ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 2 - 2026

# Design of a Visual Thinking Training System via the Synergistic Integration of Cognitive Diagnostic Models and Generative Adversarial Networks

*Zixuan Jing*[1] *and Boon Keat Ooi*[2,*]

[1] Shazhou Professional Institute of Technology, Zhangjiagang City, Jiangsu Province, 215600, China

[2] Graduate School of Management, Postgraduate Centre, Management and Science University, University Drive, Off Persiaran Olahraga, 40100 Shah Alam, Selangor, Malaysia

**SUMMARY:** *Visual thinking represents a multifaceted cognitive ability critical to learning in scientific, engineering, design, and data-centric disciplines. Existing online learning platforms measure visual thinking using overall accuracy indices without providing insights into the fine-grained cognitive factors contributing to learners' learning problems. This paper proposes a visual thinking training platform that integrates Cognitive Diagnostic Modeling (CDM) and Generative Adversarial Network (GAN) technologies to solve this problem. It disaggregates visual thinking into five diagnosable cognitive factors—visual perception, spatial relationship understanding, pattern abstraction, visual inference, and representation transformation—and matches training tasks to cognitive attributes using a Q matrix. The neural network-based CDM model provides individual-level attribute proficiency profiles, and the conditional GAN produces and supplements visual training tasks according to diagnosed weak cognitive abilities constrained by attribute labels, difficulty levels, and domain experts' evaluation. A quasi-experiment was carried out among 124 college students in six weeks. The findings suggest that the integrated CDM-GAN framework significantly outperforms the traditional fixed-task-based approach on all cognitive factors ($p < .001$; partial eta-squared between .22 and .31), with enhanced transfer task performance ($F = 42.67$, $p < .001$; partial eta-squared = .29) and lower cognitive load (effect size = 0.79). Domain expert evaluations indicated the acceptability of the quality of generated tasks concerning attribute matching and instructional value. Overall, our research demonstrates the significant advantages of an evidence-based adaptive platform supported by automatic task generation in visual thinking training.*

***KEYWORDS:*** *visual thinking; cognitive diagnostic models; generative adversarial networks; adaptive learning systems; Q-matrix*

## 1 Introduction

Visual thinking constitutes a multidimensional cognitive ability that underlies competent performance in a broad range of academic and professional domains, including science, engineering, mathematics, design, data analysis, and other increasingly visual and multimodal learning contexts. Recent reviews of artificial intelligence in education have shown that intelligent learning environments are moving from general content delivery toward personalized diagnosis, adaptive intervention, learning analytics, and AI-supported feedback

*zx.jing@outlook.com
https://doi.org/10.65102/is20261029 Published: 30 April, 2026

[1, 2]. In this broader shift, visual thinking has become especially important because learners are required to perceive salient visual features, recognize spatial relationships among structural components, abstract recurring patterns from visual stimuli, draw inferences from diagrammatic and graphical evidence, and transform visual representations into verbal, symbolic, or formal structures. Research on AI-supported complex problem solving also indicates that cognitive, metacognitive, affective, and representational processes need to be supported together rather than treated as isolated outcomes [3]. Empirical work on visual thinking and cooperative learning further suggests that visual thinking tools can improve competence acquisition when they are embedded in structured learning, assessment, and reflection activities [4]. However, visual thinking is not a unitary skill. Its sub-processes are cognitively distinct and may develop unevenly across learners. Consequently, a learner may demonstrate proficiency in visual perception while simultaneously exhibiting systematic difficulty in visual inference or representational transformation. This heterogeneity renders aggregate performance metrics, such as total scores or task completion rates, insufficient for guiding precise instructional intervention.

Despite growing interest in intelligent and adaptive educational systems, many digital visual thinking platforms continue to rely on outcome-based evaluation paradigms. Such systems usually record whether a learner provides a correct response to a visual task, but they rarely decompose performance into interpretable cognitive dimensions or provide evidence-based feedback specifying the nature of the learner's difficulty. This limitation is consequential because adaptive learning depends not only on selecting different tasks, but also on understanding why a learner fails and which cognitive attribute should be targeted next. Visual learning analytics research has similarly emphasized that visual representations of learner data are useful only when they support interpretation, explanation, and pedagogical decision-making [5]. Recent experimental evidence further shows that learners often need scaffolding to interpret complex visual learning analytics, and that generative AI agents can enhance such comprehension when they provide proactive guidance rather than passive responses [6]. These findings imply that effective visual thinking training should combine diagnostic interpretability, adaptive task sequencing, and learner-facing feedback. However, current systems often lack a mechanism for connecting these components in a coherent instructional loop.

Cognitive Diagnostic Models (CDMs) offer a principled psychometric framework for addressing this diagnostic gap. Unlike classical test theory or unidimensional item response theory, CDMs classify learners' mastery status across multiple discrete cognitive attributes required to complete a task. In a CDM-based architecture, each assessment item is linked to one or more cognitive attributes through a Q-matrix, and the model estimates the probability that a given learner has mastered each attribute conditional on their observed response pattern. Recent research has extended CDMs through neural networks, relation-aware modeling, hierarchical Bayesian structures, context-aware feature modeling, and knowledge-sensed representation learning [7-12]. For example, relation map-driven cognitive diagnosis models explicitly model student-exercise-concept interactions [8], while hierarchical cognitive diagnosis frameworks capture latent dependencies among knowledge concepts and cognitive states [10]. NeuralCD further provides a general neural cognitive diagnosis framework that improves modeling flexibility while retaining diagnostic interpretation [12]. These developments suggest that CDMs are well suited to support fine-grained learner profiling in intelligent education. When applied to visual thinking assessment, CDMs can identify whether a learner's errors reflect deficits in spatial relation recognition, pattern abstraction, visual inference, or other specific sub-processes rather than merely signaling an incorrect answer.

Parallel to advances in cognitive diagnosis, generative artificial intelligence has introduced new possibilities for educational data augmentation, content generation, adaptive feedback, and

learner modeling. Recent studies have noted that large language models and generative AI can support educational tasks such as feedback generation, learner profiling, recommendation, and learning analytics, while also raising concerns regarding transparency, privacy, reliability, and pedagogical validity [13, 14]. Learning analytics research has further argued that generative AI should be understood through the learning analytics cycle, where data collection, analysis, interpretation, feedback, and action are connected in a closed loop [15-17]. However, generative models should not be introduced into educational systems as unconstrained content generators. Instead, they should be embedded in human-centered and pedagogically governed workflows that ensure generated outputs are aligned with learning goals, learner needs, and ethical requirements. This point is particularly relevant for visual thinking training because generated visual tasks must not only look coherent, but also reliably engage the intended cognitive attributes and maintain stable task difficulty.

Generative Adversarial Networks (GANs) provide a specific generative mechanism that is relevant to this study. GANs consist of a generator trained to synthesize data resembling real samples and a discriminator trained to distinguish real from synthetic instances. Recent reviews have documented the continuing importance of GANs for synthetic data generation, data augmentation, and high-fidelity content synthesis across domains [18-21]. In educational technology research, GAN-based synthetic data generation has been explored as a way to address limited datasets, privacy restrictions, data imbalance, and model evaluation challenges. Bethencourt-Aguilar et al. demonstrated the use of GANs in educational technology research and examined the equivalence between synthetic and real educational data [22]. Vie et al. proposed privacy-preserving synthetic educational data generation for learning traces [23], while later studies further examined synthetic data generators for learning analytics and student outcome prediction [24, 25]. Recent work on privacy-preserving synthetic educational data also emphasizes that synthetic data must be evaluated jointly in terms of utility, privacy, and fairness rather than treated as automatically safe or valid [26]. These studies provide a methodological foundation for GAN-assisted educational data and task augmentation. Nevertheless, their direct application to visual thinking training remains underdeveloped, especially when generated tasks must be constrained by cognitive attributes, task difficulty, and expert validation.

Despite these developments, three limitations remain in current research. First, visual thinking training systems often lack a fine-grained attribute model that can diagnose learners' visual-cognitive strengths and weaknesses. Second, CDM-based adaptive learning research has mainly focused on subject knowledge diagnosis, computerized adaptive testing, and knowledge tracing, while its application to visual-cognitive training remains limited [7-12], [27]. Third, GAN-based educational research has primarily emphasized data augmentation, privacy preservation, or outcome prediction, but has rarely linked generated educational tasks directly to CDM-derived diagnostic profiles [22-26]. As a result, current systems may either diagnose learners without generating sufficient personalized training resources, or generate learning resources without knowing which cognitive weakness they are intended to address. This separation limits the educational value of both cognitive diagnosis and generative modeling.

This study addresses this gap by proposing and empirically evaluating a visual thinking training system that synergistically integrates Cognitive Diagnostic Models and Generative Adversarial Networks. The proposed system decomposes visual thinking into five cognitively grounded attributes: visual perception (VP), spatial relation recognition (SRR), pattern abstraction (PA), visual inference (VI), and representational transformation (RT). Each training task is linked to the required attributes through a Q-matrix. A neural CDM module estimates learner-specific attribute mastery profiles, and these profiles are then used to condition a GAN-based generation module that produces or augments visual tasks targeting diagnosed

weaknesses. The design also adopts a human-in-the-loop quality-control mechanism, including expert review, difficulty calibration, attribute alignment checks, and task-pool validation. This design is consistent with recent calls for interpretable, human-centered, and ethically governed AI systems in education [26-29]. The system operates through a four-stage adaptive loop: pre-test diagnosis, diagnosis-driven training with generated task augmentation, immediate feedback delivery, and post-training re-diagnosis.

The contribution of this study is not the development of a novel CDM algorithm or a new GAN architecture. Rather, it lies in the system-level design logic of a diagnosis-driven and generation-supported training architecture and its empirical validation in a visual thinking training context. Specifically, this study makes four contributions. First, it constructs and validates a visual thinking cognitive attribute model suitable for CDM-based assessment. Second, it designs and implements a CDM-GAN integrated training system architecture. Third, it proposes a diagnosis-to-generation adaptive loop that links interpretable learner profiles with constrained task generation. Fourth, it provides multi-dimensional empirical evidence on diagnostic performance, generated task quality, learning effectiveness, transfer performance, cognitive load, engagement, and learner experience. In doing so, this study responds to the need for AI-supported educational systems that are not only technically adaptive, but also diagnostically interpretable, pedagogically controlled, and empirically validated.

# 2 Methodology

## 2.1 Research Design

### 2.1.1 Design-Based Research Approach

The present work follows a DBR methodology comprising four phases, namely: design and prototyping of the system; evaluation by experts and modifications; piloting; and experimental assessment. In the process of design, the visual thinking attributes model, Q-matrix, configuration of the CDM, training pipeline for GANs, and recommendation algorithm were defined in parallel. The expert panel verified the attributes' validity, feasibility of task-attributes mapping, and consistency of the system architecture. Data from piloting with 18 participants during two weeks provided the information to initialize CDM parameters, adjust conditioning vectors in the GANs, and define the thresholds for the adaptive component.

### 2.1.2 Quasi-Experimental Design

The quasi-experimental research design adopted a pretest-posttest approach with a nonequivalent control group. Subjects were assigned either to the experimental group (EG; N = 63), where the CDM-GAN combined training system was implemented, or to the control group (CG; N = 61), where a traditional fixed-task visual training system was used. Both systems had similar designs in terms of interface layout, content database, and training time span. Random assignment could not be conducted due to practical limitations, but subjects belonged to equivalent pretest visual thinking ability levels and demographic backgrounds. After six weeks of training, a posttest immediately took place, followed by a delayed test after four weeks.

### 2.1.3 Research Context and Participants

The participants were 124 undergraduate students who were studying in STEM-related subjects, educational technology, and design at a comprehensive research university. The inclusion criteria were that the participants should not have any previous experience with courses on

visual thinking and adaptive learning systems. The demographic data and pre-test equivalency indices are shown in Table 1 below. All variables showed pre-test equivalency ($p > .05$).

*Table 1: Participant Demographic Characteristics and Pre-test Equivalence*

| Variable | Category | EG (n = 63) | CG (n = 61) | p |
|---|---|---|---|---|
| Age (years) | M ± SD | 21.34 ± 1.82 | 21.17 ± 1.94 | .673 |
| Gender | Female (%) | 52.4% | 49.2% | .718 |
| Discipline | STEM / Design | 38 / 25 | 37 / 24 | .931 |
| Pre-test VT Score | M ± SD | 48.61 ± 6.74 | 48.29 ± 7.03 | .794 |
| Digital Tool Exp. (yrs) | M ± SD | 3.21 ± 1.14 | 3.08 ± 1.27 | .552 |

Note. EG = Experimental Group; CG = Control Group. Pre-test VT Score is on a 100-point scale. Digital Tool Experience is self-reported years of use.

## 2.2 Visual Thinking Attribute Construction

### 2.2.1 Attribute Identification

Visual Thinking was defined by a set of five cognitive traits based on theories of spatial cognition, visual literacy, and diagrammatic reasoning. These five cognitive traits include: (A1) Visual Perception (VP) which involves the detection and discrimination of salient features in complex stimuli; (A2) Spatial Relation Recognition (SRR), which refers to the comprehension of relations between visual objects in 2D and 3D structures; (A3) Pattern Abstraction (PA), involving generalizing and identifying structure in visual sequences and arrays; (A4) Visual Inference (VI), meaning inferring valid logic or analogical conclusions from visual information; and (A5) Representational Transformation (RT), which denotes the conversion between visual representation and other modes of representation like language, symbolic notation, schema, and numbers. Definitions for these five cognitive traits have been provided in Table 2.

*Table 2: Visual Thinking Cognitive Attribute Definitions*

| Attr. ID | Cognitive Attribute | Abbrev. | Operational Definition |
|---|---|---|---|
| A1 | Visual Perception | VP | Identification and discrimination of salient visual features within complex stimuli |
| A2 | Spatial Relation Recognition | SRR | Interpretation of structural relations among visual objects in 2D/3D configurations |
| A3 | Pattern Abstraction | PA | Generalization of structural rules and schemas from visual sequences or arrays |
| A4 | Visual Inference | VI | Drawing of valid logical or analogical conclusions from visual evidence |
| A5 | Representational Transformation | RT | Translation between visual and non-visual representational formats |

Note. Attributes were operationalized through structured specification protocols including cognitive requirements, example stimuli, expected learner operations, and prototypical error patterns.

### 2.2.2 Task-Attribute Mapping via Q-Matrix

Each visual training task was mapped to one or more cognitive attributes through a binary Q-matrix $\mathbf{Q} \in \{0,1\}^{J\times K}, \quad q_{ji} \in \{0,1\}$, where $q_{ji} = 1$ if attribute i is required by task j and $q_{ji} = 0$ otherwise. Formally:

$$\mathbf{Q} \in \{0,1\}^{J\times K}, \quad q_{ji} \in \{0,1\}, \quad \sum_{i=1}^{K} q_{ji} \geq 1 \quad \forall j \in \{1,\ldots,J\}$$

where J = 56 tasks and K = 5 attributes. The construction of the Q-matrix involved an extensive process of expert consensus. Six domain experts individually classified all items, and their consistency was calculated using Fleiss' $\kappa = 0.81$ (95% CI [0.74, 0.88]), representing substantial agreement. Any disagreements were sorted out using structured discussion sessions until consensus was achieved. In order to avoid floor and ceiling effects in the diagnosis process, the number of attributes per item was restricted from one to four.

### 2.2.3 Expert Validation of the Attribute Model

The assessment panel comprised twelve experts in educational technology, cognitive psychology, visual design, and subject-specific teaching. Experts used an objective rating procedure based on a scale of five for measuring clarity, discriminability, educational significance, and diagnostic tractability of attributes. Ratings by experts for all four criteria were more than 4.2 (SD < 0.6). All attributes passed the assessment and did not require exclusion from the set of attributes. Only slight modifications of the boundaries of A2 and A5 were made.

## 2.3 Data Collection

The three data sources included performance metrics, behavioral process data, and self-reporting scales. Performance metrics involved per-task correctness (true/false), response times in milliseconds, attempts before achieving the right answer, hint use frequency, partial completeness, and scores based on task difficulty (scaled from 0 to 10). Behavioral process data involved time-stamped click events (granularity: 100 ms), task sequences, revisions made, and time spent on each sub-task. Post-test, participants completed validated self-report scales measuring NASA Task Load Index (NASA-TLX; six subscales, $\alpha = 0.83$), representing cognitive workload (0-100 point composite); behavioral and emotional engagement (12-item scale, $\alpha = 0.87$, 5-point Likert); System Usability Scale (SUS; 10 item scale, 0-100 point scale); Technology Acceptance Model perceived usefulness scale (6-item subscale, $\alpha = 0.89$); and Learner Satisfaction Scale (eight item questionnaire, $\alpha = 0.91$). A semi-structured interview was conducted after the intervention with a purposeful sample of 24 participants (12 each in both groups).

## 2.4 Data Analysis Strategy

CDMs were chosen through comparison of DINA, G-DINA, GDINA-NPC, and Neural CDMs based on criteria like AIC, BIC, RMSEA, $G^2$, and 10-fold cross-validation accuracy of classification of attributes. The posterior probability of the mastery of attribute $k$ by learner $n$ is as shown in Equation (1):

$$\hat{\alpha}_{nk} = P(\alpha_{nk} = 1 \mid \mathbf{X}_n, \mathbf{Q}, \hat{\theta}), \quad n \in \{1,\ldots,N\}, \quad k \in \{1,\ldots,K\} \tag{1}$$

where $\mathbf{X}_n$ is the response vector and $\hat{\theta}$ the estimated CDM parameter set. For the statistical analysis of learning outcomes, Analysis of Covariance (ANCOVA) was used with the pre-test results as the covariates. The design adopted in modeling the longitudinal pattern of attribute mastery included a mixed-factorial ANOVA design, where group represented the between-subject factor and the training session the within-subject factor. Paired sample t-test was used

to assess the differences between pre- and post-conditions within groups. Partial $\eta^2$ and Cohen's d were calculated as effect sizes in between-group and pairwise comparisons, respectively. Thematic analysis with constant comparison was used to code the interviews. Interrater reliability was measured by Cohen's $\kappa = 0.78$ on 20%.

# 3 System Design

## 3.1 Overall System Architecture

The CDM-GAN combined visual thinking training system consists of six functional layers which together form a closed adaptive teaching loop. Every layer has a clear list of responsibilities and interfaces with other layers via defined data interfaces, and thus it can be deployed modularly and information flow can be interpreted.

Learner Interaction Layer is the frontend interface of the system, which is delivered as a browser-based application that can be accessed on both desktop and tablet computers. It shows the visual thinking problems in succession during timed intervals, stores binary and partial answers, logs hints, and displays feedback panels after responses (the source of feedback and recommendation layer). Navigation controls enable students to re-examine finished tasks without editing their replies to ensure that the response history remains intact. To ensure that adaptive task selection could be separated in terms of interface-related usability differences, the interface layout, colour scheme, typography, and navigation structure remained constant in experimental and control conditions.

The Data Acquisition Layer runs alongside the learner interaction layer, where all interaction events are recorded into a PostgreSQL relational database with 100 millisecond temporal resolution using a persistent WebSocket connection. Per-task response accuracy (binary), response latency (milliseconds between stimulus onset and response submission), attempt count before a correct or final response, frequency and timing of hints requested, partial completion indicators, and sequences of clickstream events (including navigation paths, revision actions, and time spent on different sub-stages of tasks) were recorded. At the end of every 45 minutes of training, the data acquisition layer will collect the response vectors of the session of each learner and send them to the cognitive diagnosis layer to estimate CDM again.

After each session, the Cognitive Diagnosis Layer takes the updated cumulative response vector of each learner as input and runs the neural CDM inference process to generate an updated posterior attribute mastery probability vector. The vector is saved in a learner-model database along with the session index and timestamp, offering a longitudinal history of each learner cognitive profile during the intervention. The new mastery vector is passed to the feedback and recommendation layer that computes the task priority score and regenerates the personalized task queue of the next session. A change in the composition of a diagnosed weak-attribute set of a learner due to a mastery update leads to the system marking such an incident and initiating reassessment of the task pool concerning the affected attributes.

The Generative Task Layer is activated by a particular triggering scenario: when the effective deployment pool of one or more diagnosed weak attributes has less than five tasks that are eligible to be completed by the learner. In such a situation, the layer triggers the trained conditional GAN with the specification of the desired attribute profile, difficulty level, and task type as conditioning inputs. The GAN produces a set of 10 candidate tasks per invocation, which are automatically structurally validated and then sent to the expert review layer. All tasks that fail structural validation are rejected without expert review. All tasks that pass the structural validation go into an asynchronous review queue and once endorsed by an expert will be added to the deployment pool and are instantly available to be adaptively recommended.

The Feedback and Recommendation Layer has two parallel roles. To learners, it converts the probability vectors of mastering an attribute into graphical radar chart summaries that are shown at the end of every session with natural-language messages at the attribute level indicating what areas of visual thinking competency are improving or are still under the mastery criteria. To teachers, it keeps up a live dashboard showing classwide distributions of probabilities of attribute mastery, individual learner profiles, and session completion rates. The next session tasks to be recommended are calculated by the layer based on the priority scoring function given in the section 3.4 and saved in the personalised task queue until the next time the learner logs in.

The Expert Review Layer offers an easy to use asynchronous quality assurance interface whereby domain reviewers assess GAN-created candidate tasks on four aspects, including structural integrity, attribute alignment, difficulty appropriateness, and pedagogical suitability. Two of four rotating domain experts independently review each task, with those that receive approval by both assigned reviewers being included in the deployment pool, and those with one or no approvals being rejected. The review interface shows the task-specific attribute profile and difficulty level, next to the rendered visual stimulus, allowing reviewers to judge the alignment directly. Average review turnaround time in the intervention was 18.3 ± 4.7 hours per batch, which was in the inter-session window and so did not disrupt adaptive delivery (see Figure 1).

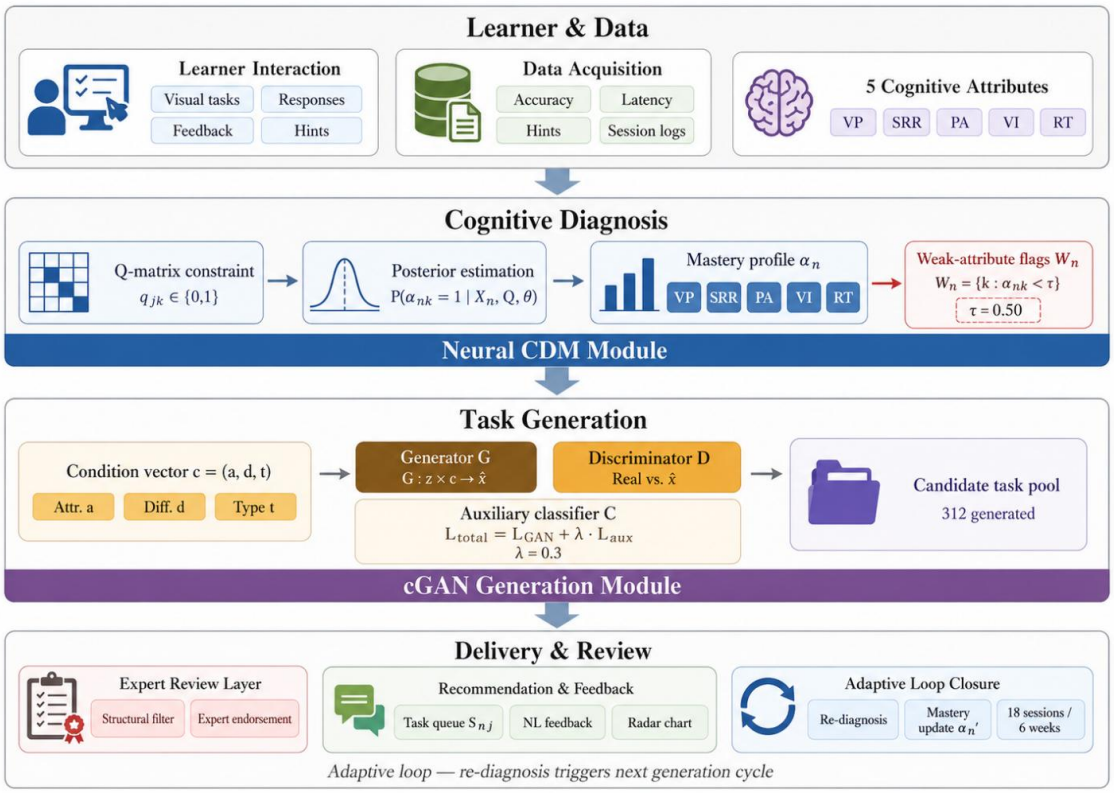


*Figure 1: Architecture of the CDM-GAN Integrated System*

## 3.2 Cognitive Diagnosis Module

The cognitive diagnosis module implements a neural CDM that takes the learner response vector $\boldsymbol{x}_n \in \{0, 1\}^J$ as input and outputs the posterior mastery probability vector $\boldsymbol{\alpha}_n \in [0, 1]^K$. The Q-matrix is embedded as a structural constraint to enforce attribute-task correspondence. The item response probability for learner n on task j is modeled as shown in Equation (2):

$$P(\boldsymbol{x}_{nj} = 1 \mid \boldsymbol{\alpha}_n, \mathbf{Q}, \theta) = \sigma\left(\sum_{k=1}^{K} q_{jk}\, W_k h(\alpha_{nk})\right) \tag{2}$$

where σ(·) is the sigmoid function, $W_k$ is a learned attribute interaction weight, and h(·) is a nonlinear transformation applied to the attribute embedding. The model is trained by minimizing the binary cross-entropy loss over all learner-task pairs as shown in Equation (3):

$$\mathcal{L} = -\frac{1}{NJ}\sum_{n=1}^{N}\sum_{j=1}^{J}\left[x_{nj}\log\hat{P}_{nj} + (1 - x_{nj})\log(1 - \hat{P}_{nj})\right] \tag{3}$$

Mastery classification was thresholded at $\alpha_{nk} \geq 0.50$. Diagnosed weak attributes were defined as shown in Equation (4):

$$\mathcal{W}_n = \{k: \hat{\alpha}_{nk} < \tau\}, \quad \tau = 0.50 \tag{4}$$

and passed as conditioning inputs to the generation module.

## 3.3 GAN-Based Visual Task Generation Module

The generation module employs a conditional GAN (cGAN) in which generator **G** and discriminator $D$ are both conditioned on the task specification vector. The generator maps noise and condition to synthesized tasks as shown in Equation (5):

$$\mathbf{G}: \mathbf{z} \times \mathbf{c} \to \tilde{\mathbf{x}}, \quad \mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad \mathbf{c} = (\mathbf{a}, d, t) \tag{5}$$

The adversarial minimax objective is as shown in Equation (6):

$$\min_G \max_D \mathbb{E}_{x \sim p_{data}}\left[\log D\,(x \mid \mathbf{c})\right] + \mathbb{E}_{\mathbf{z} \sim p_z}[\log(1 - D(G(\mathbf{z} \mid \mathbf{c}) \mid \mathbf{c}))] \tag{6}$$

An auxiliary attribute classifier C is jointly trained with G to enforce attribute alignment. The combined training loss is as shown in Equation (7):

$$\mathcal{L}_{\text{aux}} = -\mathbb{E}_{\tilde{x} \sim G}[\log C\,(\mathbf{a} \mid \tilde{x})] \Rightarrow \mathcal{L}_{\text{total}} = \mathcal{L}_{\text{GAN}} + \lambda \mathcal{L}_{\text{aux}}, \quad \lambda = 0.3 \tag{7}$$

The GAN training was performed using a dataset of 2,184 tasks labeled by experts with 800 epochs employing the Adam optimizer with learning rate $\beta_1 = 0.5$, $\beta_2 = 0.999$. Those tasks meeting the criteria for structural validation, including an aspect ratio of [0.75, 1.33], element density of [4, 24], and attribute tag consistency of at least 0.90, were submitted to experts for evaluation. Those tasks gaining acceptance from two of four experts were added to the deployment pool. The effective number of tasks per attribute after augmentation can be expressed as shown in Eq. (8):

$$\left|T_k^{\text{eff}}\right| = \left|T_k^{\text{real}}\right| + \left|T_k^{\text{GAN}}\right|\rho_k, \quad \rho_k = P(\text{expert endorsement} \mid \text{attribute}k) \tag{8}$$

where $\rho_k$ denotes the empirically estimated endorsement rate per attribute from pilot validation.

## 3.4 Diagnosis-to-Generation Adaptive Loop

At each diagnostic cycle, the system computes a priority recommendation score for task j to be presented to learner n as shown in Equation (9):

$$S_{nj} = \sum_{k \in \mathcal{W}_n} q_{jk} \, (\tau - \hat{\alpha}_{nk}) + \gamma \Delta d_{nj}, \quad \gamma = 0.2 \tag{9}$$

The first term of Eq. 9 accumulates mastery deficits across all weak attributes required by task $j$, weighted by the Q-matrix entries $q_{jk}$, so that tasks requiring multiple diagnosed weak attributes receive proportionally higher scores. The second term $\Delta d_{nj}$ is the signed difference between the calibrated difficulty of task $j$ and the learner's current estimated ability level, derived from item response theory calibration of the task pool; this term penalizes tasks that are substantially easier than the learner's current ability and rewards tasks in the proximal zone of difficulty. The weight $\gamma = 0.2$ was set to ensure that attribute-deficit targeting dominates task selection while difficulty calibration provides a secondary tiebreaker among tasks matched on attribute requirements. Tasks are ranked in descending order of $S_{nj}$, and the highest-scoring task matching at least one element of $\mathcal{W}_n$ is selected for delivery. If the pool is exhausted for a given weak attribute, the generative task layer is invoked to replenish supply as described in Section 3.3.

Following each session, the learner's cumulative response vector is updated to include the session's new responses, and the CDM re-estimation procedure is executed to produce an updated mastery vector. The updated is stored, the weak-attribute set $\mathcal{W}_n$ is recomputed, and the task queue for the subsequent session is regenerated accordingly. This session-level iteration constitutes the primary adaptive cycle of the system, repeated across all 18 training sessions of the six-week intervention. The loop is initialized at the start of the intervention using the pre-test response vector as the first input to CDM estimation, ensuring that adaptive task selection begins from the first training session rather than requiring a separate cold-start period (see Figure 2).

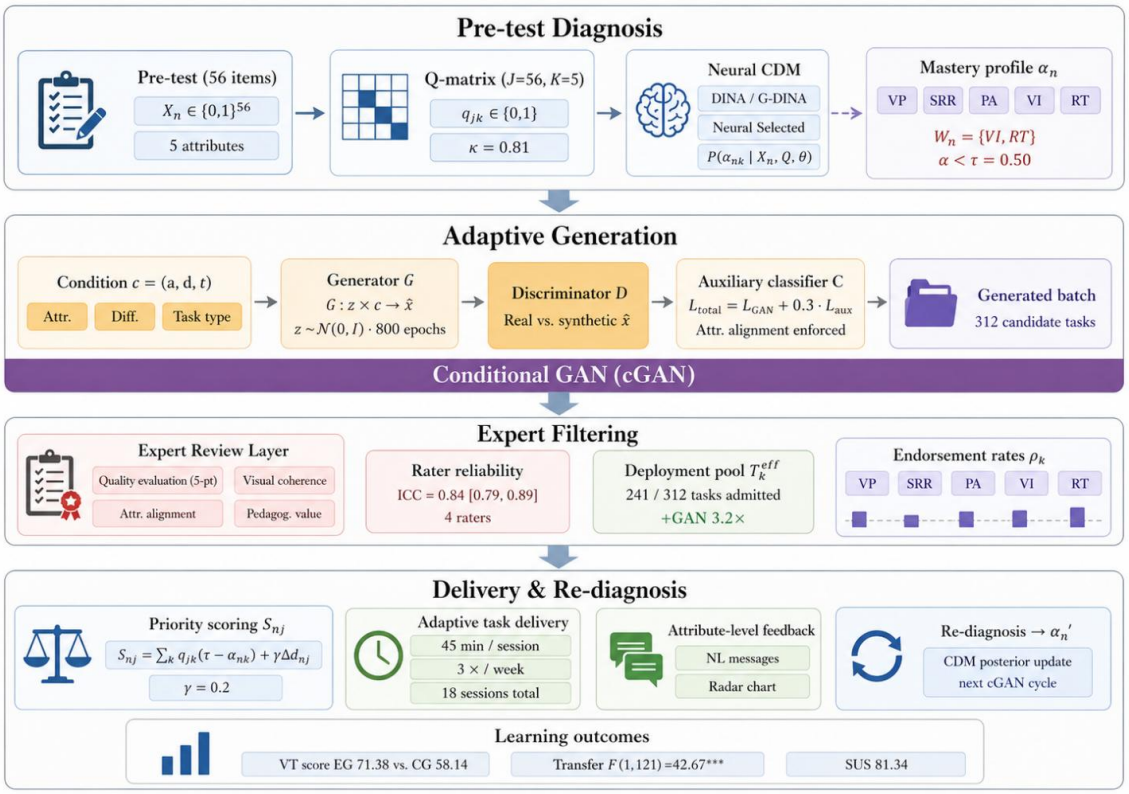


*Figure 2: Diagnosis-to-Generation Adaptive Loop*

# 4 Experimental Design and Evaluation

## 4.1 Experimental Procedure

This experiment was carried out through four successive stages. Stage 1 (Week 0, pretest): Pretest with all the participants involved was performed using a test consisting of 56 items measuring the initial level of visual thinking skills in five characteristics, as well as pretest measures of NASA-TLX and engagement. Stage 2 (Weeks 1-6, experimental intervention): EG used the system of CDM-GAN combination for three 45-minute sessions each week (overall total 18 sessions) with task selection based on Equation 10 and weekly updating of CDM estimates. CG used the system with identical sessions with fixed task selection based on Equation 10 and without CDM-GAN generation; ordering of tasks was randomized within each level of difficulty. The interface, color scheme, navigation and layout of both systems were identical; the only difference was in the task selection process (CDM-GAN adaptation and randomization for the fixed base pool). Stage 3 (Week 6, post-test): Post-test with identical 56 questions similar to pretest and repeated measures of NASA-TLX and engagement were performed. Stage 4 (Week 10, delay test): Delayed test containing 14 transfer tasks using newly learned visual thinking skills in novel contexts (meteorological map interpretation, architect floor plans recognition, biological cell structures identification, engineering diagrams decoding) was given to participants without warning under standard supervision.

## 4.2 Diagnostic Evaluation

A comparison of cognitive diagnostic models (CDMs) was conducted based on the following metrics: Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), root mean square error of approximation (RMSEA), G-squared ($G^2$), and 10-fold cross-validation of attribute classification accuracy. The interpretability of diagnostic outcomes by subject matter experts was measured using four educational researchers' evaluation of five dimensions, namely profile transparency, usefulness of feedback, discriminative power of attributes, profile stability, and instructional actions. The per-attribute agreement ratio $\rho_k$ obtained from the pilot validation was presented alongside the findings of the primary diagnostic analysis.

## 4.3 Generative Task Quality Evaluation

The GAN-generated tasks were assessed on five dimensions: visual coherence, computational attribute alignment, difficulty consistency, pedagogical value, and overall task quality. For each dimension, raters gave scores on a five-point Likert scale (1 = very poor, 5 = excellent). Eight raters (four content experts and four visual design experts) were used for this study, without knowing which tasks were machine-generated. To evaluate inter-rater reliability, the intraclass correlation coefficient (ICC) was calculated using a two-way mixed model and absolute agreement. The ICC for all dimensions ranged from 0.79 to 0.89 (average ICC = 0.84; 95% confidence interval [CI] = 0.79-0.89), showing high reliability among raters. Computational attribute alignment was tested based on the number of generated tasks that could be correctly predicted by an independently trained attribute classifier, a ResNet-18 model, fine-tuned on the annotated set of tasks designed by humans. The accuracy per attribute was defined as the percentage of tasks with correct predictions for the corresponding target attribute. The difficulty consistency was computed as the standard deviation of difficulty ratings provided by five different pilot raters on a randomly sampled subset of 20% of the generated tasks.

## 4.4 Learning Outcome Evaluation

Between-group differences on post-test and delayed tests were analyzed using ANCOVAs that included pre-test performance as a covariate. All analyses were performed in R (v. 4.3.1) with the use of ez and effectsize packages. Homogeneity of regression slopes was checked in all ANCOVA models before interpretation (interaction Fs < 1.12, all p > .29). Dependent variable in attribute-level ANCOVAs was posterior mastery probability per attribute (continuous measure ranging from 0 to 1) at post-test stage. Correction for multiple comparisons (Bonferroni) was made across five attribute contrasts (alpha = .010). Longitudinal trends in attribute mastery from six biweekly measurements were analyzed using mixed-design ANOVA with group and session as between- and within-subject factors respectively; Greenhouse-Geisser adjustments were made if Mauchly's test suggested violations of sphericity assumption. Performance on transfer test was analyzed in separate ANCOVA with post-test VT performance being added as an additional covariate in order to control for the trained-task familiarity effect. Effect sizes were calculated as partial eta-squared for all between-group contrasts (conventions: .01 = small, .06 = medium, .14 = large) and Cohen's d for all pre-to-post differences within groups.

## 4.5 Learner Experience Evaluation

Learner experience was compared between the two groups for six dimensions through the independent-samples t-test. Before performing each comparison, the homogeneity of variances was tested through Levene's test. If variance inequality was detected, the Welch's t-test was used. The Bonferroni-adjusted significance level was set at alpha = .008 in the six tests. Practical significance was evaluated using Cohen's d. Semi-structured interviews were analyzed using NVivo 14.0 through thematic analysis performed independently by two researchers. Agreement between the coders was calculated from the coded data of 20% of the transcripts using Cohen's kappa, which was 0.78. The thematic codes were then matched with the six self-reported dimensions to verify the results from the statistical test, and examples of the codes were provided.

# 5 Results

## 5.1 Diagnostic Performance Results

### 5.1.1 CDM Model Comparison and Selection

Fit indices and classification accuracy results for the four CDMs are reported in Figure 3 and Table 3 using the response data from the full pre-test (N = 124, J = 56, K = 5). The neural CDM produced consistently better results across all the indices compared to the other three models. Specifically, the neural CDM showed the smallest values of AIC (4541.87 ± 24.6, p < .001 vs. DINA), BIC (4718.49 ± 26.3, p < .001), and RMSEA (0.028 ± 0.003, p < .001). The $G^2$ index obtained its lowest value (217.4, p < .001) for the neural CDM, demonstrating the best fit of the observed data to the model's structure. In addition, attribute classification accuracy was the largest in case of the neural CDM (91.2% ± 1.3%, p < .001 vs. DINA).

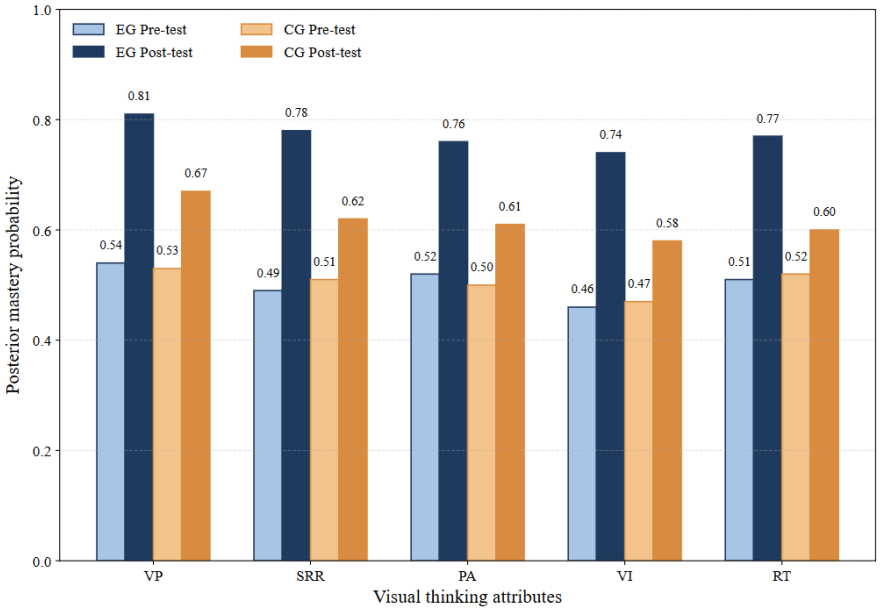


*Figure 3: Pre-to-Post Attribute Mastery Profiles*

*Table 3: CDM Model Fit Indices and Cross-Validated Attribute Classification Accuracy*

| Model | AIC | BIC | RMSEA | Attr. Acc. (%) | $G^2$ |
|---|---|---|---|---|---|
| DINA | 4812.34 ± 31.7 | 4934.18 ± 29.4 | .048 ± .006** | 78.3 ± 2.1* | 342.5 |
| G-DINA | 4623.11 ± 28.3* | 4789.56 ± 31.1* | .037 ± .005** | 86.7 ± 1.8** | 278.3* |
| GDINA-NPC | 4589.04 ± 26.9** | 4761.23 ± 28.8** | .031 ± .004*** | 89.4 ± 1.5*** | 241.6** |
| Neural CDM | 4541.87 ± 24.6*** | 4718.49 ± 26.3*** | .028 ± .003*** | 91.2 ± 1.3*** | 217.4*** |

Note. Values are M ± SD across 10-fold cross-validation. Lower AIC, BIC, RMSEA, and $G^2$ indicate better fit; higher Attr. Acc. indicates better classification. *p < .05. **p < .01. ***p < .001 vs. DINA baseline (Bonferroni-corrected).

### 5.1.2 Attribute Mastery Profile Distribution

Normal distribution of mastery probability of all the five visual-cognitive attributes was evident in the pre-test. No evidence of ceiling and floor effects were found in the pre-test scores of all the visual-cognitive attributes. Average pre-test mastery probability scores of all the visual-cognitive attributes were VP = 0.54 ± 0.09, SRR = 0.49 ± 0.11, PA = 0.52 ± 0.10, VI = 0.46 ± 0.12, RT = 0.51 ± 0.11

In the wake of the post-test stage, the distribution of mastery probability of the Essential Goals (EG) was highly skewed towards more mastery values for all the five attributes. The percentage of learners that had achieved master level ($\hat{\alpha}_{nk}\} \geq 0.50$) had increased by 27% to 34% per attribute (VP: 29.2%; SRR: 34.1%; PA: 28.6%; VI: 33.4%; RT: 30.2%). It is important to note that VI and SRR had the highest skewness of distributions, considering that they were the attributes frequently diagnosed with poor mastery and hence GAN task exposure.

### 5.1.3 Expert Interpretability Ratings of Diagnostic Profiles

Interpretability scores assigned by expert raters to profiles from neural CDM-based diagnosis are reported below (means ± SD across four raters): clarity of the profile 4.38 ± 0.42; usefulness of feedback 4.21 ± 0.51; ability to discriminate attributes 4.44 ± 0.38; consistency of the diagnosis across sessions 4.17 ± 0.49; and feasibility for instructional planning 4.29 ± 0.45. All factors scored above the pre-set minimum criterion of 4.00 (SD < 0.60). The inter-class

correlation coefficient (ICC) among raters was 0.87 (95% confidence interval: [0.81, 0.92]). Reviewer comments suggested that probability vectors for each attribute, plotted as a radar chart, provided interpretable visualization of results even for teachers unfamiliar with C

## 5.2 GAN-Generated Task Results

### 5.2.1 Expert Quality Ratings of Generated Tasks

The quality assessment of tasks created using GAN is summarized in Table 4. The analysis of quality indicators demonstrates that GAN-created tasks achieved significantly lower scores in terms of their visual coherence (4.31 ± 0.52 vs. 4.56 ± 0.41, d = 0.53, p = .038), difficulty consistency (4.22 ± 0.58 vs. 4.49 ± 0.44, d = 0.52, p = .041), and educational value (4.17 ± 0.61 vs. 4.53 ± 0.39, d = 0.71, p = .012). It is crucial to note that there was no significant difference between the attribute alignment of GAN-generated tasks and human-written tasks (4.48 ± 0.46 vs. 4.61 ± 0.38, d = 0.31, p = .094).

*Table 4: Expert Quality Ratings of GAN-Generated vs. Human-Designed Tasks (5-Point Scale)*

| Evaluation Dimension | GAN Tasks (M ± SD) | Human Tasks (M ± SD) | Cohen's d | p |
|---|---|---|---|---|
| Visual Coherence | 4.31 ± 0.52 | 4.56 ± 0.41 | 0.53 | .038* |
| Attribute Alignment | 4.48 ± 0.46 | 4.61 ± 0.38 | 0.31 | .094 ns |
| Difficulty Consistency | 4.22 ± 0.58 | 4.49 ± 0.44 | 0.52 | .041* |
| Pedagogical Value | 4.17 ± 0.61 | 4.53 ± 0.39 | 0.71 | .012* |
| Overall Expert Rating | 4.29 ± 0.49 | 4.55 ± 0.40 | 0.59 | .027* |

Note. Ratings on a 5-point Likert scale (1 = very poor, 5 = excellent). ns = not significant. *p < .05 (Bonferroni-corrected). Cohen's d reflects EG-standardized mean difference (GAN minus Human, negative = GAN lower).

### 5.2.2 Attribute Alignment of Generated Tasks

The per-attribute classification accuracies for the GAN-generated stimuli according to the ResNet-18 classifier trained independently from the GAN model were: VP = 92.4 ± 3.1%, SRR = 88.7 ± 3.8%*, PA = 90.1 ± 3.4%, VI = 86.3 ± 4.2%*, and RT = 89.5 ± 3.6%. All values are above 85%, whereas the relatively lower accuracies for VI (86.3%) and SRR (88.7%) correlate with higher visual complexities of tasks that demand inferential reasoning and understanding of spatial configuration.

### 5.2.3 Task Pool Augmentation and Diversity

GAN Augmentation increased the task pools available for adaptive delivery in terms of specific attributes. Before augmentation, the number of tasks available for each attribute in human-created pools was from 7 to 13, on average equal to 9.8 ± 2.1 tasks/attribute.

## 5.3 Learning Outcome Results

*Table 5: Pre-test and Post-test Visual Thinking Performance by Group and Attribute (ANCOVA, Covariate: Pre-test)*

| Outcome Measure | EG Pre (M±SD) | EG Post (M±SD) | CG Pre (M±SD) | CG Post (M±SD) | ANCOVA F / $\eta^2$ |
|---|---|---|---|---|---|
| VT Total Score | 48.61 ± 6.74 | 71.38 ± 5.82*** | 48.29 ± 7.03 | 58.14 ± 6.51** | F(1,121) = 47.32***, $\eta^2$ = .31 |
| A1: VP Mastery | 0.54 ± 0.09 | 0.81 ± 0.07*** | 0.53 ± 0.10 | 0.67 ± 0.08** | F(1,121) = 29.11***, $\eta^2$ = .22 |
| A2: SRR Mastery | 0.49 ± 0.11 | 0.78 ± 0.08*** | 0.51 ± 0.12 | 0.62 ± 0.10* | F(1,121) = 34.87***, $\eta^2$ = .26 |
| A3: PA Mastery | 0.52 ± 0.10 | 0.76 ± 0.09*** | 0.50 ± 0.11 | 0.61 ± 0.09* | F(1,121) = 31.42***, $\eta^2$ = .24 |
| A4: VI Mastery | 0.46 ± 0.12 | 0.74 ± 0.09*** | 0.47 ± 0.13 | 0.58 ± 0.11* | F(1,121) = 38.56***, $\eta^2$ = .28 |
| A5: RT Mastery | 0.51 ± 0.11 | 0.77 ± 0.08*** | 0.52 ± 0.10 | 0.60 ± 0.09* | F(1,121) = 33.19***, $\eta^2$ = .25 |
| Transfer Task Score | 41.32 ± 7.28 | 65.74 ± 6.11*** | 40.87 ± 7.54 | 51.63 ± 6.89** | F(1,121) = 42.67***, $\eta^2$ = .29 |

Note. EG = Experimental Group; CG = Control Group. Attribute mastery values are posterior mean probabilities (0-1). ANCOVA F-values reported with pre-test score as covariate. $\eta^2$ = partial eta-squared. *p < .05. **p < .01. ***p < .001 (Bonferroni-corrected).

### 5.3.1 Between-Group Overall Visual Thinking Score Comparison

The post-test VT total scores for the experimental group were statistically significantly greater compared to those of the control group, after controlling for the influence of the pre-test VT total scores using the analysis of covariance (ANCOVA) method. The results showed a large effect size with $F(1, 121) = 47.32$, $p < .001$, $\eta^2 = .31$. The adjusted mean difference between groups was 13.24 (95% CI [9.87, 16.61]). Thus, the study proved that there was a larger increase in the level of visual thinking in students who learned using the CDM-GAN integrated system, compared to the students who used the fixed-task system. Significant pre-to-post test differences within groups were also found. The improvement was more pronounced among participants in the experimental group, $t(62) = 19.47$, $p < .001$, $d = 3.45$, than in the control group, $t(60) = 11.23$, $p < .001$, $d = 1.44$. However, the difference in effect size was larger in the experimental group by 2.39 standard deviation units. This indicates a larger increase in the effectiveness of learning due to adaptive targeting. The post-test scores for the EG mean (71.38) were more than two standard deviations above the CG mean (58.14).

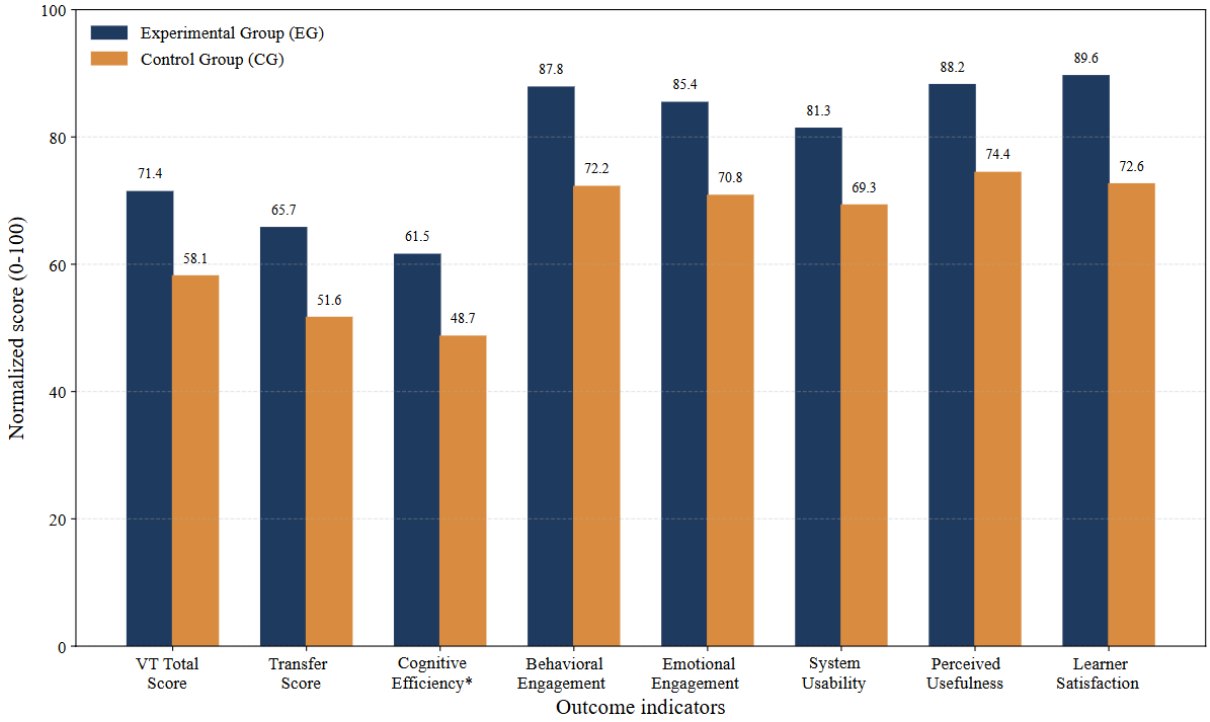


*Figure 4: Comparative Learning and Experience Outcomes between EG and CG*

### 5.3.2 Attribute-Level Mastery Improvement

ANOVA results of attribute-level analysis (Table 5) showed significant between-group differences in all five visual thinking attributes with substantial to large effect sizes ($\eta^2$ = .22-.31). Greatest between-group discrepancy was detected in terms of attribute A4 – Visual Inference (post-test EG mean = 0.74 ± 0.09 vs. post-test CG mean = 0.58 ± 0.11; $F(1,121) = 38.56$, $p < .001$, $\eta^2 = .28$) and attribute A2 – Spatial Relation Recognition ($F(1,121) = 34.87$, $p < .001$, $\eta^2 = .26$), which is consistent with the fact that these two attributes were classified as the weakest at the pre-test stage and were associated with the greatest number of GAN-supported task training sessions. Mixed ANOVA with six measurements per week detected significant Group × Session interactions on all visual thinking attributes (all $F > 6.41$, all $p < .001$, all $\eta^2 > .18$), which means that there was a monotonic divergence of trajectories between groups starting from Session 3 and reaching Session 18.

The results of between-groups analysis performed in the EG show that the most substantial absolute gains in mastery were obtained in A4: VI ($\Delta = 0.28 \pm 0.07$, $t(62) = 21.34$, $p < .001$, $d = 3.54$) and A2: SRR ($\Delta = 0.29 \pm 0.08$, $t(62) = 19.88$, $p < .001$, $d = 3.41$), while A5: RT had a third largest gain ($\Delta = 0.26 \pm 0.07$, $t(62) = 18.92$, $p < .001$, $d = 3.19$). In turn, within-group analyses of mastery gains in the CG revealed considerably lower gains that were more consistent across different attributes (mean $\Delta = 0.10 \pm 0.03$ across attributes).

## 5.4 Learner Experience Results

Cognitive load levels in the EG were found to be significantly lower than those in the CG in terms of the composite measure of NASA-TLX (EG: M = 38.47, SD = 7.62; CG: M = 51.33, SD = 9.14; $t(122) = 8.37$, $p < .001$, Cohen's d = 0.79). Specifically, in the subscales, between-group differences reached statistical significance for the variables of mental demand (EG: M = 44.21, SD = 9.34; CG: M = 58.76, SD = 11.22; $t(122) = 7.93$, $p < .001$, $d = 0.73$), effort (EG: M = 39.84, SD = 8.91; CG: M = 52.47, SD = 10.33; $t(122) = 7.21$, $p < .001$, $d = 0.68$), and frustration (EG: M = 31.62, SD = 9.74; CG: M = 47.83, SD = 12.61; $t(122) = 7.78$, $p < .001$, d

= 0.73). No statistically significant difference was found between the groups regarding the performance subscale (p = .147), which suggests that the decrease in the cognitive load among the EG was a result of less extraneous cognitive demand caused by a mismatch in difficulty of the tasks and not due to the perception of poorer performance. The results of the thematic analysis of interview

*Table 6: Learner Experience Evaluation: Between-Group Comparison of Self-Report Measures at Post-test*

| Scale / Subscale | EG (M ± SD) | CG (M ± SD) | t(122) | p / Cohen's d |
|---|---|---|---|---|
| Cognitive Load (NASA-TLX) | 38.47 ± 7.62*** | 51.33 ± 9.14 | 8.37*** | p < .001, d = 0.79 |
| Behavioral Engagement | 4.39 ± 0.58*** | 3.61 ± 0.74 | 6.51*** | p < .001, d = 0.61 |
| Emotional Engagement | 4.27 ± 0.61*** | 3.54 ± 0.72 | 5.93*** | p < .001, d = 0.57 |
| System Usability (SUS) | 81.34 ± 7.41*** | 69.28 ± 9.63 | 7.82*** | p < .001, d = 0.73 |
| Perceived Usefulness (TAM) | 4.41 ± 0.53*** | 3.72 ± 0.68 | 6.28*** | p < .001, d = 0.59 |
| Learner Satisfaction | 4.48 ± 0.49*** | 3.63 ± 0.71 | 7.44*** | p < .001, d = 0.70 |

Note. EG = Experimental Group; CG = Control Group. Cognitive Load: NASA-TLX composite (0-100); lower = lower load. Engagement, Usefulness, Satisfaction: 5-point Likert. SUS: 0-100. All p-values Bonferroni-corrected across six comparisons. ***p < .001.

# 6 Conclusion

The proposed study aimed to design, execute and evaluate experimentally a visual thinking training system based on Cognitive Diagnostic Models and Generative Adversarial Networks in a diagnosis-directed, generation-assisted adaptive framework. Five diagnosable cognitive attributes related to training tasks through an expert-validated Q-matrix were operationalized as a visual thinking process. A neural CDM predicted per-learner attribute mastery profiles, and a conditional GAN enlarged the task pool with regard to the specific attributes under the constraint of labels, difficulty level, and endorsement by experts. The adaptive loop which was created, pre-test diagnosis, targeted task delivery, immediate feedback and repeated re-diagnosis ensured that the instructional cycle was brought back to the individual learner profiles.

The quasi-experimental evaluation using 124 university students showed that the overall VT performance ($\eta^2$ = .31), as well as all the five attributes ($\eta^2$ = .22-.31), and transfer performance ($\eta$ 2 =.29) improved much more significantly compared to a traditional fixed task system. Even though the learning results were better than those of the control group, cognitive load was significantly lowered (d = 0.79) and the engagement, usability, perceived usefulness, and satisfaction were significantly greater in the experimental group. The given system is not based on a new algorithm but rather on an architectural approach to integration that makes use of the complementary advantages of interpretable diagnosis and controlled generative augmentation in a verified instructional model.

The future research must be focused on Q-matrix misspecification robustness, the creation of automatic GAN quality screening pipelines and the expansion of the framework to other visual reasoning fields such as data science, medical imaging, and engineering design. Some bright technical perspectives are integrating eye-tracking and mouse trajectory data in dynamic real-time CDM estimation, large language model-based generation of natural-language attribute-level feedback, and investigation of diffusion model structures to synthesize higher-fidelity visual tasks with better calibrated difficulty.

# About the Author

Jing Zixuan, Shazhou Professional Institute of Technology, Zhangjiagang 215600, Jiangsu, China. Jing Zixuan, Lecturer at Shazhou Professional Institute of Technology, Zhangjiagang, Suzhou, Jiangsu Province, China. Currently a PhD candidate at the Malaysian Institute of Science and Management, engaged in teaching and academic research, committed to integrating teaching practice with academic research to continuously improve professional competence.

Ooi Boon Keat. Graduate School of Management, Postgraduate Centre, Management and Science University, University Drive, Off Persiaran Olahraga 40100 Shah Alam Selangor Malaysia.Assoc. Prof. Dr. Ooi Boon Keat, PhD Supervisor, Management and Science University, Malaysia

# References

[1] Chiu, T. K. F., Xia, Q., Zhou, X., Chai, C. S., & Cheng, M. (2023). Systematic literature review on opportunities, challenges, and future research recommendations of artificial intelligence in education. *Computers and Education: Artificial Intelligence, 4*, Article 100118.

[2] Wang, S., Wang, F., Zhu, Z., Wang, J., Tran, T., & Du, Z. (2024). Artificial intelligence in education: A systematic literature review. *Expert Systems with Applications, 252*, Article 124167.

[3] Joksimović, S., Ifenthaler, D., Marrone, R., De Laat, M., & Siemens, G. (2023). Opportunities of artificial intelligence for supporting complex problem-solving: Findings from a scoping review. *Computers and Education: Artificial Intelligence, 4*, Article 100138.

[4] Maldonado López, B., Ledesma Chaves, P., & Gil Cordero, E. (2023). Visual thinking and cooperative learning in higher education: How does its implementation affect marketing and management disciplines after COVID-19? *The International Journal of Management Education, 21*(2), Article 100797.

[5] Mohseni, Z., Masiello, I., Martins, R. M., & Nordmark, S. (2024). Visual learning analytics for educational interventions in primary and secondary schools: A scoping review. *Journal of Learning Analytics, 11*(2), 91–111.

[6] Yan, L., Martinez-Maldonado, R., Jin, Y., Echeverria, V., Milesi, M., Fan, J., Zhao, L., Alfredo, R., Li, X., & Gašević, D. (2025). The effects of generative AI agents and scaffolding on enhancing students' comprehension of visual learning analytics. *Computers & Education, 234*, Article 105322.

[7] Zhuang, Y., Liu, Q., Huang, Z., Li, Z., Shen, S., & Ma, H. (2022). Fully adaptive framework: Neural computerized adaptive testing for online education. *Proceedings of the AAAI Conference on Artificial Intelligence, 36*(4), 4734–4742.

[8] Gao, W., Liu, Q., Huang, Z., Yin, Y., Bi, H., Wang, M.-C., Ma, J., Wang, S., & Su, Y. (2021). RCD: Relation map driven cognitive diagnosis for intelligent education systems. In *Proceedings of the 44th International ACM SIGIR Conference on Research and*

*Development in Information Retrieval* (pp. 501–510). Association for Computing Machinery.

[9] Zhou, Y., Liu, Q., Wu, J., Wang, F., Huang, Z., Tong, W., Xiong, H., Chen, E., & Ma, J. (2021). Modeling context-aware features for cognitive diagnosis in student learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (pp. 2420–2428). Association for Computing Machinery.

[10] Li, J., Wang, F., Liu, Q., Zhu, M., Huang, W., Huang, Z., Chen, E., Su, Y., & Wang, S. (2022). HierCDF: A Bayesian network-based hierarchical cognitive diagnosis framework. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (pp. 904–913). Association for Computing Machinery.

[11] Ma, H., Li, M., Wu, L., Zhang, H., Cao, Y., Zhang, X., & Zhao, X. (2022). Knowledge-sensed cognitive diagnosis for intelligent education platforms. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management* (pp. 1451–1460). Association for Computing Machinery.

[12] Wang, F., Liu, Q., Chen, E., Huang, Z., Yin, Y., Wang, S., & Su, Y. (2023). NeuralCD: A general framework for cognitive diagnosis. *IEEE Transactions on Knowledge and Data Engineering, 35*(8), 8312–8327.

[13] Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., & Gašević, D. (2024). Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology, 55*(1), 90–112.

[14] Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., Stadler, M., Weller, J., Kuhn, J., & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, Article 102274.

[15] Khosravi, H., Viberg, O., Kovanović, V., & Ferguson, R. (2023). Generative AI and learning analytics. *Journal of Learning Analytics, 10*(3), 1–6.

[16] Yan, L., Martinez-Maldonado, R., & Gašević, D. (2024). Generative artificial intelligence in learning analytics: Contextualising opportunities and challenges through the learning analytics cycle. In *Proceedings of the 14th Learning Analytics and Knowledge Conference* (pp. 101–111). Association for Computing Machinery.

[17] Khosravi, H., Shibani, A., Jovanović, J., Pardos, Z. A., & Yan, L. (2025). Generative AI and learning analytics: Pushing boundaries, preserving principles. *Journal of Learning Analytics, 12*(1), 1–11.

[18] Misiejuk, K., López-Pernas, S., Kaliisa, R., & Saqr, M. (2025). Mapping the landscape of generative artificial intelligence in learning analytics: A systematic literature review. *Journal of Learning Analytics, 12*(1), 12–31.

[19] Saxena, D., & Cao, J. (2022). Generative adversarial networks (GANs): Challenges, solutions, and future directions. *ACM Computing Surveys, 54*(3), Article 63.

[20] Jabbar, A., Li, X., & Omar, B. (2022). A survey on generative adversarial networks: Variants, applications, and training. *ACM Computing Surveys, 54*(8), Article 157.

[21] Vaz, B., & Figueira, Á. (2024). GANs in the panorama of synthetic data generation methods: Application and evaluation. *ACM Transactions on Multimedia Computing, Communications, and Applications, 21*(1), Article 3.

[22] Bethencourt-Aguilar, A., Castellanos-Nieves, D., Sosa-Alonso, J. J., & Area-Moreira, M. (2023). Use of generative adversarial networks (GANs) in educational technology research. *Journal of New Approaches in Educational Research, 12*(1), 153–171.

[23] Vie, J.-J., Rigaux, T., & Minn, S. (2022). Privacy-preserving synthetic educational data generation. In *Educating for a New Future: Making Sense of Technology-Enhanced Learning Adoption: EC-TEL 2022* (pp. 393–406). Springer.

[24] Zhan, C., Deho, O. B., Zhang, X., Joksimović, S., & De Laat, M. (2023). Synthetic data generator for student data serving learning analytics: A comparative study. *Learning Letters, 1*, Article 5.

[25] Farhood, H., Joudah, I., Beheshti, A., & Muller, S. (2024). Advancing student outcome predictions through generative adversarial networks. *Computers and Education: Artificial Intelligence, 7*, Article 100293.

[26] Liu, Q., Shakya, R., Jovanović, J., Khalil, M., & de la Hoz-Ruiz, J. (2025). Ensuring privacy through synthetic data generation in education. *British Journal of Educational Technology, 56*(3), 1053–1073.

[27] Tao, J., Chen, H., & Li, Y. (2024). Cognitive diagnosis method via Q-matrix-embedded neural networks. *Applied Sciences, 14*(22), Article 10380.

[28] Winne, P. H. (2021). Open learner models working in symbiosis with self-regulating learners: A research agenda. *International Journal of Artificial Intelligence in Education, 31*(3), 446–459.

[29] Hardaker, G., & Glenn, L. E. (2025). Artificial intelligence for personalized learning: A systematic literature review. *International Journal of Information and Learning Technology, 42*(1), 1–14.