorrect and 1 for correct answers.
- Visuospatial Attention: 3 items. The task require the individual to focus their attention to specific visual elements in space, by responding to specific cues while ignoring distractions. The data is continuous in a range between 0 and 1.
- Motion Perception: 5 items. In the current task, participants has to recognize directions of moving stimuli, obtained from white dots moving against a black background. This task allow to assess the subject's motion perception skills. The data is binary, representing 0 for incorrect and 1 for correct answers.
- Phonological Awareness: 13 items. In the tasks, the test-taker is presented with two auditory stimuli, and the task requires selecting the word that corresponds to a word that actually exist, while disregarding the non-word counterpart. The data is binary, representing 0 for incorrect and 1 for correct answers.

- Verbal Comprehension: 17 items. The test involves the presentation of spoken sentences or phrases to the test-taker, who is then required to select a corresponding picture that accurately represents the meaning of the presented linguistic content. The data is binary, representing 0 for incorrect and 1 for correct answers.
- Working Memory: 2 items. The test involves presenting a participant with a sequence of numbers, and then asking them to recall the items in reverse order. The length of the sequence increases proportionally with the participant's performance improvement. The test is interrupted after two consecutive errors. The data is continuous, ranging from a minimum value of 0.

3.2.3.2 Cluster Tendencies

The challenge of ascertaining the presence of clusters, as a preliminary step before the execution of the clustering process, is termed as the assessment of clustering tendency. As expressed by Bezdek and Hathaway (2002), diverse formal techniques grounded in statistics, as well as informal methods for assessing clustering tendency, have been elaborated in literature. In the study we employed both statistical and graphical measures to display the cluster tendencies of our data. The Hopkins statistic (Hopkins & Skellam, 1954) can serve to assess the spatial uniformity of data and identify patterns of clustering within the dataset. A value closer to or exceeding 0.5 signifies data uniformity, implying a diminished degree of clusterability within the dataset, whereas values approaching 0 suggest heightened clusterability. Other measures have been developed to assess graphically the cluster tendency as the Visual Assessment of Tendency (VAT) (Bezdek & Hathaway, 2002), and the improved Visual Assessment of Tendency (iVAT) (Pham et al., 2018; L. Wang et al., 2010). The measures of graphical tendencies, VAT and iVAT, are located in the Supplementary Materials folder available on the OSF repository https://osf.io/4vueh/. The evaluation of cluster tendency has served as a

benchmark metric for discerning the existence of clusters following the implementation of various preprocessing steps, ranging from linear to deep clustering-based algorithms.

3.2.3.3 Internal cluster validation metrics and stability

In unsupervised machine learning, it is not feasible to reference an output variable or output layer for estimating the performance of a classification/regression algorithm. In situations where external information is unavailable, internal validation methods offer a means to assess the quality of the clustering structure (Ezugwu et al., 2022; Palacio-Niño & Galiano, 2019). The underlying concept behind internal validation metrics revolves around two key aspects: cohesion and separation. Cohesion assesses the proximity of elements within the same cluster, whereas separation quantify the degree of demarcation between clusters (Palacio-Niño & Galiano, 2019). A good clustering is when it has high separation between clusters and high cohesion within clusters. In the present work we applied three different internal validity measures: the Calisnki-Harabasz index (CH) (Caliński & Harabasz, 1974), the Davies-Bouldin index (DB) (Davies & Bouldin, 1979), and the Silhouette score (Rousseeuw, 1987). These three metrics are suitable for the type of clustering techniques we adopted (K-means, Gaussian Mixture Models, Fuzzy clustering) (Ezugwu et al., 2022; Palacio-Niño & Galiano, 2019). The Calinski–Harabasz index assesses the compactness or proximity of clusters by computing the distances between the points in a cluster and their respective centroids. For better results CH is maximized (Caliński & Harabasz, 1974). The Davies-Bouldin index assesses the average inter-cluster similarity between any two clusters and their nearest neighbors. To achieve optimal results, minimizing the Davies-Bouldin index is desirable (Davies & Bouldin, 1979). The Silhouette score for a data point is calculated by using the average distance from other points in the same cluster $a(i)$ and the average distance from the nearest neighboring cluster $b(i)$. The silhouette score ranges from -1 to 1. A high silhouette score indicates that the object is well

matched to its own cluster and poorly matched to neighboring clusters. A score around 0 indicates overlapping clusters, and a negative score suggests that the data point might be assigned to the wrong cluster. A more in-depth exploration of internal cluster validation metrics is available in the works of Ezugwu et al. (2022), and Palacio-Nino and Berzal (2019). A pivotal consideration in assessing cluster validity is stability, which entails that a meaningful and valid cluster should exhibit resilience and not readily dissipate under non-essential alterations to the dataset (Hennig, 2007; T. Liu et al., 2022). Building on the clustering model identified as the winner through validation indices, we employed a bootstrap resampling method for assessing stability. The cluster stability of each individual cluster in the initial clustering configuration is quantified by calculating the average Jaccard coefficient across all iterations of the bootstrap resampling process (n = 50) (Garcia-Rudolph et al., 2020; Hennig, 2007). According to Hennig (Hennig, 2008), a valid and stable cluster is expected to exhibit a mean Jaccard similarity value of 0.75 or higher. Within the range of 0.6 to 0.75, clusters may suggest patterns in the data, yet the uncertainty regarding the precise inclusion of points in these clusters is present. Clusters with Jaccard values below 0.6 should be approached with caution and not readily trusted. For clusters to be deemed" highly stable", they should demonstrate average Jaccard similarities of 0.85 and above.

3.2.3.4 K-means, Gaussian Mixture Models, Fuzzy C-means

We focused on three different partitioning clustering algorithms: K-means (MacQueen, 1967), Gaussian Mixture Models (D. Reynolds, 2009), Fuzzy C-means (Hashemi et al., 2023; Nayak et al., 2015b). According to Ezugwu et al. (2022), in a partitional clustering algorithm, data is systematically organized into a structure, with no inherent hierarchy. The dataset, consisting of n objects, undergoes iterative partitioning into a predefined number, *k*, of distinct subsets. This partitioning process is guided by the optimization of a criterion function. The overarching goal is to identify the partition that minimizes the error, with a fixed number of clusters. The

algorithm starts with an initial dataset partition and iteratively assigns data points or patterns to clusters, strategically minimizing the overall error.

Partitioning clustering algorithms can be categorized into three distinct topologies: Hard/Crisp Clustering, Fuzzy Clustering, and Mixture Resolving Clustering (Ezugwu et al., 2022). The K-means algorithm is classified under the Hard/Crisp topology. In this category, each data object is assigned to a single cluster exclusively. The primary goal of K-means clustering is to segment the space based on predefined k centroids. These centroids, which represent the mean of a cluster, determine the number of clusters. The algorithm iteratively segments the space with the aim of maximizing the similarities among objects within the same cluster (intra-cluster similarity), while ensuring these are greater than the similarities with objects in different clusters (inter-cluster similarity). The Fuzzy C-means clustering (FCM) belongs to the Fuzzy clustering family.

It is one soft clustering algorithms where the clusters are defined in fuzzy sets. In soft clustering, each data object simultaneously belongs to more than one cluster (Hashemi et al., 2023). In this approach, clusters are permitted to intersect, a phenomenon referred to as Fuzzy overlap (Ezugwu et al., 2022). This overlap mirrors the ambiguity of cluster boundaries by enumerating the data points with substantial membership in the intersecting clusters. This clustering technique proves advantageous for data point clusters with indistinct and poorly separated boundaries (Kaufman & Rousseeuw, 1990). As the K-means algorithm, the number of data clusters k needs to be specified beforehand. Moreover, the U-matrix of the model get values between 0 and 1. The number of rows are equal to the number of clusters, and the number of columns correspond to the number of data points. The sum of the elements for each column must be 1. When all the components of the U-matrix are either 0 or 1, than FCM become the K-means (Hashemi et al., 2023).

According to Reynolds (D. A. Reynolds, 2018), Gaussian Mixture Models (GMMs) are parametric probability density functions expressed as a weighted sum of Gaussian component densities. Within this modeling framework, the model parameters are determined through the iterative application of the expectation-maximization (EM) algorithm or maximum a posteriori (MAP) estimation, both of which involve the available data (Ezugwu et al., 2022; Reynolds, 2009; Reynolds, 2018). In the clustering perspective, a set of data objects is assumed to originate from a combination of instances across multiple probabilistic clusters. The selection of a specific probabilistic cluster is guided by the probabilities associated with each cluster for generating the observed objects. Subsequently, a sample is selected based on the probability density function of the chosen cluster. It is presumed that the dataset is a composite of a specified number of distinct cluster groups, each contributing in varying proportions during the clustering process (Ezugwu et al., 2022). Detail about the hyperparameter used in the analyses can be found in Section 3-2-4.

3.2.3.5 Principal Component Analysis

Principal Component Analysis (PCA) serves as a multivariate method employed for the examination and simplification of datasets by reducing dimensionality while retaining pertinent information (Abdi & Williams, 2010; Wold et al., 1987). It emerges as a valuable instrument for discerning systematic variation from noise within datasets (Kurita, 2014). Operating through a linear transformation, PCA generates new variables known as principal components, which are uncorrelated and adept at encapsulating the maximum variance within the data. This technique proves valuable for exploratory analysis, accommodating both quantitative and qualitative variables (Abdi & Williams, 2010). PCA relies on eigen decomposition and singular value decomposition for computational efficacy (Wold et al., 1987). Furthermore, within the context of the Manifold Hypothesis and representation learning, PCA enables the linear discovery of sample representations (Schuster & Krogh, 2021). In our research, we utilized

Principal Component Analysis (PCA) to select the optimal number of principal components for each cognitive domain based on their ability to explain at least 80% of the variance in the original dataset. The selection criteria were guided by the robustness measure, which involves calculating the ratio of the sum of eigenvalues associated with the chosen principal components to the sum of all eigenvalues. This robustness measure ensures that the selected principal components effectively capture a substantial proportion of the total variance in the data (David & Jacobs, 2014). More detail about the number of components used in the analyses can be found in Section 3-2-4.

3.2.3.6 Variational Autoencoders

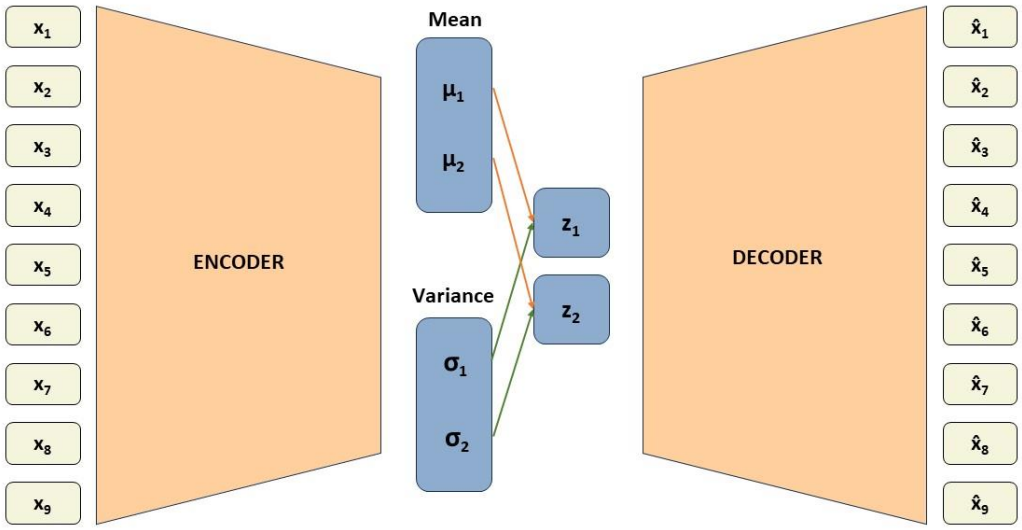

Figure 3-2-1 Graphical illustration of the structure of a Variational Autoencoder model.

As suggested by Kingma and Welling (2013) Variational Autoencoders (VAE) are the principled framework for learning deep latent-variable models. In recent years, they have found application across a diverse range of uses, including generative modeling, semi-supervised learning, and representation learning (Kingma & Welling, 2019). The VAE method proposed

here is based on the approach presented in Yan and colleagues (2023). The encoder, following the structure of the recognition model (Kingma & Welling, 2019), is designed to capture the latent representation of input features and generate samples from the decoder or generator model (Figure 3-2-1). The objective is to minimize the error between the reconstructed data $\hat{x}_i$ and the original input x_i. The encoder function $q(z|x)$ takes an input x (features related to cognitive domains) and maps it to a multivariate Gaussian distribution in the latent space. This distribution is characterized by two parameters, the mean $\mu(x)$ and the standard deviation $\sigma(x)$. The formulation of the encoder function is as follows 14:

$$\text{Encoder: } q(z|x) = \mathcal{N}(\mu(x), \sigma(x)^2) \quad (14)$$

Where $q(z|x)$ represents the conditional probability distribution of the latent variable z given the input x. $\mathcal{N}(\mu(x), \sigma(x)^2)$ denotes a multivariate Gaussian distribution with mean $\mu(x)$ and the standard deviation $\sigma(x)^2$. The decoder function or generative model (Kingma & Welling, 2019), $p(x|z)$ models the conditional distribution of the reconstructed input x given the latent variable z. It assumes a Gaussian distribution in the data space, characterized by the mean $\hat{x}_i$ and a diagonal covariance matrix with a constant variance $\sigma^2 I$, where I is the identity matrix. The formulation of the decoder function is as follows 15:

$$\text{Decoder: } p(x|z) = \mathcal{N}(\hat{x}_i, \sigma^2 I) \quad (15)$$

A forward propagation of the VAE can now be described as the encoder q takes x as input and outputs the means $\mu(x)$ and variances σ_z^2 of the normal distributions; then the latent representation z is sampled from the distribution $\mathcal{N}(\mu(z), \sigma_z^2 I)$; and lastly the decoder p takes z as input to reconstruct the input x. The aim of VAE models is two-fold: first, to minimize the reconstruction error between the input data x_i and the reconstructed data $\hat{x}_i$; and second, to

constrain the learned latent distribution to a predetermined distribution (in this case, we presume a multivariate Gaussian distribution). The reconstruction loss gauges the VAE's proficiency in reconstructing the input data, whereas the KL divergence quantifies the extent to which the latent distribution of the VAE diverges from a standard normal distribution. In our study, given the diverse input data types, we employed VAE models and utilized BCEWithLogitsLoss, which integrates a Sigmoid layer with Binary Cross Entropy Loss within a single class when the input data was binary (0 or 1) or multiclass. We utilized the reparameterization trick to ensure the differentiability of the optimization function (Kingma et al., 2015). This approach avoids direct sampling of z from the posterior and instead generates z using the formula $z = \mu_z + \sigma_z^2 + \varepsilon$, where ε is sampled from the standard normal distribution allowing the backward propagation of the gradient and the model parameters updating. The reconstruction error is then calculated as the minimization of the KL loss function 16, and the minimization of BCEWithLogitsLoss 17. We applied BCEWithLogitsLoss also to handle real-valued data normalized between 0 and 1 during the reconstruction process. BCEWithLogitsLoss is well-suited for scenarios where the targets are values within the range of 0 to 1. Finally, the loss error is composed of three components: the sum of KL divergence error, BCEWithLogitsLoss, and the multiplication of the beta β value with the KL divergence error. The beta parameter serves as a crucial hyperparameter, striking a balance between latent channel capacity, independence constraints, and reconstruction accuracy (Higgins et al., 2016). In the context of the Beta-VAE, a variant of the standard VAE designed to unveil disentangled latent factors (Higgins et al., 2016), we specifically assign a value of 0.0001 to this β term.

$$D_{KL}[q(z|\hat{x}_i)||p(z|x_i)] = E[\log\{q(z|\hat{x}_i)\} - \log\{p(z|x_i)\}] \quad (16)$$

$$\mathrm{l}(x,\hat{x}) = L = \sum_{n=1}^{N} l_n = -\sum_{n=1}^{N} w_n\left[\widehat{x_i} \cdot \log\left(\sigma(x_i)\right) + (1-\widehat{x_i}) \cdot \log\left(1-\sigma(x_i)\right)\right]$$

(17)

$$TotalLoss = \mathrm{l}(x,\hat{x}) + D_{KL} * \beta$$

(18)

3.2.3.7 Bayesian Networks for item-cluster structure

Bayesian Networks (BNs) are graphical models that illustrate dependencies among a set of variables using a directed acyclic graph (DAG) structure (Koller & Friedman, 2009; M Scutari & Denis, 2021; Marco Scutari et al., 2017). In DAG, each node corresponds to a variable, and edges depict their (conditional) dependence relationships. Nodes without any connecting edges are considered independent within the network. Together, these variables form a joint distribution $p(X)$ known as the global distribution. BNs facilitate the breakdown of this global distribution into local distributions for each variable X_i, conditioned on its parent variables $Pa(X_i)$:

$$p(X) = \prod_{i=1}^{N} p\left(X_i \mid Pa(X_i)\right)$$

Learning a Bayesian Network (BN) involves two key stages: structure learning and parameter learning. In the structure learning phase, Kitson and colleagues (2023) emphasized two main algorithm categories: Constraint based learning (CBL) and Score-based learning (SBL). In Constraint-based learning (CBL), Conditional Independence (CI) tests on the dataset are employed to identify conditional independence relationships between variables. In contrast, Score-based learning (SBL) algorithms explore different graphs by maximizing goodness-of-fit scores and a specified objective function. Moreover, the tabu search algorithm (SBL) was found by Scutari and colleagues (2019) to achieve the lowest Structural Hamming Distance

(SHD) for large sample sizes. Given these premises, in this study, we employed BNs to uncover the cluster-item structure of the PROFFILO assessment tool. This approach enabled us to explore the conditional relationships between clusters and items and to identify any hierarchical patterns within the test. Subsequently, we employed two tabu search algorithms (SBL): one focused on a single learning structure, and the second involved averaging results from 1000 bootstrap resampling to obtain a consensus Bayesian Network (BN) (Briganti et al., 2023). The comparison between these algorithms to determine the most suitable one relied on conducting a test using graph posterior probabilities and Bayes factors, as outlined in Scutari et al. (2019).

3.2.4 Hyper-parameter search

In our current investigation, various hyperparameters have been employed. This section is organized based on the components of the study, including clustering algorithms, PCA, VAE, and cluster stability. For the clustering algorithms, the following hyperparameter was applied:

- K-means: The n_init parameter was configured to 50, determining the number of times the KMeans algorithm is executed with different centroid seeds.
- GMM: The n_init parameter was set to 10, representing the number of initializations to be performed. Additionally, the covariance type was configured as “full”, indicating that each component has its own general covariance matrix.
- FCM: we left the default parameters here.

Concerning the PCA, the number of principal components have been chosen to get a robustness value > 80%. In line with this, the number of components for each cognitive domain was the following:

- Logical Reasoning: n components = 9. Explained variance = 84%.
- Visuospatial Attention: n components = 2. Explained variance = 86%.
- Motion Perception: n components = 3. Explained variance = 85%.

- Phonological Awareness: n components = 8. Explained variance = 83%.
- Verbal Comprehension: n components = 9. Explained variance = 83%.

Due to the yet low dimensionality of Working Memory domain, we did not applied any dimensionality reduction procedure. The VAE algorithm was carried out by setting several hyperparameters:

- Encoder/Decoder layers dimensions: Two hidden layers with the dimension of 512 and 256 neurons.
- Number of epochs: 1500
- Batch size: 512
- Learning rate: 1e-3
- Dimensions: It refers to the number of latent dimensions. The value is 2 for Visuospatial Attention and Motion Perception domains, while is 3 for Logical Reasoning, Phonological Awareness, and Verbal Comprehension.
- Seed: 42

The stability of the clusters was assessed through the utilization of 50 bootstrap resampling iterations.

3.2.5 Software and packages

The analyses were conducted using a system equipped with an NVIDIA GeForce RTX 2060 graphics card and an Intel i7-10750H CPU operating at 2.60GHz. Python v3.9.16 (Van Rossun & Drake, 2009), and R v4.03 (R Core Team, 2020) were used for the analyses. The clustering algorithms and the PCA were performed on Python by using the *scikit-learn* library (Pedregosa et al., 2012). The *Pytorch* library v.2.0.0 (Paszke et al., 2019) was used for VAE model computation. On R, the *bnlearn* package (Scutari, 2010) was used for Bayesian Networks implementation.

3.2.6 Results

3.2.6.1 Cluster evaluation tendencies

Table 3-2-1 provides a summary of the Hopkins statistic, spanning from the original data without any preprocessing steps to data subjected to Principal Component Analysis (PCA), Variational Autoencoder (VAE), and the combined application of PCA and VAE across the investigated cognitive functions.

	Original	PCA	VAE	PCA+VAE
Logical Reasoning	0.17	0.09	0.06	0.07
Visuospatial Attention	0.28	0.14	0.12	na
Motion Perception	0.14	0.04	0.06	na
Phonological Awareness	0.15	0.09	0.07	0.04
Verbal Comprehension	0.15	0.08	0.06	0.05
Working Memory	0.00	na	na	na

Table 3-2-1 An examination of cluster tendency through Hopkins statistic values, comparing data with no preprocessing to those subjected to Principal Component Analysis (PCA), Variational Autoencoder (VAE), and the combined application of PCA and VAE across the investigated cognitive functions.

The analysis reveals that the Hopkins statistic yielded higher values for the original data without preprocessing in comparison to the preprocessed datasets. Notably, the application of Variational Autoencoder (VAE) and PCA+VAE preprocessing strategies minimized the Hopkins statistic, indicating an improved cluster tendency. This improved tendency is further corroborated by the observed clustering quality, as detailed in the subsequent presentation.

3.2.6.2 Clustering validation and stability

Table 3-2-2 summarize the Silhouette, Calinski–Harabasz, and Davies-Bouldin validation metrics corresponding to each cognitive domain. The findings indicate enhanced cluster quality when employing dimensionality reduction techniques in contrast to the original unprocessed data. Notably, a discernible advancement across all three metrics within each cognitive domain is evident, transitioning from linear preprocessing involving PCA dimensionality reduction to non-linear methods such as VAE, and ultimately, a combination of both PCA and VAE. These results are more evident for cognitive functions that have a higher dimensionality than those with a lower one.

These findings hold practical implications. Specifically, a Silhouette value of 0.21 for K-means clustering in the Logical Reasoning task indicates that the clustering is not well-defined, implying a lack of clear separation among the objects within distinct clusters. Conversely, following preprocessing and employing the combination of PCA+VAE with K-means as the clustering algorithm, a Silhouette score of 0.61 is observed. This higher score signifies enhanced evidence of clusterizability, indicating improved separation of the data into distinct clusters.

Overall, our observations consistently identify the PCA+VAE+K-means combination as the most effective across all three cluster validation metrics for the high dimensionality cognitive domains (Logical Reasoning, Phonological Awareness, Verbal Comprehension). Specifically, in the case of Logical Reasoning, the Silhouette Score indicated a value of 0.61, corresponding to 21 distinct clusters.

The Phonological Awareness domain exhibited a Silhouette Score of 0.80 with 20 unique clusters. Lastly, Verbal Comprehension achieved a Silhouette Score of 0.77 with a total of 7 unique clusters. In contrast, for cognitive domains characterized by lower dimensionality, such as Visuospatial Attention and Motion Perception, the application of dimensionality reduction techniques, in particular the VAE+K-means reduction, led to enhanced cluster quality.

However, the distinction in cluster validation metrics between clustering with PCA or VAE and without is less pronounced compared to the previous scenarios involving high-dimensional domains. In the context of Visuospatial Attention, the Silhouette score displayed a value of 0.49, indicating the presence of 17 unique clusters. The Motion Perception domain showed a Silhouette score of 0.64, corresponding to 4 distinct clusters. For the Working Memory domain, where no dimensionality reduction technique was applied, K-means implementation was utilized. The resulting Silhouette score was 0.96, revealing the presence of 24 unique clusters. In general, over a total of 1142 complete observations, the solution proposed allowed us to discover 1142 unique profiles.

The cluster stability inspection by using fifty bootstrap resample and the average of the Jaccard similarity index of the Silhouette score revealed a good or high stability across cognitive domain with the cluster solution proposed. According to Hennig (2008), the average Jaccard index for the Logical Reasoning revealed a substantial clustering stability with a value of 0.80. The Visuospatial Attention exhibited exceptional stability with an average Jaccard index of 0.86. Similarly, the Motion Perception displayed sufficient stability with an average Jaccard index of 0.8. The Phonological Awareness also showed substantial cluster stability with an average Jaccard index of 0.83. On the other hand, the Verbal Comprehension, with an average Jaccard index of 0.70, indicated satisfactory cluster stability, albeit with some degree of uncertainty. Finally. the Working Memory solution displayed perfect stability with an average Jaccard index of 1. In general, the clustering stability for all the cognitive domain is either good, high, or perfect, suggesting that the clustering method used is effective and reliable for these data.

However, there might be some uncertainty in the clusters of Verbal Comprehension as indicated by their relatively lower Jaccard index values.

	Logical R.			Visuospatial Att.			Motion Perc.			Phonological Aw.			Verbal Comp.			Working Memory		
	SH	CH	DB	SH	CH	DB	SH	CH	DB	SH	CH	DB	SH	CH	DB	SH	CH	DB
K-means	0.21	416.31	1.82	0.41	1105.11	0.96	0.48	1359.6	0.82	0.51	311.49	1.57	0.42	283.4	1.66	0.96	62925.91	0.21
GMM	0.19	171.59	2.31	0.27	673.29	1.2	0.38	995.18	1.18	0.47	228.44	1.8	0.38	171.99	1.99	0.96	51534.93	0.21
FCM	0.21	412.48	1.72	0.41	1102.98	0.97	0.47	1337.35	0.89	0.28	272.05	1.88	0.24	248.89	2.02	0.96	19594.81	0.20
PCA+K-means	0.28	522.70	1.48	0.47	1488.29	0.78	0.54	1869.64	0.67	0.59	406.5	1.22	0.52	370.86	1.26	na	na	na
PCA+GMM	0.22	332.63	1.84	0.45	1317.21	0.81	0.54	1860.46	0.67	0.55	262.72	1.5	0.49	208.86	1.55	na	na	na
PCA+FCM	0.23	519.62	1.48	0.47	1487.32	0.83	0.53	1851.75	1.03	0.31	351.26	1.22	0.26	315.65	1.28	na	na	na
VAE+K-means	0.58	1092.68	0.78	0.49	1699.38	0.64	0.64	3739.94	0.53	0.77	1533.56	0.75	0.77	1627.76	0.72	na	na	na
VAE+GMM	0.45	979.59	0.97	0.35	879.20	0.73	0.55	2411.48	0.68	0.56	651.33	1.21	0.58	726.11	1.04	na	na	na
VAE+FCM	0.54	1090.17	0.84	0.46	1377.34	0.73	0.63	3736.96	0.53	0.75	1350.15	0.84	0.76	1403.94	0.72	na	na	na
PCA+VAE+K-means	0.61	1589.78	0.72	na	na	na	na	na	na	0.8	5645.51	0.55	0.77	3822.46	0.52	na	na	na
PCA+VAE+GMM	0.49	865.53	0.95	na	na	na	na	na	na	0.58	2357.32	0.92	0.68	2452.79	0.55	na	na	na
PCA+VAE+FCM	0.58	1416.00	0.80	na	na	na	na	na	na	0.79	3883.04	0.57	0.69	3546.05	0.55	na	na	na

Table 3-2-2. Analyzing cognitive dimensions through clustered internal evaluation metrics: Silhouette (SH), Calinski–Harabasz (CH), and Davies-Bouldin (DB). Highlighted in bold are the optimal models maximizing metrics within each cognitive domain (Logical Reasoning, Visuospatial Attention, Motion Perception, Phonological Awareness, Verbal Comprehension, Working Memory). Nas referring to models that are not been executed due to low dimensionality of input data.

3.2.6.3 Cluster-item structure

Based on the Log Bayes Factor, the Single DAG structure outperformed the average structure across 1000 bootstrap resampling iterations, with a logBF value of 135.70. Figure 2 illustrates the graphical representation of the cluster-item structure based on the learned Single DAG structure. The cluster variables for each cognitive domain are shown in orange, while the connection variables facilitating the interconnection of different cognitive domains are highlighted in red. Overall, a hierarchical structure of assessment is evident, with Phonological Awareness at the apex, followed by Motion Perception, Visuospatial Attention, Logical Reasoning, Working Memory, and finally, Verbal Comprehension. This hierarchy aligns with the Hierarchy of Cognition (Harvey, 2019). From a bottom-up perspective, sensory/multisensory and perception abilities form

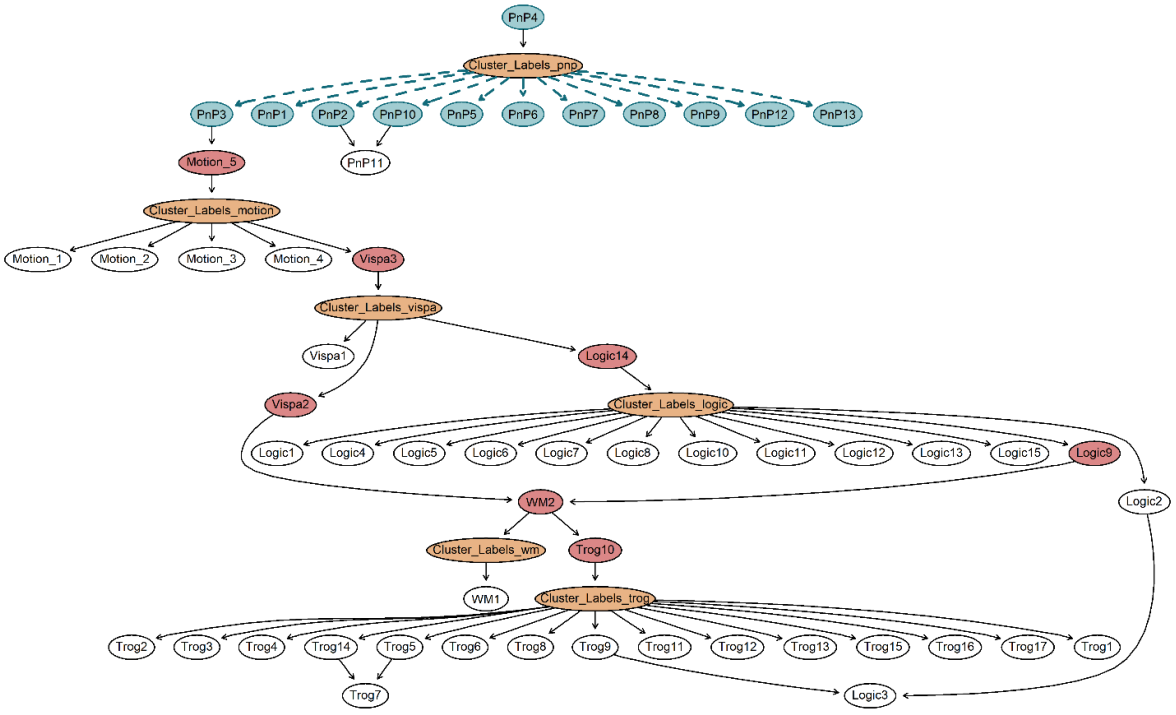

Figure 3-2-2: The Bayesian Network cluster-item structure of the PROFFILO assessment. The Markov blanket of the primary cognitive domain is highlighted in blue, clusters for each cognitive domain are shown in orange, and red denotes variables connecting and indicating conditional dependencies between different cognitive domains.

the base of cognition, followed by attention mechanisms, memory (Working Memory), executive functioning (such as Logical reasoning), and ultimately, language and verbal skills.

3.2.7 Discussion

This study centers on applying and evaluating dimensionality reduction techniques, ranging from linear Principal Component Analysis (PCA) to deep dimensionality based Variational Autoencoders (VAE), and the joint combination of PCA and VAE within the context of clustering. The primary focus is on assessing their impact on clustering students' cognitive profiles over a series of six cognitive domain (Logical Reasoning, Visuospatial Attention, Motion Perception, Phonological Awareness, Verbal Comprehension, and Working Memory). The emphasis of this research lies in recognizing the significance of obtaining a robust representation of the observed heterogeneity in cognitive profiles. This is pivotal for fostering students' metacognitive skills, raising awareness of their learning processes (Webster, 2002), and facilitating the adoption of personalized learning experiences (Altun, 2016). The results suggest that when working with data marked by high heterogeneity and dimensionality, as seen in Logical Reasoning, Phonological Awareness, and Verbal Comprehension, the combined use of PCA and VAE (PCA+VAE) surpasses the individual application of PCA and VAE in learning latent representations. Conversely, for sub-tests with lower dimensionality, specifically Visuospatial Attention and Motion Perception, applying VAE to the test resulted in improved cluster quality. Meanwhile, for the Working Memory test, which comprises only two dimensions, straightforward clustering techniques proved effective in achieving excellent results. Furthermore, the utilization of Bayesian Networks (BNs) enabled the derivation of a cluster-item structure for the test. This approach not only facilitated the identification of relationships among variables but also offered insights into the hierarchical arrangement of cognitive domains within the test.

Nevertheless, certain limitations need to be taken into account. Primarily, the analyses were conducted by encompassing the entire sample of students, spanning from 6 to 16 years old. While this approach may not be particularly pertinent to the methodological scope of the current study, it could gain significance in clinical applications. In such cases, more nuanced models tailored to the specific age groups of students could be developed to enhance precision and relevance. Furthermore, it's important to note that the clusters have undergone internal validation exclusively. This limitation may impact the broader applicability of the results. Subsequent studies should consider validating the proposed approach in varied educational settings or across different cognitive functions to enhance the robustness of the findings.

In summary, the application of dimensionality reduction techniques demonstrates promising results in the realm of student cognitive profiling, particularly for data characterized by high dimensionality and heterogeneity. The insights gained have implications for advancing personalized learning experiences and understanding the intricate relationships within students' cognitive domains.

3.2.8 Conclusion

This study initiates opportunities for enhancing methodologies employed in cognitive profiling, thereby making a substantial contribution to the overarching framework of personalized learning experiences. The encouraging outcomes emphasize the efficacy of dimensionality reduction techniques in elucidating the complexities inherent in students' cognitive domains, thereby laying the foundation for prospective advancements in both educational research and practical applications. As we gain a deeper understanding of the nuances within these cognitive domains, the feasibility of customizing educational approaches for individual students becomes more tangible. This not only caters to the varied learning needs but also establishes a foundation for targeted interventions aimed at enhancing students' skills. Moreover, the investigation into

Bayesian Networks contributes an additional layer of significance to the study by introducing a structured approach for elucidating relationships among variables within cognitive domains. This not only assists in unraveling intricate cluster structures but also provides teachers, psychology and educators with a systematic framework to navigate the hierarchical organization of cognitive skills. In conclusion, this study plays a key role in the ongoing improvement of personalized learning, recognizing and addressing the different ways individual students think and learn.

3.2.9 Supplementary Materials

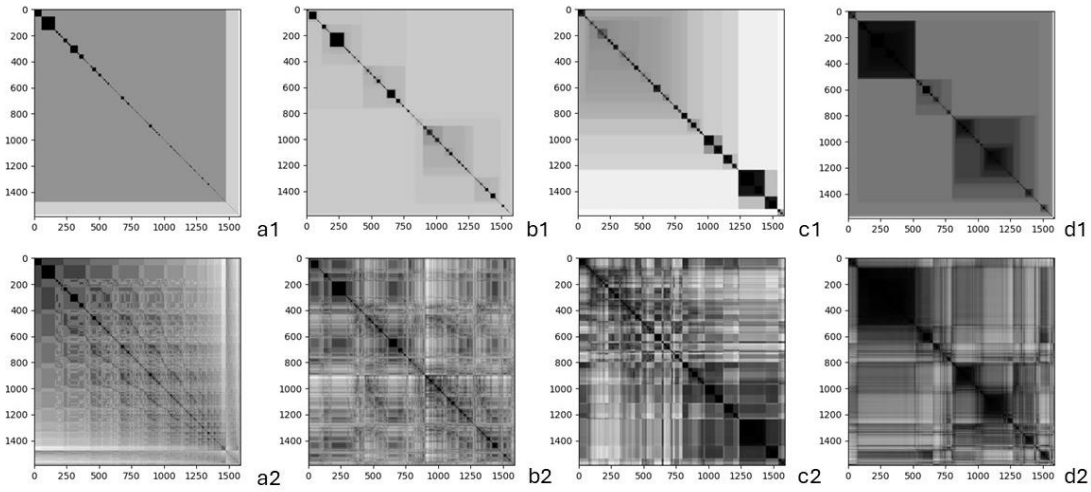

Figure 3-2-3: iVAT and VAT applied to the logical reasoning cognitive function. Panels a1 to d1 depict the iVAT results for raw data, PCA, VAE, and PCA+VAE dimensional reduction approaches, respectively. Correspondingly, panels a2 to d2 are associated with the inspection of VAT cluster tendencies.

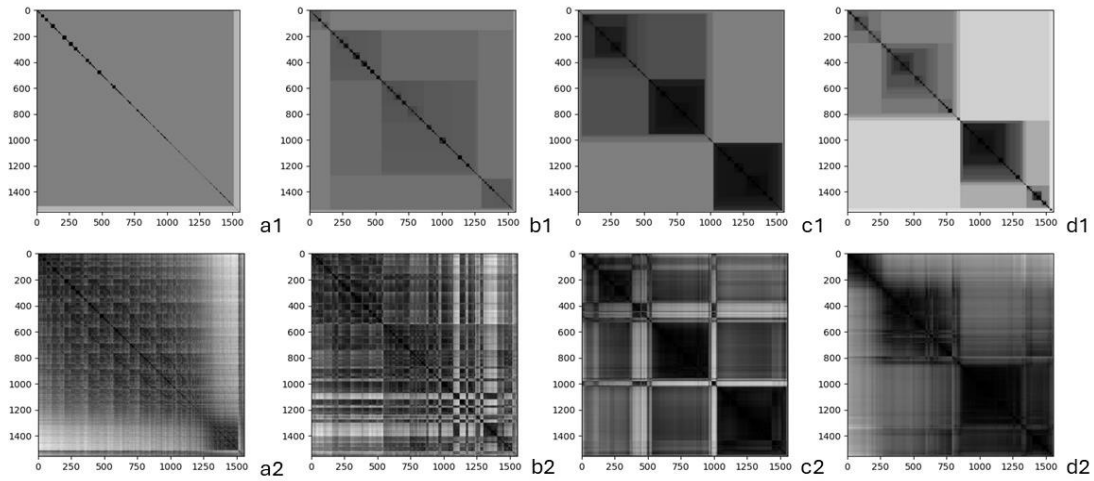

Figure 3-2-4: iVAT and VAT applied to the motion perception cognitive function. Panels a1 to d1 depict the iVAT results for raw data, PCA, VAE, and PCA+VAE dimensional reduction approaches, respectively. Correspondingly, panels a2 to d2 are associated with the inspection of VAT cluster tendencies.

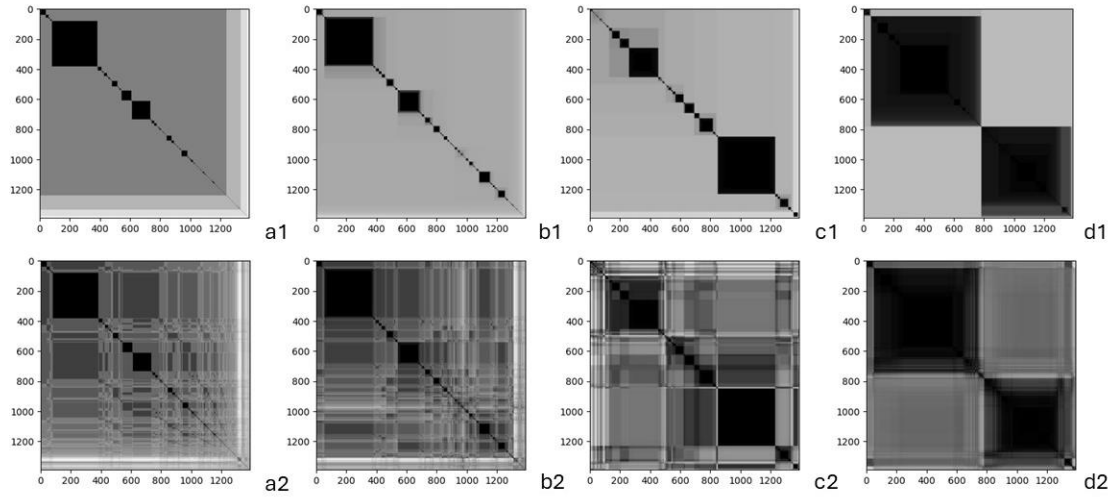

Figure 3-2-5: iVAT and VAT applied to the verbal comprehension cognitive function. Panels a1 to d1 depict the iVAT results for raw data, PCA, VAE, and PCA+VAE dimensional reduction approaches, respectively. Correspondingly, panels a2 to d2 are associated with the inspection of VAT cluster tendencies.

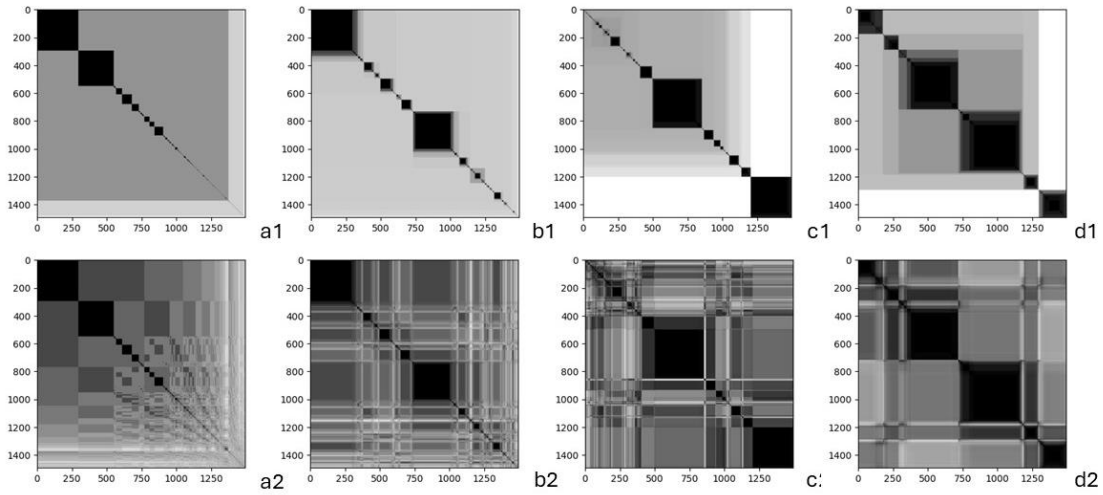

Figure 3-2-5: iVAT and VAT applied to the phonological awareness cognitive function. Panels a1 to d1 depict the iVAT results for raw data, PCA, VAE, and PCA+VAE dimensional reduction approaches, respectively. Correspondingly, panels a2 to d2 are associated with the inspection of VAT cluster tendencies.

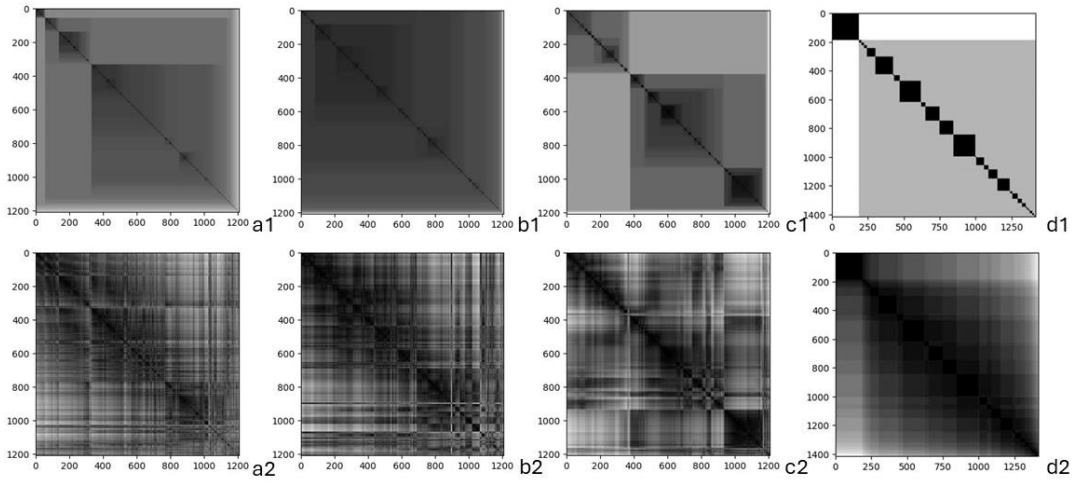

Figure 3-2-6: iVAT and VAT applied to the visuospatial attention and working memory cognitive functions. Panels a1 to c1 depict the iVAT results for raw data, PCA, VAE, and dimensional reduction approaches, respectively of the visuospatial attention function. Correspondingly, panels a2 to c2 are associated with the inspection of VAT cluster tendencies. Panels d1 and d2, are the iVAT and VAT of the working memory raw data function.

3.3

Information Theory, Machine Learning, and Bayesian Networks in the Analysis of Dichotomous and Likert Responses for Questionnaire Psychometric Validation.

Orsoni, M., Benassi, M., & Scutari, M. (2024, January 25). Information Theory, Machine Learning, and Bayesian Networks in the Analysis of Dichotomous and Likert Responses for Questionnaire Psychometric Validation. Retrieved from osf.io/preprints/psyarxiv/r4y68 (under review)

3.3 Information Theory, Machine Learning, and Bayesian Networks in the Analysis of Dichotomous and Likert Responses for Questionnaire Psychometric Validation.

3.3.1 Abstract

The validation of questionnaires, crucial for discriminating between diverse populations, is a standard practice in psychology and medicine. While latent factor models have conventionally dominated psychometric questionnaire validation, recent developments have introduced alternative methodologies such as Network Analysis. This study presents a pioneering approach that integrates information theory, machine learning (ML), and Bayesian networks (BNs) into questionnaire validation. This novel perspective shifts the emphasis from latent factors to individual items. We used the Jensen-Shannon Divergence (JSDd) for item selection, employing three machine learning algorithms (Decision Trees, Random Forests, and Support Vector Machines with a linear kernel) to identify the items with optimal discriminative power. The selection process balanced the number of items against model accuracy in a data-driven manner. Bayesian Networks (BNs) were employed to uncover conditional dependences between items, offering insights into the complex systems underlying the psychological construct. We validated the proposed method on two simulated data sets, one with dichotomous and the other with Likert-scale data. Results show the efficacy of the proposed method in identifying the most discriminative items, thereby enhancing the instrument's discriminative power. Simultaneously, it mitigated respondent burden by minimizing the required number of administered items and providing insights into the criterion validity, content validity, and construct validity of the instrument.

3.3.2 Introduction

The replication crisis in psychology is an indisputable reality. One fundamental factor contributing to it is the prevalence of theories that exhibit a weak logical connection with the empirical hypotheses used for evaluation, precipitating a crisis in psychological theory as well (Oberauer & Lewandowsky, 2019). Recent publications have acknowledged this issue and shed light on its underlying causes (Borsboom et al., 2021; Eronen, 2020; Eronen & Bringmann, 2021; Oberauer & Lewandowsky, 2019). Eronen and Bringmann (2021) delineates three challenges in developing robust psychological theories: the absence of rigorous constraints on theories by empirical phenomena, issues with the validity of psychological constructs, and the impediments encountered in uncovering causal relationships between psychological variables. Moreover, Fried (2020) stated that theories in psychology do not explicitly specify the functional relationship between two variables, the conditions necessary for the hypothesized effect to manifest, or the magnitude of the proposed effect. Many psychological constructs are commonly assessed through the use of questionnaires. They play a crucial role because they enable the identification of specific symptoms and indicators associated with psychological disorders or behaviors and are routinely incorporated into comprehensive assessment protocols (Demetriou et al., 2015; Rosellini & Brown, 2021). Indeed, assessing the validity of scales intended for distinguishing between different clinical populations is a commonplace procedure frequently carried out in psychology and medicine (Trognon et al., 2022). Most questionnaires use questions with binary or dichotomous (such as true/false; presence/absence) and ordinal Likert-type response scales (Rosellini & Brown, 2021) which serve as guidance for clinicians during initial evaluations of individual patients, providing insights about the population they belonging to and offering a quantitative assessment of possible symptoms (Demetriou et al., 2015). The latent variable approach is one of the most commonly used frameworks for validating psychometric questionnaires. It uses the covariation between observed items that

measure behaviour, cognition or affect to determine whether the item relationships align with the established definition of a psychopathology or psychological construct (Rosellini & Brown, 2021). The latent variable approach comprises a first exploratory item pool assessment and an initial construct validity stage, followed by a confirmatory step in which the exploratory results should be replicated with a more restrictive model, usually Confirmatory Factor Analysis (CFA) or a Structural Equations Models (SEM) (Rosellini & Brown, 2021). According to Taherdoost and colleagues (2014), Factor Analysis (FA) serves four distinct functions in data analysis: identifying latent variables that account for the correlation between observed variables, isolating their shared variance from the respective error variances, revealing the underlying factor structure (Costello & Osborne, 2005) and providing the construct validity evidence (Taherdoost et al., 2014). Furthermore, FA can summarize a substantial number of variables (factors) into a more manageable and concise set, thus enabling the formulation and fine-tuning of theoretical frameworks. For these reasons, Exploratory Factor Analysis (EFA) has emerged as the preferred approach for questionnaire interpretation in contemporary research (Kishore et al., 2021; Williams et al., 2010). However, constructing an EFA model is a complex undertaking that requires researchers to make important preliminary assumptions throughout the entire analysis process (Watkins, 2018). Crucially, EFA assumes that items are normally distributed to produce reliable results: deviations from normality and linearity impact the Pearson correlation coefficients (r) that results are derived from. This is problematic because, as mentioned earlier, self-reported questionnaires commonly employ dichotomous or ordinal measurement scales. Furthermore, EFA results are subjective because the researcher's choices for the number of factors and rotational scheme are predominantly guided by pragmatic considerations rather than theoretical principles (Rosellini & Brown, 2021; Watkins, 2018; Williams et al., 2010). Furthermore, within the confirmatory framework, it is imperative to consider the local independence assumption stipulated by structural equation modelling (SEM).

This assumption posits that the residuals associated with observed variables must exhibit no intercorrelation or any form of mutual dependence (Guyon et al., 2017; Sobel, 1997). Unfortunately, this assumption can restrict the exploration of causal relationships among items, thereby limiting the acquisition of crucial insights necessary for theoretical investigations (Guyon et al., 2017). Furthermore, it is important to recognize that latent variables and psychological constructs are distinct entities, and we should only equate them when all requisite causal assumptions have been explicitly delineated (Fried, 2020). Therefore, we should consider alternatives to FA that address some of these limitations. Information theory offers a systematic approach to quantifying and examining information Yeung (2008), typically with Shannon's information measures as (conditional) entropy and (conditional) mutual information. In the context of this paper, we will use information theory to validate and optimize questionnaires in terms of item selection and produce a more concise and efficient questionnaire, thus reducing respondent burden (Brockett et al., 1981). As previously stated, item selection constitutes a crucial aspect of applying EFA. Shannon's information measures can identify the items that offer the greatest amount of information, which helps researchers select those with the greatest discriminatory power across different levels of the construct. As a result, we can improve the overall discriminative power and the analytical accuracy of the study and, at the same time, mitigate respondent burden by identifying redundant or low-information items that can be removed from the questionnaire without substantial information loss. Shannon's entropy (SHE) and the Kullback-Leibler (KL) divergence were the first indices used as item selection methods in the literature (Yigit et al., 2018), followed by Mutual Information (Peng et al., 2005; Ross, 2014) and the Jensen-Shannon divergence (JSD) (Wang et al., 2020). Machine learning (ML) is a field of computer science where algorithms and models are developed to enable computers to learn without the need for explicit programming (M. I. Jordan & Mitchell, 2015). Its primary objective is to analyse data, identify their patterns,

and use these patterns to make accurate predictions or decisions. In recent years, there has been an increasing number of machine learning applications in psychology (Jacobucci & Grimm, 2020; Orsoni, et al., 2023). In this paper, we use supervised machine learning algorithms to determine the optimal balance between the number of questionnaire items and model performance from the data. Supervised learning learns a mapping function, denoted $f(x)$, that produces an output $\hat{y}$ for each input x and can be used to generate predictions. As previously stated, one of the critical tasks in questionnaire validation is structure learning. EFA, CFA and SEM are traditionally used to show the relationships and strengths between the variables and the underlying latent factors in graphical form. Here, we will replace them with Bayesian Networks (BNs) (Koller & Friedman, 2009) because they are more flexible in integrating prior knowledge, updating beliefs using observed data, and learning the structure of questionnaire data to investigate the underlying psychological construct. This amounts to incorporating questionnaire validation in the "network framework" (Briganti et al., 2022), in which psychological constructs are viewed as complex systems linking items from psychometric measurements, symptoms and traits (Borsboom, 2017; Briganti et al., 2022; Fried, 2020; McNally, 2016). This framework provides an alternative perspective on questionnaire validation by integrating principles from Information Theory, supervised machine learning, and Bayesian networks. It encompasses multiple stages, including item selection, theory formulation or refinement and questionnaire construct validation. In contrast to traditional methods that rely on latent factors, this approach emphasizes the individual items themselves. The remainder of the paper is structured as follows. In Section 3-3-3, we will provide some background and a detailed description of our proposed approach. We will then present the results of an empirical validation of the proposed method on two simulated data sets in Section 3-3-4 for both dichotomous (Section 3-3-4-1) and Likert-scale (Section 3-3-4-2) data, followed by a brief discussion (Section 3-3-5) and conclusions (Section 3-3-6).

3.3.3 Methods

In this section, we will introduce the required notions of information theory (Section 3-3-3-1), machine learning (Section 3-3-3-2) and BNs (Section 3-3-3-3) followed by a discussion of the differences from FA (Section 3-3-3-4), item selection with the JSD for dichotomous and Likert-scale variables (Sections 3-3-3-5 and 3-3-3-6), and BN structure learning (Section 3-3-3-7).

3.3.3.1 Information Theory

For item selection, we chose to use the Jensen-Shannon divergence (JSD) to measure the dissimilarity between two probability distributions because its square root represents a metric distance that satisfies triangular inequality (Lin, 1991; Nielsen, 2010) and because it is defined in the range [0, 1]. A value of 0 implies a perfect overlap (dependence), while a value of 1 implies complete separation (independence). In the subsequent sections of the article, we shall denote the distance of the Jensen-Shannon Divergence (JSD) as "JSDd". The JSDd between two probability distributions P and Q is defined as:

$$JSDd(P \parallel Q) = \frac{1}{2}\left(D_{KL}(P \parallel M) + D_{KL}(Q \parallel M)\right) \quad (19)$$

where $D_{KL}(\mathrm{P} \parallel \mathrm{Q})$ represents the Kullback-Leibler divergence between P and Q, and M is the average distribution given by:

$$M = \frac{1}{2}\,(P + Q) \quad (20)$$

3.3.3.2 Supervised Machine Learning

We applied three different machine learning (ML) algorithms: Decision Trees (CART), Random Forests (RF), and Support Vector Machines (SVM) with a linear kernel.

We used 10-fold cross-validation both to learn the models and to tune their hyperparameters, which is crucial in evaluating their performance. We refer the reader to Kuhn and Johnson (2013) for an introduction to these models and to their evaluation. The values we considered for each of the hyperparameters are as follows:

- the maximum number of levels or splits allowed in a tree in CART and RF: 1, 2, 3, 4, 5, 6, 7 and unlimited.
- the number of trees in RF: 100, 200.
- the penalty parameter of SVM, which controls the balance between smoothness and minimizing the training error: 0.0001, 0.001, 0.01, 0.1, 1.0, 10.0, 100.0, 1000.0.

The model evaluation and hyperparameters selection have been conducted by comparing accuracy, precision, recall and the F1 score (Orsoni et al., 2023). These metrics allow for selecting the best-performing model, which exhibits the most effective learned mapping function, through a comparative analysis of various aspects of its performance.

3.3.3.3 Bayesian Networks

BNs are graphical models that represent dependencies between a set of variables with a directed acyclic graph (DAG) structure (Koller & Friedman, 2009; Scutari & Denis, 2021; Scutari et al., 2017). Each node corresponds to a variable within the DAG, and edges represent their (conditional) dependence relationships. Isolated nodes not touched by any edges are deemed independent within the network. Collectively, the variables have a joint distribution $p(X)$, which is called the *global distribution* in this context. BNs facilitate the decomposition of this global distribution into a local distribution for each X_i that is conditional on the parent variables $Pa(X_i)$:

$$p(X) = \prod_{i=1}^{N} p\big(X_i \mid Pa(X_i) \big)$$

Using Bayesian Networks (BNs) for causal inference requires several strong assumptions (Briganti et al., 2022): the presence of a DAG as the underlying structure for the data and the assumptions of causal faithfulness and causal sufficiency (Kitson et al., 2023). Causal faithfulness states that variables in the network manifest probabilistic dependencies if and only if they are linked by edges, whereas causal sufficiency requires that we observe all causal factors affecting the variables. This implies the absence of latent variables, selection bias and systematic patterns of missing values.

Learning a BN involves two fundamental stages: structure learning and parameter learning. In a recent comprehensive review, Kitson et al. (2023) explored the structure learning step, which encompasses various algorithmic approaches grouped into two primary categories: Constraint-based learning (CBL) and Score-based learning (SBL). Constraint-based learning (CBL) applies Conditional Independence (CI) tests to the data to ascertain the conditional independence relationships between variables. Within the realm of CBL, "global discovery" algorithms aim at comprehensively learning the entire graph structure, and "local discovery" algorithms focus on developing the local skeleton of the graph, elucidating relationships involving each individual variable. Score-based learning (SBL) algorithms instead revolve around exploring various graphs while maximizing the goodness-of-fit scores and a designated objective function. The technical details of these algorithms are outside the scope of this work; we refer the interested reader to the survey papers from Briganti, Scutari, and McNally (2022); Kitson et al. (2023); Scanagatta and colleagues (2015); Scanagatta, Salmeròn, and Stella (2019); Scutari et al. (2017).

Various works (Cowell, 2001; Scutari et al., 2019) have explored which type of algorithm can effectively learn the most accurate graph structure from data. Cowell (2001) demonstrated the equivalence between CBL and SBL when we assume a fixed, known topological ordering and use log-likelihood and G2 as matching statistical criteria. Scutari et al. (2019) found that

constraint-based algorithms are more accurate than score-based algorithms for discrete BNs and small sample sizes, while tabu search (SBL) attains the lowest Structural Hamming Distance (SHD) for large sample sizes. Moreover, Colombo and Maathuis (2014) showed in high dimensional space how the PC-Stable learnt graphs with lower Structural Hamming distance (SHD) (Tsamardinos et al., 2006) from the true graph.

Following from these findings, we tested both CBL (PC-Stable) and SBL (tabu search) structure learning algorithms to select the most appropriate one. We combined them with the test built from Bayes factors and graph posterior probabilities presented in Scutari et al. (2019). A good BN practice is learning only statistically significant edges to obtain a sparse and interpretable DAG. To achieve this, we followed the approach described in Briganti, Scutari, and McNally (2022): we performed bootstrap resampling 1000 times, applied structure learning to each bootstrap sample and averaged the resulting DAGs to obtain a consensus BN. We use the data driven threshold from Scutari and Nagarajan (2013) to establish significance, which in turn increases the precision and robustness and helps in assessing the strength of the connections in the network.

3.3.3.4 Contrasting the Current Approach with Factor Analysis

Using BNs instead of FA introduces several key differences in questionnaire analysis and in how we interpret results.

Variable selection and reduction. FA primarily focuses on reducing the dimensionality of a set of observed variables by identifying the underlying latent factors that make them correlated. Its primary objective is to account for the shared variance among the observed variables. In contrast, information theory and BNs adopt a more comprehensive approach by considering the information content and relationships among all observed variables, encompassing potential dependencies that extend beyond the latent factors. These approaches provide a broader

perspective that takes into account the full range of dependencies and information exchange among the variables.

Assumptions about linearity. Classical FA and SEM assume that the relationships between latent constructs and observed variables are linear. Information theory and BNs adopt a more flexible approach by not imposing explicit assumptions about linearity. This allows for the modelling of non-linear and complex relationships, as well as capturing different types of probabilistic dependencies.

Model specification and hypothesis testing. Researchers often adopt a confirmatory perspective, defining a specific model structure and proposing hypotheses regarding the relationships between latent constructs and observed variables. SEM are then employed to test these predefined hypotheses and assess the model's fit to the data. In contrast, BNs are more exploratory, offering the flexibility to derive insights directly from the data and to model probabilistic relationships without rigid assumptions. In addition, we can also examine a priori (theory-driven) hypotheses concerning the relationships between items and compare them with the relationships derived from the data. Furthermore, BNs can easily integrate available contextual information and prior knowledge in the modelling process (Kitson et al., 2023; Zhang & Schuster, 2021), thus enabling the incorporation of prior distributions and the updating of beliefs based on observed data.

Sample size. Classical FA and SEM typically require larger sample sizes to produce stable parameter estimates with adequate statistical power (Wolf et al., 2013). In contrast, information theory and BNs are more robust at smaller sample sizes (Ameur et al., 2022). This is due to their emphasis on modelling probabilistic relationships rather than estimating individual parameters.

3.3.3.5 Items Selection via Bayesian Networks and Jensen-Shannon Divergence for Dichotomous Data

When a questionnaire aims to recognise the presence/absence of certain symptoms or features that are representative of specific populations, it is common to measure them with dichotomous items (Ising et al., 2012; Teng et al., 2010). We used BNs and the JSDd to analyse dichotomous items and, in particular, as an item selection technique for binary classification. The procedure is summarised in Algorithm 1. Firstly, we estimated the marginal probability distribution of group membership for each item in the questionnaires. We then computed the root-mean-squared JSD from the probabilities to obtain the JSDd. The JSDd quantifies the separation between the distributions of the two groups for each item. By generating a distribution of distances (Figure 3-3-1), we can identify items with higher values, specifically higher than the median, indicating a greater distinction between the two populations. Subsequently, we assessed the number of items to include in the model by iteratively eliminating 5% of them in decreasing order of JSDd until only 5% were left. The model that achieves the best balance between accuracy and the number of selected items, thus optimizing the item selection process, is the most parsimonious model that can still effectively distinguish between the populations of interest.

Algorithm 1 Items Selection via Bayesian Networks and Distance of the Jensen-Shannon Divergence

Require: Data
  for $i \leftarrow 1$ to $ncol(Data)$ **do**
    Compute Bayesian Network for $X_i \rightarrow Group$ for all $i \in N, i = 1, ..., N$
    Estimate the marginal probability according to $Group$
    Compute $M = \frac{1}{2}(P + Q)$
    Compute $\mathrm{D_{JS}}(P_i \parallel Q_i) = \frac{1}{2}\mathrm{D_{KL}}(P_i \parallel M) + \frac{1}{2}\mathrm{D_{KL}}(Q_i \parallel M)$
    Compute $JSDd = \sqrt{\mathrm{D_{JS}}(P_i \parallel Q_i)}$
  end for

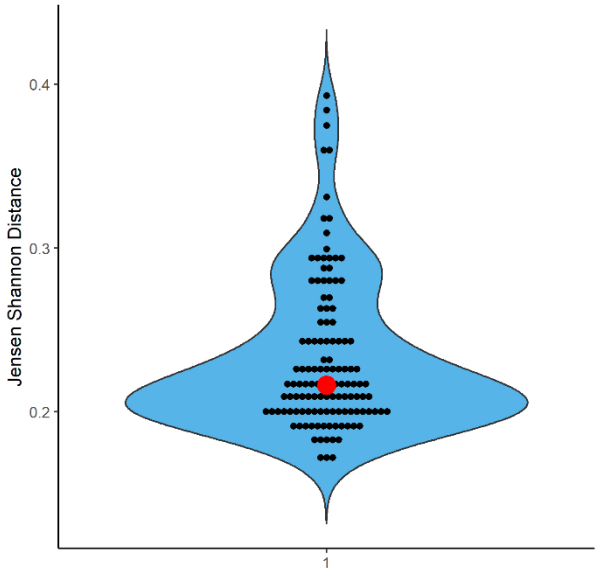

Figure 3-3-1 An example of the level of similarity between two populations as measured by the distribution of the Jensen-Shannon distance over items. Larger values indicate more pronounced differences between the groups, whereas smaller values indicate a higher degree of similarity. The median is shown as a red dot.

3.3.3.6 Items Selection via a Consensus Measure and Jensen-Shannon Divergence for ordinal Likert-scale data

Ordinal data measured on Likert scales comprise ordered categories to convey approximate ordering (such as strongly agree, agree, neither agree nor disagree, disagree, and strongly disagree). The numeric values assigned to the categories have no semantic meaning, rendering calculations of averages or other numeric summary statistics meaningless (Tastle & Wierman, 2006; 2007). To account for this, we incorporated a consensus measure to evaluate the level of consensus or dissensus between items and populations in the estimation of the Jensen-Shannon

distance (JSD) for each item. A consensus measure for ordinal data, based on information theory, has been proposed in Tastle and Wierman (2006, 2007); Wierman and Tastle (2005):

$$Consensus(X_i) = 1 + \sum_{j=1}^{|X_i|} p_j \, i \log_2 \left(1 - \frac{|X_{i,j} - \mu_x|}{d_x}\right) \quad (21)$$

where:

- X_i is the ordinal variable.
- $X_{ij}$is the jth attribute in the Likert scale for X_i
- $|X_i|$ is the number of attributes in the Likert scale for X_i
- p_{ij} represents the probability or frequency associated with each X_{ij}, indicating the relative proportion of individuals who selected that specific response.
- d_x is the width of X_i, defined as $d_x = X_{max} - X_{min}$
- μ_x is the mean of X_i, defined as $\mu_x = \sum_{i=1}^{n} p_i X_i$

Consensus refers to the collective opinion or position reached by a group of individuals acting together; it is characterised by a state of general agreement among the group members. Dissensus can then be understood as its complement, expressed as $Dissensus(X) = 1 - Consensus(X)$.

Algorithm 2 summarises the iterative procedure to compute the JSDd distribution given the Consensus measure. After obtaining the JSDd distribution, the rest of the analysis follows that presented earlier in Section 3-3-3-5.

Algorithm 2 Items Selection via Consensus measure and Jensen-Shannon Divergence

Require: Data
  for $i \leftarrow 1$ to $ncol(Data)$ **do**
    Compute Consensus(X_i) as in (21) for $X_i \rightarrow$ Group for all $i \in N, i = 1, \ldots, N$
    Compute $M = \frac{1}{2}(P + Q)$
    Compute $\mathrm{D_{JS}}(P_i \parallel Q_i) = \frac{1}{2}\mathrm{D_{KL}}(P_i \parallel M) + \frac{1}{2}\mathrm{D_{KL}}(Q_i \parallel M)$
    Compute $JSDd = \sqrt{\mathrm{D_{JS}}(P_i \parallel Q_i)}$
  end for

3.3.3.7 Discovering Questionnaire Construct by using Bayesian Networks

In a recent study, Briganti, Decety, et al. (2022) found BNs to be a valuable tool for researchers using psychometric data. These networks help identify causal relationships, enabling the discovery of novel insights into psychological constructs. Furthermore, they facilitate the generation of new hypotheses and provide supportive evidence for pre-existing ones. With the growing popularity and application of network theory to psychological constructs in recent years, emerging studies such as those by Briganti, Scutari, and McNally (2022), Briganti, Scutari, and Linkowski (2021), and Scutari and Denis (2021) provide valuable guidance and support to researchers in obtaining accurate results from such analyses.

In this study, we employed a BN to discover the underlying structure of the questionnaire and infer the psychological construct after applying the item selection procedure discussed in the previous sections.

3.3.4 Results

We will now apply the approach we proposed in Section 2 to simulated data with dichotomous and ordinal variables. The analyses were carried out using R v4.0.3 (R Core Team, 2020) and Python v3.9 (Van Rossum & Drake, 2009). We used the *bnlearn* R package (Scutari, 2010) to implement BNs the *scikit-learn* Python library (Pedregosa et al., 2011) to implement the other ML models.

3.3.4.1 Example on dichotomous data

To evaluate the effectiveness of the proposed method with dichotomous data, we conducted experiments using three separate data sets which are described in detail in Appendix A. The disparities among these data sets are reflected in the degree of agreement among the responses to 50 dichotomous items within two distinct samples, each comprising 200 individuals. Such

data are similar to those arising from a questionnaire consisting of dichotomous items that would discriminate between populations.

We established three tiers of group similarity for the item responses between the generated samples, denoted as “low”, “average”, and “high”. The items in the “low” data show a significant disparity in subjects’ responses. In contrast, the “average” and “high” data successively mitigated this level of dissimilarity.

Further information on the data generation process can be found in Appendix A (Section 3-3-3-8). Figure 3-3-2 shows the violin plots of the JSDd obtained from Algorithm 1 and illustrates the item distances between respondents from the two groups. The sample within the “low” data has a median JSDd value of 0.06 (minimum: 0.02; maximum: 0.12). The “average” data has a median JSDd value of 0.26 (minimum: 0.11; maximum: 0.35). Lastly, the “high” similarity data has a median JSDd value of 0.36 (minimum: 0.16; maximum: 0.50). In each sample, items above the median display larger differences between the groups, while items below the median display smaller differences.

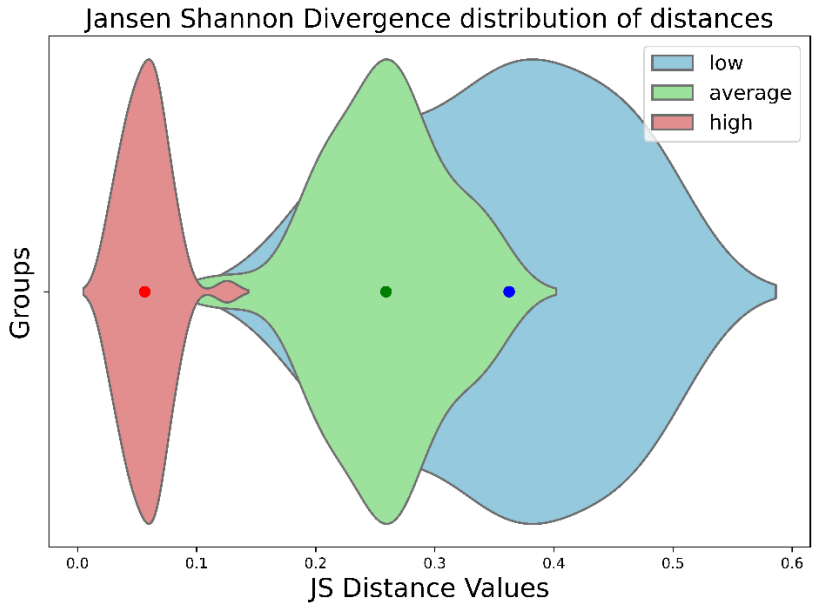

Figure 3-3-2: The JSDd ($\cdot \parallel \cdot$) in a dichotomous example. The dots correspond to the median value of the distance.

We then applied and compared the three ML algorithms from Section 3-3-3-2 (CART, RF, SVM with a linear kernel) and measured their performance according to Accuracy as a function of the proportion of removed variables (Tables 3-3-3, 3-3-5, 3-3-5 and Figure 3-3-3). We can see that reducing the number of variables, as discussed in Section 3-3-3-5, improves, or maintains model performance for the test set for all three algorithms in all three tiers of similarity. In the "low" data, characterized by relatively consistent variability in participants' responses to the task, a reduction in data noise facilitated a gradual improvement in algorithmic performance on the test set, eventually resulting in a decline attributed to a substantial loss of information. In this scenario, the CART algorithm emerged as the top performer. With an 85% reduction in the number of items, the model achieved an accuracy of 70.0% on the training set and 66.3% on the test set. In the "average" data, RF and SVM performed similarly, whereas the test set performance showed a gradual improvement with variable reduction for CART. By reducing the number of items to 75%, the CART model performed best, achieving a training set Accuracy performance of 93.8% and a test performance of 93.8%. All models behaved similarly as we introduced more variability in the "high" data. Notably, Figure 3-3-3, Panel c shows how the Accuracy of the models remained relatively consistent for the test set as the number of items decreased. In this particular scenario, the model with the most favorable trade-off between performance and item reduction was RF. By removing 85% of the items, it demonstrated an impressive training accuracy of 99.4% while maintaining a strong 98.8% accuracy on the test set. Additional performance information can be found in Appendix B (Section 3-3-8-4). After comparing the models across the three distinct scenarios, we brought forward only the winning model from the "high" data set to the BN analysis of the questionnaires.

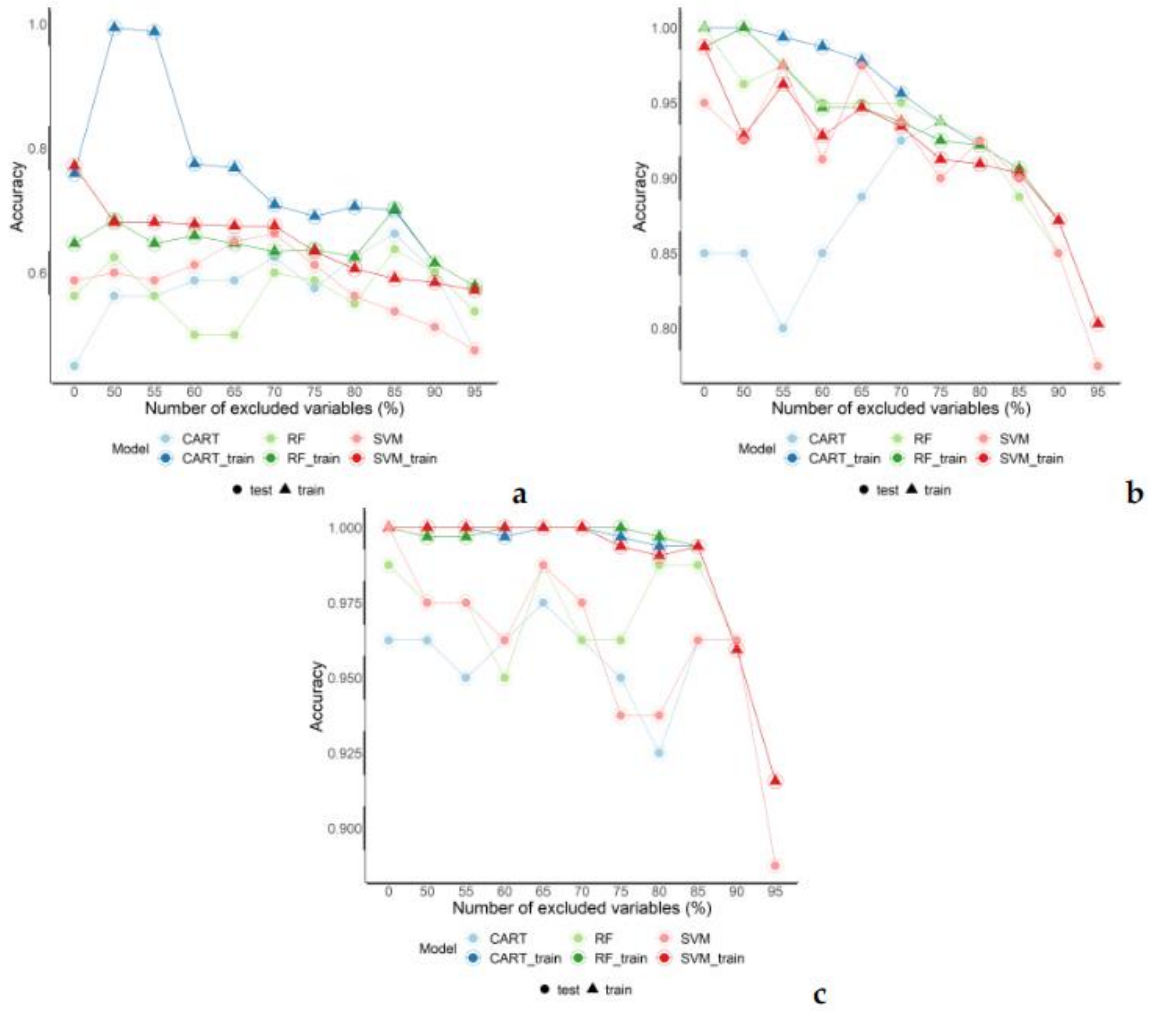

Figure 3-3-3 Comparing Model Performance: CART, RF, and SVM algorithms evaluated by Accuracy and variable depletion according to JSDd in the three data sets. Panel a displays the Accuracy of the three models for the training and test sets as a function of the percentage of excluded variables under "low" dissimilarity. Panel b illustrates the performance under "average" dissimilarity, while panel c presents the performance under "high" dissimilarity.

As previously mentioned, BNs enable researchers to uncover the optimal relationships among features using various metrics and learning algorithms while also allowing for testing their own hypotheses. We employed a data-driven approach to discover the questionnaire construct in this scenario. We conducted a comparative analysis of two Bayesian Network (BN) algorithms, PC-stable and tabu search, to determine their performance in terms of the log-Bayes factor (logBF). Initially, we compared the individual DAG structures obtained from both algorithms to an

averaged structure after bootstrap resampling, for which we estimated an appropriate threshold based on the data. For the tabu search algorithm, our analysis revealed a logBF of 10.4, favoring the single structure learned by the algorithm. However, for PC-stable, such a direct comparison was not feasible due to the presence of undirected arcs in the single DAG structure. Consequently, we selected the averaged network as the appropriate point of comparison against the single DAG obtained from the tabu search algorithm. The logBF computed between the single DAG structure derived from the tabu search and the averaged structure from PC-stable strongly favored the former, with a logBF of 49.5. A graphical comparison of these BN structures is provided in Appendix C, Figure 3-3-7. Figure 3-3-4 shows the item relationships discovered by the model. We have highlighted the Markov Blanket relationships starting from Feature11 in light blue. The arrows represent the direction of the relationship. Feature6 and Feature40 in light beige are dependent on Feature41 and Feature37 and Feature41 and Feature35, respectively. At this stage, it is possible to query the model to better understand the relationships among the interconnected nodes, as shown in Briganti and colleagues (2022).

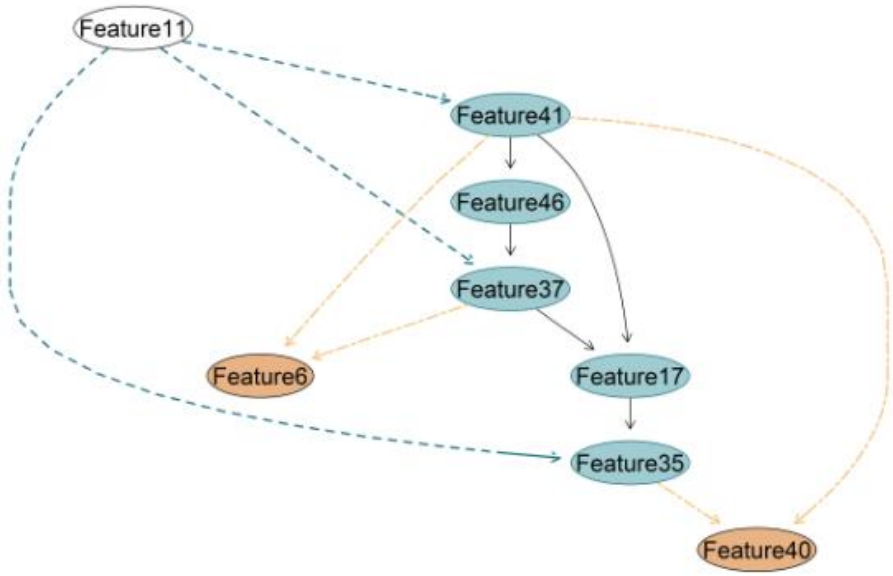

Figure 3-3-4: Directed Acyclic Graph (DAG) estimated from tabu search algorithm on high distance data set. The arrows reflect the direction of the feature relationship.

Furthermore, it is feasible to test one's own hypothesis concerning the relationship between items and compare it against the data-driven approach. Furthermore, by incorporating the Group variable into the model, we conducted a comparison between a single run of the tabu search algorithm and averaging it over 1000 bootstrapped samples. Subsequently, we assessed the two learned structures using the logBF, which indicated no significant difference between the two structures (value: $5.68 \times 10{-}14$). We chose the averaged DAG structure which simply consists of the Markov Blanket for our Group, signifying the individuals' affiliation with their respective groups. This selection also enables us to evaluate the relevance of the features chosen for the classification task. Indeed, literature already showed how the Markov blanket comprises a concise set of pertinent features that achieves optimal classification performance (Lee et al., 2020). The graphical representation of this graph structure has been included in Appendix C, Figure 3-3-8.

3.3.4.2 Example on ordinal data

We studied ordinal data similarly to dichotomous data. Specifically, we simulated two samples answering five ordinal Likert scale responses. Each sample consisted of two groups, each comprising 200 subjects and 50 features. The distinction between the samples lay in the distribution of responses across the Likert scale bins. In the "reversed" sample, we simulated a group with a left-skewed distribution for the answer process and another with a right-skewed distribution while maintaining identical but reversed probabilities. In the second simulated sample, which we will refer to as "uniform," we simulated a left-skewed group and a group with a uniform distribution of answers among the five possibilities. Details about the sample characteristics are presented in Appendix A, Section 3-3-8-3. We subsequently employed Algorithm 2 on both the "reversed" and "uniform" samples. Similar to the dichotomous approach, we compared the performance of three machine learning algorithms, Section 3-3-3-2 (CART, RF, SVM with linear kernel). The evaluation was based on accuracy, considering the

proportion of removed variables. The results are summarized in Tables 3-3-3, 3-3-7, and Figure 3-3-5. From the Algorithm 2, we obtained the JSDd of the "reversed" sample (median: 0.03; minimum: 2.79 × 10−5; maximum: 0.08) and of the "uniform" sample (median: 0.15; minimum: 0.07; maximum 0.21). The reduction in the number of variables, which we discussed in Section 2.6, enhances or sustains model performance across all algorithms for all samples on the test set. In the "reversed" sample, the performance of the algorithms demonstrated stability as variables were progressively excluded. Notably, the RF algorithm exhibited superior performance, achieving 100% accuracy on the training set and maintaining 100% on the test set with an 80% reduction in the number of items. Similarly, the algorithms displayed consistently high performance for the "uniform" data. As in the "reversed" data, RF achieved the highest accuracy on the test set. Specifically, with an 80% reduction in variables, the algorithm reached 100% accuracy on the training set and 96.3% on the test set. Model performance comparisons are illustrated in Figure 3-3-5 and described in more detail in Appendix B.

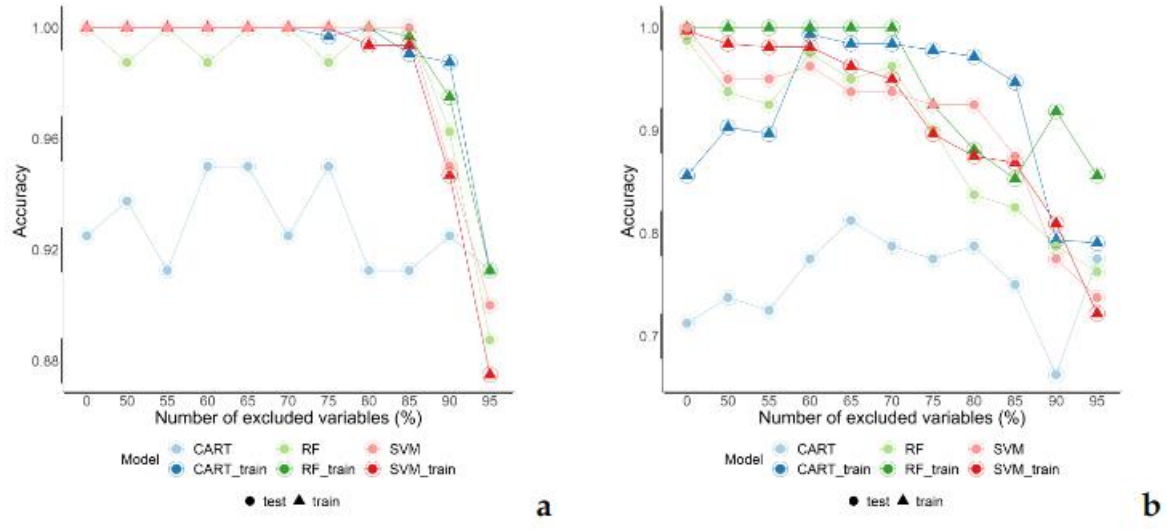

Figure 3-3-5: Comparing Model Performance. The figure presents an evaluation of CART, RF, and SVM algorithms based on accuracy and variable depletion assessed by JSDd in two data sets. In Panel a, accuracy on both training and test sets is shown as a function of the percentage of excluded variables for the "reversed" data. Panel b illustrates the performance for the "uniform" data.

Following the identification of significant deviations from normality in the distributions using the Shapiro-Wilk test (W = 0.87; p-value < 0.001), we applied the variable discretisation method proposed by Hartemink (2001) to the items selected by RF. Subsequently, we learned a discrete BN structure rather than a Gaussian BN. The structure comprised three factors for each variable, excluding the dependent variable (Group), which featured two factors. As in the dichotomous case, we compared the PC-stable and tabu search algorithms in terms of the logBF. Initially, we compared the individual DAG structures obtained from both algorithms to an averaged structure after bootstrap sampling, for which we estimated an appropriate threshold based on the data. The logBF for tabu search was 61.5 in favour of the single structure. In the context of the PC-stable algorithm, the averaged DAG showed conditional independence between all the node variables. We graphically compared the single structure tabu search and the averaged one from PC-stable in Appendix C, Figure 3-3-9. Figure 3-3-6 shows the item relationships discovered by the model. We highlighted the Markov Blanket relationships starting from Feature_8 in light blue. The arrows represent the direction of the relationship.

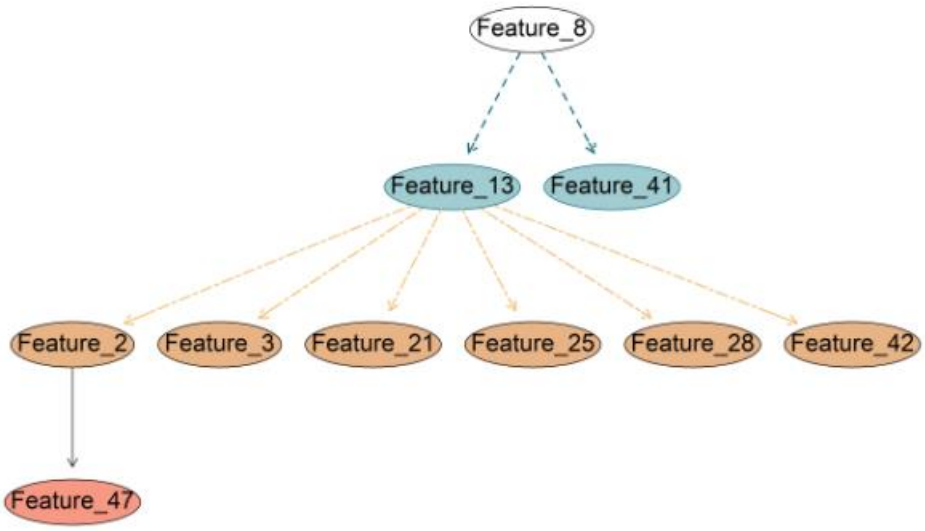

Figure 3-3-6: Directed Acyclic Graph (DAG) estimated from tabu search on the "reversed" data set. The arrows reflect the direction of the feature relationships.

Moreover, we integrated the Group variable into our model as we did in the dichotomous case. We compared a single execution of the tabu search algorithm with another instance averaged across 1000 bootstrapped samples. The logBF results indicated no significant difference between the two structures, with a value of $-5.68\times10-14$ in favour of the average structure. Again, the entire learned structure constitutes the Markov Blanket for our Group variable, indicating the individual affiliation to their respective groups. This choice also allows us to assess the relevance of the selected features for the classification task. The graphical representation of this graph structure has been included in Appendix C, Figure 3-3-10.

Finally, the model facilitates querying to better comprehend the relationships among interconnected nodes and allows for testing additional hypotheses on the item relationships and contrasting them with the data-driven results.

3.3.5 Discussion

Questionnaires are essential tools in many disciplines and for many practitioners like psychologists, mental health therapists, and researchers. Conventionally, questionnaire validation relies on latent trait models like EFA, CFA and SEM. In this paper, we introduced a novel methodology that integrates information theory, machine learning, and Bayesian networks for questionnaire validation. Diverging from the latent traits framework, our approach is rooted in Network Analysis theory, specifically emphasising item-level analysis. This perspective enables us to conceptualise psychological constructs as manifestations of interactions between variables rather than latent attributes (Guyon et al., 2017). We used the Jensen-Shannon divergence as an entropy measure for item selection, quantifying the dissimilarity between the item probability distributions according to the group membership. We used this learned distance distribution as a feature selection tool to select those items that exhibited greater dissimilarity between the groups with several ML models. We then selected

the best-performing, parsimonious model and learned the structure of the selected items in the questionnaire using BNs and conditional independence. Recent work from Trognon et al. (2022) emphasized that within the current validity framework, validity is contingent upon the level of evidence and theoretical justification supporting the interpretation and application of scores obtained from a scale. However, this formulation focuses not on the scale but on interpreting the generated scores.

In this study, we addressed this limitation. Notably, the item selection procedure and the application of machine learning have enabled us to initially assess the criterion validity of the questionnaire. Criterion validity measures how accurately a test reflects the intended outcome. The algorithm's performance directly evaluates the efficacy of selected items in representing the discriminative capacity of the construct, aligning directly with another validity measure, content validity. Content validity examines item representativeness and the instrument's appropriateness for measuring the targeted construct. For instance, a questionnaire revealing low discriminative capacity during the validation process suggests a correspondingly low content and criterion validity. This observation provides evidence that there may be an opportunity to enhance these two validity measures by defining more effective items under the construct of interest.

We applied the method to dichotomous and ordinal Likert scale data across various simulated data sets. Our results confirmed the promise of our methodology for item selection and the validation of psychological constructs, offering valuable insights into the internal structure of the questionnaire. The proposed approach selectively retrieves pertinent items, thus contributing to a reduction in questionnaire length. This is crucial because existing literature suggests that lengthy questionnaires impose a higher response burden, potentially resulting in decreased response rates and compromised data quantity and quality (Eisele et al., 2022; Rolstad et al., 2011). Moreover, it avoids common criticisms of EFA and CFA/SEM in

questionnaire validation. Firstly, it does not require any assumption of normality in item distributions. The literature suggests using polychoric correlations over Pearson correlations when analyzing dichotomous and ordinal data in FA (Holgado-Tello et al., 2010). Still, they are rarely used in practice (Kiwanuka et al., 2022) and make the interpretation of factor loadings more difficult Coenders and Saris (1995). Secondly, it avoids the subjectivity in factor loadings interpretation (Fairweather, 2001; Jordan & Spiess, 2019; Tracy, 1990; Wrigley, 1958). Our approach overcomes this aspect by focusing on items instead of latent factors.

However, it is important to acknowledge certain limitations of our study. We relied on simulated dichotomous and ordinal data as the basis for our work. In future research, we will explore this approach further by examining its applicability in real-world scenarios. We focused specifically on the implementation of three distinct ML algorithms. Nonetheless, other algorithms may be more suitable in specific circumstances, and the selection of the most appropriate model may also be influenced by its interpretability and performance metrics.

We should also be wary of potential violations of the causal inference assumptions in BN. For instance, as suggested by Briganti, Decety, et al. (2022), the presence of latent factors that are derived from other factors not included in the questionnaire but that underlie the interaction of items within the network violates causal sufficiency. However, there are Causal Bayesian Learning algorithms that can handle latent variables, thereby eliminating the necessity for the causal sufficiency assumption. Subsequent research should consider the incorporation of

these algorithms and their application in the realm of psychological data, where the violation of the causal sufficiency assumption is highly likely. For more details, see Kitson et al. (2023). Finally, the coexistence of interactive manifestations does not exclude the possibility of a common underlying cause Guyon et al. (2017). This introduces a novel perspective briefly explored by (Epskamp et al., 2016) in explaining psychological constructs. Here, latent variable

and network analysis models are perceived not as competing but as complementary models. This conceptualization gives rise to a distinctive class of Latent Network Models (LNMs) models. For an in-depth analysis, see Epskamp et al. (2016); Guyon et al. (2017).

3.3.6 Conclusions

In this paper, we have introduced a novel perspective to the validation of psychometric questionnaires, distinctively diverging from classical approaches based on Factor Analysis (FA). However, both have their strengths and limitations. FA provides a confirmatory and hypothesis-driven framework, while our approach offers a more exploratory and data-driven perspective, allowing for flexible modelling and discovery of relationships. Researchers can consider the specific requirements and characteristics of their study when selecting the most appropriate approach for validating their psychometric questionnaires.

3.3.7 Data availability statement

The data associated with this study has been deposited at the OSF repository: https://osf.io/u84qv

3.3.8 Appendices

3.3.8.1 Appendix A - Focus on Data Set Simulation

We present the binary data generation process written in Python using pandas and numpy libraries.

3.3.8.2 Binary Data Generation

We created a class containing a set of methods and attributes that define the behaviour of the binary data generator.

Constructor method:

It takes in several parameters: num features, num subjects, noise level, and seed. These parameters are used to initialize the attributes of the class with the given values.

- *num_features*: Represents the number of features in the binary data.
- *num_subjects*: Represents the number of subjects in the binary data.
- *noise_level*: Represents the noise level in the binary data, expressed in a value between 0 and 1.
- *seed*: Represents the seed value for reproducibility.
- *data*: Represents the binary data generated by the class. It is initially set to None.

Generate data method:

This method is used to generate the binary data based on the specified parameters:

- If a *seed* was provided during initialisation, it sets the random seed to ensure reproducibility.
- If *feature_probs* is not provided (defaults to *None*), it generates random probabilities for each feature. Using these probabilities, it generates binary data with shape (num subjects, num features) using *np.random.binomial*.
- If a *noise_level* greater than 0.0 is specified, it introduces noise to the data by performing a logical XOR operation between the generated binary data and additional binary noise data. The noise level determines the probability of noise in each feature.
- Finally, it creates a Pandas DataFrame from the generated data, with column names like “Feature1”, “Feature2”, etc., and assigns it to the data attribute of the class.

manipulate probabilities method:

This method allows you to change the probabilities of the binary features in the generated data:

- It checks if the data has been generated (*self.data* is not *None*) and whether the length of new probs matches the number of features. For each feature, it generates new binary data with the updated probability and replaces the corresponding column in the data DataFrame.

Add group column method:

This method allows you to add a new column to the generated data, used for grouping or labelling purposes: It checks if the data has been generated (*self.data* is not *None*) and adds a new column named "Group" with the specified group name to the data DataFrame.

Get data method:

This method returns the generated binary data as a Pandas DataFrame. The class described in this passage aims to manipulate the similarity between responses from two distinct groups to influence the method's behaviour. Three data sets labelled as "low", "average", and "high" are used, each representing different levels of similarity in response probabilities between the two groups. The process involves establishing arbitrary initial probability values for features within all three groups, generating samples for group1. Subsequently, for group2, these probabilities are systematically adjusted using a controlled random element, stored in a list called "modified probs." The iterative process introduces stochastic changes in probabilities, simulating variations in the likelihood of a feature having the value "1". The calculated changes are added to the initial probability while ensuring the resulting probability remains within the [0, 1] range. Table 1 presents the parameters employed in the generation of binary data in the study.

Similarity	**Binary data generation parameters**				
	num_samples	**num_features**	**noise level**	**seed**	**range_change**
Low	400	50	0.2	42	± .40
Average	400	50	0.2	42	± .25
High	400	50	0.2	42	± .05

Table 3-3-1: Binary data generation parameters

3.3.8.3 Ordinal Data Generation

We created a Generation class containing a set of methods and attributes that define the behaviour of the binary data generator.

Constructor method:

It takes in several parameters: num features, num subjects, noise level, and seed. These parameters are used to initialize the attributes of the class with the given values.

- *n_features*: Represents the number of features in the binary data.
- *n_subjects*: Represents the number of subjects in the binary data.
- *range_list*: The range of possible integer values in each feature
- *random_seed*: Represents the seed value for reproducibility.
- *limit*: The limit for the weights.
- *custom_weights*: Custom weights for the features. If it is set to *None*, weights for the features will be generated randomly.

The method generates the subjects by iterating over the number of subjects and, for each subject, generating a list of feature values. If custom weights is not *None*, it uses the custom weights. Otherwise, it generates random weights within the limit and then normalizes them. If

none of the weights are greater than 0, it assigns equal probability to two randomly selected features.

Other two methods have been implemented:

- *dataset*: This method generates a data set from the generated subjects. It returns a tuple containing the list of subjects and a pandas DataFrame representing the subjects. The DataFrame has columns named 'Feature 1', 'Feature 2', etc., according to the number of features.
- *plot*: This method plots histograms for each feature in the generated subjects. It generates a histogram for each feature, showing the distribution of the feature values across all subjects.

Table 2 presents the parameters employed in the generation of ordinal data in the study.

Type	**Group**	**Ordinal data generation parameters**					
		num_samples	num_features	Range list	seed	limit	Custom weights
Reversed	Group1	200	50	5	42	.2	[.05, .05, .2, .3, .4]
	Group2	200	50	5	42	.2	[.4, .3, .2, .05, .05]
Uniform	Group1	200	50	5	42	.5	[.05, .05, .2, .3, .4]
	Group2	200	50	5	42	.5	[.2, .2, .2, .2, .2]

Table 3-3-2: Ordinal data generation parameters

3.3.8.4 Appendix B - Details about ML algorithms performances

3.3.8.5 Performances on dichotomous data

	Proportion of removed variables (%)										
	0	50	55	60	65	70	75	80	85	90	95
Models	Accuracy on train and test set (%)										
CART	75.9–45	99.4–56.3	98.8–56.3	77.5–58.8	76.9–58.8	70.9–62.5	69.1–57.5	70.6–62.5	70–66.3	61.6–60	57.8–47.5
RF	64.7–56.3	68.4–62.5	64.7–56.3	65.9–50	64.7–50	63.4–60	63.8–58.8	62.5–55	70.3–63.8	61.6–60	57.8–53.8
SVM	77.2–58.8	68.1–60	68.1–58.8	67.8–61.3	67.5–65	67.5–66.3	63.4–61.3	60.6–56.3	59.1–53.8	58.4–51.3	57.2–47.5

Table 3-3-3: Model performances according to Accuracy metric and JS distance for dichotomous data in train and test set for the three algorithms: Decision Tree (CART), Random Forest (RF), and Support Vector Machines (SVM) for the low data set.

	Proportion of removed variables (%)										
	0	50	55	60	65	70	75	80	85	90	95
Models	Accuracy on train and test set (%)										
CART	100–85	100–85	99.4–80	98.8–85	97.8–88.8	95.6–92.5	93.8–93.8	92.2–92.5	90.6–90	87.2–85	80.3–77.5
RF	98.8–100	100–96.3	97.5–97.5	94.7–95	94.7–95	93.8–95	92.5–93.8	92.2–92.5	90.6–88.8	87.2–85	80.3–77.5
SVM	98.8–95	92.8–92.5	96.3–97.5	92.8–91.3	94.7–97.5	93.4–93.4	91.3–90	90.9–92.5	90.3–90	87.2–85	80.3–77.5

Table 3-3-4: Model performances according to Accuracy metric and JS distance for dichotomous data in train and test set for the three algorithms: Decision Tree (CART), Random Forest (RF), and Support Vector Machines (SVM) for the average data set.

	Proportion of removed variables (%)										
	0	50	55	60	65	70	75	80	85	90	95
Models	Accuracy on train and test set (%)										
CART	100–96.3	100–96.3	100–95	99.7–96.3	100–97.5	100–96.3	99.7–95	99.4–92.5	99.4–96.3	95.9–96.3	91.6–88.8
RF	100–98.8	99.7–97.5	99.7–97.5	100–95	100–98.8	100–96.3	100–96.3	99.7-98.8-	99.4–98.8	95.9–96.3	91.6–88.8
SVM	100–100	100–97.5	100–97.5	100–96.3	100–98.8	100–97.5	99.4–93.8	99.1–93.8	99.4–96.3	95.9–96.3	91.6–88.8

Table 3-3-5: Model performances according to Accuracy metric and JS distance for dichotomous data in train and test set for the three algorithms: Decision Tree (CART), Random Forest (RF), and Support Vector Machines (SVM) for the high data set.

3.3.8.6 Performances on ordinal data

	Proportion of removed variables (%)										
	0	50	55	60	65	70	75	80	85	90	95
Models	Accuracy on train and test set (%)										
CART	100–92.3	100–93.8	100–91.3	100–95	100–95	100–92.5	99.7–95	100–91.3	99.1–91.3	98.8–92.5	91.3–91.3
RF	100–100	100–98.8	100–100	100–98.8	100–100	100–100	100–98.8	100–100	99.7–100	97.5–96.3	91.3–88.8
SVM	100–100	100–100	100–100	100–100	100–100	100–100	100-100-	99.4–100	99.4–100	94.7–95	87.5–90

Table 3-3-6: Model performances according to Accuracy metric and JS distance for ordinal data in train and test set for the three algorithms: Decision Tree (CART), Random Forest (RF), and Support Vector Machines (SVM) for the reversed data set.

	Proportion of removed variables (%)										
	0	50	55	60	65	70	75	80	85	90	95
Models	Accuracy on train and test set (%)										
CART	85.6–71.3	90.3–73.8	89.7–72.5	99.4–77.5	98.4–81.3	98.4–78.8	97.8–77.5	97.2–78.8	94.7–75	79.4–66.3	79.1–77.5
RF	100–98.8	100–93.8	100–92.5	100–97.5	100–95	100–96.3	92.5–90	88.1–83.8	85.3–82.5	91.9–78.8	85.6–76.3
SVM	99.7–100	98.4–95	98.1–95	98.1–96.3	96.3–93.8	95–93.8	89.7–92.5	87.5–92.5	86.9–87.5	80.9–77.5	72.2–73.8

Table 3-3-7: Model performances according to Accuracy metric and JS distance for ordinal data in train and test set for the three algorithms: Decision Tree (CART), Random Forest (RF), and Support Vector Machines (SVM) for the uniform data set.

3.3.8.7 Appendix C - Bayesian Networks details

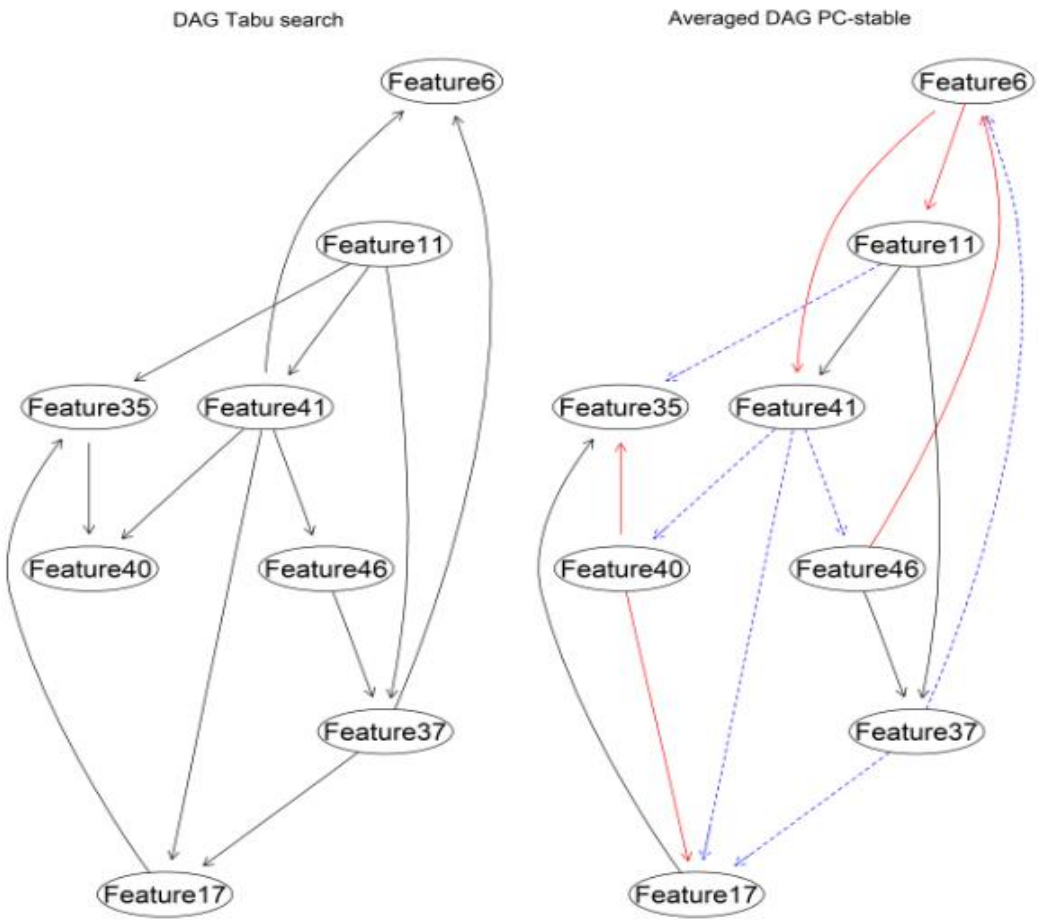

Figure 3-3-7: Comparative graphical analysis of Directed Acyclic Graphs (DAGs) depicting distinctions between the tabu search and the averaged PC-stable algorithm. Red arrows indicate connections that are either absent or reversed in relation to the DAG generated by the tabu search. In contrast, blue arrows signify connections identified in the tabu Search DAG but overlooked in the context of the PC-stable algorithm.

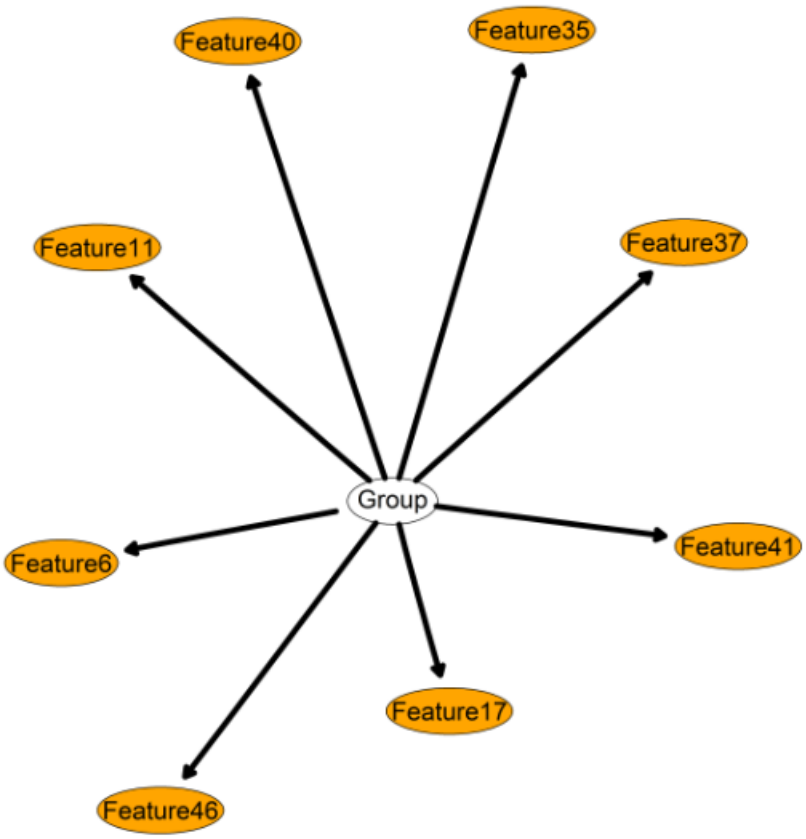

Figure 3-3-8: Averaged bootstrapped Directed Acyclic Graph (DAG) with a threshold value estimated from data. The Markov Blanket of the model, determined by the group membership variable (DV), is highlighted in orange. The size of the arrows connecting the items to the DV variable reflects the strength and direction of the relationship.

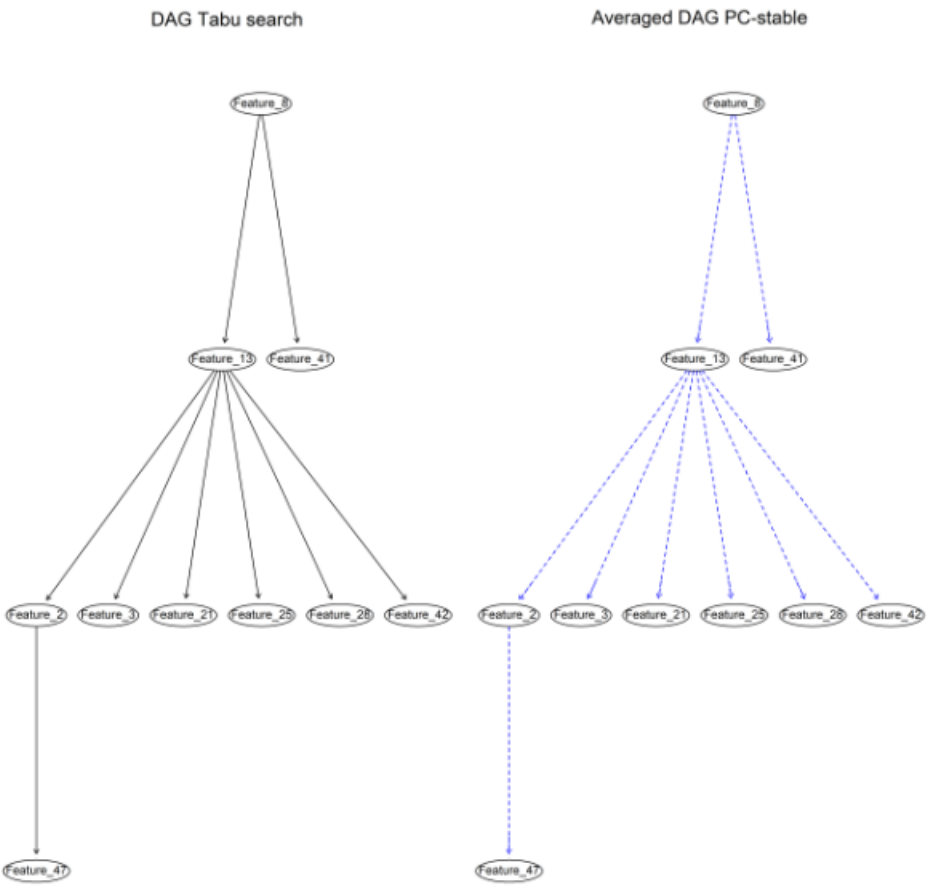

Figure 3-3-9: Comparative graphical analysis of Directed Acyclic Graphs (DAGs) depicting distinctions between the tabu search and the averaged PC-stable algorithm. Blue arrows signify connections identified in the tabu Search DAG but overlooked in the context of the PC-stable algorithm.

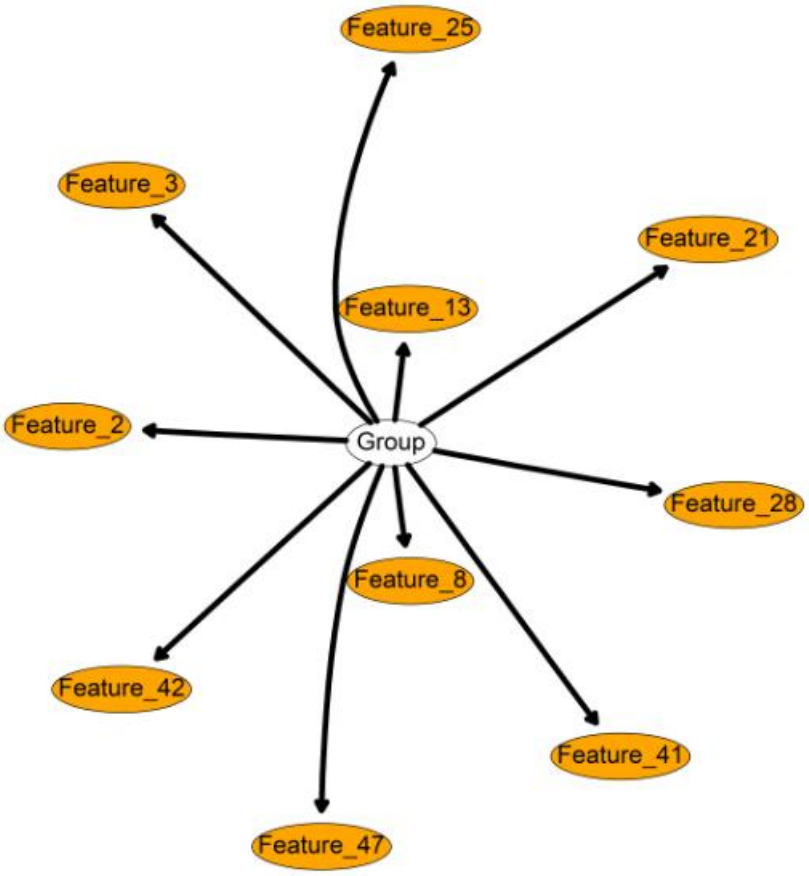

Figure 3-3-10: Averaged bootstrapped Directed Acyclic Graph (DAG) with a threshold value estimated from data. The Markov Blanket of the model, determined by the group membership variable, is highlighted in orange. The size of the arrows connecting the items to the DV variable reflects the strength and direction of the relationship.

Part III
Applications on Training

4.1

Recommending Mathematical Tasks Based on Reinforcement Learning and Item Response Theory.

Orsoni, M., Pögelt, A., Duong-Trung, N., Benassi, M., Kravcik, M., Grüttmüller, M. (2023). Recommending Mathematical Tasks Based on Reinforcement Learning and Item Response Theory. In: Frasson, C., Mylonas, P., Troussas, C. (eds) Augmented Intelligence and Intelligent Tutoring Systems. ITS 2023. Lecture Notes in Computer Science, vol 13891. Springer, Cham. https://doi.org/10.1007/978-3-031-32883-1_2

4.1 Recommending Mathematical Tasks Based on Reinforcement Learning and Item Response Theory.

4.1.1 Abstract

Recommending challenging and suitable exercises to students in an online learning environment is important, as it helps to stimulate their engagement and motivation. This requires considering their individual goals to improve learning efficiency on one side and on the other to provide tasks with an appropriate difficulty for the particular person. Apparently, this is not a trivial issue, and various approaches have been investigated in the areas of adaptive assessment and dynamic difficulty adjustment. Here, we present a solution for the domain of mathematics that rests on two pillars: Reinforcement Learning (RL) and Item Response Theory (IRT). Specifically, we investigated the effectiveness of two RL algorithms in recommending mathematical tasks to a sample of 125 first year Bachelor's students of computer science. Our recommendation was based on the Estimated Total Score (ETS) and item difficulty estimates derived from IRT. The results suggest that this method allowed for personalized and adaptive recommendations of items within the userselected threshold while avoiding those with an already achieved target score. Experiments were performed on a real data set to demonstrate the potential of this approach in domains where task performance can be rigorously measured.

4.1.2 Introduction

Conventional university teaching methods usually provide uniform learning exercises for the study groups. Depending on the level of knowledge, exercises can differ in the perception of difficulty by students. For optimal support and challenge of students, an individual selection of tasks is needed, which can be made based on various metrics, e.g. the level of knowledge or the desired final grade. Individualized learning tries to stimulate the motivation and engagement of students, taking into account theories like the zone of proximal development (Vygotskij, 1978) and flow (Csikszentmihalyi, 1990). The first provides students with tasks beyond their current ability to scaffold the learning process. The second aims to avoid boredom and frustration if the chosen difficulty level does not correspond with the student's ability. Dynamic Difficulty Adjustment (DDA) mechanism, which originated from computer games, is a technique used to automatically adjust the difficulty of online tasks according to the abilities of the user (Constant & Levieux, 2019; Xue et al., 2017), with the goal of keeping the user's attention and engagement. The DDA concept (Arey & Wells, 2001) emphasizes the importance of three aspects: the task difficulty (static or dynamic), the user's status (this can include performance or engagement, but also personality and emotions), and the adaptation method, which can be based on rules or data-driven approaches (e.g. probabilistic models, reinforcement learning). Physiologically, user involvement is driven by discovering new knowledge, learning patterns, ideas, and excitement while achieving a particular learning goal (Lopes & Lopes, 2022). In educational contexts, DDA can ensure that students are presented with tasks suitable for their current level of proficiency, leading to more engaging learning experiences.

One approach to implementing a DDA mechanism is using the Item Response Theory (IRT), a statistical model that estimates an individual's proficiency at a particular task by analyzing their responses to a set of items (Embretson & Reise, 2013). This enables to a recommendation

of appropriately challenging tasks for the student. However, recommending tasks based on IRT estimates can be suboptimal, as it does not consider the student's learning progress. Therefore, we propose the IRT integration with Reinforcement Learning (RL), which allows for optimizing task recommendations based on the student's past performance.

This study presents a system that utilizes IRT and RL to recommend tasks to first-semester bachelor's degree computer science students taking a mathematics module. Using our proposed method, which employs and compares the Proximal Policy Optimization (PPO) (Schulman, Wolski, Dhariwal, Radford, & Klimov, 2017), and the synchronous, deterministic variant of the Asynchronous Advantage Actor-Critic (Mnih et al., 2016) algorithm (called A2C), we aim to demonstrate the benefits of personalized task recommendation in the educational settings. In more detail, we incorporated the learner's goals into our recommender system. Literature suggested that specific interventions to set personal academic goals and exam preparation are essential factors contributing to the student's success while in the university (Stelnicki et al., 2015). Moreover, goal setting can help students develop a sense of agency, intrinsic motivation, and the ability to manage their learning (Surr, 2018). We compared the performance of our proposed method to a random baseline, using data from 125 students. The results of our study will provide insight into the effectiveness of using IRT and RL for recommending items in line with the learner's past performance and goals.

In the following, we first reference some related work and background information. Then we present our experiments thoroughly, including the results. Finally, we discuss the outcomes and conclude the paper.

4.1.3 Related Work

Computerized adaptive assessment methods in well-structured domains like mathematics have a long tradition of selecting tasks according to the student's ability (Tvarožek et al., 2008),

where structured task description schemes allowed for a detailed analysis of student's errors and on-demand generation of task instances facilitated independent student work. During the recent Corona crisis, professional rule-based adaptive learning systems like bettermarks (https://bettermarks.com/) were very popular.

Recent machine learning approaches address the DDA issue also in other domains if there is a significant question bank and users with different competencies (Zhang & Goh, 2021), considering even individual difficulty levels. This method can be applied when three conditions are met: a discrete action space exists, a feedback signal is a quantitative measure of difficulty, and a target performance value is selected.

DDA can be achieved using statistical models such as IRT (Embretson & Reise, 2013). IRT estimates a learner's proficiency based on their responses to a set of items and has been applied in various educational contexts (Hori et al., 2020). However, traditional recommendation approaches may not be suitable in educational settings where a student's learning potential changes over time. Reinforcement Learning (RL) addresses this issue by optimizing task recommendations based on the student's past performance and progress (Sutton & Barto, 1998). In recent years, the combination of IRT and RL has been proposed as a solution for recommendation in mathematics and cognitive domains. For example, the authors in (Leite et al., 2022) suggested using an RL system to recommend items based on the student's ability estimates from an IRT model to improve algebra abilities. Also, the study mentioned earlier (Zhang & Goh, 2021) used IRT to estimate the student's knowledge and RL to adjust task difficulty.

This work is distinct from the previous approaches in recommender systems that combine RL and IRT. It utilizes IRT to estimate the difficulty of items based on the student's past performance and uses this information to compute the expected total score threshold distribution for mathematical modules. This relevant information allowed to integrate into an

RL system of the learner's goal to make recommendations that align with the student's objectives.

4.1.4 Background

In Reinforcement Learning (RL), an agent learns to make decisions by interacting with its environment and receiving feedback through rewards or penalties. The agent's goal is to learn a policy mapping from states to actions that maximize the expected cumulative reward over time (Sutton & Barto, 1998). In the present work, we used and compared the performances of two popular RL algorithms: the Proximal Policy Optimization (PPO) (Schulman, Wolski, Dhariwal, Radford, & Klimov, 2017), and the synchronous, deterministic variant (A2C) of the Asynchronous Advantage Actor Critic (A3C) (Mnih et al., 2016). PPO is designed to improve the stability and efficiency of policy gradient methods. It is an actor-critic algorithm that uses a value function to estimate the expected cumulative reward for a given policy, and it uses a trust region method to optimize the policy. The basic idea of PPO is to optimize the procedure so that the new policy is close to the previous one but with improved expected cumulative reward (Schulman, Wolski, Dhariwal, Radford, & Klimov, 2017). The variant of A3C combines the actor-critic method with the advantage function. The actor-critic process separates the policy, which generates the actions, from the value function, which estimates the expected cumulative reward for a given policy. The advantage function estimates the improvement of taking a given action compared to the average action. The term "synchronous" refers to the method of updating the parameters of the actor and critic networks. All agents update their parameters simultaneously using the same synchronous data. In contrast, in the original asynchronous version, each agent updates its parameters independently using its data (Mnih et al., 2016).

4.1.5 Experiments

4.1.5.1 Experimental Dataset

This study analyzes a data set collected at Leipzig University of Applied Sciences starting from the winter semester of 2021/22. The data set includes the results of weekly exercises from a mathematics module taken by 125 Bachelor first-year computer science students. To pass the module, students must solve at least 35% of the weekly exercises over the semester. Each weekly practice includes several tasks specific to the topic covered in that week's lecture. The data set also includes solution attempts made after the semester. The tasks differ slightly for each attempt and student but are assumed to have equivalent difficulty and be based on the same concept. To practice the subject matter, students can work on the exercises and subtasks multiple times. Only the most successful attempt will be counted toward the final grade. The assignments are provided through the OPAL learning management system and ONYX testing software, and some tasks allow using the computer algebra system MAXIMA. The data set is separated into tables for student results and task information. To encourage reproducibility and further investigation, we publish the dataset with the implementation codes on our GitHub repository https://github.com/MatteoOrsoni/ITS2023_Recommending-Math-Tasks .

4.1.5.2 Result Features

Participant: An ascending number that anonymously references students.

Test id: References the weekly exercise (test).

Test attempt: Attempt in which the student solves the weekly exercise.

Test score: Points scored by the student test pass score Points to pass the weekly exercise.

Test max score: Maximum points of the weekly exercise.

Test pass: Status whether the student has passed the weekly exercise.

Item id: References the actual subtask in a weekly exercise.

Item attempt: Attempt in which the student solves the subtask.

Item datestamp: Timestamp in which the student completed the subtask.

Item sessionStatus: Represents the status of the subtask. (final - The student has solved the task and submitted his/her answers; pendingSubmission - The student has viewed the assignment but has not responded to it; pendingResponseProcessing - The student has entered answers but has not submitted them; initial - The student has not viewed the assignment)

Item duration: Time spent on the subtask.

Item score: Points on the subtask scored by the student.

Item max score: Maximum points of the subtask.

Item candidate responses: Answers from the student.

Item correct responses: Correct answers of the subtask.

Item candidate responses score: Scores of the student's answers.

Item correct responses score: (Maximum)-point scores of the subtask

Item variables: Variable assignments of the subtask execution

4.1.5.3 Task Features

Item id: (Equivalent to the result table) references the subtask.

Is test: Status whether the item is a test (tests are groupings of subtaks and are usually equivalent to weekly exercises).

Test name: Folder name in which the test file is located.

Item description: Tasks description in HTML format.

All in all, there are 18576 solutions from a total of 99 different items in a total of 14 modules (including tests for exam preparation) in the data set. Due to the low number of attempts inside some modules and excluding tests for exam preparation, in this study, the analysis focused on 10 modules. On average, the students needed 464 seconds and achieved an average of 2.18

points per item, with an average maximum score of 3.37 points. Furthermore, students practiced a single item on average 1.85 times, with a maximum of 72 times.

4.1.5.4 Framework and Baselines

The IRT models have been implemented by using the mirt: a Multidimensional Item Response Theory Package in R (Chalmers, 2012), while the RL solutions in Python by using the Stable Baseline 3 (Raffin et al., 2021) library. We compared the two RL solutions (PPO, A2C) with a random baseline procedure. According to this, we ran the environment for 1000 episodes, collecting each reward and averaging at the end. For each episode, the actions were taken randomly into the set of those possible. The averaged reward was then taken as baseline values to be compared to the average reward after 1000 episodes estimated by implementing PPO and A2C algorithms. In the following, we will delve deeper into constructing the item difficulty estimation model and the environment in which the RL algorithms were implemented.

4.1.5.5 Difficulty Level

	1PL			2PL			3PL		
Module	AIC	BIC	LL	AIC	BIC	LL	AIC	BIC	LL
1	3452	3550	-1692	3379	3569	-1623	3432	3717	-1617
2	995	1021	-488	932	980	-450	941	1012	-447
3	533	545	-262	533	552	-261	NA*	NA*	NA*
4	1157	1183	-569	1167	1213	-568	1182	1251	-567
5	515	529	-253	521	543	-252	529	562	-252
6	711	728	-350	704	733	-342	713	756	-342
7	844	868	-413	842	884	-405	854	917	-403
8	717	746	-348	711	763	-336	731	809	-336
9	1059	1096	-516	1066	1133	-507	1073	1174	-498

10	664	698	-318	673	735	-310	682	775	-302

Table 4-1-1 AIC (Akaike Information Criteria), BIC (Bayesian Information Criterion), and LL (Log Likelihood). In bold, the models for each module that reached the significant level p<.05 among others. * It has not been possible to estimate the parameters due to too few degrees of freedom.

In the present study, an IRT approach is used to estimate the difficulty of items presented to students in a course each week and to create different thresholds based on the sigmoid distribution of the estimated total score (ETS) of the winning IRT model. It allows us to consider the learners' objectives for that particular module. IRT is a statistical procedure that allows for the discovery of a learner's latent trait for a specific concept and the estimation of different parameters (difficulty, discrimination, and guessing) embedded within the item according to the chosen IRT model. Three other IRT models (1PL, 2PL, 3PL) were compared, and the best one was selected using metrics such as the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and loglikelihood. The values of these metrics are summarized in Table 4-1-1. The winner over the three possible models was then selected based on the p-value obtained. Only the significantly different model ($p < .05$) from the others was used in further analysis. The estimated total score of the winner IRT model has been then used to estimate the θ value, the correspondent threshold difficulty for that specific module. The thresholds have been chosen arbitrarily except for the first, which was the one that allowed the student to pass the module. Two to four thresholds have been set into the RL solution for each module, according to the number of items (number of possible actions) and the steepness of the sigmoid distribution underlying the estimated total score. Moreover, the IRT solution gave us the values of items' difficulty for each module. These values have been used in the RL environment configuration.

4.1.6 Reinforcement Learning Environment

The recommender system has been developed as a Markov Decision Process (MDP), consisting of a tuple (*S,A,R,P*) of elements. The tuple defines the MDP completely, where the agent interacts with the environment. The goal is to find a policy (mapping from states to actions) that maximizes the expected cumulative reward over time. A specific recommender system has a similar MDP representation for each module created. It has been summarized as follows:

1. State Space *S*: It represents all possible states of the system. It is related to the answering process of the student according to the item presented in the module. Each state or item in the module has been described as a tuple of five elements (*d,s,m,dt,t*), where:
   (a) *d*: The difficulty of the module items according to the IRT difficulty estimation.
   (b) *s*: The score obtained by the student for each item.
   (c) *m*: The maximum possible score for that specific item.
   (d) *dt*: The difficulty threshold. This parameter does not change until the end of each episode.
   (e) *t*: It is the threshold. This parameter does not change until the end of each episode and is strictly related to the difficulty threshold. It is a numeric value corresponding to the score the student has to obtain by summing the score items.
2. Action Space *A*: It represents all possible actions that can be taken in each state.
3. Reward Function *R*: It is a function that assigns a numerical reward to each state-action pair (*s,a*) and is used to evaluate the quality of different policy choices. We included three different numerical rewards in the environment. A reward is related to the Difficulty, Actions, and Episode.
   (a) *Difficulty*: For every action chosen by the agent, that is, for every next item chosen, we wanted to create a function that gave a positive reward to the agent if the selected action

was in line with the difficulty threshold of the item estimated by the IRT model and the threshold chosen by the user. In this way, we wanted to favor items that had difficulties equal to or lower than the user's needs to reach a certain threshold, discouraging items that were too difficult to achieve the goal.

$$RD = \begin{cases} k_1 \; if \; s_t \leq \text{dt}, \forall \text{a} \in \text{A} \\ k_2 \; otherwise \end{cases}$$

In this function, if the action selected by the agent is in line with the IRT estimate and is less than or equal to the user's threshold, the agent will receive a positive reward k_1. If not, the agent will receive a reward of zero k_2.

(b) *Actions*: For every action taken by the agent, this reward function was constructed to track the actions taken and to avoid items for which the student has received a score equal to the highest possible from being presented again.

$$RA = \{k_3 \; if \; a_t \in action\_used, \forall a \in A \}$$

where *actions_used* means the set of actions/items for which the student has already achieved the highest possible score. If the agent recommended an action in the *actions_used* it received a negative reward.

(c) *Episode*: The last reward function was related to the episode conclusion.

Each episode was set to have a maximum duration related between (54% - 150%) longer than the number of possible actions, to allow the agent to present the items again for which the subject had not reached the highest possible score and to reach the thresholds with the items with higher difficulty. If the agent could reach the established threshold within the maximum length of the episode, it received a positive reward; otherwise, it did not receive any reward.

$$RE = \begin{cases} k_4 \; if \; s_t + s_{t+1} + \cdots + s_{t+n} \geq t \\ k_5 \; otherwise \end{cases}$$

At the end of each episode, the overall reward function was created based on the three functions. If the agent achieved a cumulative score on the items equal to or higher than the set threshold, then the reward function R included *RD* + *RA* + *RE*. Otherwise, it only had *RD* + *RA*. *RD* and *RA* are considered intermediate rewards that should guide the agent in its choice of future actions.

4. Transition Probability Function *P*: It defines the probability of transitioning from one state to another after taking a specific action.

4.1.7 Hyperparameters

	RL configuration			
Module	**policy**	**Custom_net**	**ts**	**lr**
1	mlp	Yes: [128, 64]	105	10−7
2	mlp	No	105	10−7
3	mlp	No	105	10−7
4	mlp	Yes: [128, 64]	105	10−7
5	mlp	No	105	10−7
6	mlp	Yes: [64, 32]	105	10−7
7	mlp	No	105	10−7
8	mlp	Yes: [64, 32]	105	10−7
9	mlp	No	105	10−7
10	mlp	Yes: [64, 32]	105	10−7

Table 4-1-2. Hyperparameters are implemented in both the PPO and A2C algorithms. lr: learning rate, ts: timesteps, Custom_net: Custom_network, policy: the policy implemented.

In this section, we summarized the hyperparameters used in each module. In Table 4-1-2, we have included the hyperparameters for configuring the reinforcement learning environment. Specifically, the PPO and A2C algorithms were trained for 105 timesteps across all modules, each for 1 hour. The learning rate was set at 10−7. Finally, the training algorithms were based on a policy object that implements an actor-critic approach, utilizing a 2-layer MLP with 64 units per layer (Raffin et al., 2021). It is true for some modules, while others utilize a custom network architecture. Table 4-1-3 summarizes the hyperparameters associated with the custom environment, including the maximum length of each episode and its relationship with the number of possible actions. It also shows the number of thresholds considered in each module and the numeric values of the threshold (t) based on the estimated total score and the corresponding θ value (dt) obtained from the winning IRT solution. In addition, it considers N as the number of complete subjects' recordings for each module. This value has been extracted using the student's first attempt for each task in each module.

	Environment configuration				
Module	**length (%)**	**n°t**	**dt**	**t**	**N**
1	20 (+82%)	4	[-2.80, -1.96, .03, .57]	[10.5, 15.5, 25.9, 28.3]	131
2	20 (+150%)	3	[-1.05, 0, 2]	[10.5, 23, 25]	129
3	8 (+100%)	2	[1.11, 2.56]	[10.5, 15.1]	132
4	20 (+150%)	3	[-0.15, 0.63, 2.20]	[10.5, 16, 26]	87
5	10 (+150%)	3	[-0.15, 1.05, 1.60]	[10.5, 15.4, 19.3]	99
6	10 (+100%)	3	[0.63, 1.05, 2.02]	[10.5, 13, 15]	100
7	20 (+150%)	3	[-0.75, 0.33, 1.17]	[10.5, 22.4, 32.2]	53
8	10 (+100%)	4	[18.7, 27, 38, 48]	[-1.24, 0.03, 0.75, 3,22]	44
9	20 (+54%)	4	[-2.68, -1, 0, 1]	[10.5, 24.3, 33, 40]	50
10	25 (+79%)	4	[-0.27, 0.33, 1.12, 2.62]	[14.525, 17.31, 23, 31.5]	21

Table 4-1-3. Hyperparameters in the environment configuration. Length (%) is related to the maximum episode length and the relative percentage compared to the number of possible actions. n°t: is the number of thresholds included in the environment for that module. dt: is the difficulty threshold. t: is the threshold value. N is the number of complete subjects' recordings for each module.

4.1.8 Experiment Results

This study evaluated the performance of two reinforcement learning solutions, PPO and A2C, and a random baseline solution in collecting average rewards after 1000 episodes. The results, as illustrated in Figure 4-1-1, demonstrate that the PPO solution outperformed both the A2C solution and the random baseline across all modules presented to subjects. A comparison of the mean improvement in collecting average cumulative rewards among the three solutions is

summarized in Table 4-1-4. Evidently, the PPO solution achieved, on average, a 22.83% increase in rewards over the random action solution. Furthermore, this advantage in collecting rewards was consistent across all modules, with an improved range of 4.50% to 78.94% compared to the baseline. In contrast, the A2C algorithm demonstrated only a moderate improvement in collecting rewards, with an average increase of 1.29% over the baseline across all modules. This improvement was inconsistent, with a range of -8.99% to 7.69%.

4.1.9 Remarks and Discussion

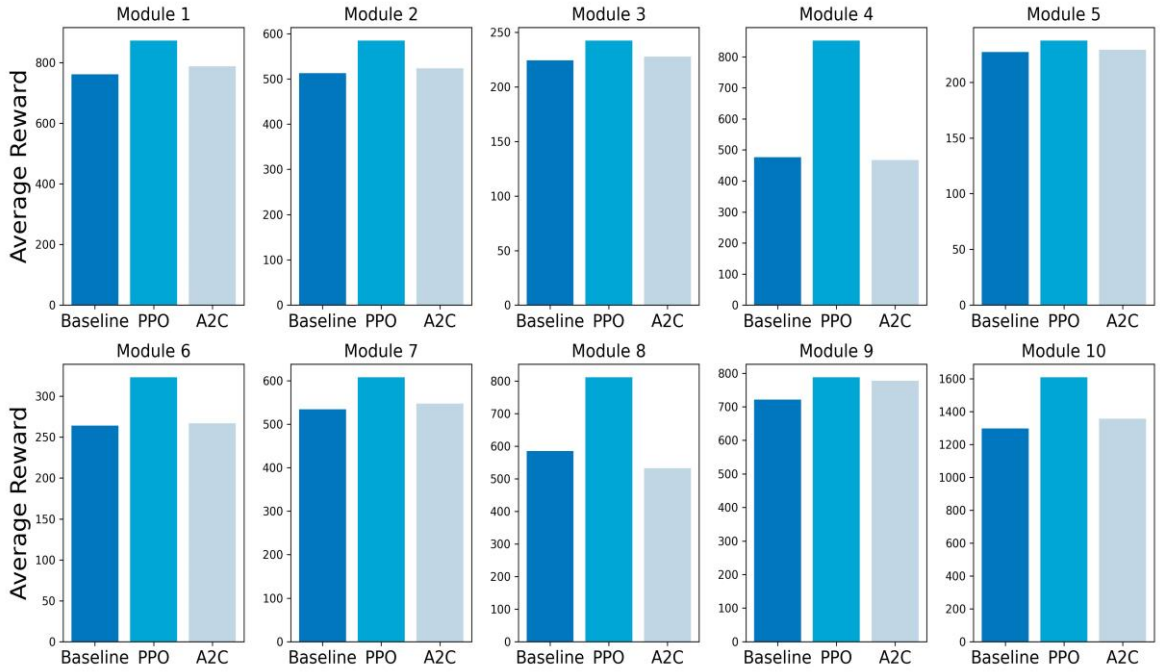

Figure 4-1-1 Comparing the Performance of RL Algorithms and Baseline Across Modules. Average reward after 1000 episodes comparing Baseline, PPO, and A2C recommendations.

The presented research centers on developing a recommender system that utilizes reinforcement learning and item response theory to enhance item recommendations for first-year bachelor's students in computer science taking a mathematics module. The integration of RL and IRT allows for personalized and adaptive recommendations based on the estimated difficulty threshold, enabling the system to suggest items within the user-selected threshold while avoiding items for which the student has already achieved the maximum possible score. In other words, the higher the threshold set by the student, the more complex the recommended items were, according to the θ value of the ETS distribution. This aspect is particularly relevant because of the significance of allowing learners to determine their own difficulty level. As previously mentioned, interventions aimed at establishing personal academic goals are a crucial component in promoting student success.

Module	**PP0/Baseline**	**A2C/Baseline**
1	+14.61	+3.44
2	+13.98	+2.08
3	+8.00	+1.47
4	+78.94	-1.87
5	+4.50	+0.9
6	+22.39	+1.1
7	+13.83	+2.53
8	+38.72	-8.99
9	+9.20	+7.69
10	+24.11	+1.27
Avg.	+22.83	+1.29

Table 4-1-4. Performances comparison in the average reward between PPO and Baseline and A2C and Baseline actions. The values are expressed in percentual terms.

Moreover, by facilitating goal setting, students can develop a stronger sense of agency, intrinsic motivation, and self-directed learning skills. The results demonstrate that incorporating RL solutions leads to improved performance, as measured by the average reward collected by the

agents over 1000 episodes. Specifically, as highlighted in the results section, the PPO algorithm outperforms the A2C algorithm in every module, achieving an average reward that is 22.83% higher than the baseline.

Nevertheless, some considerations have to be mentioned. Firstly, while we have seen an improvement in the average reward collected, we need to determine if the recommendations benefit students. A future study should investigate this aspect more thoroughly. Secondly, our study used offline students' data for which we had complete answers for a module. It allowed us to use each episode as a new user and the answers as a transition over time for a specific user for that episode. This approach led to a policy strictly dependent on the answers collected, the students who answered all the items in each module, and the sample size and the possible transitions it learned. We only had a few dozen subjects for some modules who answered the entire set of items. In future studies, we plan to use this policy as a starting point and enhance it by incorporating online interaction between the user and the system. In addition, we used arbitrary thresholds derived from the estimated total score of the IRT solution, but there may be better options for achieving better results on test evaluations. In a future study, we plan to integrate this aspect by finding the best possible thresholds for each module that can provide the most informative guide for students to succeed on test evaluations. Lastly, we focused on item difficulty rather than the student's ability to solve a specific task. A future study should include this aspect in the RL environment to suggest items that also consider the student's ability to solve them.

4.1.10 Conclusion

This study presented a system for enhancing item recommendations for first year bachelor's computer science students taking a mathematics module. The integration of Reinforcement Learning (RL) and Item Response Theory (IRT) allowed for personalized and adaptive

recommendations based on the estimated difficulty threshold, enabling the system to suggest items within the user-selected scale while avoiding items for which the student has already achieved the maximum possible score. Results showed that incorporating RL solutions improved performance as measured by the average reward collected by the agents over 1000 episodes. Specifically, the proximal policy optimization algorithm outperformed the A2C algorithm in every module, achieving an average reward that is 22.83% higher than the baseline. Overall, this study provides valuable insight into the effectiveness of using IRT and RL for dynamic difficulty adjustment and the benefits of personalized task recommendation in educational settings. The proposed method can potentially improve learning outcomes and engagement in the domain of mathematics as well as other areas.

Discussion & Conclusions

5.1 General Discussion and Conclusions

This dissertation aims to explore the theoretical and practical applications of Artificial Intelligence and gamification within the realms of Education, Learning, and Neuropsychological Assessment, from a psychometric perspective. The work can be segmented into three primary sections. The first delves into theoretical studies related to the integration of artificial intelligence in educational and learning contexts. The second encompasses three studies focusing on the application of AI and gamification in Neuropsychological assessment and discriminative assessment through questionnaires. Lastly, the third one centers around applications in training, particularly in the implementation of an AI-based recommender system for mathematical learning purposes.

The growth of Information Communication Technologies (ICT) has provided fresh opportunities to integrate innovative technologies into educational, learning, and psychological practices. In the first study, the focus is on highlighting how Artificial Intelligence (AI) and gamification can enhance the learning experiences of future EU students in alignment with 21st century skills, specifically emphasizing creativity and problem-solving (Benvenuti et al., 2023). The 21st century necessitates individuals to acquire a diverse set of skills for effective work, personal growth, and overall development. Schools and educators play pivotal roles in shaping students' education by prioritizing skills like computational thinking, critical thinking, problem-solving, creativity, and other essential competencies in our increasingly digital world. This paper delves into the ways in which creativity, critical thinking, and problem-solving can be fostered by AI serving as a valuable tool to support teachers in their practices. However, the literature indicates that the application of AI in education, including chatbots and tutoring systems, is currently limited compared to other sectors. Resistance from decision-makers and the need for greater AI knowledge among stakeholders, including students, contribute to this

limitation. The studies included in this work, underscore positive student responses to AI, especially when they are familiar with its concepts or applications. This open the way to future directions, where the importance of integrating new technologies into education is fundamental, with a specific emphasis on uniform dissemination across European schools. Moreover, collaborative efforts and lifelong learning initiatives for teachers are encouraged, utilizing existing networks, and promoting AI literacy in education is deemed crucial for its effective integration and utilization as a reliable educational tool.

Directing its focus towards examining the impact of gamification on learning and education, the second study (Orsoni et al., 2023) set out to construct a checklist protocol, as suggested by Metwally et al. (2021) for researcher and practitioners; an extensive literature review formed this foundational step. The aim is to equip both with an invaluable tool for their work in practice. The checklist guides individuals to concentrate on key aspects emphasized in the literature, nudging them towards embracing optimal methodological practices. It comprises crucial elements; moderators and mediators that influence gamification's success, along with methodological considerations and essential study design components This study, based on a systematic analysis of 72 studies, focuses on seven primary aspects that should be considered while developing gamified applications for educational and learning purposes. These include study design; theoretical foundations; personalization and motivation for engagement in learning processes; game elements that enhance interest and immersion, which subsequently inform optimal game design, ultimately culminating in measurable learning outcomes. This comprehensive checklist illuminates diverse perspectives on key characteristics, revealing not only mixed sentiments but also considerations and methodological constraints. The checklist, comprising 24 items, some featuring a 4-point quantitative Likert scale; guides researchers to implement contextually relevant and effective elements. It employs a structured assessment of methodological rigor and learning impact using values that range from -1 to 3. Moreover, due

to the high-cost nature of implementing a gamified learning environment, involving various professionals, the checklist aims to streamline the production process, reducing costs by providing a starting point of what is more effective and what is less. As an innovative approach for evaluating the quality of researchers' work, this proposal suggests a point-based system. Scores range from 0 to 20, with higher points awarded for exemplary execution or adherence to standards established during the project initiation phase. A higher score indicates a superior incorporation of factors that enhance research quality and methodological rigor, while a lower score suggests potential oversights. The criteria concerning study design advocate for experimental or quasi-experimental approaches, pre-post assessments, control groups are also essential, and the consideration of covariates is paramount. Items 10-16 explore personalization aspects: these encompass individual behaviors and characteristics vital in determining appropriate gamification systems. We consider player types, including a variety of gaming personality models; this emphasizes the necessity for future research to establish the critical role that gamer personalities play in gamification. We categorize motivation and engagement outcomes separately (item 17) underscoring our stance on assessing motivation using psychometrically validated measures. The checklist underscores an essential point: evaluating one's own motivations before and after implementing gamified interventions is crucial for achieving higher methodological rigor. Moreover, we have considered the impacts of game elements (items 18-21). We underscore the necessity for researchers to meticulously choose specific components and ponder their influence on learning outcomes. We place particular importance on feedback (item 19), urging researchers to comprehensively incorporate indicators of its presence and type in their designs. Another critical factor is game design (item 22). This emphasizes the need to incorporate game-design principles for effective learning through gamification; indeed, it is a pivotal consideration. The study culminates with a summary of evidence derived from reviews and meta-analyses providing insights into

behavioral changes, motivational influences or affective responses, cognitive enhancements and overall learning outcomes that are influenced by gamification. Results underscore an arraying effect size: this varies based on three primary factors: interventions; populations involved, and fundamental elements of the game design itself.

In the second section, two studies adopting AI have been applied to discover cognitive profiles among secondary school students by using a neuropsychological, gamified assessment tool called PROFFILO. The first work (Orsoni et al., 2023), introduced a novel clustering method that use the joint combination of Kohonen's Self-Organizing Maps (SOMs) and k-means clustering in the determination of cognitive profiles over six different cognitive functions (logical reasoning, visual perception, visuospatial attention, phonological awareness, verbal comprehension, and working memory). This approach demonstrates enhanced replicability of clustering among typically developing students; an Artificial Neural Network (ANN) algorithm validates the efficiency of the profiling technique, exhibiting its effectiveness with new user profiles. The clustering solution found uncovers nine distinct groups, each exhibiting varying levels of cognitive capabilities. In more detail, it pinpoints challenges mostly in visual perception, visuospatial attention and working memory. One group struggled with visual perception (vhLR-lVP), another faced issues related to visuospatial attention (vhLR-vlVA). Two groups exhibited difficulties in working memory: aALL-lWM and vhLR-lWM, and the vlVP-lVA exhibited deficits in both visuospatial and visual perception domains. This outcome partially aligns with Yokota et al. 's (2014) study which underscores not only the importance of perceptual organization, but also attention when clustering typically developing children. Additionally, the solution facilitated differentiation between clusters in Specific Learning Difficulty (SLD) and non-SLD groups. Notably, examination of the aALL-lWM and vhLR-lWM clusters revealed 2.5- and 4.16-times higher instances for SLD students compared to non-SLD, affirming prior research associating low working memory with cognitive risk factors for

dyslexia and dyscalculia (Geary & Hoard, 2001; McLean & Hitch, 1999; Moll et al., 2016). In general, this study underscores the novelty of utilizing machine learning within psychological experiments; it also explores potential practical implications for clinical practice, specifically underlining personalized interventions rooted in validated clustering models. However, limitations such as sample size constraints set the stage for future recommendations, including the validation of external validity and exploration of alternative clustering algorithms. To face with some of these limitations, another work (Orsoni et al., under review), explores and evaluates dimensionality reduction techniques, including linear Principal Component Analysis (PCA) and deep dimensionality reduction based Variational Autoencoders (VAE), as well as their combined use (PCA+VAE) in clustering students' cognitive profiles across the six cognitive domains investigated by using the PROFFILO software. The primary objective is to capture the heterogeneity in cognitive profiles, crucial for enhancing students' metacognitive skills and enabling personalized learning experiences. Differently than the previous study (Orsoni et al., 2023), data from more than 1600 students have been used in the analysis. The findings indicate that for highly heterogeneous and dimensional data, the combined PCA+VAE outperforms individual PCA and VAE applications, especially in logical reasoning, phonological awareness, and verbal comprehension. Conversely, VAE alone get more ability in enhancing cluster quality for lower-dimensional sub-tests like visuospatial attention and motion perception compared to the linear approach of PCA. Straightforward clustering techniques are effective for the working memory test that contains only two dimensions. Moreover, a Bayesian Networks (BNs) provide insights into the hierarchical arrangement of cognitive domains within the test. Nevertheless, limitations arise from the wide age range within the sample and the sole reliance on internal validation for clusters. This underscores the necessity for more refined models in clinical contexts and external validation across diverse educational settings or cognitive functions. To summarize, dimensionality reduction techniques

exhibit potential in student cognitive profiling, with implications for personalized learning and a deeper understanding of intricate relationships within cognitive domains.

The last work of this section is related to the application of artificial intelligence into the practice of questionnaire psychometric validation. In this study (Orsoni et al., 2024), the idea is to present an innovative approach to questionnaire validation, departing from traditional latent trait models and embracing Network Analysis theory with a specific emphasis on item-level analysis. The methodology integrates information theory, machine learning, and Bayesian networks, redefining psychological constructs as interactions between variables rather than latent attributes (Guyon et al., 2017). The Jensen-Shannon divergence acts as an entropy measure for item selection, gauging dissimilarity between item probability distributions based on group membership. This learned distance distribution is then employed as a feature selection tool with various ML models, and the most effective, parsimonious model is selected. The structure of the chosen items in the questionnaire is subsequently learned using BNs and conditional independence. In addressing limitations of current validity frameworks, we evaluate criterion validity by assessing the algorithm's ability to represent the discriminative capacity of the construct. This aligns with content validity, scrutinizing item representativeness and appropriateness. Notably, a questionnaire with low discriminative capacity indicates diminished content and criterion validity, highlighting the potential to enhance these measures through the definition of more effective items. The application of our method to dichotomous and ordinal Likert scale data across simulated datasets affirms its potential for item selection and validation of psychological constructs. The approach selectively identifies relevant items, leading to a reduction in questionnaire length, a critical factor in alleviating response burden and maintaining data quality. Furthermore, it addresses common criticisms of EFA and CFA/SEM by sidestepping assumptions of normality in item distributions and eliminating subjectivity in factor loadings interpretation. While acknowledging study limitations,

particularly the reliance on simulated data, we advocate for further exploration in real-world scenarios. While our focus is on three ML algorithms, future research should consider alternative models based on specific circumstances, considering interpretability and performance metrics. In addition, caution regarding potential violations of causal inference assumptions due to the presence of latent variables in BN must be considered. One possible idea to handle this problem is incorporating Causal Bayesian Learning algorithms. Indeed, the presence of interactive manifestations alongside a common underlying cause doesn't dismiss the possibility of a shared perspective, where latent variable and network analysis models converge, creating a unique category known as Latent Network Models (LNMs). Further investigation into this conceptualization is recommended for a thorough understanding of psychological constructs (Epskamp et al., 2016; Guyon et al., 2017).

In the third and last section we implemented a Reinforcement Learning (RL) based Item Response Theory (IRT) recommender system with the aim to provide adaptive items according to the learner's objective on ten mathematical first-year university modules. The integration of RL and IRT enables personalized and adaptive recommendations based on estimated difficulty thresholds, allowing the system to propose items within the user-selected threshold of difficulty while avoiding those for which the student has already achieved the maximum score. This user-driven difficulty setting is crucial for promoting student success, fostering personal academic goals, and developing a sense of agency, intrinsic motivation, and self-directed learning skills. The results indicate that incorporating RL solutions leads to improved performance, as evidenced by the average reward collected by agents over 1000 episodes. Specifically, the Proximal Policy Optimization (PPO) algorithm outperforms the Advantage Actor-Critic (A2C) algorithm in every module, achieving an average reward 22.83% higher than the baseline. However, certain considerations have been presented. Firstly, while we observe an enhancement in the average reward collected, further investigation is needed to determine if the

recommendations effectively benefit students. Subsequent studies should delve deeper into this aspect. Secondly, our study utilized offline students' data, limiting the policy to transitions dependent on the answers collected from a subset of students who completed all items in each module. Future studies plan to incorporate online interactions to refine the policy. Additionally, arbitrary thresholds derived from IRT solutions were used, but a more nuanced approach to finding optimal thresholds is envisioned in future studies. Lastly, our focus on item difficulty, rather than the student's ability, suggests a direction for future research to include the student's ability in the RL environment for more comprehensive item recommendations.

In conclusion, the integration of artificial intelligence and gamification presents a transformative potential for reshaping the landscape of education, learning, assessment, and training within the fields of psychology and neuropsychology. These technologies offer exciting possibilities to enhance engagement, motivation, and the overall effectiveness of educational processes. However, as discussed in previous sections, the application of gamification in learning and education is not without its challenges. A major limitation lies in the methodological aspects related to implementation and study design. Overcoming these challenges requires careful consideration of how gamification is integrated into educational frameworks and ensuring that appropriate research methodologies are employed to assess its impact. As gamification continues to evolve, its integration into education and learning holds the potential to create dynamic, enjoyable, and effective learning environments that cater to the diverse needs of students.

On the other hand, artificial intelligence has demonstrated remarkable progress in recent years, unlocking innovative solutions and applications that were once deemed unimaginable. The rapid advancement of AI technology holds considerable promise poised to revolutionize traditional educational paradigms and enhance our understanding of cognitive processes. The potential benefits are vast, spanning from personalized learning experiences to more insightful

and targeted neuropsychological assessments. The integration of AI promises to transform education into a more personalized and adaptive endeavor, tailored learning paths and content recommendations have the potential to engage learners at a more individualized level fostering a deeper understanding of subject matter. Intelligent tutoring systems powered by AI stand to revolutionize the way students receive instruction real-time feedback, and adaptive tutoring approaches have the potential to address individual learning styles and accelerate the learning process. Moreover, AI-driven tools in neuropsychological assessment offer the prospect of uncovering nuanced patterns and correlations in cognitive processes. This deeper understanding can lead to more accurate assessments and targeted interventions for cognitive disorders, and a promising starting point for more personalized rehabilitations. The application of AI in education can contribute to making learning more inclusive by providing tailored support for students with diverse needs, especially those with specific learning difficulties or neurodevelopmental disabilities. In addition, AI's role in continuous learning aligns with the evolving nature of the modern workforce. This is related for example to personalized recommendations for skill development throughout one's career, and to adapt to changing industry demands. Finally, ethical considerations are crucial when integrating AI, especially when the affected groups include children or vulnerable populations. The European Commission is advancing clear ethical standards to govern the development of AI systems. Building upon the principles outlined in the General Data Protection Regulation (GDPR), which acknowledges the need for special protections for children's personal data, the Commission has introduced provisions to safeguard the processing of their data and ensure that children comprehend and can assert their data protection rights. Since 2018, the European Commission has launched various initiatives focused on AI, such as the European AI strategy, the White Paper on Artificial Intelligence, and most recently, the newly approved AI Act (European et al., 2022; Jobin et al., 2019). These endeavors aim to establish a clear roadmap for the development

of AI systems that fully prioritize ethical considerations possible. Nowadays, several principles are considered crucial in the AI systems implementation. First and foremost, it's essential for AI applications to prioritize human welfare and minimize harm across individuals, society, and the environment. This involves conducting thorough evaluations to identify and address potential negative consequences effectively. Additionally, these AI systems should adhere to the principle of justice, ensuring fairness and equity in their deployment. Furthermore, transparency is crucial in establishing clear requirements for AI systems, particularly in high-risk applications. This not only fosters accountability but also improves understanding of the capabilities and limitations of AI systems. Moreover, accountability is paramount, ensuring that responsibility is assigned for the actions and outcomes of AI systems. Ethical principles such as autonomy and responsibility play a vital role in safeguarding individuals' rights to make informed choices and holding stakeholders accountable for the continuous impact of AI applications, thereby facilitating ongoing improvement. The ethical principle of collective beneficence dictates the establishment of a governance structure at both European and national levels. This encourages collaboration and shared responsibility to ensure that AI serves the common good and upholds human values effectively. Finally, striking a balance between leveraging data-driven insights and preserving individual privacy is crucial to ensure the responsible deployment of these technologies. The use of transparent algorithms and the establishment of clear accountability measures are essential elements for fostering trust in AI-driven educational systems and neuropsychological tools. Continuous monitoring and adaptation of AI applications are necessary to identify and address any biases that may arise, ensuring that these technologies evolve in accordance with ethical standards. As we move towards a future driven by AI, navigating ethical considerations carefully and maintaining a commitment to responsible AI deployment become paramount. Despite the substantial benefits, it is imperative to approach these advancements with a thoughtful and discerning mindset to

shape a future where technology enhances the educational experience while respecting the rights and privacy of individuals.

6.1 References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., … Zheng, X. (2016). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems*. http://arxiv.org/abs/1603.04467

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*. https://api.semanticscholar.org/CorpusID:122379222

Ackermann, E. (2001). *Piaget ' s Constructivism , Papert ' s Constructionism : What ' s the difference ?* https://api.semanticscholar.org/CorpusID:17875194

Adams, C. D. (1982). Variations in the Sensitivity of Instrumental Responding to Reinforcer Devaluation. *The Quarterly Journal of Experimental Psychology Section B*, *34*(2b), 77–98. https://doi.org/10.1080/14640748208400878

Ahmad, S. F., Rahmat, M. K., Mubarik, M. S., Alam, M. M., & Hyder, S. I. (2021). Artificial Intelligence and Its Role in Education. In *Sustainability* (Vol. 13, Issue 22). https://doi.org/10.3390/su132212902

Alam, A. (2022). *Employing Adaptive Learning and Intelligent Tutoring Robots for Virtual Classrooms and Smart Campuses: Reforming Education in the Age of Artificial Intelligence BT - Advanced Computing and Intelligent Technologies* (R. N. Shaw, S. Das, V. Piuri, & M. Bianchini (eds.); pp. 395–406). Springer Nature Singapore.

Aljabali, R. N., & Ahmad, N. (2019). A Review on Adopting Personalized Gamified Experience in the Learning Context. *2018 IEEE Conference on E-Learning, e-*

Management and e-Services, IC3e 2018, 61–66. https://doi.org/10.1109/IC3e.2018.8632635

Allan, N. P., Hume, L. E., Allan, D. M., Farrington, A. L., & Lonigan, C. J. (2014). Relations between inhibitory control and the development of academic skills in preschool and kindergarten: A meta-analysis. *Developmental Psychology*, *50*(10), 2368–2379. https://doi.org/10.1037/a0037493

Alloghani, M., Al-Jumeily Obe, D., Mustafina, J., Hussain, A., & Aljaaf, A. (2020). *A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science* (pp. 3–21). https://doi.org/10.1007/978-3-030-22475-2_1

Alloway, T. P., & Elsworth, M. (2012). An investigation of cognitive skills and behavior in high ability students. *Learning and Individual Differences*, *22*(6), 891–895. https://doi.org/https://doi.org/10.1016/j.lindif.2012.02.001

Alomari, I., Al-Samarraie, H., & Yousef, R. (2019). THE ROLE OF GAMIFICATION TECHNIQUES IN PROMOTING STUDENT LEARNING: A REVIEW AND SYNTHESIS. *Journal of Information Technology Education: Research*, *18*, 395–417.

Alsawaier, R. S. (2018). The effect of gamification on motivation and engagement. *International Journal of Information and Learning Technology*, *35*(1), 56–79. https://doi.org/10.1108/IJILT-02-2017-0009

Altman, N., & Krzywinski, M. (2018). The curse(s) of dimensionality. *Nature Methods*, *15*(6), 399–400. https://doi.org/10.1038/s41592-018-0019-x

Altun, A. (2016). *Understanding Cognitive Profiles in Designing Personalized Learning Environments BT - The Future of Ubiquitous Learning: Learning Designs for Emerging Pedagogies* (B. Gros, Kinshuk, & M. Maina (eds.); pp. 259–271). Springer Berlin

Heidelberg. https://doi.org/10.1007/978-3-662-47724-3_14

Ameur, H., Njah, H., & Jamoussi, S. (2022). Merits of Bayesian networks in overcoming small data challenges: a meta-model for handling missing data. *International Journal of Machine Learning and Cybernetics*, *14*, 229–251. https://api.semanticscholar.org/CorpusID:250102271

Angeli, C., & Giannakos, M. (2020). Computational thinking education: Issues and challenges. *Computers in Human Behavior*, *105*, 106185. https://doi.org/10.1016/j.chb.2019.106185

Anowar, F., Sadaoui, S., & Selim, B. (2021). Conceptual and empirical comparison of dimensionality reduction algorithms (PCA, KPCA, LDA, MDS, SVD, LLE, ISOMAP, LE, ICA, t-SNE). *Computer Science Review*, *40*, 100378. https://doi.org/https://doi.org/10.1016/j.cosrev.2021.100378

Antonaci, A., Klemke, R., & Specht, M. (2019). The effects of gamification in online learning environments: A systematic literature review. *Informatics*, *6*(3), 1–22. https://doi.org/10.3390/informatics6030032

Antonietti, A., & Molteni, S. (2014). *Educare al pensiero creativo. Modelli e strumenti per la scuola, la formazione e il lavoro*. Erikson.

Aparicio, A. F., Vela, F. L. G., Sánchez, J. L. G., & Montes, J. L. (2012). Analysis and application of gamification. *Interacción*. https://api.semanticscholar.org/CorpusID:108996499

Arey, D., & Wells, E. (2001). Balancing act:«the art and science of dynamic difficulty adjustment». *Game Developers Conference*.

Asimov, I. (1942). “Runaround.” In Doubleday (Ed.), *I, Robot (The Isaac Asimov Collection*

ed.). (p. 40).

Bacon, P.-L., Harb, J., & Precup, D. (2016). The Option-Critic Architecture. *Proceedings of the AAAI Conference on Artificial Intelligence*, *31*. https://doi.org/10.1609/aaai.v31i1.10916

Baddeley, A. D., & Hitch, G. (1974). *Working Memory* (G. H. B. T.-P. of L. and M. Bower (ed.); Vol. 8, pp. 47–89). Academic Press. https://doi.org/https://doi.org/10.1016/S0079-7421(08)60452-1

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate. *CoRR*, *abs/1409.0*. https://api.semanticscholar.org/CorpusID:11212020

Bai, S., Hew, K. F., & Huang, B. (2020). Does gamification improve student learning outcome? Evidence from a meta-analysis and synthesis of qualitative data in educational contexts. *Educational Research Review*, *30*(June 2019), 100322. https://doi.org/10.1016/j.edurev.2020.100322

Ball, P. J., Smith, L., Kostrikov, I., & Levine, S. (2023). *Efficient Online Reinforcement Learning with Offline Data*.

Baram, A. B., Muller, T. H., Nili, H., Garvert, M. M., & Behrens, T. E. J. (2021). Entorhinal and ventromedial prefrontal cortices abstract and generalize the structure of reinforcement learning problems. *Neuron*, *109*(4), 713-723.e7. https://doi.org/10.1016/j.neuron.2020.11.024

Bartle, R. (1996). *Hearts, clubs, diamonds, spades: Players who suit MUDs*.

Batiibwe, M. S. K. (2019). Using Cultural Historical Activity Theory to understand how emerging technologies can mediate teaching and learning in a mathematics classroom: a

review of literature. *Research and Practice in Technology Enhanced Learning*, *14*(1), 12. https://doi.org/10.1186/s41039-019-0110-7

Becker, D. R., Miao, A., Duncan, R. J., & McClelland, M. M. (2014). Behavioral self-regulation and executive function both predict visuomotor skills and early academic achievement. *Early Childhood Research Quarterly*, *29*, 411–424. https://api.semanticscholar.org/CorpusID:144056452

Behl, A., Jayawardena, N., Pereira, V., Islam, N., Giudice, M. Del, & Choudrie, J. (2022). Gamification and e-learning for young learners: A systematic literature review, bibliometric analysis, and future research agenda. *Technological Forecasting and Social Change*, *176*(December 2021), 121445. https://doi.org/10.1016/j.techfore.2021.121445

Benassi, M., Garofalo, S., Ambrosini, F., Sant'Angelo, R. P., Raggini, R., De Paoli, G., Ravani, C., Giovagnoli, S., Orsoni, M., & Piraccini, G. (2020). Using Two-Step Cluster Analysis and Latent Class Cluster Analysis to Classify the Cognitive Heterogeneity of Cross-Diagnostic Psychiatric Inpatients. *Frontiers in Psychology*, *11*(November). https://doi.org/10.3389/fpsyg.2020.01085

Bennani, S., Maalel, A., & Ben Ghezala, H. (2021). Adaptive gamification in E-learning: A literature review and future challenges. *Computer Applications in Engineering Education*, *30*(2), 628–642. https://doi.org/10.1002/cae.22477

Benvenuti, M., Cangelosi, A., Weinberger, A., Mazzoni, E., Benassi, M., Barbaresi, M., & Orsoni, M. (2023). Artificial intelligence and human behavioral development: A perspective on new skills and competences acquisition for the educational context. *Computers in Human Behavior*, *148*, 107903. https://doi.org/https://doi.org/10.1016/j.chb.2023.107903

Benvenuti, M., & Mazzoni, E. (2020). Enhancing wayfinding in pre-school children through

robot and socio-cognitive conflict. *Br. J. Educ. Technol.*, *51*, 436–458. https://api.semanticscholar.org/CorpusID:199015673

Bernik, A., Vusić, D., & Wattanasoontorn, V. (2022). Computer Game Elements and its Impact on Higher Education. *Tehnički Glasnik*, *16*(4), 566–571. https://doi.org/10.31803/tg-20220126221837

Bers, M. U., González-gonzález, C., Belén, M., Torres, A., Study, C., Development, H., & Science, C. (2019). Coding as a playground : Promoting positive learning experiences in childhood classrooms. *Computers & Education*, *138*(April), 130–145. https://doi.org/10.1016/j.compedu.2019.04.013

Bezdek, J. C., & Hathaway, R. J. (2002). VAT: a tool for visual assessment of (cluster) tendency. *Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02 (Cat. No.02CH37290)*, *3*, 2225–2230 vol.3. https://api.semanticscholar.org/CorpusID:124919037

Bocconi, S., Chioccariello, A., Dettori, G., Ferrari, A., & Engelhardt, K. (2016). *Developing Computational Thinking in Compulsory Education*. https://doi.org/10.2791/792158

Bocconi, S., Chioccariello, A., & Earp, J. (2018). *The nordic approach to introducing computational thinking and programming in compulsory education.* 1–42. https://doi.org/doi.org/10.17471/54007

Bocconi, S., Chioccariello, A., Kampylis, P., Wastiau, P., Engelhardt, K., Earp, J., Horvath, M., Malagoli, C., Cachia, R., Giannoutsou, N., & Punie, Y. (2022). *Reviewing Computational Thinking in Compulsory Education* (A. I. dos Santos, R. Cachia, N. Giannoutsou, & Y. Punie (eds.)).

Borsboom, D. (2017). A network theory of mental disorders. *World Psychiatry : Official*

Journal of the World Psychiatric Association (WPA), *16*(1), 5–13. https://doi.org/10.1002/wps.20375

Borsboom, D., van der Maas, H. L. J., Dalege, J., Kievit, R. A., & Haig, B. D. (2021). Theory Construction Methodology: A Practical Framework for Building Theories in Psychology. *Perspectives on Psychological Science : A Journal of the Association for Psychological Science*, *16*(4), 756–766. https://doi.org/10.1177/1745691620969647

Botvinick, M. M. (2012). Hierarchical reinforcement learning and decision making. *Current Opinion in Neurobiology*, *22*(6), 956–962. https://doi.org/10.1016/j.conb.2012.05.008

Botvinick, M., Wang, J. X., Dabney, W., Miller, K. J., & Kurth-Nelson, Z. (2020). Deep Reinforcement Learning and Its Neuroscientific Implications. *Neuron*, *107*(4), 603–616. https://doi.org/10.1016/j.neuron.2020.06.014

Bozkurt, A., & Durak, G. (2018). A systematic review of gamification research: In pursuit of homo ludens. *International Journal of Game-Based Learning*, *8*(3), 15–33. https://doi.org/10.4018/IJGBL.2018070102

Braun, A., März, A., Mertens, F., & Nisser, A. (2020). *Rethinking Education in the Digital Age*. https://doi.org/10.2861/84330

Breiman, L. (1984). *Classification and Regression Treesle* (1st ed.). Routledge. https://doi.org/https://doi.org/10.1201/9781315139470

Breiman, L. (2001). Random forests. *MACHINE LEARNING*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Briganti, G., Decety, J., Scutari, M., McNally, R. J., & Linkowski, P. (2022). Using Bayesian Networks to Investigate Psychological Constructs: The Case of Empathy. *Psychological Reports*, 332941221146711. https://doi.org/10.1177/00332941221146711

Briganti, G., Scutari, M., & Linkowski, P. (2021). Network Structures of Symptoms From the Zung Depression Scale. *Psychological Reports*, *124*(4), 1897–1911. https://doi.org/10.1177/0033294120942116

Briganti, G., Scutari, M., & McNally, R. J. (2023). A tutorial on bayesian networks for psychopathology researchers. *Psychological Methods*, *28*(4), 947–961. https://doi.org/10.1037/met0000479

Brockett, P., Haaland, P., & Levine, A. (1981). Information theoretic analysis of questionnaire data. *IEEE Transactions on Information Theory*, *27*(4), 438–446. https://doi.org/10.1109/TIT.1981.1056360

Brooks, B. L., Sherman, E. M. S., & Strauss, E. (2010). NEPSY-II: A Developmental neuropsychological assessment, second edition. *Child Neuropsychology*, *16*(1), 80–101. https://doi.org/10.1080/09297040903146966

Brooks, R., Brooks, S., & Goldstein, S. (2012). The Power of Mindsets: Nurturing Engagement, Motivation, and Resilience in Students. In *Handbook of Research on Student Engagement* (pp. 541--562).

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T. J., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language Models are Few-Shot Learners. *ArXiv*, *abs/2005.1*. https://api.semanticscholar.org/CorpusID:218971783

Buckley, S. (2012). *The Role of Computational Thinking and Critical Thinking in Problem Solving in a Learning Environment*.

Cadieu, C. F., Hong, H., Yamins, D. L. K., Pinto, N., Ardila, D., Solomon, E. A., Majaj, N. J.,

& DiCarlo, J. J. (2014). Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLoS Computational Biology*, *10*(12), e1003963. https://doi.org/10.1371/journal.pcbi.1003963

Cai, X., Kim, S., & Lee, D. (2011). Heterogeneous coding of temporally discounted values in the dorsal and ventral striatum during intertemporal choice. *Neuron*, *69*(1), 170–182. https://doi.org/10.1016/j.neuron.2010.11.041

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics*, *3*(1), 1–27. https://doi.org/10.1080/03610927408827101

Callaway, F., Jain, Y. R., van Opheusden, B., Das, P., Iwama, G., Gul, S., Krueger, P. M., Becker, F., Griffiths, T. L., & Lieder, F. (2022). Leveraging artificial intelligence to improve people's planning strategies. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(12), 1–11. https://doi.org/10.1073/pnas.2117432119

Callaway, F., Lieder, F., Krueger, P. M., & Griffiths, T. L. (2017). Mouselab-MDP : A new paradigm for tracing how people plan. *The 3rd Multidisciplinary Conference on Reinforcement Learning and Decision Making*, *June*, 1–6.

Caponetto, I., Earp, J., & Ott, M. (2014). Gamification and education: A literature review. *Proceedings of the European Conference on Games-Based Learning*, *1*(October), 50–57.

Caporarello, L., Magni, M., & Pennarola, F. (2019). One Game Does not Fit All. Gamification and Learning: Overview and Future Directions. In *Lecture Notes in Information Systems and Organisation* (Vol. 27). Springer International Publishing. https://doi.org/10.1007/978-3-319-90500-6_14

Carlson, A. G., Rowe, E., & Curby, T. W. (2013). Disentangling Fine Motor Skills' Relations to Academic Achievement: The Relative Contributions of Visual-Spatial Integration and

Visual-Motor Coordination. *The Journal of Genetic Psychology*, *174*(5), 514–533. https://doi.org/10.1080/00221325.2012.717122

Casaletto, K. B., & Heaton, R. K. (2017). Neuropsychological Assessment: Past and Future. *Journal of the International Neuropsychological Society*, *23*(9–10), 778–790. https://doi.org/DOI: 10.1017/S1355617717001060

Catts, H. W., Compton, D., Tomblin, J. B., & Bridges, M. S. (2012). Prevalence and nature of late-emerging poor readers. *Journal of Educational Psychology*, *104*(1), 166–181. https://doi.org/10.1037/a0025323

Cavalcanti, A. P., Barbosa, A., Carvalho, R., Freitas, F., Tsai, Y.-S., Gašević, D., & Mello, R. F. (2021). Automatic feedback in online learning environments: A systematic literature review. *Computers and Education: Artificial Intelligence*, *2*, 100027. https://doi.org/https://doi.org/10.1016/j.caeai.2021.100027

Celik, I., Dindar, M., Muukkonen, H., & Järvelä, S. (2022). The Promises and Challenges of Artificial Intelligence for Teachers: a Systematic Review of Research. *TechTrends*, *66*(4), 616–630. https://doi.org/10.1007/s11528-022-00715-y

Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, *48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Chassignol, M., Khoroshavin, A., Klimova, A., & Bilyatdinova, A. (2018). Artificial Intelligence trends in education: a narrative overview. *Procedia Computer Science*, *136*, 16–24. https://doi.org/https://doi.org/10.1016/j.procs.2018.08.233

Chawla, N. V, Bowyer, K. W., Hall, L. O., & Kegelmeyer, P. W. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–

357.

Chen, C. L. P. (2015). Deep learning for pattern learning and recognition. *2015 IEEE 10th Jubilee International Symposium on Applied Computational Intelligence and Informatics*, 17. https://api.semanticscholar.org/CorpusID:6265469

Chen, L., Chen, P., & Lin, Z. (2020). Artificial Intelligence in Education: A Review. *IEEE Access*, *8*, 75264–75278. https://doi.org/10.1109/ACCESS.2020.2988510

Chiang, F.-K., & Wallis, J. D. (2018). Neuronal Encoding in Prefrontal Cortex during Hierarchical Reinforcement Learning. *Journal of Cognitive Neuroscience*, *30*(8), 1197–1208. https://doi.org/10.1162/jocn_a_01272

Clatworthy, J., Buick, D., Hankins, M., Weinman, J., & Horne, R. (2005). The use and reporting of cluster analysis in health psychology: A review. *British Journal of Health Psychology*, *10*(3), 329–358. https://doi.org/10.1348/135910705X25697

Coenders, G., & Saris, W. (1995). Categorization and measurement quality. The choice between Pearson and Polychoric correlations. In *Acta Linguistica Hungarica - ACTA LINGUIST HUNG* (pp. 125–144).

Cohen, J. (1992). Statistical Power Analysis. *Current Directions in Psychological Science*, *1*(3), 98–101. https://doi.org/10.1111/1467-8721.ep10768783

Colombo, D., & Maathuis, M. H. (2014). Order-independent constraint-based causal structure learning. *J. Mach. Learn. Res.*, *15*(1), 3741–3782.

Constant, T., & Levieux, G. (2019). Dynamic Difficulty Adjustment Impact on Players' Confidence. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12. https://doi.org/10.1145/3290605.3300693

Cornoldi, C., Giofrè, D., & Baldi, A. P. (2017). *PROVE MT AVANZATE - 3 - CLINICA.*

Cortes, Corinna, & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018

Cortese, A. (2022). Metacognitive resources for adaptive learning⋆. *Neuroscience Research*, *178*(March 2021), 10–19. https://doi.org/10.1016/j.neures.2021.09.003

Costello, A. B., & Osborne, J. (2005). Best Practices in Exploratory Factor Analysis: Four Recommendations for Getting the Most From Your Analysis. *Practical Assessment, Research & Evaluation*, *10*, 1–9.

Cowell, R. G. (2001). Conditions Under Which Conditional Independence and Scoring Methods Lead to Identical Selection of Bayesian Network Models. *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, 91–97.

Csikszentmihalyi, M. (1990). Flow: The Psychology of Optimal Experience. *Harper and Row*. https://doi.org/10.2307/258925

Dabney, W., Kurth-Nelson, Z., Uchida, N., Starkweather, C. K., Hassabis, D., Munos, R., & Botvinick, M. (2020). A distributional code for value in dopamine-based reinforcement learning. *Nature*, *577*(7792), 671–675. https://doi.org/10.1038/s41586-019-1924-6

David, C. C., & Jacobs, D. J. (2014). Principal component analysis: a method for determining the essential dynamics of proteins. *Methods in Molecular Biology (Clifton, N.J.)*, *1084*, 193–226. https://doi.org/10.1007/978-1-62703-658-0_11

Davies, D. L., & Bouldin, D. W. (1979). A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-1*(2), 224–227. https://doi.org/10.1109/TPAMI.1979.4766909

De Sousa Borges, S., Durelli, V. H. S., Reis, H. M., & Isotani, S. (2014). A systematic

mapping on gamification applied to education. *Proceedings of the ACM Symposium on Applied Computing*, *Icmc*, 216–222. https://doi.org/10.1145/2554850.2554956

Demetriou, C., Ozer, B. U., & Essau, C. A. (2015). Self-Report Questionnaires. In *The Encyclopedia of Clinical Psychology* (pp. 1–6). https://doi.org/https://doi.org/10.1002/9781118625392.wbecp507

Denden, M., Tlili, A., Chen, N. S., Abed, M., Jemni, M., & Essalmi, F. (2022). The role of learners' characteristics in educational gamification systems: a systematic meta-review of the literature. *Interactive Learning Environments*, 1–23. https://doi.org/10.1080/10494820.2022.2098777

Dessureault, J.-S., & Massicotte, D. (2022). DPDRC, a Novel Machine Learning Method about the Decision Process for Dimensionality Reduction before Clustering. In *AI* (Vol. 3, Issue 1, pp. 1–21). https://doi.org/10.3390/ai3010001

Deterding, S., Khaled, R., Nacke, L. E., & Dixon, D. (2011). Gamification: Toward a Definition. Gamification Workshop. *CHI 2011 Gamification Workshop Proceedings*, *January 2011*, 12–15.

Devendren, A., & Nasri, N. M. (2022). Systematic Review: Students' Perceptions of the Use of Gamification. *International Journal of Academic Research in Business and Social Sciences*, *12*(8), 144–164. https://doi.org/10.6007/ijarbss/v12-i8/14268

Devers, C., & Gurung, R. (2014). Critical Perspective on Gamification in Education. In *Gamification in Education and Business* (pp. 1–14).

Dichev, C., & Dicheva, D. (2017). Gamifying education: what is known, what is believed and what remains uncertain: a critical review. In *International Journal of Educational Technology in Higher Education* (Vol. 14, Issue 1). International Journal of Educational

Technology in Higher Education. https://doi.org/10.1186/s41239-017-0042-5

Dicheva, D., & Dichev, C. (2015). Gamification in Education: Where Are We in 2015? *E-Learn 2015 - Kona, Hawaii, United States, July 2014*, 1445–1454.

Dickinson, A. (1985). Actions and habits: the development of behavioural autonomy. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, *308*(1135), 67–78. https://doi.org/10.1098/RSTB.1985.0010

Dikmen, M. (2021). Does gamification affect academic achievement? A meta-analysis of studies conducted in Turkey. *International Journal of Curriculum and Instruction Dikmen / International Journal of Curriculum and Instruction*, *13*(3), 3001–3020.

Domenech, P., Rheims, S., & Koechlin, E. (2020). Neural mechanisms resolving exploitation-exploration dilemmas in the medial prefrontal cortex. *Science (New York, N.Y.)*, *369*(6507). https://doi.org/10.1126/science.abb0184

Dong, Z., Yang, D., Reindl, T., & Walsh, W. M. (2015). A novel hybrid approach based on self-organizing maps, support vector regression and particle swarm optimization to forecast solar irradiance. *Energy*, *82*, 570–577. https://doi.org/10.1016/j.energy.2015.01.066

Drachen, A., Canossa, A., & Yannakakis, G. N. (2009). Player Modeling using Self-Organization in Tomb Raider: Underworld.le. *IEEE Symposium on Computational Intelligence and Games*, 1–8.

Dubé, A. K., & Wen, R. (2021). Identification and evaluation of technology trends in K-12 education from 2011 to 2021. *Education and Information Technologies*, *0123456789*. https://doi.org/10.1007/s10639-021-10689-8

Dulaney, A., Vasilyeva, M., & O'Dwyer, L. (2015). Individual differences in cognitive

resources and elementary school mathematics achievement: Examining the roles of storage and attention. *Learning and Individual Differences*, *37*, 55–63. https://doi.org/https://doi.org/10.1016/j.lindif.2014.11.008

Eisele, G., Vachon, H., Lafit, G., Kuppens, P., Houben, M., Myin-Germeys, I., & Viechtbauer, W. (2022). The Effects of Sampling Frequency and Questionnaire Length on Perceived Burden, Compliance, and Careless Responding in Experience Sampling Data in a Student Population. *Assessment*, *29*(2), 136–151. https://doi.org/10.1177/1073191120957102

Embretson, S. E., & Reise, S. P. (2013). *Item Response Theory*. Taylor \& Francis. https://books.google.it/books?id=9Xm0AAAAQBAJ

Engeström, Y. (2014). *Activity Theory and Learning at Work BT - Tätigkeit - Aneignung - Bildung: Positionierungen zwischen Virtualität und Gegenständlichkeit* (U. Deinet & C. Reutlinger (eds.); pp. 67–96). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-02120-7_3

Engeström, Y. (2015). *Learning by Expanding: An Activity-Theoretical Approach to Developmental Research* (2nd ed.). Cambridge University Press. https://doi.org/DOI: 10.1017/CBO9781139814744

Epskamp, S., Rhemtulla, M., & Borsboom, D. (2016). Generalized Network Psychometrics: Combining Network and Latent Variable Models. *Psychometrika*, *82*, 904–927. https://api.semanticscholar.org/CorpusID:35732614

Eronen, M. I. (2020). Causal discovery and the problem of psychological interventions. *New Ideas in Psychology*, *59*, 100785. https://doi.org/https://doi.org/10.1016/j.newideapsych.2020.100785

Eronen, M. I., & Bringmann, L. F. (2021). The Theory Crisis in Psychology: How to Move Forward. *Perspectives on Psychological Science*, *16*(4), 779–788. https://doi.org/10.1177/1745691620970586

European, C., Centre, J. R., Charisi, V., Chaudron, S., Di Gioia, R., Vuorikari, R., Escobar-Planas, M., Sanchez, I., & Gomez, E. (2022). *Artificial intelligence and the rights of the child – Towards an integrated agenda for research and policy*. Publications Office of the European Union. https://doi.org/doi/10.2760/012329

Ezugwu, A. E., Ikotun, A. M., Oyelade, O. O., Abualigah, L., Agushaka, J. O., Eke, C. I., & Akinyelu, A. A. (2022). A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. *Engineering Applications of Artificial Intelligence*, *110*, 104743. https://doi.org/https://doi.org/10.1016/j.engappai.2022.104743

Facey-Shaw, L., Specht, M., Van Rosmalen, P., Borner, D., & Bartley-Bryan, J. (2017). Educational functions and design of badge systems: A conceptual literature review. *IEEE Transactions on Learning Technologies*, *11*(4), 536–544. https://doi.org/10.1109/TLT.2017.2773508

Fadel, C., & Trilling, B. (2009). *21st Century Skills: Learning for Life in Our Times*. https://api.semanticscholar.org/CorpusID:145152570

Fadhli, M., Brick, B., Setyosari, P., Ulfa, S., & Kuswandi, D. (2020). A meta-analysis of selected studies on the effectiveness of gamification method for children. *International Journal of Instruction*, *13*(1), 845–854. https://doi.org/10.29333/iji.2020.13154a

Fagin, B., Harper, J. M., Baird, L. C., Hadfield, S. M., & Sward, R. E. (2006). Critical Thinking and Computer Science: Implicit and Explicit Connections. *Journal of Computing Sciences in Colleges*.

Faiella, F., & Ricciardi, M. (2015). Gamification and Learning : a Review Of issues And research. *Journal of E-Learning and Knowledge Society*, *11*(3), 13–21. https://www.learntechlib.org/p/151920/

Fairweather, J. R. (2001). Factor Stability, Number of Significant Loadings, and Interpretation: Evidence from Three Studies and Suggested Guidelines. *Operant Subjectivity*, *25*(1). https://ojs.library.okstate.edu/osu/index.php/osub/article/view/8925

Fan, C., Yao, L., Zhang, J., Zhen, Z., & Wu, X. (2023). Advanced Reinforcement Learning and Its Connections with Brain Neuroscience. *Research (Washington, D.C.)*, *6*, 64. https://doi.org/10.34133/research.0064

Fan, J., Fang, L., Wu, J., Guo, Y., & Dai, Q. (2020). From Brain Science to Artificial Intelligence. *Engineering*, *6*(3), 248–252. https://doi.org/https://doi.org/10.1016/j.eng.2019.11.012

Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex (New York, N.Y. : 1991)*, *1*(1), 1–47. https://doi.org/10.1093/cercor/1.1.1-a

Fenwick, M. E., Kubas, H. A., Witzke, J. W., Fitzer, K. R., Miller, D. C., Maricle, D. E., Harrison, G. L., Macoun, S. J., & Hale, J. B. (2016). Neuropsychological Profiles of Written Expression Learning Disabilities Determined by Concordance-Discordance Model Criteria. *Applied Neuropsychology: Child*, *5*(2), 83–96. https://doi.org/10.1080/21622965.2014.993396

Ferrari, A., Punie, Y., & Bre, B. N. (2013). *DIGCOMP : A Framework for Developing and Understanding Digital Competence in Europe* . https://doi.org/10.2788/52966

Ferrari, A., Punie, Y., & Redecker, C. (2012). *Understanding Digital Competence in the 21st*

Century: An Analysis of Current Frameworks BT - 21st Century Learning for 21st Century Skills (A. Ravenscroft, S. Lindstaedt, C. D. Kloos, & D. Hernández-Leo (eds.); pp. 79–92). Springer Berlin Heidelberg.

Figol, N., Faichuk, T., Pobidash, I., Trishchuk, O., & Teremko, V. (2021). Application fields of gamification. *Artificial Intelligence*, *10*, 93–100. https://api.semanticscholar.org/CorpusID:233790423

Firdaus, N., & Dixit, M. (2018). *Deep Learning Technique*. https://api.semanticscholar.org/CorpusID:212495631

Foley-Nicpon, M., Assouline, S. G., & Stinson, R. D. (2012). Cognitive and academic distinctions between gifted students with autism and Asperger syndrome. *Gifted Child Quarterly*, *56*(2), 77–89. https://doi.org/10.1177/0016986211433199

François-Lavet, V., Henderson, P., Islam, R., Bellemare, M. G., & Pineau, J. (2018). An Introduction to Deep Reinforcement Learning. *Found. Trends Mach. Learn.*, *11*, 219–354. https://api.semanticscholar.org/CorpusID:54434537

Freund, Y., & Schapire, R. E. (1996). Experiments with a New Boosting Algorithm. *Proceedings of the 13th International Conference on Machine Learning*, 148–156. https://doi.org/10.1.1.133.1040

Fried, E. I. (2020). Lack of Theory Building and Testing Impedes Progress in The Factor and Network Literature. *Psychological Inquiry*, *31*(4), 271–288. https://doi.org/10.1080/1047840X.2020.1853461

Frydenberg, E., Ainley, M., & Russell, V. (2005). *Schooling Issue Digest: Student Motivation and Engagement.*

García-Madruga, J. A., Vila, J. O., Gómez-Veiga, I., Duque, G., & Elosúa, M. R. (2014).

Executive processes, reading comprehension and academic achievement in 3th grade primary students. *Learning and Individual Differences*, *35*, 41–48. https://doi.org/https://doi.org/10.1016/j.lindif.2014.07.013

Garcia-Rudolph, A., Garcia-Molina, A., Opisso, E., & Tormos Muñoz, J. (2020). Personalized Web-Based Cognitive Rehabilitation Treatments for Patients with Traumatic Brain Injury: Cluster Analysis. *JMIR Med Inform*, *8*(10), e16077. https://doi.org/10.2196/16077

Gareth James Trevor Hastie, Robert Tibshirani, D. W. (2013). *An introduction to statistical learning : with applications in R*. Springer. https://search.library.wisc.edu/catalog/9910207152902121

Geary, D. C., & Hoard, M. K. (2001). Numerical and arithmetical deficits in learning-disabled children: Relation to dyscalculia and dyslexia. *Aphasiology*, *15*(7), 635–647. https://doi.org/10.1080/02687040143000113

Gerber, H. R. (2014). Problems and Possibilities of Gamifying Learning: A Conceptual Review. *Internet Learning*, *1*. https://doi.org/10.18278/il.3.2.4

Gershman, S. J. (2018). The Successor Representation: Its Computational Logic and Neural Substrates. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *38*(33), 7193–7200. https://doi.org/10.1523/JNEUROSCI.0151-18.2018

Gershman, S. J., & Daw, N. D. (2017). Reinforcement Learning and Episodic Memory in Humans and Animals: An Integrative Framework. *Annual Review of Psychology*, *68*, 101–128. https://doi.org/10.1146/annurev-psych-122414-033625

Gillath, O., Ai, T., Branicky, M. S., Keshmiri, S., Davison, R. B., & Spaulding, R. (2021). Attachment and trust in artificial intelligence. *Computers in Human Behavior*, *115*,

106607. https://doi.org/https://doi.org/10.1016/j.chb.2020.106607

Giri, N., Saini, T., Bhole, K., Bhosale, A., Shetty, T., Subramanyam, A., & Shelke, S. (2020). Detection of Dyscalculia Using Machine Learning. *2020 5th International Conference on Communication and Electronics Systems (ICCES)*, 1–6. https://doi.org/10.1109/ICCES48766.2020.9137871

Gkintoni, E., Halkiopoulos, C., Antonopoulou, H., & Πετροπουλος, N. (2021). Gamification of Neuropsychological Tools as a Multi-Sensory Approach to Education. Stroop's Paradigm. *Technium Romanian Journal of Applied Sciences and Technology*, *3*, 92–102. https://doi.org/10.47577/technium.v3i8.4798

Glover, I. (2013). Play As You Learn : Gamification as a Technique for Motivating Learners. Proceedings of World Conference on Educational Multimedia. *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications*, 1999–2008.

Gohel, P., Singh, P., & Mohanty, M. (2021). Explainable AI: current status and future directions. *ArXiv*, *abs/2107.0*.

Göksel, N., & Bozkurt, A. (2019). *Artificial Intelligence in Education: Current Insights and Future Perspectives* (pp. 224–236).

Goldman-Rakic, P. S. (1991). Cellular and circuit basis of working memory in prefrontal cortex of nonhuman primates. *Progress in Brain Research*, *85*, 325–336.

González Mariño, J. C., Cantú Gallegos, M. D. L., Camacho Cruz, H. E., & Rosales Camacho, J. A. (2018). Redesigning the bartle test of gamer psychology for its application in gamification processes of learning. *IMSCI 2018 - 12th International Multi-Conference on Society, Cybernetics and Informatics, Proceedings*, *1*(July), 35–40.

Gotoh, Y. (2004). *DIMENSIONALITY REDUCTION TECHNIQUES FOR SEARCH RESULTS CLUSTERING*. https://api.semanticscholar.org/CorpusID:58963759

Grassé, P.-P. (1959). La reconstruction du nid et les coordinations interindividuelles chez Bellicositermes natalensis et Cubitermes sp. la théorie de la stigmergie: Essai d'interprétation du comportement des termites constructeurs. *Insectes Sociaux*, *6*(1), 41–80. https://doi.org/10.1007/BF02223791

Graves, A., Wayne, G., Reynolds, M., Harley, T., Danihelka, I., Grabska-Barwińska, A., Colmenarejo, S. G., Grefenstette, E., Ramalho, T., Agapiou, J., Badia, A. P., Hermann, K. M., Zwols, Y., Ostrovski, G., Cain, A., King, H., Summerfield, C., Blunsom, P., Kavukcuoglu, K., & Hassabis, D. (2016). Hybrid computing using a neural network with dynamic external memory. *Nature*, *538*(7626), 471–476. https://doi.org/10.1038/nature20101

Greenberg, D. L., & Verfaellie, M. (2010). Interdependence of episodic and semantic memory: evidence from neuropsychology. *Journal of the International Neuropsychological Society : JINS*, *16*(5), 748–753. https://doi.org/10.1017/S1355617710000676

Guilford, J. . (1950). Creativity. *American Psychologist*, 151–158.

Guyon, H., Falissard, B., & Kop, J.-L. (2017). Modeling Psychological Attributes in Psychology – An Epistemological Discussion: Network Analysis vs. Latent Variables. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.00798

Haenlein, M., & Kaplan, A. (2019). A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence. *California Management Review*, *61*(4), 5–14.

Halassa, M. M., Fellin, T., Takano, H., Dong, J.-H., & Haydon, P. G. (2007). Synaptic islands defined by the territory of a single astrocyte. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *27*(24), 6473–6477. https://doi.org/10.1523/JNEUROSCI.1419-07.2007

Hallifax, S., Serna, A., Marty, J.-C., & Lavoué, E. (2019). Adaptive Gamification in Education: A Literature Review of Current Trends and Developments. *European Conference on Technology Enhanced Learning (EC-TEL)*, 3–16.

Hamari, J., Koivisto, J., & Sarsa, H. (2014). Does Gamification Work? — A Literature Review of Empirical Studies on Gamification. In *Proceedings of the Annual Hawaii International Conference on System Sciences*. https://doi.org/10.1109/HICSS.2014.377

Harry, A. (2023). Role of AI in Education. *Interdiciplinary Journal and Hummanity (INJURITY)*, *2*, 260–268. https://doi.org/10.58631/injurity.v2i3.52

Hartemink, & John, A. (2001). *Principled computational methods for the validation discovery of genetic regulatory networks*.

Harvey, P. D. (2019). Domains of cognition and their assessment. *Dialogues in Clinical Neuroscience*, *21*(3), 227–237. https://doi.org/10.31887/DCNS.2019.21.3/pharvey

Hashemi, S. E., Gholian-Jouybari, F., & Hajiaghaei-Keshteli, M. (2023). A fuzzy C-means algorithm for optimizing data clustering. *Expert Systems with Applications*, *227*, 120377. https://doi.org/https://doi.org/10.1016/j.eswa.2023.120377

Hassan, L., Dias, A., & Hamari, J. (2019). How motivational feedback increases user's benefits and continued use: A study on gamification, quantified-self and social networking. *International Journal of Information Management*, *46*(July 2018), 151–162. https://doi.org/10.1016/j.ijinfomgt.2018.12.004

Hebb, O. (1950). The Organization of Behavior: A Neuropsychological Theory. *Journal of the American Medical Association*, *143*(12), 1123.

Heim, S., Tschierse, J., Amunts, K., Wilms, M., Vossel, S., Willmes, K., Grabowska, A., & Huber, W. (2008). Cognitive subtypes of dyslexia. *Acta Neurobiologiae Experimentalis*, *68*(1), 73–82.

Hennig, C. (2007). Cluster-wise assessment of cluster stability. *Computational Statistics & Data Analysis*, *52*(1), 258–271. https://doi.org/https://doi.org/10.1016/j.csda.2006.11.025

Hennig, C. (2008). Dissolution point and isolation robustness: Robustness criteria for general cluster analysis methods. *Journal of Multivariate Analysis*, *99*(6), 1154–1176. https://doi.org/https://doi.org/10.1016/j.jmva.2007.07.002

Heylighen, F. (2016). Stigmergy as a universal coordination mechanism I: Definition and components. *Cognitive Systems Research*, *38*, 4–13. https://doi.org/https://doi.org/10.1016/j.cogsys.2015.12.002

Higgins, I., Matthey, L., Pal, A., Burgess, C. P., Glorot, X., Botvinick, M. M., Mohamed, S., & Lerchner, A. (2016). beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. *International Conference on Learning Representations*. https://api.semanticscholar.org/CorpusID:46798026

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hodges, S., Scott, J., Sentance, S., Miller, C., Villar, N., Schwiderski-Grosche, S., Hammil, K., & Johnston, S. (2013). .NET gadgeteer: a new platform for K-12 computer science education. *Proceeding of the 44th ACM Technical Symposium on Computer Science Education*, 391–396. https://doi.org/10.1145/2445196.2445315

Holgado-Tello, F., Chacón-Moscoso, S., Barbero–García, I., & Vila, E. (2010). Polychoric versus Pearson correlations in Exploratory and Confirmatory Factor Analysis with ordinal variables. *Quality and Quantity*, *44*, 153–166. https://doi.org/10.1007/s11135-008-9190-y

Hopkins, B., & Skellam, J. G. (1954). A New Method for determining the Type of Distribution of Plant Individuals. *Annals of Botany*, *18*(70), 213–227. http://www.jstor.org/stable/42907238

Hori, K., Fukuhara, H., & Yamada, T. (2020). Item response theory and its applications in educational measurement Part I: Item response theory and its implementation in R. *Wiley Interdisciplinary Reviews: Computational Statistics*, *14*. https://api.semanticscholar.org/CorpusID:228967589

Howard-Jones, P. A., & Jay, T. (2016). Reward, learning and games. *Current Opinion in Behavioral Sciences*, *10*, 65–72. https://doi.org/10.1016/j.cobeha.2016.04.015

Huang, R., Ritzhaupt, A. D., Sommer, M., Zhu, J., Stephen, A., Valle, N., Hampton, J., & Li, J. (2020). The impact of gamification in educational settings on student learning outcomes: a meta-analysis. *Educational Technology Research and Development*, *68*(4), 1875–1901. https://doi.org/10.1007/s11423-020-09807-z

Hubel, D. H., & Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *The Journal of Physiology*, *148*(3), 574–591. https://doi.org/10.1113/jphysiol.1959.sp006308

Hung, A. C. Y. (2017). A critique and defense of gamification. *Journal of Interactive Online Learning*, *15*(1), 57–72.

Indriasari, T. D., Luxton-Reilly, A., & Denny, P. (2020). Gamification of student peer review

in education: A systematic literature review. *Education and Information Technologies*, *25*(6), 5205–5234. https://doi.org/10.1007/s10639-020-10228-x

Inocencio, F. D. C. (2018). Using gamification in education: A systematic literature review. *International Conference on Information Systems 2018, ICIS 2018*, 1–17.

Ising, H. K., Veling, W., Loewy, R. L., Rietveld, M. W., Rietdijk, J., Dragt, S., Klaassen, R. M. C., Nieman, D. H., Wunderink, L., Linszen, D. H., & van der Gaag, M. (2012). The validity of the 16-item version of the Prodromal Questionnaire (PQ-16) to screen for ultra high risk of developing psychosis in the general help-seeking population. *Schizophrenia Bulletin*, *38*(6), 1288–1296. https://doi.org/10.1093/schbul/sbs068

Israel-Fishelson, R., & Hershkovitz, A. (2022). Computers & Education Studying interrelations of computational thinking and creativity : A scoping review (2011 – 2020). *Computers & Education*, *176*(October 2021), 104353. https://doi.org/10.1016/j.compedu.2021.104353

Jacobucci, R., & Grimm, K. J. (2020). Machine Learning and Psychological Research: The Unexplored Effect of Measurement. *Perspectives on Psychological Science*, *15*(3), 809–816. https://doi.org/10.1177/1745691620902467

James, W. (2007). *The principles of psychology* (Vol. 1). Cosimo, Inc.

Jayanth Krishnan, K., & Mitra, K. (2022). A modified Kohonen map algorithm for clustering time series data. *Expert Systems with Applications*, *201*, 117249. https://doi.org/https://doi.org/10.1016/j.eswa.2022.117249

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*, 389–399. https://api.semanticscholar.org/CorpusID:201827642

Johnson-Laird, P. (1998). *The Computer and the Mind*. Collins.

Johnson, E. S. (2014). Understanding Why a Child Is Struggling to Learn. *Topics in Language Disorders*, *34*(1), 59–73. https://doi.org/10.1097/TLD.0000000000000007

Jonides, J., Smith, E. E., Koeppe, R. A., Awh, E., Minoshima, S., & Mintun, M. A. (1993). Spatial working memory in humans as revealed by PET. *Nature*, *363*(6430), 623–625. https://doi.org/10.1038/363623a0

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science (New York, N.Y.)*, *349*(6245), 255–260. https://doi.org/10.1126/science.aaa8415

Jordan, P., & Spiess, M. (2019). Rethinking the Interpretation of Item Discrimination and Factor Loadings. *Educational and Psychological Measurement*, *79*(6), 1103–1132. https://doi.org/10.1177/0013164419843164

Juntunen, P., Liukkonen, M., Lehtola, M., & Hiltunen, Y. (2013). Cluster analysis by self-organizing maps : An application to the modelling of water quality in a treatment process. *Applied Soft Computing Journal*, *13*(7), 3191–3196. https://doi.org/10.1016/j.asoc.2013.01.027

Kaisar, S. (2020). Developmental dyslexia detection using machine learning techniques : A survey. *ICT Express*, *6*(3), 181–184. https://doi.org/https://doi.org/10.1016/j.icte.2020.05.006

Kalogiannakis, M., Papadakis, S., & Zourmpakis, A. I. (2021). Gamification in science education. A systematic review of the literature. *Education Sciences*, *11*(1), 1–36. https://doi.org/10.3390/educsci11010022

Kapp, K. (2012). *The Gamification of Learning and Instruction: Game-based Methods and*

Strategies for Training and Eduction.

Kaufman, J. C., & Glaveanu, V. P. (2019). A Review of Creativity Theories. In *The Cambridge Handbook of Creativity* (pp. 27–43). Cambridge University Press.

Kaufman, L., & Rousseeuw, P. (1990). Finding Groups in Data: An Introduction To Cluster Analysis. In *Wiley, New York. ISBN 0-471-87876-6.* https://doi.org/10.2307/2532178

Khaldi, A., Bouzidi, R., & Nader, F. (2023). Gamification of e-learning in higher education: a systematic literature review. *Smart Learning Environments*, *10*(1). https://doi.org/10.1186/s40561-023-00227-z

Khaleghi, A., Aghaei, Z., & Mahdavi, M. A. (2021). A Gamification Framework for Cognitive Assessment and Cognitive Training: Qualitative Study. *JMIR Serious Games, 9*. https://api.semanticscholar.org/CorpusID:233035809

Kim, J., & Castelli, D. M. (2021). Effects of gamification on behavioral change in education: A meta-analysis. *International Journal of Environmental Research and Public Health*, *18*(7). https://doi.org/10.3390/ijerph18073550

Kingma, Diederik P, & Welling, M. (2013). Auto-Encoding Variational Bayes. *CoRR*, *abs/1312.6*. https://api.semanticscholar.org/CorpusID:216078090

Kingma, Diederik P, & Welling, M. (2019). *An Introduction to Variational Autoencoders*. now. http://ieeexplore.ieee.org/document/9051780

Kingma, Durk P, Salimans, T., & Welling, M. (2015). Variational Dropout and the Local Reparameterization Trick. In C Cortes, N. Lawrence, D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 28). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2015/file/bc7316929fe1545bf0b98d114

ee3ecb8-Paper.pdf

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., & Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, *114*(13), 3521–3526. https://doi.org/10.1073/pnas.1611835114

Kishore, K., Jaswal, V., Kulkarni, V., & De, D. (2021). Practical Guidelines to Develop and Evaluate a Questionnaire. *Indian Dermatology Online Journal*, *12*(2), 266–275. https://doi.org/10.4103/idoj.IDOJ_674_20

Kitson, N. K., Constantinou, A. C., Guo, Z., Liu, Y., & Chobtham, K. (2023). A survey of Bayesian Network structure learning. *Artificial Intelligence Review*, *56*(8), 8721–8814. https://doi.org/10.1007/s10462-022-10351-w

Kiwanuka, F., Kopra, J., Sak-Dankosky, N., Nanyonga, R. C., & Kvist, T. (2022). Polychoric Correlation With Ordinal Data in Nursing Research. *Nursing Research*, *71*(6), 469–476. https://doi.org/10.1097/NNR.0000000000000614

Kocakoyun, S., & Ozdamli, F. (2018). A Review of Research on Gamification Approach in Education. In *Socialization - A Multidimensional Perspective* (pp. 51–72).

Kohonen, T. (1990). The Self-organizing Map. *PROCEEDINGS OF THE IEEE*, *78*(9), 1464–1480.

Koivisto, J., & Hamari, J. (2019). The rise of motivational information systems: A review of gamification research. *International Journal of Information Management*, *45*(December 2018), 191–210. https://doi.org/10.1016/j.ijinfomgt.2018.10.013

Koller, D., & Friedman, N. (2009). *Probabilistic Graphical Models: Principles and*

Techniques. MIT Press. https://books.google.it/books?id=7dzpHCHzNQ4C

Kozachkov, L., Kastanenka, K. V, & Krotov, D. (2023). Building transformers from neurons and astrocytes. *Proceedings of the National Academy of Sciences of the United States of America*, *120*(34), e2219150120. https://doi.org/10.1073/pnas.2219150120

Krath, J., Schürmann, L., & von Korflesch, H. F. O. (2021). Revealing the theoretical basis of gamification: A systematic review and analysis of theory in research on gamification, serious games and game-based learning. *Computers in Human Behavior*, *125*(August), 106963. https://doi.org/10.1016/j.chb.2021.106963

Kraus, J. M., Müssel, C., Palm, G., & Kestler, H. A. (2011). Multi-objective selection for collecting cluster alternatives. *Computational Statistics*, *26*(2), 341–353. https://doi.org/10.1007/s00180-011-0244-6

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In F. Pereira, C. J. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems* (Vol. 25). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

Kudo, M. F., Lussier, C. M., & Swanson, H. L. (2015). Reading disabilities in children: A selective meta-analysis of the cognitive literature. *Research in Developmental Disabilities*, *40*, 51–62. https://doi.org/10.1016/j.ridd.2015.01.002

Kuga, N., Sasaki, T., Takahara, Y., Matsuki, N., & Ikegaya, Y. (2011). Large-Scale Calcium Waves Traveling through Astrocytic Networks In Vivo. *Journal of Neuroscience*, *31*(7), 2607–2614. https://doi.org/10.1523/JNEUROSCI.5319-10.2011

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling* (Springer (ed.)). Springer.

Kurita, T. (2014). *Principal Component Analysis (PCA) BT - Computer Vision: A Reference Guide* (K. Ikeuchi (ed.); pp. 636–639). Springer US. https://doi.org/10.1007/978-0-387-31439-6_649

Kurita, T. (2019). *Principal Component Analysis (PCA) BT - Computer Vision: A Reference Guide* (pp. 1–4). Springer International Publishing. https://doi.org/10.1007/978-3-030-03243-2_649-1

Laine, T. H., & Lindberg, R. S. N. (2020). Designing Engaging Games for Education: A Systematic Literature Review on Game Motivators and Design Principles. *IEEE Transactions on Learning Technologies*, *13*(4), 804–821. https://doi.org/10.1109/TLT.2020.3018503

Landers, R. N. (2014). Developing a Theory of Gamified Learning: Linking Serious Games and Gamification of Learning. *Simulation and Gaming*, *45*(6), 752–768. https://doi.org/10.1177/1046878114563660

Langer, M., & Landers, R. N. (2021). The future of artificial intelligence at work: A review on effects of decision automation and augmentation on workers targeted by algorithms and third-party observers. *Computers in Human Behavior*, *123*, 106878. https://doi.org/https://doi.org/10.1016/j.chb.2021.106878

Langer, N., Weber, M., Vieira, B. H., Strzelczyk, D., Wolf, L., Pedroni, A., Heitz, J., Müller, S., Schultheiss, C., Tröndle, M., Lasprilla, J. C. A., Rivera, D., Scarpina, F., Zhao, Q., Leuthold, R., Wehrle, F., Jenni, O. G., Brugger, P., Zaehle, T., … Zhang, C. (2022). The AI Neuropsychologist: Automatic scoring of memory deficits with deep learning. *BioRxiv*, 2022.06.15.496291. https://doi.org/10.1101/2022.06.15.496291

Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, *1*(4), 541–551. https://doi.org/10.1162/neco.1989.1.4.541

Lee, J., Jeong, J.-Y., & Jun, C.-H. (2020). Markov blanket-based universal feature selection for classification and regression of mixed-type data. *Expert Syst. Appl.*, *158*, 113398. https://api.semanticscholar.org/CorpusID:219425536

Lee, S. W., Shimojo, S., & O'Doherty, J. P. (2014). Neural computations underlying arbitration between model-based and model-free learning. *Neuron*, *81*(3), 687–699. https://doi.org/10.1016/j.neuron.2013.11.028

Leite, W. L., Roy, S. G., Chakraborty, N., Michailidis, G., Huggins-Manley, A. C., D'Mello, S. K., Faradonbeh, M. K. S., Jensen, E., Kuang, H., & Jing, Z. (2022). A novel video recommendation system for algebra: An effectiveness evaluation study. *LAK22: 12th International Learning Analytics and Knowledge Conference*. https://api.semanticscholar.org/CorpusID:247222410

Leong, Y. C., Radulescu, A., Daniel, R., DeWoskin, V., & Niv, Y. (2017). Dynamic Interaction between Reinforcement Learning and Attention in Multidimensional Environments. *Neuron*, *93*, 451–463. https://api.semanticscholar.org/CorpusID:30453355

Levine, S., Kumar, A., Tucker, G., & Fu, J. (2020). *Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems*.

Liapis, A., Yannakakis, G. N., Alexopoulos, C., & Lopes, P. (2016). Can Computers Foster

Human Users' Creativity? Theory and Praxis of Mixed-Initiative Co-Creativity. *Digital Culture and Education*, *8*(2), 136–153. https://www.um.edu.mt/library/oar/handle/123456789/29476%0A

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., & Wierstra, D. (2016). *CONTINUOUS CONTROL WITH DEEP REINFORCEMENT LEARNING*. https://goo.gl/J4PIAz

Limantara, N., Meyliana, Hidayanto, A. N., & Prabowo, H. (2019). The elements of gamification learning in higher education: A systematic literature review. *International Journal of Mechanical Engineering and Technology*, *10*(2), 982–991.

Lin, E., Mukherjee, S., & Kannan, S. (2020). A deep adversarial variational autoencoder model for dimensionality reduction in single-cell RNA sequencing analysis. *BMC Bioinformatics*, *21*(1), 64. https://doi.org/10.1186/s12859-020-3401-5

Lin, J. (1991). Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, *37*(1), 145–151. https://doi.org/10.1109/18.61115

Liu, M., & Lu, H. (2002). A STUDY ON THE CREATIVE Problem-Solving Process in Computer Programming. *Proceeding of the International Conference on Engineering Education.* https://pdfs.semanticscholar.org/057e/f657236382b17b7b3e9865178709def3296b.pdf.

Liu, T., Yu, H., & Blair, R. H. (2022). Stability estimation for unsupervised clustering: A review. *WIREs Computational Statistics*, *14*(6), e1575. https://doi.org/https://doi.org/10.1002/wics.1575

Lopes, J. C., & Lopes, R. P. (2022). *A Review of Dynamic Difficulty Adjustment Methods for Serious Games BT - Optimization, Learning Algorithms and Applications* (A. I.

Pereira, A. Košir, F. P. Fernandes, M. F. Pacheco, J. P. Teixeira, & R. P. Lopes (eds.); pp. 144–159). Springer International Publishing.

Lumsden, J., Edwards, E. A., Lawrence, N. S., Coyle, D., & Munafo, M. R. (2016). Gamification of Cognitive Assessment and Cognitive Training: A Systematic Review of Applications and Efficacy. *JMIR Serious Games*, *4*. https://api.semanticscholar.org/CorpusID:23779301

Lumsden, J., Edwards, E. A., Lawrence, N. S., Coyle, D., & Munafò, M. R. (2016). Gamification of cognitive assessment and cognitive training: A systematic review of applications and efficacy. *JMIR Serious Games*, *4*(2). https://doi.org/10.2196/games.5888

Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

Luque, A., Carrasco, A., Martín, A., & De, A. (2019). The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition*, *91*, 216–231. https://doi.org/10.1016/j.patcog.2019.02.023

Maaten, L. J. P. van der, & Hinton, G. E. (2008). Visualizing High-Dimensional Data Using t-SNE. *Journal of Machine Learning Research*, *9*(nov), 2579–2605. https://research.tilburguniversity.edu/en/publications/visualizing-high-dimensional-data-using-t-sne

MacQueen, J. (1967). *Some methods for classification and analysis of multivariate*

observations. https://api.semanticscholar.org/CorpusID:6278891

Maher, M. Lou, Brady, K., & Fisher, D. H. (2013). Computational Models of Surprise in Evaluating Creative Design AI Approaches for Assessing Surprise. *Proceedings of the Fourth International Conference on Computational Creativity (ICCC-2013)*, 147–151.

Maia, T. V. (2009). Reinforcement learning, conditioning, and the brain: Successes and challenges. *Cognitive, Affective & Behavioral Neuroscience*, *9*(4), 343–364. https://doi.org/10.3758/CABN.9.4.343

Majuri, J., Koivisto, J., & Hamari, J. (2018). Gamification of education and learning: A review of empirical literature. *CEUR Workshop Proceedings*, *2186*, 11–19.

Mamekova, A. T., Toxanbayeva, N. K., Naubaeva, K. T., Ongarbayeva, S. S., & Akhmediyeva, K. N. (2021). A Meta-Analysis on the Impact of Gamification over Students' Motivation. *Journal of Intellectual Disability - Diagnosis and Treatment*, *9*(4), 417–422. https://doi.org/10.6000/2292-2598.2021.09.04.9

Manzano-León, A., Camacho-Lazarraga, P., Guerrero, M. A., Guerrero-Puerta, L., Aguilar-Parra, J. M., Trigueros, R., & Alias, A. (2021). Between level up and game over: A systematic literature review of gamification in education. *Sustainability (Switzerland)*, *13*(4), 1–14. https://doi.org/10.3390/su13042247

Marczewski, A. (2015). *User Types HEXAD* (pp. 65–80).

Matsuo, Y., LeCun, Y., Sahani, M., Precup, D., Silver, D., Sugiyama, M., Uchibe, E., & Morimoto, J. (2022). Deep learning, reinforcement learning, and world models. *Neural Networks*, *152*, 267–275. https://doi.org/https://doi.org/10.1016/j.neunet.2022.03.037

Mazzoni, E., & Benvenuti, M. (2015). A Robot-Partner for Preschool Children Learning English Using Socio-Cognitive Conflict. *Journal of Educational Technology & Society*,

18(4), 474–485. http://www.jstor.org/stable/jeductechsoci.18.4.474

Mazzoni, E., Benvenuti, M., & Orsoni, M. (2022). Robotica e tecnologie per lo sviluppo: come costruire le competenze del futuro. In *La Società dei Robot* (pp. 215–226). Mondadori.

McCarthy, G., Puce, A., Constable, T., Krystal, J. H., Gore, J. C., & Goldman-Rakic, P. (1996). Activation of Human Prefrontal Cortex during Spatial and Nonspatial Working Memory Tasks Measured by Functional MRI. *Cerebral Cortex*, *6*(4), 600–611. https://doi.org/10.1093/cercor/6.4.600

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955. *AI Magazine*, *27*(4), 1–17. https://doi.org/https://doi.org/10.1609/aimag.v27i4.1904

McLean, J. F., & Hitch, G. J. (1999). Working Memory Impairments in Children with Specific Arithmetic Learning Difficulties. *Journal of Experimental Child Psychology*, *74*(3), 240–260. https://doi.org/10.1006/jecp.1999.2516

McNally, R. J. (2016). Can network analysis transform psychopathology? *Behaviour Research and Therapy*, *86*, 95–104. https://doi.org/https://doi.org/10.1016/j.brat.2016.06.006

Menghini, D., Finzi, A., Benassi, M., Bolzani, R., Facoetti, A., Giovagnoli, S., Ruffino, M., & Vicari, S. (2010). Different underlying neurocognitive deficits in developmental dyslexia: A comparative study. *Neuropsychologia*, *48*(4), 863–872. https://doi.org/10.1016/j.neuropsychologia.2009.11.003

Metwally, A. H. S., Nacke, L. E., Chang, M., Wang, Y., & Yousef, A. M. F. (2021). Revealing the hotspots of educational gamification: An umbrella review. *International*

Journal of Educational Research, *109*(August), 101832. https://doi.org/10.1016/j.ijer.2021.101832

Mijwil, M. M., Aggarwal, K., Mutar, D. S., Mansour, N., & Singh, R. S. S. (2022). The Position of Artificial Intelligence in the Future of Education: An Overview. *Asian Journal of Applied Sciences*. https://api.semanticscholar.org/CorpusID:248630963

Miller, G. A. (2003). The cognitive revolution: a historical perspective. *Trends in Cognitive Sciences*, *7*(3), 141–144. https://doi.org/10.1016/S1364-6613(03)00029-9

Miller, L. D., Soh, L.-K., Chiriacescu, V., Ingraham, E., Shell, D. F., Ramsay, S., & Hazley, M. P. (2013). Improving learning of computational thinking using creative thinking exercises in CS-1 computer science courses. *2013 IEEE Frontiers in Education Conference (FIE)*, 1426–1432. https://doi.org/10.1109/FIE.2013.6685067

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., & Kavukcuoglu, K. (2016). Asynchronous Methods for Deep Reinforcement Learning. In M. F. Balcan & K. Q. Weinberger (Eds.), *Proceedings of The 33rd International Conference on Machine Learning* (Vol. 48, pp. 1928–1937). PMLR. https://proceedings.mlr.press/v48/mniha16.html

Mohammed, Y. B., & Ozdamli, F. (2021). Motivational Effects of Gamification Apps in Education: A Systematic Literature Review. *Brain. Broad Research in Artificial Intelligence and Neuroscience*, *12*(2). https://doi.org/10.18662/brain/12.2/196

Molin, F., Haelermans, C., Cabus, S., & Groot, W. (2020). The effect of feedback on metacognition - A randomized experiment using polling technology. *Computers and Education*, *152*(October 2019), 103885. https://doi.org/10.1016/j.compedu.2020.103885

Moll, K., Göbel, S. M., Gooch, D., Landerl, K., & Snowling, M. J. (2016). Cognitive Risk

Factors for Specific Learning Disorder. *Journal of Learning Disabilities*, *49*(3), 272–281. https://doi.org/10.1177/0022219414547221

Momennejad, I., Russek, E. M., Cheong, J. H., Botvinick, M. M., Daw, N. D., & Gershman, S. J. (2017). The successor representation in human reinforcement learning. *Nature Human Behaviour*, *1*(9), 680–692. https://doi.org/10.1038/s41562-017-0180-8

Mora, A., Riera, D., González, C., & Arnedo-Moreno, J. (2017). Gamification: a systematic review of design frameworks. *Journal of Computing in Higher Education*, *29*(3), 516–548. https://doi.org/10.1007/s12528-017-9150-4

Mubin, O., Stevens, C. J., Shahid, S., Mahmud, A. Al, & Dong, J.-J. (2013). *A review of the applicability of robots in education*. https://api.semanticscholar.org/CorpusID:7969426

Muller, T. H., Butler, J. L., Veselic, S., Miranda, B., Wallis, J. D., Dayan, P., Behrens, T. E. J., Kurth-Nelson, Z., & Kennerley, S. W. (2024). Distributional reinforcement learning in prefrontal cortex. *Nature Neuroscience*. https://doi.org/10.1038/s41593-023-01535-w

Nacke, L. E., Bateman, C., & Mandryk, R. L. (2011). BrainHex: Preliminary results from a neurobiological gamer typology survey. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *6972 LNCS*, 288–293. https://doi.org/10.1007/978-3-642-24500-8_31

Nadi-Ravandi, S., & Batooli, Z. (2022). Gamification in education: A scientometric, content and co-occurrence analysis of systematic review and meta-analysis articles. In *Education and Information Technologies* (Vol. 27, Issue 7). Springer US. https://doi.org/10.1007/s10639-022-11048-x

Nah, F. F. H., Telaprolu, V. R., Rallapalli, S., & Venkata, P. R. (2013). Gamification of Education Using Computer Games. In *Human Interface and the Management of*

Information (pp. 99–107).

Nair, S., & Mathew, J. (2021). Learning through Play: Gamification of Learning A Systematic Review of Studies on Gamified Learning. *Journal of Information Technology Management*, *14*(1), 113–126. https://doi.org/10.22059/JITM.2021.322193.2779

Nayak, J., Naik, B., & Behera, H. S. (2015a). Fuzzy C-Means (FCM) Clustering Algorithm: A Decade Review from 2000 to 2014. In L. C. Jain, H. S. Behera, J. K. Mandal, & D. P. Mohapatra (Eds.), *Computational Intelligence in Data Mining - Volume 2* (pp. 133–149). Springer India.

Nayak, J., Naik, B., & Behera, H. S. (2015b). *Fuzzy C-Means (FCM) Clustering Algorithm: A Decade Review from 2000 to 2014 BT - Computational Intelligence in Data Mining - Volume 2* (L. C. Jain, H. S. Behera, J. K. Mandal, & D. P. Mohapatra (eds.); pp. 133–149). Springer India.

Nesayan, A., Amani, M., & Gandomani, R. A. (2018). Cognitive Profile of Children and its Relationship With Academic Performance. *Basic and Clinical Neuroscience*, *10*, 165–174. https://api.semanticscholar.org/CorpusID:134234235

Nguyen, V. Q., Nguyen, V. H., Cao, V. L., Khac, N. A. Le, & Shone, N. (2021). A Robust PCA Feature Selection To Assist Deep Clustering Autoencoder-Based Network Anomaly Detection. *2021 8th NAFOSTED Conference on Information and Computer Science (NICS)*, 335–341. https://api.semanticscholar.org/CorpusID:246682472

Nielsen, F. (2010). A family of statistical symmetric divergences based on Jensen's inequality. *Computing Research Repository - CORR*.

Nurtanto, M., Kholifah, N., Ahdhianto, E., Samsudin, A., & Isnantyo, F. D. (2021). A Review of Gamification Impact on Student Behavioral and Learning Outcomes. *International*

Journal of Interactive Mobile Technologies, *15*(21), 22–36. https://doi.org/10.3991/ijim.v15i21.24381

O'Reilly, R. C., & Frank, M. J. (2006). Making Working Memory Work: A Computational Model of Learning in the Prefrontal Cortex and Basal Ganglia. *Neural Comput.*, *18*(2), 283–328. https://doi.org/10.1162/089976606775093909

Oberauer, K., & Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychonomic Bulletin & Review*, *26*(5), 1596–1618. https://doi.org/10.3758/s13423-019-01645-2

Ofosu-Ampong, K. (2020). The Shift to Gamification in Education: A Review on Dominant Issues. *Journal of Educational Technology Systems*, *49*(1), 113–137. https://doi.org/10.1177/0047239520917629

Okonkwo, C. W., & Ade-Ibijola, A. (2021). Chatbots applications in education: A systematic review. *Computers and Education: Artificial Intelligence*, *2*, 100033. https://doi.org/https://doi.org/10.1016/j.caeai.2021.100033

Oliveira, W., Hamari, J., Shi, L., Toda, A. M., Rodrigues, L., Palomino, P. T., & Isotani, S. (2022). Tailored gamification in education: A literature review and future agenda. In *Education and Information Technologies*. Springer US. https://doi.org/10.1007/s10639-022-11122-4

OpenAI. (2022). *ChatGPT: Optimizing language models for dialogue*. https://openai.com/blog/

Orrù, G., Monaro, M., Conversano, C., Gemignani, A., & Sartori, G. (2020). Machine Learning in Psychometrics and Psychological Research. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.02970

Orsini, A., Pezzuti, L., & Picone, L. (2012). Wechsler Intelligence Scale for Children IV Edizione Italiana. In *Giunti O.S.*

Orsoni, M, Benassi, M., & Scutari, M. (2024). Information Theory, Machine Learning, and Bayesian Networks in the Analysis of Dichotomous and Likert Responses for Questionnaire Psychometric Validation. *PsyArXiv*. https://doi.org/10.31234/osf.io/r4y68

Orsoni, M, Giovagnoli, S., Garofalo, S., Mazzoni, N., Spinoso, M., & Benassi, M. (n.d.). *Unlocking Cognitive Patterns: A Comparative Exploration of Linear and Deep Dimensionality Reduction Approaches in Clustering Students' Cognitive Profiles.*

Orsoni, Matteo, Benvenuti, M., Giovagnoli, S., Mazzoni, E., Magri, S., Bartolini, L., Bertani, S., & Benassi, M. (2021). PROFFILO: a new digital assessment tool to evaluate learning difficulties in secondary school. In *BOOK OF ABSTRACTS: 25th Annual International CyberPsychology, CyberTherapy & Social Networking Conference (CYPSY25)* (p. 49).

Orsoni, Matteo, Dubé, A., Prandi, C., Giovagnoli, S., Benassi, M., Mazzoni, E., & Benvenuti, M. (2023). Learning Landscape in Gamification: The Need for a Methodological Protocol in Research Applications. *Perspectives on Psychological Science*, 17456916231202488. https://doi.org/10.1177/17456916231202489

Orsoni, Matteo, Giovagnoli, S., Garofalo, S., Magri, S., Benvenuti, M., Mazzoni, E., & Benassi, M. (2023). Preliminary evidence on machine learning approaches for clusterizing students' cognitive profile. *Heliyon*, *9*(3), e14506. https://doi.org/https://doi.org/10.1016/j.heliyon.2023.e14506

Orsoni, Matteo, Pögelt, A., Duong-Trung, N., Benassi, M., Kravcik, M., & Grüttmüller, M. (2023). Recommending Mathematical Tasks Based on Reinforcement Learning and Item Response Theory. In C. Frasson, P. Mylonas, & C. Troussas (Eds.), *Augmented Intelligence and Intelligent Tutoring Systems* (pp. 16–28). Springer Nature Switzerland.

Ortiz-Rojas, M., Chiluiza, K., & Valcke, M. (2017). Gamification and learning performance: A systematic review of the literature. *Proceedings of the 11th European Conference on Games Based Learning, ECGBL 2017, October*, 515–522.

Ortiz, M., Chiluiza, K., & Valcke, M. (2016). Gamification in Higher Education and Stem: a Systematic Review of Literature. *EDULEARN16 Proceedings*, *1*(August), 6548–6558. https://doi.org/10.21125/edulearn.2016.0422

Ought. (2023). *Elicit: The AI Research Assistant*. https://elicit.org

Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan-a web and mobile app for systematic reviews. *Systematic Reviews*, *5*(1), 210. https://doi.org/10.1186/s13643-016-0384-4

Oyama, K., Hernádi, I., Iijima, T., & Tsutsui, K.-I. (2010). Reward prediction error coding in dorsal striatal neurons. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *30*(34), 11447–11457. https://doi.org/10.1523/JNEUROSCI.1719-10.2010

Özgenel, M. (2018). Modeling the relationships between school administrators' creative and critical thinking dispositions with decision making styles and problem solving skills. *Kuram ve Uygulamada Egitim Bilimleri*, *18*(3), 673–700. https://doi.org/10.12738/estp.2018.3.0068

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *The BMJ*, *372*. https://doi.org/10.1136/bmj.n71

Palacio-Niño, J.-O., & Galiano, F. B. (2019). Evaluation Metrics for Unsupervised Learning Algorithms. *ArXiv*, *abs/1905.0*. https://api.semanticscholar.org/CorpusID:153313022

Palamara, F., Piglione, F., & Piccinini, N. (2011a). Self-Organizing Map and clustering algorithms for the analysis of occupational accident databases. *Safety Science*, *49*(8–9), 1215–1230. https://doi.org/10.1016/j.ssci.2011.04.003

Palamara, F., Piglione, F., & Piccinini, N. (2011b). Self-Organizing Map and clustering algorithms for the analysis of occupational accident databases. *Safety Science*, *49*(8–9), 1215–1230. https://doi.org/10.1016/j.ssci.2011.04.003

Panigrahi, C. . A. (2020). Use of Artificial Intelligence in Education. *The Management Accountant Journal*. https://api.semanticscholar.org/CorpusID:230703163

Papadakis, S., & Kalogiannakis, M. (2018). Evaluating a Course for Teaching Advanced Programming Concepts with Scratch to Preservice Kindergarten Teachers: A Case Study in Greece. *Early Childhood Education*. https://api.semanticscholar.org/CorpusID:70006046

Papert, S. (1980). *Mindstorms: children, computers, and powerful ideas*. Basic Books, Inc.

Papert, S. (1993). *The children's machine: rethinking school in the age of the computer*. Basic Books, Inc.

Park, S., & Kim, S. (2019). A badge design framework for a gamified learning environment: Cases analysis and literature review for badge design. *JMIR Serious Games*, *7*(2), 1–12. https://doi.org/10.2196/14342

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019). PyTorch: An

Imperative Style, High-Performance Deep Learning Library. *Neural Information Processing Systems*. https://api.semanticscholar.org/CorpusID:202786778

Pavithra, M. S., & Parvathi, D. R. (2017). *A Survey on Clustering High Dimensional Data Techniques*. https://api.semanticscholar.org/CorpusID:53486355

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Müller, A., Nothman, J., Louppe, G., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2012). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. http://arxiv.org/abs/1201.0490

Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(8), 1226–1238. https://doi.org/10.1109/TPAMI.2005.159

Perea, G., Navarrete, M., & Araque, A. (2009). Tripartite synapses: astrocytes process and control synaptic information. *Trends in Neurosciences*, *32*(8), 421–431. https://doi.org/10.1016/j.tins.2009.05.001

Perez-Poch, A., Olmedo-Torre, N., Sánchez, F., Salán, N., & López, D. (2016). On the influence of creativity in basic programming learning in a first-year engineering course. *International Journal of Engineering Education*, *32*(5), 2302–2309.

Pham, N. Van, Pham, L. T., Nguyen, T. D., & Ngo, L. T. (2018). A new cluster tendency assessment method for fuzzy co-clustering in hyperspectral image analysis. *Neurocomputing*, *307*, 213–226. https://doi.org/https://doi.org/10.1016/j.neucom.2018.04.022

Piaget, J. (1962). The relation of affectivity to intelligence in the mental development of the child. *Bulletin of the Menninger Clinic*, *26*, 129–137.

Piray, P., Toni, I., & Cools, R. (2016). Human Choice Strategy Varies with Anatomical Projections from Ventromedial Prefrontal Cortex to Medial Striatum. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *36*(10), 2857–2867. https://doi.org/10.1523/JNEUROSCI.2033-15.2016

Poletti, M., Carretta, E., Bonvicini, L., & Giorgi-Rossi, P. (2018). Cognitive Clusters in Specific Learning Disorder. *Journal of Learning Disabilities*, *51*(1), 32–42. https://doi.org/10.1177/0022219416678407

Polson, N. G., & Sokolov, V. O. (2018). Deep Learning. *ArXiv*, *abs/1807.0*. https://api.semanticscholar.org/CorpusID:49902641

Pradeep Kumar Gupta, & Dr. Vibha Sharma. (2017). Working Memory and Learning Disabilities: A Review. *International Journal of Indian Psychology*, *4*(4). https://doi.org/10.25215/0404.013

R Core Team. (2020). *R: A language and environment for statistical computing*. https://www.r-project.org/.

Raffin, A., Hill, A., Gleave, A., Kanervisto, A., Ernestus, M., & Dormann, N. (2021). Stable-Baselines3: Reliable Reinforcement Learning Implementations. *J. Mach. Learn. Res.*, *22*, 268:1-268:8. https://api.semanticscholar.org/CorpusID:246432496

Rao, K. K., Inguva, R., & Rao, A. S. (2020). A Medical AI Agent as a Tool for Neuropsychiatric Diagnoses. *2020 23rd International Symposium on Measurement and Control in Robotics (ISMCR)*, 1–6. https://api.semanticscholar.org/CorpusID:227222043

Rapp, A., Hopfgartner, F., Hamari, J., Linehan, C., & Cena, F. (2019). Strengthening

gamification studies: Current trends and future opportunities of gamification research. *International Journal of Human Computer Studies*, *127*(November 2018), 1–6. https://doi.org/10.1016/j.ijhcs.2018.11.007

Raschka, S., & Mirjalili, V. (2019). *Python Machine Learning* (3rd ed.). Packt Publishing.

Raven, J. (1989a). The Raven Progressive Matrices: A Review of National Norming Studies and Ethnic and Socioeconomic Variation Within the United States. *Journal of Educational Measurement*, *26*(1), 1–16. https://doi.org/10.1111/j.1745-3984.1989.tb00314.x

Raven, J. (1989b). The Raven Progressive Matrices: A Review of National Norming Studies and Ethnic and Socioeconomic Variation Within the United States. *Journal of Educational Measurement*, *26*(1), 1–16. https://doi.org/10.1111/j.1745-3984.1989.tb00314.x

Reed, S., Zhang, Y., Zhang, Y., & Lee, H. (2015). Deep visual analogy-making. *Advances in Neural Information Processing Systems*, *2015-Janua*, 1252–1260.

Rehman, I., Mahabadi, N., Sanvictores, T., & Rehman, C. I. (2023). *Classical Conditioning.*

Reynolds, C. R., & Shaywitz, S. E. (2009). Response to Intervention: Ready or not? Or, from wait-to-fail to watch-them-fail. *School Psychology Quarterly*, *24*(2), 130–145. https://doi.org/10.1037/a0016158

Reynolds, D. (2009). Gaussian Mixture Models. In S. Z. Li & A. Jain (Eds.), *Encyclopedia of Biometrics* (pp. 659–663). Springer US. https://doi.org/10.1007/978-0-387-73003-5_196

Reynolds, D. A. (2018). Gaussian Mixture Models. *Encyclopedia of Biometrics*. https://api.semanticscholar.org/CorpusID:1063711

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the

Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778

Ritzhaupt, A. D., Huang, R., Sommer, M., Zhu, J., Stephen, A., Valle, N., Hampton, J., & Li, J. (2021). A meta-analysis on the influence of gamification in formal educational settings on affective and behavioral outcomes. *Educational Technology Research and Development*, *69*(5), 2493–2522. https://doi.org/10.1007/s11423-021-10036-1

Rolstad, S., Adler, J., & Rydén, A. (2011). Response burden and questionnaire length: is shorter better? A review and meta-analysis. *Value in Health : The Journal of the International Society for Pharmacoeconomics and Outcomes Research*, *14*(8), 1101–1108. https://doi.org/10.1016/j.jval.2011.06.003

Rong, Q., Lian, Q., & Tang, T. (2022). Research on the Influence of AI and VR Technology for Students' Concentration and Creativity. *Frontiers in Psychology*, *13*(March), 1–9. https://doi.org/10.3389/fpsyg.2022.767689

Rosellini, A. J., & Brown, T. A. (2021). Developing and Validating Clinical Questionnaires. *Annual Review of Clinical Psychology*, *17*(1), 55–81. https://doi.org/10.1146/annurev-clinpsy-081219-115343

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386–408. https://doi.org/10.1037/h0042519

Ross, B. C. (2014). Mutual Information between Discrete and Continuous Data Sets. *PLOS ONE*, *9*(2), e87357. https://doi.org/10.1371/journal.pone.0087357

Rothenberg, A. (1979). Einstein's creative thinking and the general theory of relativity: a

documented report. *American Journal of Psychiatry*, *136*(1), 38–43. https://doi.org/10.1176/ajp.136.1.38

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. https://doi.org/https://doi.org/10.1016/0377-0427(87)90125-7

Rozi, F., Rosmansyah, Y., & Dabarsyah, B. (2019). A Systematic Literature Review on Adaptive Gamification: Components, Methods, and Frameworks. *Proceedings of the International Conference on Electrical Engineering and Informatics*, *2019-July*(July), 187–190. https://doi.org/10.1109/ICEEI47359.2019.8988857

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. Parallel Distributed Processing: Exploration in the Microstructure of Cognition, vol. 1. *Foundations*, 318–362.

Rumelhart, D. E., McClelland, J. L., & PDP Research Group, C. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition, Vol. 1: Foundations*. MIT press.

Rummery, G., & Niranjan, M. (1994). On-Line Q-Learning Using Connectionist Systems. *Technical Report CUED/F-INFENG/TR 166*.

Russell, S., & Norving, P. (2021). *Artificial Intelligence: A Modern Approach* (4th ed.). Hoboken: Pearson.

Ryan, R. M., & Deci, E. L. (2000). Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions. *Contemporary Educational Psychology*, *25*(1), 54–67. https://doi.org/10.1006/ceps.1999.1020

Sailer, M., & Homner, L. (2020). The Gamification of Learning: a Meta-analysis.

Educational Psychology Review, *32*(1), 77–112. https://doi.org/10.1007/s10648-019-09498-w

Saleem, A. N., Noori, N. M., & Ozdamli, F. (2022). Gamification Applications in E-learning: A Literature Review. *Technology, Knowledge and Learning*, *27*(1), 139–159. https://doi.org/10.1007/s10758-020-09487-x

Salem, N., & Hussein, S. (2019). Data dimensional reduction and principal components analysis. *Procedia Computer Science*, *163*, 292–299. https://doi.org/https://doi.org/10.1016/j.procs.2019.12.111

Sanmugam, M., Zaid, N. M., Mohamed, H., Abdullah, Z., Aris, B., & Suhadi, S. M. (2015). Gamification as an educational technology tool in engaging and motivating students; An analyses review. *Advanced Science Letters*, *21*(10), 3337–3341. https://doi.org/10.1166/asl.2015.6489

Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D., Lillicrap, T., & Deepmind, G. (2016). *One-shot Learning with Memory-Augmented Neural Networks Google DeepMind*.

Santos, A. C. G., Oliveira, W., Hamari, J., Rodrigues, L., Toda, A. M., Palomino, P. T., & Isotani, S. (2021). The relationship between user types and gamification designs. In *User Modeling and User-Adapted Interaction* (Issue 0123456789). Springer Netherlands. https://doi.org/10.1007/s11257-021-09300-z

Saputro, R. E., Salam, S. B., & Zakaria, M. H. (2017). A review of intrinsic motivation elements in gamified online learning. *Journal of Theoretical and Applied Information Technology*, *95*(19), 4934–4948.

Saxena, M., & Mishra, D. K. (2021). Gamification and gen Z in higher education: A systematic review of literature. *International Journal of Information and Communication*

Technology Education, *17*(4). https://doi.org/10.4018/IJICTE.20211001.oa10

Scanagatta, M., de Campos, C. P., Corani, G., & Zaffalon, M. (2015). Learning Bayesian Networks with Thousands of Variables. In C Cortes, N. Lawrence, D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 28). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2015/file/2b38c2df6a49b97f706ec9148ce48d86-Paper.pdf

Scanagatta, M., Salmerón, A., & Stella, F. (2019). A survey on Bayesian network structure learning from data. *Progress in Artificial Intelligence*, 1–15. https://api.semanticscholar.org/CorpusID:192587394

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal Policy Optimization Algorithms. *ArXiv*, *abs/1707.0*. https://api.semanticscholar.org/CorpusID:28695052

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Openai, O. K. (2017). *Proximal Policy Optimization Algorithms*.

Schuster, V., & Krogh, A. (2021). A Manifold Learning Perspective on Representation Learning: Learning Decoder and Representations without an Encoder. In *Entropy* (Vol. 23, Issue 11). https://doi.org/10.3390/e23111403

Scolari, M., Seidl-Rathkopf, K. N., & Kastner, S. (2015). Functions of the human frontoparietal attention network: Evidence from neuroimaging. *Current Opinion in Behavioral Sciences*, *1*, 32–39. https://doi.org/https://doi.org/10.1016/j.cobeha.2014.08.003

Scutari, M, & Denis, J. B. (2021). *Bayesian Networks: With Examples in R* (C. and Hall/CRC.

(ed.); 2nd ed.). https://doi.org/https://doi.org/10.1201/9780429347436

Scutari, Marco. (2010). Learning Bayesian Networks with the bnlearn R Package. *Journal of Statistical Software*, *35*(3), 1–22. https://doi.org/10.18637/jss.v035.i03

Scutari, Marco, Auconi, P., Caldarelli, G., & Franchi, L. (2017). Bayesian Networks Analysis of Malocclusion Data. *Scientific Reports*, *7*(1), 15236. https://doi.org/10.1038/s41598-017-15293-w

Scutari, Marco, Graafland, C. E., & Gutiérrez, J. M. (2019). Who learns better Bayesian network structures: Accuracy and speed of structure learning algorithms. *International Journal of Approximate Reasoning*, *115*, 235–253. https://doi.org/https://doi.org/10.1016/j.ijar.2019.10.003

Scutari, Marco, & Nagarajan, R. (2013). Identifying significant edges in graphical models of molecular networks. *Artificial Intelligence in Medicine*, *57*(3), 207–217. https://doi.org/https://doi.org/10.1016/j.artmed.2012.12.006

Seaborn, K., & Fels, D. I. (2015). Gamification in theory and action: A survey. *International Journal of Human Computer Studies*, *74*, 14–31. https://doi.org/10.1016/j.ijhcs.2014.09.006

Semyanov, A., & Verkhratsky, A. (2021). Astrocytic processes: from tripartite synapses to the active milieu. *Trends in Neurosciences*, *44*(10), 781–792. https://doi.org/https://doi.org/10.1016/j.tins.2021.07.006

Shamseer, L., Moher, D., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., & Stewart, L. A. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. *BMJ (Clinical Research Ed.)*, *350*, g7647. https://doi.org/10.1136/bmj.g7647

Shute, V. J., Sun, C., & Asbell-clarke, J. (2017). Demystifying computational thinking. *Educational Research Review*, *22*, 142–158. https://doi.org/10.1016/j.edurev.2017.09.003

Smutny, P., & Schreiberova, P. (2020). Chatbots for learning: A review of educational chatbots for the Facebook Messenger. *Computers & Education*, *151*, 103862. https://doi.org/https://doi.org/10.1016/j.compedu.2020.103862

So, H.-J., & Seo, M. (2018). A SYSTEMATIC LITERATURE REVIEW OF GAME-BASED LEARNING AND GAMIFICATION RESEARCH IN ASIA. In *ROUTLEDGE INTERNATIONAL HANDBOOK OF SCHOOLS AND SCHOOLING IN ASIA* (pp. 396–418).

Sobel, M. E. (1997). *Measurement, Causation and Local Independence in Latent Variable Models BT - Latent Variable Modeling and Applications to Causality* (M. Berkane (ed.); pp. 11–28). Springer New York.

Stelnicki, A. M., Nordstokke, D., & Saklofske, D. H. (2015). Who Is the Successful University Student? An Analysis of Personal Resources. *Canadian Journal of Higher Education*, *45*, 214–228. https://api.semanticscholar.org/CorpusID:143399539

Stevens, C., & Bavelier, D. (2012). The role of selective attention on academic foundations: A cognitive neuroscience perspective. *Developmental Cognitive Neuroscience*, *2*, S30–S48. https://doi.org/10.1016/j.dcn.2011.11.001

Stott, A., & Neustaedter, C. (2013). Analysis of Gamification in Education. *Carmster.Com*, 1–8. http://carmster.com/clab/uploads/Main/Stott-Gamification.pdf

Suarez-Alvarez, M. M., Pham, D. T., Prostov, M. Y., & Prostov, Y. I. (2012). Statistical approach to normalization of feature vectors and clustering of mixed datasets.

Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, *468*(2145), 2630–2651. https://doi.org/10.1098/rspa.2011.0704

Subramanian, A., Chitlangia, S., & Baths, V. (2022). Reinforcement learning and its connections with neuroscience and psychology. *Neural Networks*, *145*, 271–287. https://doi.org/https://doi.org/10.1016/j.neunet.2021.10.003

Subramanyam, A., Jyrwa, S., Bansinghani, J. M., Dadhakar, S. J., Dhingra, T. V, Ramchandani, U. R., & Sengupta, S. (2019). *Dyscalculia Detection Using Machine Learning BT - Pattern Recognition and Machine Intelligence* (B. Deka, P. Maji, S. Mitra, D. K. Bhattacharyya, P. K. Bora, & S. K. Pal (eds.); pp. 111–120). Springer International Publishing.

Sul, J. H., Kim, H., Huh, N., Lee, D., & Jung, M. W. (2010). Distinct roles of rodent orbitofrontal and medial prefrontal cortex in decision making. *Neuron*, *66*(3), 449–460. https://doi.org/10.1016/j.neuron.2010.03.033

Sun, L., Hu, L., & Zhou, D. (2022). Programming attitudes predict computational thinking: Analysis of differences in gender and programming experience. *Computers and Education*, *181*(27), 104457. https://doi.org/10.1016/j.compedu.2022.104457

Sun, Y., Kamel, M. S., Wong, A. K. C., & Wang, Y. (2007). Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognition*, *40*(12), 3358–3378. https://doi.org/10.1016/j.patcog.2007.04.009

Surendeleg, G., Murwa, V., Yun, H., & Kim, Y. S. (2014). The role of gamification in education - a literature review. *Contemporary Engineering Sciences*, *7*, 1609–1616. https://doi.org/10.12988/ces.2014.411217

Surr, W. (2018). Student Goal-Setting: An Evidence-Based Practice. *American Institutes for*

Research.

Sutton, R. S., & Barto, A. G. (1998). Reinforcement Learning: An Introduction. *IEEE Transactions on Neural Networks*, *9*(5), 1054. https://doi.org/10.1109/TNN.1998.712192

Świechowski, M., Godlewski, K., Sawicki, B., & Mańdziuk, J. (2023). Monte Carlo Tree Search: a review of recent modifications and applications. *Artificial Intelligence Review*, *56*(3), 2497–2562. https://doi.org/10.1007/s10462-022-10228-y

Taherdoost, H., Sahibuddin, S. Bin, & Jalaliyoon, N. (2014). *Exploratory Factor Analysis ; Concepts and Theory*. https://api.semanticscholar.org/CorpusID:46711710

Takamiya, A., Tazawa, Y., Kudo, K., & Kishimoto, T. (2019). [Artificial Intelligence in Psychiatry]. *Brain and Nerve = Shinkei Kenkyu No Shinpo*, *71 1*, 15–23. https://api.semanticscholar.org/CorpusID:58638513

Tarnavsky, E. A., Smolyansky, E., & Knaan Harpaz, I. (2020). *Connected Papers*. Connected Papers. About. https://www.connectedpapers.com

Tastle, W J, & Wierman, M. J. (2006). An information theoretic measure for the evaluation of ordinal scale data. *Behavior Research Methods*, *38*(3), 487–494. https://doi.org/10.3758/bf03192803

Tastle, William J, & Wierman, M. J. (2007). Consensus and dissention: A measure of ordinal dispersion. *International Journal of Approximate Reasoning*, *45*(3), 531–545. https://doi.org/https://doi.org/10.1016/j.ijar.2006.06.024

Team, J. (2022). *JASP (Version 0.16.1)*.

Teng, E., Becker, B. W., Woo, E., Knopman, D. S., Cummings, J. L., & Lu, P. H. (2010). Utility of the functional activities questionnaire for distinguishing mild cognitive impairment from very mild Alzheimer disease. *Alzheimer Disease and Associated*

Disorders, *24*(4), 348–353. https://doi.org/10.1097/WAD.0b013e3181e2fc84

Thorndike, E. L. (1898). Animal intelligence: An experimental study of the associative processes in animals. *The Psychological Review: Monograph Supplements*, *2*(4), i–109. https://doi.org/10.1037/H0092987

Thorndike, R. L. (1953). Who belongs in the family? *Psychometrika*, *18*(4), 267–276. https://doi.org/10.1007/BF02289263

Touré-Tillery, M., & Fishbach, A. (2014). How to Measure Motivation: A Guide for the Experimental Social Psychologist. *Social and Personality Psychology Compass*, *8*(7), 328–341. https://doi.org/10.1111/spc3.12110

Tracy, L. (1990). Treating factor interpretations as hypotheses. *Social Behavior and Personality*, *18*, 309–326. https://api.semanticscholar.org/CorpusID:145739241

Trognon, A., Cherifi, Y. I., Habibi, I., Demange, L., & Prudent, C. (2022). Using machine-learning strategies to solve psychometric problems. *Scientific Reports*, *12*(1), 18922. https://doi.org/10.1038/s41598-022-23678-9

Trung, B. D., Son, N. T., Tung, N. D., Son, K. A., Anh, B. N., & Lam, P. T. (2023). Educational Data Mining: A Systematic Review on the Applications of Classical Methods and Deep Learning Until 2022. *2023 IEEE Symposium on Industrial Electronics & Applications (ISIEA)*, 1–15. https://doi.org/10.1109/ISIEA58478.2023.10212273

Tsamardinos, I., Brown, L. E., & Aliferis, C. F. (2006). The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, *65*(1), 31–78. https://doi.org/10.1007/s10994-006-6889-7

Tsutsui, K.-I., Grabenhorst, F., Kobayashi, S., & Schultz, W. (2016). A dynamic code for

economic object valuation in prefrontal cortex neurons. *Nature Communications*, *7*, 12554. https://doi.org/10.1038/ncomms12554

Tu, C.-H., Yen, C.-J., Sujo-Montes, L., & Roberts, G. A. (2015). Gaming personality and game dynamics in online discussion instructions. *Educational Media International*, *52*(3), 155–172. https://doi.org/10.1080/09523987.2015.1075099

Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind*, 433–460.

Tvarožek, J., Kravčík, M., & Bieliková, M. (2008). *Towards Computerized Adaptive Assessment Based on Structured Tasks BT - Adaptive Hypermedia and Adaptive Web-Based Systems* (W. Nejdl, J. Kay, P. Pu, & E. Herder (eds.); pp. 224–234). Springer Berlin Heidelberg.

U.K. Department for Education. (2013). The National Curriculum in England: Framework document. *The Stationery Office.*, *September*.

Uddin, M., Wang, Y., & Woodbury-Smith, M. (2019). Artificial intelligence for precision medicine in neurodevelopmental disorders. *Npj Digital Medicine*, *2*(1), 112. https://doi.org/10.1038/s41746-019-0191-0

Ultsch, A., & Herrmann, L. (2007). The architecture of emergent self-organizing maps to reduce projection errors. *ESANN 2005 Proceedings - 13th European Symposium on Artificial Neural Networks*, 1–6.

van Laar, E., van Deursen, A. J. A. M., van Dijk, J. A. G. M., & de Haan, J. (2019). Determinants of 21st-century digital skills: A large-scale survey among working professionals. *Computers in Human Behavior*, *100*, 93–104. https://doi.org/https://doi.org/10.1016/j.chb.2019.06.017

Van Rossun, G., & Drake, F. . (2009). *Python 3 Reference Manual*.

Vandenbroucke, L., Verschueren, K., & Baeyens, D. (2017). The development of executive functioning across the transition to first grade and its predictive value for academic achievement. *Learning and Instruction*, *49*, 103–112. https://doi.org/https://doi.org/10.1016/j.learninstruc.2016.12.008

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, \Lukasz, & Polosukhin, I. (2017). Attention is All You Need. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 6000–6010.

Verkhratsky, A., & Butt, A. (2007). Astrocytes. In *Glial Neurobiology* (pp. 93–123). https://doi.org/https://doi.org/10.1002/9780470517796.ch7

Vicari, S. (2007). *PROMEA. Prove di Memoria e Apprendimento per l'Età Evolutiva*. Giunti O.S.

Visalakshi, N. K., & Thangavel, K. (2009). Impact of normalization in distributed K-means clustering. In *International Journal of Soft Computing* (Vol. 4, Issue 4, pp. 168–172).

Vock, M., Preckel, F., & Holling, H. (2011). Mental abilities and school achievement: A test of a mediation hypothesis. *Intelligence*, *39*(5), 357–369. https://doi.org/10.1016/j.intell.2011.06.006

Vygotskij, L. (1978). *Mind in society: The development of higher psychological processes*. Harvard university Press.

Walesiak, M., & Dudek, A. (2020). The Choice of Variable Normalization Method in Cluster Analysis. *Education Excellence and Innovation Management: A 2025 Vision to Sustain Economic Development During Global Challenges*, *June*, 325–340.

Wang, L., Nguyen, U. T. V, Bezdek, J. C., Leckie, C. A., & Ramamohanarao, K. (2010). *iVAT and aVAT: Enhanced Visual Analysis for Cluster Tendency Assessment BT* -

Advances in Knowledge Discovery and Data Mining (M. J. Zaki, J. X. Yu, B. Ravindran, & V. Pudi (eds.); pp. 16–27). Springer Berlin Heidelberg.

Wang, W., Song, L., Wang, T., Gao, P., & Xiong, J. (2020). A Note on the Relationship of the Shannon Entropy Procedure and the Jensen–Shannon Divergence in Cognitive Diagnostic Computerized Adaptive Testing. *SAGE Open*, *10*(1), 2158244019899046. https://doi.org/10.1177/2158244019899046

Wanick, V., & Bui, H. T. M. (2019). Gamification in Management: a systematic review and research directions. *Int. J. Serious Games*, *6*, 57–74. https://api.semanticscholar.org/CorpusID:195856942

Watkins, C. J. C. H., & Dayan, P. (1992). Q-learning. *Machine Learning*, *8*(3), 279–292. https://doi.org/10.1007/BF00992698

Watkins, M. W. (2018). Exploratory Factor Analysis: A Guide to Best Practice. *Journal of Black Psychology*, *44*(3), 219–246. https://doi.org/10.1177/0095798418771807

Webster, W. R. (2002). *Metacognition and the autonomous learner: student reflections on cognitive profiles and learning environment development*. https://api.semanticscholar.org/CorpusID:353240

Werbach, K. (2014). (Re)Defining Gamification: A Process Approach. *International Conference on Persuasive Technology*. https://api.semanticscholar.org/CorpusID:44295071

Wierman, M. J., & Tastle, W. J. (2005). Consensus and dissention: theory and properties. *NAFIPS 2005 - 2005 Annual Meeting of the North American Fuzzy Information Processing Society*, 75–79. https://doi.org/10.1109/NAFIPS.2005.1548511

Willert, N. (2021). A systematic literature review of gameful feedback in computer science

education. *International Journal of Information and Education Technology*, *11*(10), 464–470. https://doi.org/10.18178/ijiet.2021.11.10.1551

Williams, B., Onsman, A., & Brown, T. (2010). Exploratory Factor Analysis: A Five-Step Guide for Novices. *Australian Journal of Paramedicine*, *8*, 1–13. https://api.semanticscholar.org/CorpusID:54541939

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, *8*(3), 229–256. https://doi.org/10.1007/BF00992696

Willingham, D. T. (2019). How to teach Critical Thinking. *Education: Future Frontiers*.

Wilson, D., Calongne, C., & Henderson, B. (2015). Gamification Challenges and a Case Study in Online Learning. *Internet Learning*, *4*(2). https://doi.org/10.18278/il.4.2.7

Wing, J. M. (2006). Computational thinking. *Commun. ACM*, *49*(3), 33–35. https://doi.org/10.1145/1118178.1118215

Wing, J. M. (2010). *Computational Thinking: What and Why?* https://api.semanticscholar.org/CorpusID:63382972

Wing, J. M. (2017). Computational thinking's influence on research and education for all. *Journal on Educational Technology*, *25*, 7–14. https://api.semanticscholar.org/CorpusID:64533042

Wingström, R., Hautala, J., & Lundman, R. (2022). Redefining Creativity in the Era of AI? Perspectives of Computer Scientists and New Media Artists. *Creativity Research Journal*, *00*(00), 1–17. https://doi.org/10.1080/10400419.2022.2107850

Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, *2*(1), 37–52.

https://doi.org/https://doi.org/10.1016/0169-7439(87)80084-9

Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample Size Requirements for Structural Equation Models: An Evaluation of Power, Bias, and Solution Propriety. *Educational and Psychological Measurement*, *76*(6), 913–934. https://doi.org/10.1177/0013164413495237

Wrigley, C. (1958). Objectivity in Factor Analysis. *Educational and Psychological Measurement*, *18*(3), 463–476. https://doi.org/10.1177/001316445801800303

Xu, J., Lio, A., Dhaliwal, H., Andrei, S., Balakrishnan, S., Nagani, U., & Samadder, S. (2021). Psychological interventions of virtual gamification within academic intrinsic motivation: A systematic review. *Journal of Affective Disorders*, *293*(July), 444–465. https://doi.org/10.1016/j.jad.2021.06.070

Xu, X., Zhao, Z., Li, R., & Zhang, H. (2018). Brain-Inspired Stigmergy Learning. *IEEE Access*, *7*, 54410–54424. https://api.semanticscholar.org/CorpusID:53742342

Xue, S., Wu, M., Kolen, J., Aghdaie, N., & Zaman, K. A. (2017). Dynamic Difficulty Adjustment for Maximized Engagement in Digital Games. *Proceedings of the 26th International Conference on World Wide Web Companion*, 465–471. https://doi.org/10.1145/3041021.3054170

Yadav, A., Mayfield, C., Zhou, N., Hambrusch, S., & Korb, J. T. (2014). Computational Thinking in Elementary and Secondary Teacher Education. *ACM Trans. Comput. Educ.*, *14*(1). https://doi.org/10.1145/2576872

Yağcı, M. (2022). Educational data mining: prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments*, *9*(1), 11. https://doi.org/10.1186/s40561-022-00192-z

Yamakawa, H. (2020). Attentional Reinforcement Learning in the Brain. *New Generation Computing*, *38*(1), 49–64. https://doi.org/10.1007/s00354-019-00081-z

Yan, J., Ma, M., & Yu, Z. (2023). bmVAE: a variational autoencoder method for clustering single-cell mutation data. *Bioinformatics*, *39*(1), btac790. https://doi.org/10.1093/bioinformatics/btac790

Yeung, R. W. (2008). *Information Theory and Network Coding*. https://api.semanticscholar.org/CorpusID:438461

Yigit, H. D., Sorrel, M. A., & de la Torre, J. (2018). Computerized Adaptive Testing for Cognitively Based Multiple-Choice Data. *Applied Psychological Measurement*, *43*(5), 388–401. https://doi.org/10.1177/0146621618798665

Yıldırım, İ., & Şen, S. (2019). The effects of gamification on students' academic achievement: a meta-analysis study. *Interactive Learning Environments*, *29*(8), 1301–1318. https://doi.org/10.1080/10494820.2019.1636089

Yokota, S., Takeuchi, H., Hashimoto, T., Hashizume, H., Asano, K., Asano, M., Sassa, Y., Taki, Y., & Kawashima, R. (2015). Individual differences in cognitive performance and brain structure in typically developing children. *Developmental Cognitive Neuroscience*, *14*, 1–7. https://doi.org/10.1016/j.dcn.2015.05.003

Zainuddin, Z., Chu, S. K. W., Shujahat, M., & Perera, C. J. (2020). The impact of gamification on learning and instruction: A systematic review of empirical evidence. *Educational Research Review*, *30*(March). https://doi.org/10.1016/j.edurev.2020.100326

Zhang, H., & Schuster, T. (2021). A methodological review protocol of the use of Bayesian factor analysis in primary care research. *Systematic Reviews*, *10*(1), 15. https://doi.org/10.1186/s13643-020-01565-6

Zhang, K., & Aslan, A. B. (2021). AI technologies for education: Recent research & future directions. *Computers and Education: Artificial Intelligence*, *2*, 100025. https://doi.org/https://doi.org/10.1016/j.caeai.2021.100025

Zhang, P., & Bai, G. (2005). An Activity Systems Theory Approach to Agent Technology. *International Journal of Knowledge and Systems Science*, *2*, 60–65. https://api.semanticscholar.org/CorpusID:17740300

Zhang, Q., & Yu, Z. (2022). Meta-Analysis on Investigating and Comparing the Effects on Learning Achievement and Motivation for Gamification and Game-Based Learning. *Education Research International*, *2022*. https://doi.org/10.1155/2022/1519880

Zhang, Y., & Goh, W.-B. (2021). Personalized task difficulty adaptation based on reinforcement learning. *User Modeling and User-Adapted Interaction*, *31*(4), 753–784. https://doi.org/10.1007/s11257-021-09292-w

Zhang, Z., Li, G., Xu, Y., & Tang, X. (2021). Application of Artificial Intelligence in the MRI Classification Task of Human Brain Neurological and Psychiatric Diseases: A Scoping Review. *Diagnostics (Basel, Switzerland)*, *11*(8). https://doi.org/10.3390/diagnostics11081402

Zhao, B., Dong, X., Guo, Y., Jia, X., & Huang, Y. (2021). PCA Dimensionality Reduction Method for Image Classification. *Neural Processing Letters*, *54*, 347–368. https://api.semanticscholar.org/CorpusID:240083675

Zhao, F., Zeng, Y., Wang, G., Bai, J., & Xu, B. (2018). A Brain-Inspired Decision Making Model Based on Top-Down Biasing of Prefrontal Cortex to Basal Ganglia and Its Application in Autonomous UAV Explorations. *Cognitive Computation*, *10*, 296–306. https://api.semanticscholar.org/CorpusID:4912990