Computers in Human Behavior Reports 18 (2025) 100678

Contents lists available at ScienceDirect

Computers in Human Behavior Reports

journal homepage: www.sciencedirect.com/journal/computers-in-human-behavior-reports

Applying cognitive theory of multimedia learning principles to augmented reality and its effects on cognitive load and learning outcomes

Vito Candido *, Alberto Cattaneo

Educational Technologies in VET, Swiss Federal University for Vocational Education and Training, Via Besso 84/86, Lugano Massagno, 6900, Switzerland

A R T I C L E I N F O

Keywords:
Augmented reality
Cognitive load theory
Cognitive theory of multimedia learning
Hand-held augmented reality
Head-mounted augmented reality

A B S T R A C T

In the last decade, the use of augmented reality as a learning support tool has been extensively researched, largely due to the proliferation of augmented reality-compatible smartphones. However, findings related to cognitive load levels remain mixed, with studies reporting both an increase and a decrease in cognitive load while using augmented reality when compared to traditional media. This study investigates the influence on cognitive load and learning outcomes of applying cognitive theory of multimedia learning principles to augmented reality applications. Application design plays a pivotal role in determining cognitive load levels, particularly extraneous load, yet cognitive theory of multimedia learning principles have been scarcely investigated in augmented reality contexts. A randomized experimental design was employed, in which 88 participants were assigned to one of three groups: augmented reality with a head-mounted display, hand-held augmented reality, or video (control group). The educational materials for all conditions were designed according to the principles of the cognitive theory of multimedia learning. Learning outcomes were assessed through retention and transfer tasks following an activity involving a Tangram game. Bayesian analyses provided evidence for no difference in extraneous load levels between conditions. Regarding performance, the findings were inconclusive, showing evidence neither for difference nor equivalence between conditions in both retention and transfer tasks.

This study supports the idea that the effectiveness of the cognitive theory of multimedia learning principles can be extended to augmented reality. The encouraging results on extraneous load levels require further investigation regarding performance.

1. Introduction

This study investigates how established cognitive theory of multimedia learning (CTML) principles, known for their capacity to optimize cognitive load (CL), influence extraneous load (EL) and learning outcomes in augmented reality environments. By comparing a video-based instruction (control group) with two augmented reality (AR) implementations, we examine whether these principles maintain their effectiveness across different media. We hypothesize that CTML principles, when properly implemented, will demonstrate comparable effectiveness in both AR and video-based learning environments, particularly in terms of managing extraneous load and supporting learning outcomes.

AR is a technology that allows information to be superimposed (e.g., 3D models, holograms, text) on the real world in real time (Azuma, 1997; Carmigniani et al., 2011; Mann et al., 2023). Users can interact with virtual elements as if they were part of their surrounding environment by using various technical solutions, both for the type of device used and for the mechanism for recognizing the real-world area where additional information is overlaid. In hand-held AR (HH-AR), a smartphone or a tablet is used to display the real world with additional information overlaid on the screen of the device. In head-mounted display AR, wearable devices like Microsoft HoloLens 2 allow for hands-free simultaneous viewing of the real world and the overlaid information. Finally, spatial AR uses specific projectors capable of accurately aligning holograms with the real world. These solutions can be designed to place the additional information based on the recognition of markers or specific shapes in the environment. Marker-based AR is a tracking and registration technique that involves the use of markers, such as printed symbols. These are used by the AR system to recognize and track an object in the real world, and consequently display related additional information. In marker-less AR, the system does not require the presence of a physical marker but uses computer vision techniques to

* Corresponding author.
E-mail address: vito.candido@suffp.swiss (V. Candido).

https://doi.org/10.1016/j.chbr.2025.100678
Received 17 October 2024; Received in revised form 8 April 2025; Accepted 25 April 2025
Available online 2 May 2025

identify and track real-world features such as angles or material textures (Zhou et al., 2008). AR can also provide different levels of realism and various possibilities for interaction between the overlaid information and the real world. In terms of realism, overlaid information can represent stereotypical versions of emulated objects or else highly faithful models (Grieves, 2019). Possibilities for interaction, can be limited to the visualization in time and space of the overlaid information or extend to its perfect synchronization with its real counterpart. Digital twins represent the highest level of realism and interaction with reality. They are digital models that replicate the real functioning of the component that they represent (Batty, 2018).

AR has caught the attention of education researchers due to the potential that its affordances offer for enhancing learning (e.g., Candido et al., 2023; Yu, 2023). However, it is still unclear whether AR use can increase or reduce CL and to what extent applying principles derived from the CTML can influence CL levels. In their review, Akçayir and Akçayir (2017) identify four studies on CL levels in AR learning tasks. Two studies raise concerns about higher CL levels (Cheng & Tsai, 2013; Dunleavy et al., 2009), while two others suggest a potential reduction in CL (Bressler & Bodzin, 2013; Santos et al., 2014). Based on this, the authors highlight the need for further studies to understand when AR use might lead to cognitive overload, considering, among other factors, interface characteristics.

1.1. Cognitive load theory

Cognitive load theory (CLT) emphasizes the limit of human cognitive resources (Sweller, 1988, 2022; Sweller et al., 1998, 2019). Our information-processing capacity is strongly dictated by what we know as well as by our working memory. This latter is limited in both capacity (Cowan, 2001; Miller, 1956) and duration (Peterson & Peterson, 1959). On average, we are able to remember 7 ± 2 new pieces of information and process about 3–4 items in parallel (combining, contrasting, working on the items) for a period of about 20 s. When processing complex information, however, fewer chunks can be handled in working memory (Baddeley, 2012). In a learning task, when the amount of information being processed exceeds an individual's cognitive capacity, cognitive overload occurs. Cognitive overload is characterized by a rapid decline in cognitive performance, resulting in the individual's inability to effectively manage new information and achieve expected learning outcomes. CLT also seeks to understand how cognitive resources are allocated during a learning task.

The most acknowledged model about CL involves a tripartite classification of cognitive load, distinguishing Intrinsic Load (IL), Extraneous Load (EL), and Germane Load (GL) (Sweller et al., 2019). IL comprises objective and subjective components. The objective component reflects task complexity, structure, and the number of elements requiring simultaneous processing. The subjective component depends on individual characteristics: innate cognitive abilities (e.g. working memory capacity) and domain-specific prior knowledge (the expertise in the specific topic). Prior knowledge differences create the most significant variations in IL levels. For example, in the context of chess, prior knowledge of various piece configurations makes it much easier for an expert player to determine the best move in a given position, far more than their innate cognitive abilities (Kalyuga, 2005; Sweller, 2011; Sweller et al., 1998, 2019). The EL concerns how the design of educational materials impacts on CL: an effective arrangement of learning materials allows for saving cognitive resources in understanding the structure of the educational material itself, allowing to concentrate attention on the learning task. Finally, the GL represents the working memory used to structure the information in mental schemes and to link them to already known information (Paas et al., 2010). Although conceptually distinct, IL and EL, when measured using validated self-report scales, have often shown to be correlated. EL levels assume particular importance when the task generates a significant amount of IL (Paas et al., 2003; Sweller, 1994). This is because it is difficult to find high levels of EL in tasks with low IL in the data, otherwise there would be no discernible impact on performance. In compliance with this conclusion, Krieglstein et al. (2022) note that in presence of very simple tasks and greater available cognitive resources, it is not possible to detect high levels of EL, as the ease of the task allows for compensating for the elements of EL, without them interfering with learning.

Defining when a task can determine good levels of IL becomes then a central point when intending to measure the levels of EL. To do this, it is foremost necessary to first consider the subjective component of IL – in terms of prior knowledge (Kalyuga, 2005) –and then to examine the objective component – namely the number of elements that must be kept in mind during the execution of the task (Miller, 1956). The complexity and interconnection of the elements to be analyzed also have a significant impact, substantially reducing the number of pieces of information we are capable of handling (Cowan, 2010). To assess a task's IL, three factors require sequential evaluation: the learner's prior knowledge of the subject serves as the starting point, followed by an analysis of the cognitive demands required for task execution, and concluding with an examination of implemented strategies for focusing attention on key information elements.

1.2. Cognitive load in AR

Studies investigating the use of AR in education have produced mixed results with respect to CL. A review by Avila-Garzon et al. (2021) has reported a reduction in CL, while others have found contradictory results (Akçayır & Akçayır, 2017). Several factors can influence CL levels in AR-based educational settings. The target population (e.g., the age of the participants and their education level), can undoubtedly be an intervening variable when looking at the impact of AR on CL levels; however, in this context, we focused on the main aspects that emerge from the literature as factors that have a significant impact and can easily impact CL levels when using AR, although being still under-investigated. The following factors will be discussed: (1) the type of AR technology used, (2) the choice of a control group in comparative studies, (3) the learning task, and (4) the CL measurement instrument employed. Although no meta-analyses have investigated the relative impact of these factors yet, we will discuss each of them, and conclude with proposing what we consider the most crucial factor – (5) the application of CTML principles - which will be addressed in detail in the following section.

Starting from the first factor, most reviews consider studies conducted with HH-AR support (Akçayır & Akçayır, 2017; Garzón & Acevedo, 2019). However, Buchner et al. (2021a) report substantial differences depending on the technological solution chosen. Comparing CL levels and performance, spatial AR seems to produce better outcomes than HMD-AR and HH-AR (Alves et al., 2019). The advantage of spatial AR in CL and learning outcomes was also demonstrated in previous usability studies (Baumeister et al., 2017; Hochreiter et al., 2018). However, spatial AR is rarely employed in educational contexts. Usability studies indicated that HMD-AR, especially using the HoloLens device, resulted in higher CL levels than a screen-based solution. This effect diminished with the HoloLens 2 (Seeliger et al., 2022). No meta-analyses have quantified the differences among these technical solutions, although these differences are evident, particularly favouring spatial AR.

Although no studies have directly investigated the effect of different types of control groups on CL levels, we examined the effects of AR on learning outcomes, based on the assumption that higher CL levels often correspond to poorer learning results (Sweller, 2019). Based on the meta-analysis conducted by Yu (2023), we extracted the reported effect sizes for learning outcomes. We excluded studies without control groups (e.g. Chin & Wang, 2021) and those with anomalous effect sizes (Gonzalez et al., 2020; Ibáñez et al., 2014). The studies were then categorized into two groups: comparisons with "learning as usual" (LaU) and comparisons with other experimental groups. The weighted average effect size was $d = 1.28$ for LaU comparisons and $d = 0.63$ for comparisons with other experimental groups. This disparity suggests that the choice of control group could have an important impact on CL levels.

Another factor influencing CL is the learning task. Roca-González et al. (2017) demonstrated AR effectiveness in tasks requiring high visuospatial skills. In such contexts, AR can reduce CL levels during task execution (Lee et al., 2020), although not always producing large effect sizes (Gecu-Parmaksiz & Delialioğlu, 2020). Exceptional results favoring AR have been observed in physics (Ibáñez et al., 2014) and nursing studies (Gonzalez et al., 2020). However, these cases lack information about CL levels.

Finally, the studies reported in recent meta-analyses on AR (Garzón & Acevedo, 2019; Yu, 2023) often have not employed measures capable of distinguishing between the different components of CL. Therefore, this could be an aspect that contributes to the unclear effects of using AR on CL, as we are unable to distinguish whether AR use might influence IL or EL levels. The application of the cognitive theory of multimedia learning (CTML) principles is finally a very important factor that can impact CL levels, particularly with regard to EL levels. For this reason, we will discuss it in detail in the next section.

1.3. Cognitive theory of multimedia learning

The Cognitive Theory of Multimedia Learning (CTML; Mayer, 2002, 2014, 2020; Mayer & Fiorella, 2021), which is influenced (inter alia) by CLT, is an evidence-based theory for managing CL in learning supported by multimedia. The CTML has demonstrated the effectiveness of a series of specific principles for reducing extraneous processing. Among these are the principles of coherence, signaling, spatial contiguity, and temporal contiguity. Additionally, principles aimed at facilitating the management of information flow have been identified, such as the principles of segmenting, modality, and closely related to this, the voice principle. While the first group of principles primarily impacts EL, the second group aims to avoid the risk of exceeding the available cognitive resource limits, even for essential processing of learning information, thereby influencing IL. Table 1 presents the definitions of the previously mentioned principles, along with their related references. Since the evidence for these principles is very robust, we have cited only the first study addressing the single principle and a recent review or meta-analysis on it. In the next section we focus on how CTML principles have been investigated in studies using AR.

Table 1
List of key principles for reducing intrinsic load (IL) and extraneous load (EL).

Principle	Description	References
Principles aimed at reducing IL		
Segmenting	People learn better when a multimedia message is presented in learner-paced segments rather than as a continuous unit.	Mayer & Chandler, 2001; Rey et al., 2019
Modality	People learn better from a multimedia message when the words are spoken rather than written.	Mousavi et al., 1995; Ginns, 2005
Voice	People learn better when the words in a multimedia lesson are spoken in an appealing human voice rather than a machine voice.	Mayer et al., 2005; no review, or meta-analysis found
Principles aimed at reducing EL		
Coherence	People learn better when extraneous information is excluded from multimedia lessons.	Harp & Mayer, 1997; Rey, 2012
Spatial Contiguity	Place printed text next to the corresponding part of the graphic.	Mayer, 1989; Schroeder & Cenkci, 2020
Temporal Contiguity	Present corresponding visual and verbal material at the same time.	Mayer & Anderson, 1991; Ginns, 2006
Signaling	People learn better when cues are added that highlight or spotlight the key information in a multimedia lesson and its organization.	Yarbus & Yarbus, 1967; Alpizar et al., 2020.

Note. While Mayer identifies the voice principle as supporting generative processing, here this principle is considered to reduce IL. A human voice was used to avoid potential issues associated with robotic voices, rather than to enhance appeal.

1.4. Related works

Research on applying CTML principles to immersive technologies is still in its infancy and has primarily focused on immersive virtual reality (IVR). Several studies have explored the effectiveness of CTML principles in virtual environments (Makransky et al., 2017b, 2019; Meyer et al., 2019), but the results have not always been consistent with those observed in traditional media. For example, the redundancy principle has not produced the same effects observed in non-immersive contexts (Baceviciute et al., 2022; Makransky et al., 2019; Moreno & Mayer, 2002). This discrepancy highlights the need to explore the applicability of CTML principles across various immersive technologies, including AR.

Indeed, research on applying CTML principles to AR is even more limited compared to IVR. Çeken and Taşkın (2022), in their review, found that none of the studies examined made explicit reference to CTML in creating educational materials for AR. Despite the absence of research on CTML, some researchers have begun to explore the optimization of AR interfaces using multimedia learning principles (Dabor et al., 2019; Rodenburg et al., 2018), suggesting a growing interest in adapting these principles to the AR context as well as in testing their validity. Elford et al. (2022) studied the effects of AR on learning outcomes and CL in chemistry education, finding no significant differences in learning outcomes or IL and EL levels between AR and traditional methods. Although their study shared the goal of studying how to reduce EL with our research, they did not explicitly reference CTML nor apply specific principles to achieve this result.

In contrast, Thees et al. (2020) conducted a study on the use of AR with HoloLens 2 in physics education, explicitly applying spatial and temporal contiguity principles to reduce EL. Their AR system integrated real-time data visualization with experimental equipment, allowing students to observe numerical values, false-color images, and graphs directly on the heated metal rod. Results showed that the AR system significantly reduced EL, as measured by a self-report scale. However, no significant effects were observed on learning outcomes nor on IL. Additionally, the study did not control for the effect of possible intervening variables, such as time spent with the different technological supports.

The lack of applying the segmentation principle might have further emphasized the effectiveness of AR in using spatial and temporal contiguity (Mayer, 2020, p. 236). Nevertheless, the fact that AR served as support for executing a preparatory activity for learning rather than for learning itself, which was obtained from textbook study, might explain why no significant differences were found in learning outcomes.

Despite these studies, a significant research gap still remains. In the latest edition of the multimedia learning handbook (Mayer, 2005), AR was not explored in depth among the technologies that can benefit from applying CTML principles. Although the immersion principle has been introduced for IVR, which could potentially apply to AR as well (Makransky, 2021, p. 296), specific studies systematically exploring the application of CTML principles to AR are still lacking. This gap in the literature underscores the need for further research to understand how CTML principles can be effectively applied and adapted in the AR context to improve learning and manage cognitive load.

1.5. Aims and research question

From the analysis of previous literature, it is clear that there is still no definitive answer regarding the effects of AR on CL, particularly in terms of EL. Moreover, educational AR applications developed so far have not consistently applied or verified the validity of CTML principles. It is plausible to hypothesize that AR applications developed following CTML principles could minimize EL levels and, consequently, improve learning outcomes. Testing individual principles of greater interest for AR – namely spatial contiguity, temporal contiguity, and signaling – would inevitably involve excluding some core principles. For example, to test the effectiveness of spatial and temporal contiguity (Küçük et al., 2016), it would be necessary to ignore segmenting, otherwise no effect would be observed (Mayer, 2021a). Similarly, testing signaling, where AR could operate directly in the real world, would involve ignoring the application of coherence (Mayer, 2021a).

Our intent, instead, was to develop an application in line with the core principles of CTML while simultaneously verifying its effectiveness. To this end, we decided to compare AR with the medium which is most similar to this technology, and for which CTML principles have been extensively tested, namely video (Mayer, 2021b). In this context, we formulated the following research questions and hypotheses:

RQ1: Do the levels of IL and EL experienced by learners differ depending on whether they use HH-AR, HMD-AR, or video instruction?
HP1a: The selected task will elicit a significantly higher level of IL compared to the familiarization phase.
HP1b: EL levels will not differ across conditions during all phases of the learning task.
RQ2: How do learning outcomes, specifically task completion rates, differ among students using HH-AR, HMD-AR, and video instruction?
HP2a: The task completion rate will be equivalent for students in the retention task.
HP2b: The task completion rate will be equivalent for students in the transfer task.

2. Materials and methods

To test our hypotheses, we conducted a quasi-experimental study comparing two AR applications (HH and HMD) and a video, each developed in accordance with multimedia learning principles. We chose to test two AR modalities since the various AR technological solutions are often not distinguished in existing studies. In particular, we chose to test HH-AR and HMD-AR because the former has been the most used in studies so far, while the latter could represent a new frontier for AR and could be particularly relevant in education, as it gives the possibility to the user to have both hands at disposal. In the following sections, we will present the participants, the procedure, and the measurements employed. We have also included two intermediate sections about the choice of the task and the specific principles we applied within the application, as these aspects are particularly relevant to the research question.

2.1. Material: tangram task

Firstly, we had to identify a task capable of inducing sufficiently high IL levels. The literature suggests two methods: the manipulation of visuospatial information with varying degrees of difficulty, and the management of the amount of information to be memorized (Vanneste et al., 2021). We opted for the former method. The tangram game proved to be a particularly suitable activity for this purpose, as it is effective in increasing CL levels in visuospatial tasks (Vanneste et al., 2021; Morton et al., 2022). The tangram involves creating various shapes using seven tiles: five triangles, one square, and one parallelogram. This game is adaptable to creativity and problem-solving exercises, depending on the instructions provided. Since our aim was to increase IL across time, we asked participants to compose two specific shapes: a square and a stylized house (Fig. 1). Traditional visuospatial tasks were not employed for two different reasons: firstly, because we wanted the task to be approached with a specific strategy for solving it; secondly, because visuospatial tasks are usually tackled through paper-based multiple-choice tests, which are less suitable for the creation of an AR application. Our goal is to select a task suitable for use with all three adopted media, ensuring that none provides specific learning advantages.

Our objective was not only to select a task that would allow us to appreciate a noticeable change in IL levels, but also to facilitate the learning of a problem-solving strategy by correctly applying the principles of multimedia learning. It is possible to enable a solver to create the desired shape with tangram tiles because the use of the right strategy allows them to optimize their cognitive resources. From a geometric analysis of tangram tiles, you find that each tile can be decomposed into minimal units in the form of isosceles right triangles, as shown in Table 2. In assembling the desired shape, the seven pieces must be placed within a wooden frame that clearly delineates the area in which the pieces must be positioned. This allows all seven pieces to be placed only in the correct configuration. In other words, if the two largest pieces are positioned inside the frame inefficiently in terms of space, it becomes immediately evident that the remaining five pieces cannot be accommodated within the frame. Our strategy, based on the analysis of the shapes and the constraints imposed by the edges of the frame, involved inserting the pieces in descending order, from largest to smallest. This

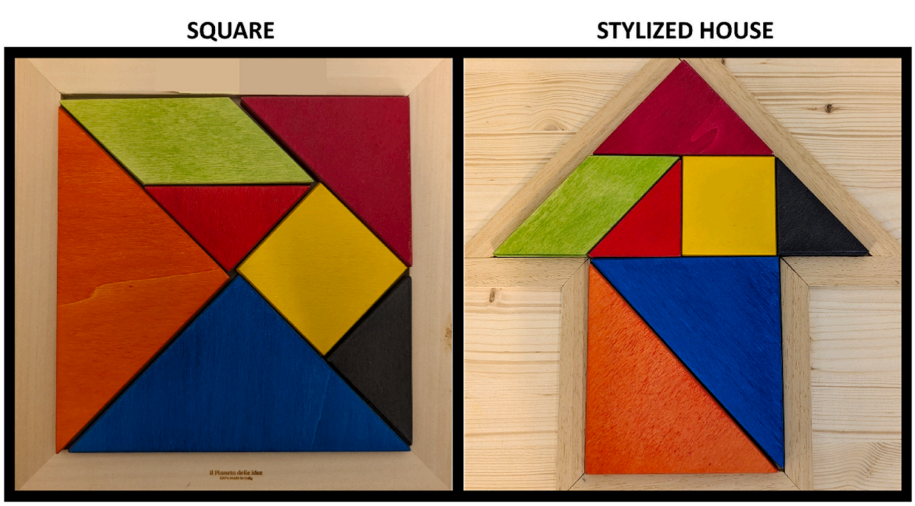
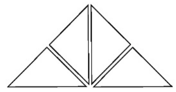
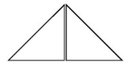
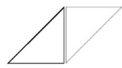
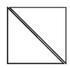
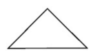


Fig. 1. Square and stylized house.
Note. The square was used in the learning phase and the retention task. The stylized house was used in the transfer task.

Table 2
Classification of the tangram tiles.

Tile	Description	units	Image
1, 2	Large Triangle	4	
3	Medium Triangle	2	
4	Parallelogram	2	
5	Square	2	
6, 7	Small Triangle	1	

method imposes limitations on the subsequent pieces, facilitating identification of the correct arrangement to create the desired shape. It is reasonable to think that the solution to this task will result from the combination of the strategy taught using multimedia materials as suggested by the CTML (Mayer, 1999) and the interaction with the tangram pieces and wooden shape which according to embodied cognition theory enhances learning through direct sensorimotor engagement with physical objects (Wilson & Golonka, 2013).

In conclusion, while acknowledging that the solution to the problem will not solely be the result of the strategy we taught, we believe that choosing the tangram game has allowed us to have reasonable confidence in determining good levels in IL levels, as well as to focus on a learning outcome supported by the principles of multimedia learning.

2.2. Developing the three applications in line with multimedia learning principles

The two AR applications and the video content were specifically developed for the study by the research team. In the case of AR, the application is based on Unity, supported by the MRTK 2 toolkit. As for the videos, iVideo.education was used, an educational platform for creating interactive videos.

To ensure comparability across conditions, the design of the applications was developed as similarly as possible, although accounting for the different affordances offered by the various tools (see Fig. 2) for detailed screenshots of the three applications). This involved overcoming challenges, particularly in the development of AR applications, to ensure that the differences between the compared media were minimal, except for the technical differences intrinsic to the individual tools. To optimize the CL levels in the three applications (see Table 3 for an overview), we chose the principles that proved their effectiveness in our baseline condition (video), according to the meta-analysis by Mayer (2021), and applied them to the other two AR applications.

In particular, for managing essential processing, we applied the segmentation and modality principles throughout the entire application. The "voice principle" was implemented too, based on evidence from six different studies indicating that the use of robotic voices can impair learning outcomes, with a median effect size of $d = 0.74$ in favor of the human voice (Mayer, 2021, p. 237). Concretely, information related to the rules of the game, strategies for solving it, and suggestions for reaching the final solution were presented in separate segments, with words spoken by an appealing human voice. For reducing extraneous processing, we implemented principles including coherence, spatial and temporal contiguity and signaling. As per the signaling, spatial contiguity, and temporal contiguity principles, these were employed only when various aids were provided during the execution of the task. Therefore, the application of these principles resulted in some discrepancies dictated by the different technological solutions (see Table 4):

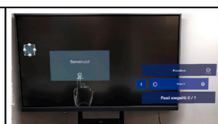
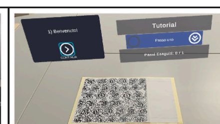
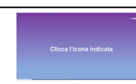
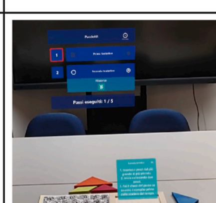
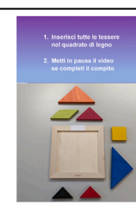
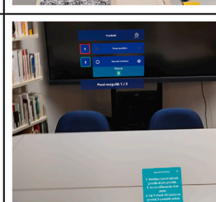
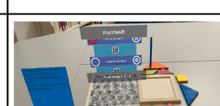
Phase	Description	Time limit	HMD-AR	HH-AR	V
Phase 1: Familiarization					
Familiarization **Reduce IL** **Segmenting** 4 steps: - Steps check - Opening sub-steps - Substep check - Displaying resources **Modality & Voice** The instructions were given by a human voice. A brief text summarized what was said to avoid the transient effect. **Reduce EL** **Coherence** Only the essential instructions for using the feature have been explained. **Spatial & temporal contiguity** Animations were shown with the voice description. A summary was displayed next to the images. **Signaling** Deictic used to indicate where to interact with the application.	Participants learn to use all the features of the application through guided tasks in the tutorial.	-			
Phase 2: Tangram Resolution Attempt					
CL and Ease of Use Scales	Participants complete the Cognitive Load (CL) scale by Klepsch et al. (2017) and the scale by Venkatesh & Davis (2000).	-			
Tangram Task – Puzzle 1 **Reduce IL** **Modality & Voice** The instructions were given by a human voice. A brief text summarized what was said to avoid the transient effect. **Segmenting** Applied to divide the learning process into different phases. **Reduce EL** **Coherence** Only the game instructions are explained at this stage. **Spatial & temporal contiguity** For the last time, an animation showed how to interact with the application to start the first attempt. A summary of the game rules was displayed near the wooden shape or on the PC screen. **Signaling** Deictic used to indicate where to interact with the application.	The participant receives instructions on the game rules. No hints or strategies are provided on how to complete the task. Audio instructions: "Try to fit the seven tiles into the wooden square as quickly as possible. All pieces must lay flat and cannot overlap each other. After two and a half minutes, the task will end. When you are ready, play the video (V); when you are ready, press the 'Start' button (AR)". Text instructions: "Place all the tiles into the wooden square".	2m30s			
Attempt 2 - Puzzle 1 **Reduce IL** **Modality & Voice** The instructions were given by a human voice. A brief text summarbed what was said to avoid the transient effect. **Segmenting** Applied to divide the learning process into different phases. **Reduce EL** **Coherence** Only the strategy for solving the problem is explained at this stage. **Spatial & Temporal contiguity** A summary of the strategy was displayed near the wooden shape or on the PC screen. **Signaling** Not applied in this stage.	Regardless of the outcome of the first attempt, the participant is asked to perform the task again. This time, they receive instructions to optimize their attempts. Audio instructions: "Try inserting the pieces from largest to smallest; this way, you'll occupy a larger area and more easily understand if the other pieces can fit into the wooden shape. Start by placing the first two: you will have one minute. When you are ready, play the video (V); when you are ready, press the 'Start' button (AR)". Text instructions: 1. Insert the pieces from largest to smallest. 2. Start by placing two pieces.	1m			

Fig. 2. Screenshots and instructions for the three conditions.

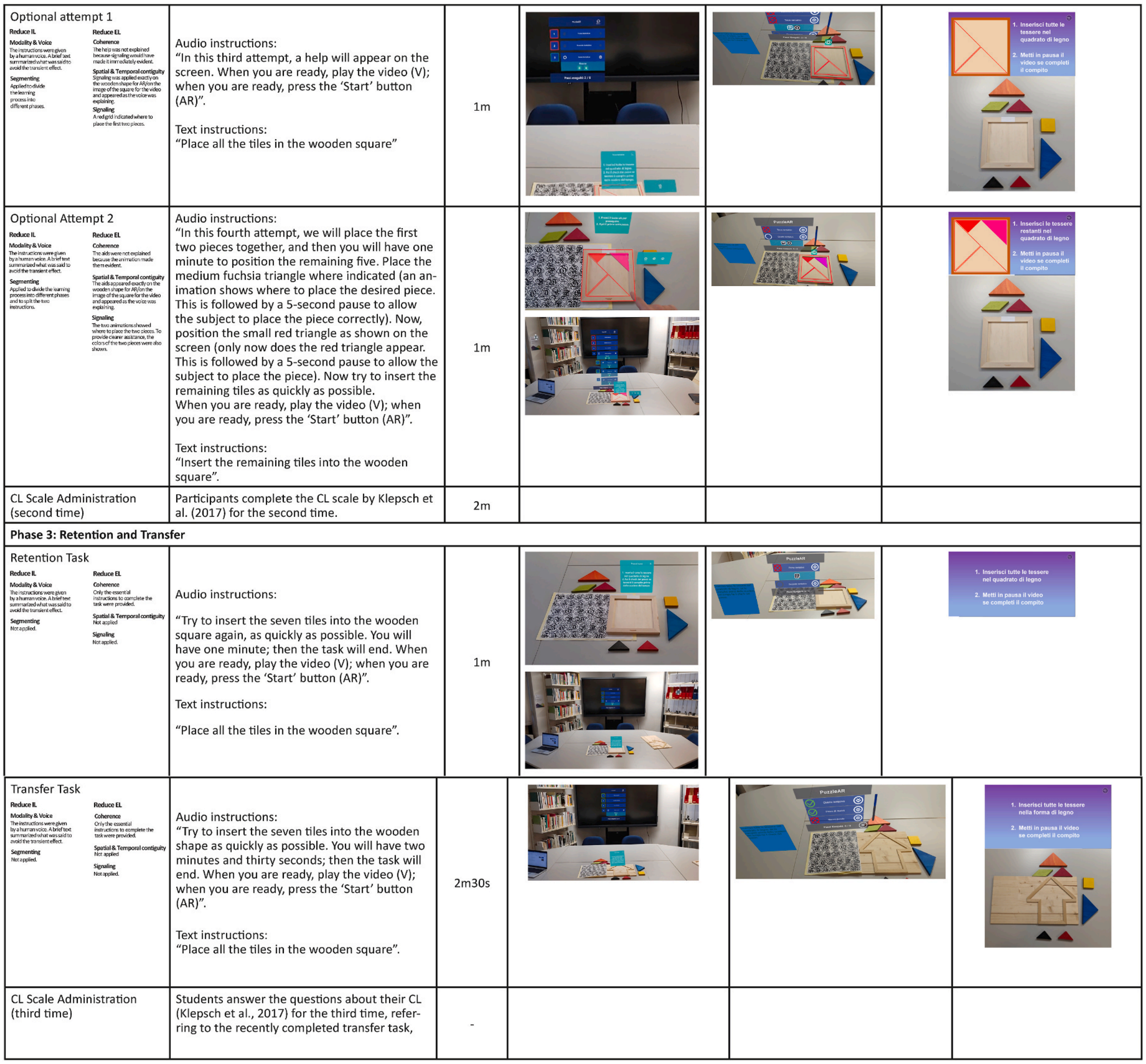


Fig. 2. (*continued*).

Table 3
Multimedia learning principles implemented in the three applications.

Principle	Application in the applications
Principles aimed at reducing IL	
Segmenting	Applied throughout the entire application
Modality	Applied throughout the entire application
Voice	Applied throughout the entire application
Principles aimed at reducing EL	
Coherence	Applied throughout the entire application
Spatial Contiguity	Applied only when visual aids were presented
Temporal Contiguity	Applied only when visual aids were presented
Signaling	Applied only when visual aids were presented

Note. IL = Intrinsic Load, EL = Extraneous Load.

while information could be displayed on real objects in AR, this was not possible within the video. Principles related to generative processes were excluded, with the aim of testing whether AR could be effective even without the implementation of these, which should compensate for the increase in CL levels (Makransky et al., 2019).

Additional differences inherent in the technological devices used in the different conditions can be mentioned with respect to usability: When displaying help in the video on a laptop, an attention shift is needed to find correspondences between the video and reality. Although the use of AR could mitigate this limitation, in the HH-AR condition it became uncomfortable to view additional elements while holding the phone and working with one hand on the wooden tiles. In contrast, in the HMD-AR version, additional information could be displayed directly

Table 4
Differences in applying multimedia learning principles across the three applications.

Differences	V	HH-AR	HMD-AR
Spatial Contiguity	Information displayed next to the image on laptop screen	Information displayed next to real objects through the screen of a smartphone	Information displayed next to real objects through the lenses of the HMD
Temporal Contiguity	Visual and verbal material presented simultaneously on laptop screen	Visual and verbal material presented simultaneously on real world through the screen of a smartphone	Visual and verbal material presented simultaneously on real world through the lenses of the HMD
Signaling	Applied on images in the video	Applied on real word through the screen of a smartphone	Applied on real world through the lenses of the HMD

Note. V = Video, HH-AR = Hand-Held Augmented Reality, HMD = Head-Mounted-Display Augmented Reality.

while proceeding with the execution of the task, facilitating interaction.

2.3. Participants

We recruited 90 students from a Swiss upper secondary level vocational school, all of whom were enrolled in the second and third year of sanitary facilities or heating systems installation programs. These programs are almost entirely attended by male students. All participants voluntarily took part in the data collection, which was performed in November 2022. Only participants who were unfamiliar with the tangram game and completed the entire procedure were included in the sample; two were excluded resulting in a final sample of 88 students (age range: 15–33 years; mean age: 19.9 years; SD = 3.62, all males). Power analysis indicates that a sample size of 84 is required to detect effects of $f = 0.25$ ($\alpha = .05$, $1 - \beta = 0.95$) when comparing three groups across three measurement points with repeated measures ANOVA, assuming a small correlation of 0.20 between repeated measures. None of the students involved had prior experience with AR. Participation was voluntary, and all participants signed a consent form before starting the experiment. Participants were randomly assigned to the three conditions (see Table 5).

2.4. Measures and instruments

This section describes the measures used, the scale used for administering CL, and the other measurements employed.

2.4.1. CL measurement

We used a self-report scale to distinguish different CL types. While physiological measurements and secondary tasks provide continuous, potentially more objective data, their interpretation still heavily relies on subjective measures (Korbach et al., 2018). Self-report, though subject to bias, avoids the intrusiveness of alternative methods. We selected the scale validated by Klepsch et al. (2017) because it consists of only 8 items (two for IL, three for both EL and GL) and can be employed in multiple administrations. Furthermore, this scale has proven to be effective in detecting levels of EL in multimedia environments (Skulmowski, 2023) and to be suitable for tasks based on the principles of multimedia learning (Krieglstein et al., 2022).

2.4.2. Timing and task completion measurements

To determine the completion times for the different tasks, we referred to Baran (2007), who reported a study on tangram tasks, where adults aged 20–30 completed two types of forms: The first had easily recognizable tile contours, while the second featured shapes in which tile contours were less distinguishable. Since our square falls into the latter category, we referred to the completion times reported in that study, which averaged 128 s and reached a maximum of 329 s. We also considered the times reported by Nakano, who indicated an average completion time of 483 s for the same task we identified as transfer. However, in Nakano's study, participants received no instructions on how to solve the task. Thus, we considered the total duration of 330 s—excluding the time required to listen to explanations—sufficient for the tasks included in our protocol and we developed our entire procedure within Baran's (2007) range.

We subsequently divided the time among the various steps we had planned, allocating 150 s for the first attempt, followed by 60 s for the second, third, and fourth attempts, during which participants were progressively guided by the instructions they received. Finally, after a break, we allotted 60 s for all participants to complete the retention task, as it was deemed simpler than tackling the task for the first time, and once again 150 s for the transfer task to verify whether, after having learned the strategy, participants were able to complete a similar task within exactly the same amount of time.

During the initial familiarization phase, we measured the time required to perform some simple actions while using the application in AR conditions or interacting with the video player. Subsequently, during the use of the application or the interaction with the player, we timed the task completion in different attempts, assessing success or failure based on completion within the established time limits. Task completion was treated as a dichotomous variable: it was considered successful only if all seven pieces were placed within the wooden shape within the established time limit, and unsuccessful otherwise. Finally, participants were asked if they were familiar with the game they had just interacted with, without mentioning its name, to reduce the risk of communication between students about the game and searches for solutions online.

Table 5
Participant overview per condition.

Condition	Instrument	Participants	Age (mean ± SD)
HMD-AR	Microsoft HoloLens 2	28	19.8 ± 3.28 (16–31)
HH-AR	Smartphone	30	19.4 ± 4.37 (15–33)
V	Laptop	30	20.5 ± 3.08 (16–30)

Note. HMD-AR = Head-Mounted-Display Augmented Reality, HH-AR = Hand-Held Augmented Reality, V = Video.

2.4.3. Ease of use

We assessed the perceived ease of use immediately after the familiarization task, given that AR, especially HMD-AR, is not widely adopted and low familiarity with these technologies can impact EL and learning outcomes. The four-item 'ease of use' subscale of the Technology Acceptance Model (TAM) by Venkatesh and Davis (2000) was employed.

2.5. Study procedure

Participants were randomly assigned to one of three conditions. Students assigned to condition V learned to complete the Tangram by watching a video, while students in the other two conditions received the same instructions using AR. In the HH-AR condition, both visual and textual instructions were displayed using a smartphone, whereas in the HMD-AR condition, instructions were provided using the wearable device HoloLens 2. The whole procedure was conducted with one researcher per participant and organized around three main phases: familiarization, tangram resolution, and retention and transfer. During the familiarization phase, participants learned how to use the application's features for the subsequent problem-task by interacting with the menus of the two AR applications or with the video player. This phase had two objectives: to allow participants to understand how to interact with the devices and to determine whether there were significant differences in CL levels among the three conditions during activities related solely to the usability of the application. Our goal was to identify a baseline about CL for the three interfaces, before introducing any task. At the end of this phase, participants completed the CL scale for the first time (Klepsch et al., 2017) and perceived ease of use scale (Venkatesh & Davis, 2000).

In the second phase, participants performed the tangram task, attempting to position the seven tiles inside the wooden frame. For the first attempt, a time limit of 2 min 30 s was given, based on Baran et al. (2007).

If the initial attempt did not lead to a solution within the set time, the participant received the following instruction: "Try to insert the pieces from the largest to the smallest, in this way, you will occupy a larger area and you will understand more easily if the other pieces can fit into the wooden form. Start by placing the first two." Afterwards, another minute was given to solve the task. If it was not solved, the placement of the two largest tiles was shown by the application, offering an extra minute to solve the task. In case of further failure, the placement of two other tiles was illustrated by the application, leaving only three elements to be mentally managed.

The division into attempts was introduced to make the task execution more inclusive, to avoid overly high frustration, and to provide support to participants. Participants had a total of 330 s available, equivalent to the 329 s used by the slowest participant in Baran et al.'s (2007) study and to prevent participants from engaging in long, frustrating exercise sessions without receiving any instructions on how to proceed in solving the task, as outlined in Nakano's (2017) protocol. Regardless of the task outcome, all participants were shown the solution by the researcher without providing any explanation on the taught strategy, and then asked to respond again to the CL scale.

After approximately 2 min of waiting, each participant faced a task of immediate retention, trying to reconstruct the same shape in a time limit of 1 min. Subsequently, students were introduced to the transfer task, which required solving a problem with characteristics similar to the previous one. The house figure, in fact, presents patterns common to the square, making it suitable for transfer. As in the initial task, participants were given 2 min 30 s to complete it. Finally, students responded again to the CL questions, referring to the transfer task. The instructions provided via audio and text are reported in Fig. 2, the entire procedure is summarized in Fig. 3.

2.6. Data analysis

Data were analyzed using JASP software (Version 0.18.2). We chose to use this software because it specifically supports Bayesian analyses while also providing the possibility to perform inferential analyses (Wagenmakers et al., 2018). For the hypotheses in which we aimed to support the null hypothesis (*HP1b*, *HP2a* e *HP2b*), indeed, we integrated inferential analyses with Bayesian analyses. We controlled for potential intervening variables, namely tutorial completion times, perceived ease of use of the applications, and age differences among participants. These

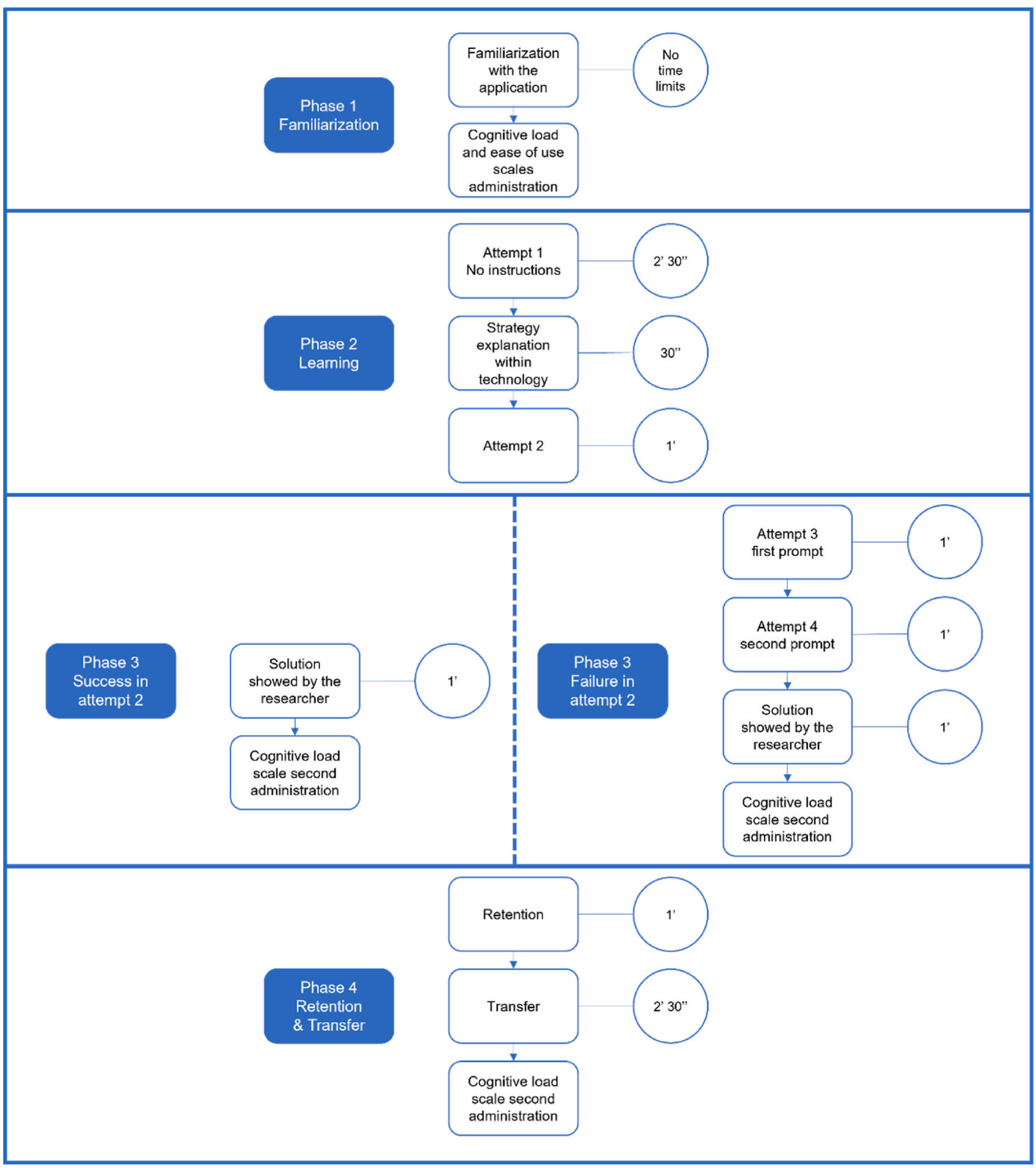


Fig. 3. Procedure summary.

variables were examined for their influence on both EL levels and performance in retention and transfer tasks. A repeated measures ANOVA was utilized to observe IL levels across the three phases of the experiment. Linear regression was applied to control for variables impacting EL, while logistic regression assessed their direct impact on performance. Lastly, ANCOVA and logistic regression were used to evaluate the effects on EL levels and performance.

To support the null hypothesis in EL levels across conditions (Jeffreys, 1939; Kass & Raftery, 1995; Oberauer, 2022), we used, in addition to inferential tests, Bayesian ANCOVA, ANOVA and logistic regression. For the inferential ANCOVA, the formula can be expressed as:

$$LogELT1\_mean = \beta_0 + \beta_1 Condition + \beta_2 EaseOfUse\_mean + \beta_3 (Condition \times EaseOfUse\_mean) + \varepsilon$$

Where *LogELT1_mean* represents extraneous load values, *Condition* refers to the experimental group (HH-AR, HMD-AR, or video), *EaseOfUse_mean* is the centered covariate, and ε represents the error term. In this case, we only considered the main effects.

For Bayesian ANCOVA, we focused solely on the interaction effect between ease of use and conditions. Below, we report the syntax used in JASP:

LogELT1_mean ~ Condition: EaseOfUse_mean

The syntax ":" indicates that only the interaction effect is analyzed, without including main effects.

For the Bayesian analysis, we used the BF_{01}, which represents the ratio between the null hypothesis (0) and the alternative hypothesis (1). The higher the value produced, the more likely the null hypothesis is compared to the alternative hypothesis. For example, a $BF_{01} = 3$ indicates that the null hypothesis is 3 times more likely than the alternative hypothesis (Goss-Sampson, 2020, pp. 39–40). To facilitate the interpretation of the results, Table 6 presents the interpretation of the bayes factor, commonly abbreviated as BF.

3. Results

In this section, we present the reliability of the scales used followed by the results following the research hypotheses order.

3.1. Reliability test

In accordance with the findings of Klepsch and colleagues (2017), who validated the scale, we tested the reliability of the subscales per each submission time: IL, GL, and EL. The reliability, assessed using McDonald's omega (Goodboy and Martin, 2020) coefficient (ω), is reported in Table 7.

Independently from the time measurement, reliability levels are always acceptable for IL and EL, but consistently too low for GL, which was therefore excluded from the subsequent analysis. The reliability of the perceived ease of use scale from the TAM was checked too and gave good results (ω = 0.88, 95 % *CI* [0.84, 0.92]).

Table 6
Bayes factor classification for H1 and H0 evidence.

BF_{10}	Evidence for H1	BF_{01}	Evidence for H0
1	No evidence	1	No evidence
1–3	Anecdotal evidence	1–3	Anecdotal evidence
3–10	Moderate evidence	3–10	Moderate evidence
10–30	Strong evidence	10–30	Strong evidence
30–100	Very strong evidence	30–100	Very strong evidence
>100	Extreme evidence	>100	Extreme evidence

Note. BF_{10} = Bayes Factor 10, H1 = Alternative hypothesis, BF_{01} = Bayes Factor 01, H0 = Null hypothesis.

Table 7
Reliability test for the Cognitive load scale.

Test	ω	95 % CI
IL-T1	0.69	[0.55, 0.83]
IL-T2	0.69	[0.56, 0.82]
IL-T3	0.62	[0.47, 0.77]
GL-T1	0.39	[0.19, 0.58]
GL-T2	0.52	[0.36, 0.67]
GL-T3	0.61	[0.49, 0.73]
EL-T1	0.65	[0.52, 0.77]
EL-T2	0.73	[0.63, 0.82]
EL-T3	0.83	[0.77, 0.89]

Note. IL = Intrinsic Load, GL = Germane Load, EL = Extraneous Load, T1 = Familiarization, T2 = Learning Task, T3 = Transfer Task.

3.2. RQ1 cognitive load

The following section addresses hypothesis *HP1a*, according to which the problem-solving task will be able to elicit significantly higher levels of IL compared to the familiarization phase, and hypothesis *HP1b*, according to which, due to the application of CTML principles across all three media, no differences in EL levels will be observed among the three conditions.

3.2.1. HP1a Intrinsic Load

Fig. 4 displays the descriptive statistics for IL across time (T1, T2, and T3) and conditions (AR-HH, AR-HMD, V).

As already evident from the descriptive statistics, IL levels significantly increased during the learning task, as well as during retention and transfer tasks. To test this hypothesis, we initially conducted a repeated-measures ANOVA. Due to violations of the normality assumption, a non-parametric Friedman test was employed, revealing a significant effect across the measurement phases ($\chi^2 = 81.6$, df = 2, $p < .001$, $W = 0.46$). Durbin-Conover pairwise comparisons further demonstrated significant differences between all measurement phases: Phase 1 and Phase 2 ($F = 11.92$, $p < .001$, $r = 1.27$), Phase 1 and Phase 3 ($F = 8.43$, $p < .001$, $r = 0.90$), and Phase 2 and Phase 3 ($F = 3.50$, $p < .001$, $r = 0.37$), indicating substantial changes in Intrinsic Load over time.

3.2.2. HP1b extraneous load

To conduct the analysis on EL, three variables were controlled: familiarization time perceived ease of use and age. Although Bayesian statistics typically demonstrate reduced sensitivity to violations of assumptions, we cautiously chose to adhere to the assumptions of normality and heteroscedasticity where necessary. This necessitated the transformation of EL data measured across three distinct time intervals. The logarithmic transformation was applied to EL1 and EL3 to address significant skewness, supported by Osborne (2010), underscoring its efficacy in skewness reduction and variance stabilization. For EL2, exhibiting lesser skewness, the square root transformation was adopted, in line with Tabachnick and Fidell (2013). Both methods were assessed across all distributions, with the optimal one chosen based on the Shapiro-Wilk test.

Table 8 displays the descriptive statistics for EL across time (T1, T2, and T3) and conditions (AR-HH, AR-HMD, V). The scale used for measuring EL ranged from 0 to 9.

Controlled variables. Among the variables analyzed (see below, Table 9), only the perceived ease of use measured after the familiarization phase, had a significant impact on EL levels. A logarithmic transformation was applied to meet the normality assumption ($W = 0.99$, $p = .645$). Both the intercept (β = 0.54, SE = 0.17, $t = 3.26$, $p = .002$) and the effect of perceived ease of use (β = −0.05, SE = 0.02, $t = -2.41$, $p = .019$) were significant, with an overall R^2 of 0.08. This suggests that the higher the perceived ease of use, the lower the perceived EL levels.

Test of nonsignificant differences. To assess EL in the first time interval

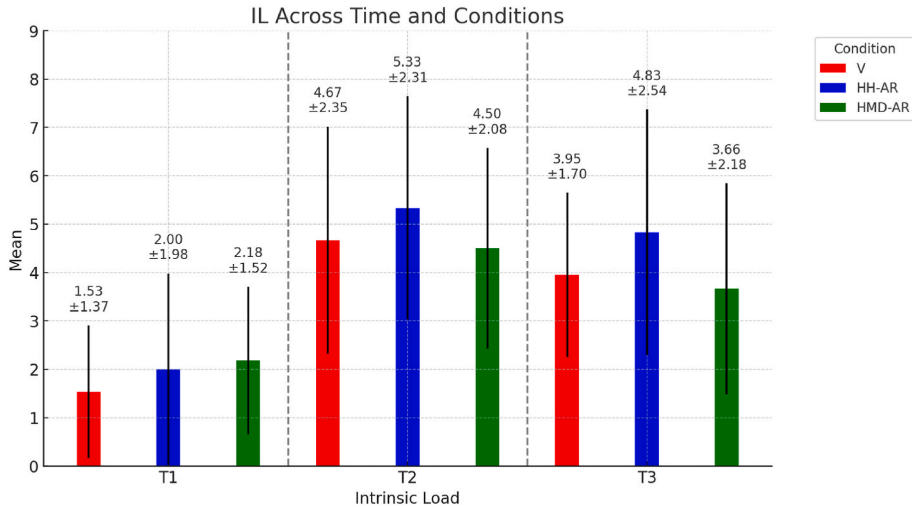


Fig. 4. Intrisic load across time and conditions.
Note. T1 = Familiarization, T2 = Learning Task, T3 = Transfer Task. Scale range: from 0 to 9. HMD-AR = Head-Mounted-Display Augmented Reality, HH-AR = Hand-Held Augmented Reality, V = Video.

Table 8
Extraneous load values (Means and Standard Deviations) per condition and measurement points.

Extraneous load	Condition	Mean	SD
T1	HH-AR	1.83	1.64
	HMD-AR	1.82	1.85
	V	1.12	1.23
T2	HH-AR	3.47	2.02
	HMD-AR	2.88	2.15
	V	3.00	2.15
T3	HH-AR	2.73	2.24
	HMD-AR	2.70	2.03
	V	2.22	1.78

Note. T1 = Familiarization, T2 = Learning Task, T3 = Transfer Task. Scale range: from 0 to 9. HMD-AR = Head-Mounted-Display Augmented Reality, HH-AR = Hand-Held Augmented Reality, V = Video.

for the three conditions, ANCOVA analysis was conducted controlling for ease of use. To meet normality assumptions, EL values were log-transformed. Homogeneity of variances and normality were confirmed by Levene's ($p = .874$) and Shapiro-Wilk tests ($p = .221$), respectively. Results showed a significant effect of ease of use on EL F(1, 82) = 16.99, $p < .001$. The condition had no significant effect $F(2, 82) = 0.72$, $p = .489$. Regarding the interaction between condition and ease of use, the analyses also showed no significant effect on EL: $F(2, 82) = 0.47$, $p = .864$. The second and third time intervals were analyzed using ANOVA to assess the effect of conditions on EL levels. The assumption analysis confirmed homogeneity of variances and normality in both time intervals through Levene's test ($p = .532$; $p = .721$) and the Shapiro-Wilk test ($p = .148$; $p = .313$), reporting non-significant results in both cases (T2: $F(2, 85) = 0.71$, $p = .494$; T3: $F(2, 85) = 0.34$, $p = .715$.

Assessing Evidence for the Null Hypothesis. To examine the interaction between experimental conditions and ease of use on EL levels, a Bayesian ANCOVA was conducted for the initial measurement. A non-informative prior was used to ensure the data were not influenced. The interaction effect between the condition and perceived ease of use showed strong evidence in favor of the null hypothesis ($BF_{01} = 8.72$, error % = 0.006). Fig. 5 shows that if there were no significant differences in the perceived ease of use across the three conditions, there would be no significant differences in EL levels. An ANOVA was then used to check for significant equalities between conditions in the second and third intervals (T2: $BF_{01} = 5.62$, error % = 0.028; T3: $BF_{01} = 7.56$, error % = 0.026). The results of the pairwise comparisons are reported in Table 10. The ease of use accounted for much of the observed differences between conditions in the first time interval, as demonstrated by the ANOVA result ($BF_{01} = 1.12$, error % = 0.023), which, conversely, reached broad significance in the ANCOVA.

3.3. RQ2 performance

In this section, we present the observed performance in the retention and transfer tasks, testing hypothesis *HP2*, according to which we expected equivalence between conditions regarding the successful completion of the retention (*HP2a*) and transfer tasks (*HP2b*).

Table 11 shows the completion percentages of the retention and transfer tasks, per condition.

Controlled variables. A binomial logistic regression revealed a significant effect of the time spent familiarizing with the application on the results obtained in the retention task. A significant intercept was found ($\beta = 2.99$, $SE = 0.83$, $Z = 3.58$, $p < .001$), and familiarization time had a negative effect ($\beta = -0.00$, $SE = 0.00$, $Z = -2.33$, $p = .020$, $R^2 = 0.10$), indicating that an increase in time decreases the likelihood of success. No significant effects were found for the transfer task.

3.3.1. HP2a test of differences in success rate – retention

A logistic Chi square test was conducted to compare the success in the retention task across the three conditions ($\chi^2(2) = 11$, $p = .004$). Post-hoc comparisons, corrected using the Bonferroni adjustment, revealed a significant difference between HMD-AR and video, favoring the video condition ($\chi^2(1) = 10.4$, $p = .003$, OR = 10.5). The results of Bayesian analyses confirmed those of inferential analyses, providing strong support for the alternative hypothesis when comparing HMD-AR and video ($BF_{10} = 25.09$). A logistic regression accounting for familiarization times did not reveal significant differences between conditions, although this result may have been influenced by an insufficient sample size to adequately incorporate this variable. Indeed, the HMD-AR condition still appeared to negatively affect student performance, even though this result was only marginally significant ($p = .075$). Results were also confirmed by the Bayesian analysis ($BF_{10} = 1.33$).

3.3.2. HP2b test of differences in success rate – transfer

The superior effectiveness of the video compared to HMD-AR was not confirmed in the transfer task. Although the video demonstrated a higher likelihood of success (63.3 %) compared to HMD-AR (46.4 %), the comparison between conditions did not show any significant differences ($\chi^2(2) = 2.79$, $p = .247$). Bayesian analyses also confirmed that

Table 9
Possible intervening variables controlled for Extraneous Load and learning outcomes (retention and transfer).

Possible intervening variables	Test Performed	Assumption Checks	Main Test Result
Familiarization time and retention outcomes	Logistic Regression	Always met	Familiarization time significantly negatively predicts retention outcomes (β = −0.05, *SE* = 0.02, *t* = −2.41, *p* = .019, R^2 of 0.08)
Familiarization time and transfer outcomes	Logistic Regression	Always met	Familiarization time do not predicts transfer outcomes significant (*p* = .268)
Familiarization time and EL	Linear Regression	T1: Shapiro-Wilk: *p* = .063, T2: Shapiro-Wilk: *p* = .211, T3: Shapiro-Wilk: *p* = .073	No significant effect of familiarization time at Time 1 (*p* = .142), Time 2 (*p* = .918), Time 3 (*p* = .487)
Age on retention and transfer	Binomial Logistic Regression	Always met	No significant of age on retention (*p* = .221) and transfer (*p* = .084)
Age and Conditions	Kruskal-Wallis	Non-significant violations of homogeneity of variances (Levene's F = 1.05, *p* = .354)	Significant difference in age groups ($\chi^2(2)$ = 6.08, *p* = .048, ε^2 = 0.07)
Age and EL	Linear Regression	T1: Shapiro-Wilk: *p* = .064, T2: Shapiro-Wilk: *p* = .083, T3: Shapiro-Wilk: *p* = .055	No significant effect of age on EL at Time 1 (*p* = .280), Time 2 (*p* = .887), Time 3 (*p* = .925)
Ease of use and Performance	Binomial Logistic Regression	Always met	No significant effect of ease of use on retention (*p* = .125) and transfer (*p* = .651)
Ease of use and Conditions	Kruskal-Wallis	Violations of normality (Shapiro-Wilk, *p* < .001)	No significant differences in ease of use between conditions (χ^2 = 1.24, *p* = .539)
Ease of use and EL (T1)	Linear Regression	Shapiro-Wilk: *p* = .645	Significant effect of ease of use on EL (β = −0.00, *SE* = 0.00, *Z* = −2.33, *p* = .020, *OR* = 1.00)

Note. T1 = Familiarization, T2 = Learning Task, T3 = Transfer Task, EL = Extraneous Load; all the significant regressions reported have also a significant intercept.

there was no evidence supporting either the null or the alternative hypothesis (BF_{10} = 1.08).

4. Discussion

This study investigated the impact of HH-AR and HMD-AR on CL levels and learning outcomes in retention and transfer tasks. Educational materials and application designs were developed following CTML principles and to minimize differences between applications. This aimed to determine if using identical instructions and applying the same principles could support the null hypothesis referred to both CL levels and performance. Given its poor reliability, GL has not been considered. The next sections discuss our hypotheses considering the reported results.

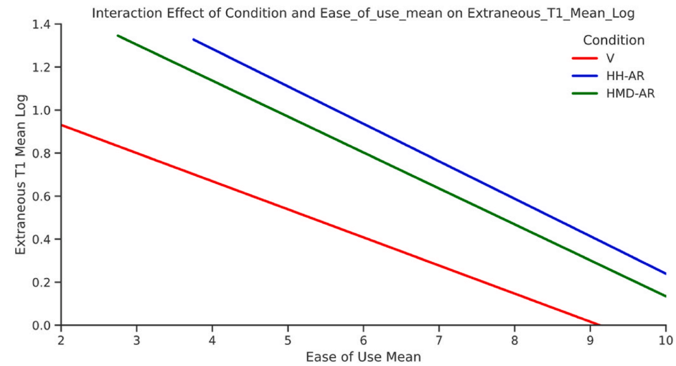


Fig. 5. Representation of the interaction effect between perceived ease of use and EL, per condition.
Note. The graph shows how EL levels vary with increasing perceived ease of use of the application. With high levels of ease of use, the null hypothesis is strongly supported.

Table 10
Bayesian post hoc comparison of EL Levels across the three time intervals.

Pair comparison	BF_{01}	Error %
T1 - ANCOVA		
HMD-AR vs V	3.69	0.002
HMD-AR vs HH-AR	3.47	0.003
HH-AR vs V	3.25	0.003
T2 - ANOVA		
HMD-AR vs V	3.75	0.009
HMD-AR vs HH-AR	2.44	0.009
HH-AR vs V	2.36	0.009
T3 - ANOVA		
HMD-AR vs V	2.94	0.009
HMD-AR vs HH-AR	3.75	0.009
HH-AR vs V	3.16	0.009

Note. T1 = Familiarization, T2 = Learning Task, T3 = Transfer Task. Scale range: from 0 to 9. HMD-AR = Head-Mounted-Display Augmented Reality, HH-AR = Hand-Held Augmented Reality, V = Video.

Table 11
Tangram completion rate for retention and transfer tasks divided per conditions.

Task	% V	% HH-AR	% HMD-AR
Retention	93.33 %	77.42 %	57.14 %
Transfer	63.33 %	67.74 %	46.43 %

Note. V = Video, HH-AR = Hand Held Augmented Reality, HMD-AR = Head Mounted Display Augmented Reality.

4.1. HP1a: the tangram task can elicit good levels of IL

Hypothesis 1 was supported. The results confirmed that the tangram game was capable of increasing IL. The introduction of the tangram task led to a significant increase in IL, with widely significant differences between T1 and T2. As expected, engaging in a similar task for the second time already knowing a strategy for dealing with it, as occurred in the transfer task (T3), produced significantly lower IL levels compared to the first attempt (T2), but still considerably higher compared to the familiarization task (T1).

These findings are entirely consistent with the literature (Vanneste et al., 2021; Morton et al., 2022). The significant increase in IL levels was plausible during the learning phase. Indeed, in the study conducted by Nakano (2017), students had to complete four sessions of 5 min each to successfully accomplish the same task employed in our transfer task. Additionally, only 60 % of the participants in that study achieved success. In other words, it is plausible that performing this task with less time and without prior knowledge of an effective strategy for solving it can be very demanding in terms of IL. Conversely, an increased understanding of the game's solution method, combined with greater experience, can explain the significant reduction in IL levels between the learning phase and the retention and transfer tasks (Kalyuga, 2005). In other words, although the task was the same, students found it easier to perform because they had better understood its underlying mechanism.

The identification of EL elements in the design of a problem-solving task is more evident in contexts where the task presents a real cognitive challenge (Paas et al., 2003; Sweller, 1994), In particular, if the participant is not familiar with the CLT. (Klepsch et al., 2017). This aspect has direct relevance to performance: the absence of adequate difficulty indeed makes it impossible to observe significant variations in performance. Consequently, even in presence of inadequate instructions, an overly simple task most likely leads to a positive performance outcome. Fulfilling this requirement was thus very relevant for verifying the second hypothesis.

4.2. HP1b: EL levels between conditions do not differ

Hypothesis 1b was supported. The inferential analysis of EL levels revealed no significant differences between conditions. Further Bayesian analyses indicated moderate to strong evidence in favor for the null hypothesis concerning EL levels between the three conditions in all the measurement times. Furthermore, a linear relationship between perceived ease of use and EL levels was identified, suggesting that an increase in the perceived ease of use of the application corresponds to a decrease in perceived EL levels. During the familiarization phase, the ANCOVA demonstrated that EL levels moderately to strongly supported the null hypothesis, accounting for differences in perceived ease of use among the three applications. This suggests that if students had the same level of familiarity with all three technologies, no significant differences in EL levels would have been observed. During the other two measurement points the ANOVA supported the null hypothesis regardless of the perceived ease of use. These results are consistent with the review conducted by Akçayir and Akçayir (2017), who suggested that the design of an application has a greater impact on CL than the tool itself. Although differences may exist due to the type of technical solution employed (Buchner et al., 2021), when AR is used in contexts where it provides no added value, it generates the same levels of EL, provided that users have equivalent levels of ease of use with the technology employed. The interaction effect with ease of use is consistent with a recent study conducted by Delgado and Mayer (2024), in which they demonstrated that explaining in advance how to use the main tools presented in the VR environment significantly reduced EL levels (d = 0.42) when compared to the control group that used the application directly.

Although especially HMD-AR can technically improve the application of CTML's signaling and spatial-temporal contiguity principles when compared to video, the combined effect of these principles on EL levels differs from their individual application. For example, the coherence principle can make signaling unnecessary (Mayer, 2020, p. 167): highlighting becomes irrelevant when only essential information is presented. Similarly, spatial-temporal contiguity becomes less important with well-segmented materials (Mayer, 2020, p. 236): the impact is minimal when small amounts of information need to be remembered. While not denying the technical differences between technologies, it is argued that, when CTML principles are applied, the choice of technology becomes secondary for EL levels. Our results suggest that AR does not significantly impact CL levels if CTML principles are applied, although this can be different in specific cases such as when AR is used as a support for transposing from 2D to 3D (Roca-González et al., 2017). AR might reduce IL levels for specific tasks, but not EL levels, despite improving the implementation of signaling and spatio-temporal contiguity. Different scenarios emerge in school learning contexts where segmenting is not applied (Thees et al., 2020) or in real-time assistance tasks with instructional material that cannot be adequately structured (Marques et al., 2023); in these cases, the advantages of AR can be exploited. The equivalent EL levels across conditions can be explained by the application of CTML principles, which optimized the learning materials and made the choice of media less relevant. This result can also be understood in light of the selected task: in the tangram solution task, the potential advantages of AR's enhanced signaling and spatial-temporal contiguity were less critical for learning the problem-solving strategy.

4.3. Hypothesis 2a and 2b: learning outcomes do not differ in retention and transfer task

Contrary to our expectations, Hypothesis 2 was not fully supported. The significant difference identified among conditions in the retention task was then reduced to non-significance when including the time spent familiarizing with the application as a covariate. However, this result may be due to the sample size. We can only state that the less time participants spent familiarizing themselves with the application, the greater their chances of success in the retention task. However, we cannot claim that the difference between conditions would have disappeared. In fact, even the Bayesian logistic regression supports, albeit only anecdotally, the alternative hypothesis rather than the null hypothesis. The comparison between groups revealed significant differences in familiarization times across conditions, thus explaining that the difference in performance could have been determined by the greater fatigue of the students in the HMD-AR condition, who took significantly more time than the other conditions for the participants to familiarize themselves with the technology.

The performances observed in the transfer task also did not confirm our hypothesis: in this case, a linear relationship between performances and the time taken to complete the familiarization task was not identified. Nevertheless, the non-significant differences did not reach the anticipated significant equality. It is possible that other uncontrolled variables, such as the visuospatial abilities of the individual subjects, may have influenced the differences (however not significant) and prevented the identification of the significant equality hypothesized.

Studies conducted to date have employed AR in contexts where its affordances could represent an added value, particularly due to the possibility of more effectively applying the principles of spatial and temporal contiguity. In the study by Thees and colleagues (2020), AR was employed in a context where it could have provided a significant advantage compared to the control group; indeed, receiving contextualized information could have facilitated the learning task. However, even though the experimental AR condition reported significantly lower levels of EL compared to the control group, no significant differences emerged in students' learning outcomes. In contrast, in the study by Altmeyer and colleagues (2020), no significant difference was observed in EL levels, yet a small but significant difference in learning outcomes emerged in favor of the AR group ($\eta_p^2 = 0.07$).

In the study by Thees and colleagues (2020), the HoloLens 2 device was used in the AR condition, whereas in the study by Altmeyer and colleagues (2020), a tablet was employed—a device with which most individuals are familiar. The difference in devices may have influenced both the lack of significant differences with the control group when the HoloLens was used and the differences observed in the tablet condition, even though both studies utilized AR to improve the application of the same principles.

Having chosen a learning context in which AR did not represent a clear added value, no advantages emerged from AR use in either EL or learning outcomes. Conversely, the video condition yielded significantly better learning outcomes compared to the HMD-AR condition in the retention task. The reduced effectiveness of HMD-AR could be explained by its limited usability and familiarity for novices. It is well known that becoming proficient in the use of augmented reality (AR)—particularly HMD-AR, as observed in our study—can also affect the outcomes of usability tests. It is therefore plausible that similar issues have been encountered when employing HMD-AR for learning support (Goad et al., 2023).

5. Limitations

Our study is not without limitations. Due to the time constraints imposed by the protocol, it was not possible to measure CL levels also in the retention task, thus leaving the levels of EL generated by the applications in this context undetermined. Faced with the need to choose between the retention and transfer tasks, preference was given to the latter (Mayer, 2020). Moreover, the duration of the protocol prevented a comprehensive examination of all possible intervening variables, including the visuospatial ability of participants, which should be taken into consideration in further studies. The reliability of the administered scales was not always optimal, which led us to exclude the GL subscale. The low reliability of GL can be explained by the fact that in self-report measurements, it is very complex to distinguish GL from IL (Kalyuga, 2011). Also, the tangram task may have influenced the poor reliability of the scale, being unrelated to the students' study program. Another limitation is related to gender balance: in our sample, female participants were entirely absent, thus placing a clear limit on the possibility of generalizing these results, particularly when dealing with visuospatial tasks where the difference between men and women is known (Castro-Alonso & Jansen, 2019). Moreover the absence of females in the sample might have influenced the reported levels of IL. Indeed, although participants in our study found the task complex, it is possible that women could have perceived it as even more complex. Nevertheless, this does not invalidate the obtained results, as the task was capable of causing a significant increase in IL levels. Similarly, the performance observed in retention and transfer tasks cannot be generalized to a female population, which might have achieved substantially lower performance on the same task. There is no evidence suggesting that perceived levels of EL differ between men and women; however, given the established relationship between IL and EL (Sweller, 1994), it is plausible that women might have perceived higher levels of EL. Result generalization is furthermore limited to our sample age. The pedagogical approach was effective for our sample age but could be counterproductive for other target groups like children (Ashman, 2020; Sinha & Kapur, 2021). Additionally, our sample came from a single school. This certainly allows us to control fewer variables but limits the generalizability of the data.

Further research including children and other target groups is therefore needed. As a study comparing different technologies, unexamined differences may exist between conditions, which could expose us to a media-comparison critique (Clark, 1994). However, we compared AR to video, the medium most closely aligned with CTML, to demonstrate that with adequate learning materials, technology becomes secondary, not to conduct a "technology race". This is also the reason why the application of the CTML principles was adapted to the single technology with the aim to optimize them with respect to their distinctive affordances. This has been demonstrated by achieving equivalent EL levels regardless of the technology used. The effect of HMD-AR on GL levels remains unexplored, and the role of pedagogical choices when using AR was not considered in this study. Future studies could address current limitations in the literature by applying CTML principles to minimize EL levels, comparing AR with various experimental conditions, and employing CL measures capable of distinguishing the three components. Research could also explore tasks where AR affordances can be leveraged, such as those involving mental rotation. Future studies could examine principles that might be more effectively applied in AR, such as signaling overcoming sample limitation of the present study. For instance, researchers could design learning tasks where the ability to see signaling directly applied in the real world could provide added value for learning. In such cases, significant differences between AR and other technologies might emerge.

6. Conclusion

Although its results should be read in light of the limitations just mentioned, this study has shown that applying the principles of CTML to multimedia learning materials can substantially eliminate differences in EL levels, regardless of the technologies used, provided that an adequate familiarization procedure is conducted. This finding contrasts with observations made for virtual reality (e.g., Skulmowski & Xu, 2022), demonstrating that AR itself does not increase EL levels when multimedia design principles are optimally applied. This aligns with Wu et al. (2013), who argue that, within the educational context, the design and use of AR have an impact on learning outcomes greater than that of the technology itself. To achieve these results, we applied several CTML principles, which have been scarcely explored in the use of AR, especially in its HMD form. We also employed Bayesian analyses to support the null hypothesis when necessary.

We believe that this study can contribute to CTML both theoretically, by demonstrating the effectiveness of applying different principles with HMD-AR and HH-AR, and methodologically, by providing an approach for contexts where the principles are likely effective, reducing the need for separate studies in AR. Moreover, studies that examine combinations of principles—more common in real learning scenarios—could further consolidate the effectiveness of this theory.

In real learning scenarios, the aim is to create the best possible instructional materials rather than applying a single principle to observe its effect. The interaction of principles significantly alters the final impact on CL levels and learning outcomes. The goal is to ensure that the medium used does not negatively influence students' learning outcomes by increasing EL. The application of CTML principles, so far under-investigated in AR, could provide a solid foundation for studying the unique affordances of AR, such as three-dimensional visualization in the real world. This would help researchers determine whether the effects are due to the design of educational materials or the affordances of the technology, thus laying a solid groundwork for future research.

However, the analysis of learning outcomes did not fully support our hypotheses, which warrants further research. In the meantime, these findings can be useful from a general pedagogical perspective, highlighting that AR per se does not necessarily have negative effects on EL levels, and from an instructional design perspective, providing evidence that consistently applying specific CTML principles can support students' learning by minimizing their EL. Additionally, this study shows new evidence on the existing differences between VR and AR. We know that VR tends to generate high levels of CL, which can impair learning outcomes, leading to the recommendation to include generative activities to facilitate learning when using VR (Makransky et al., 2019). However, our findings suggest that this approach may not be necessary for AR and highlight differences between these technologies, at least with respect to EL.

This study represents only the starting point for examining the effectiveness of CTML principles with AR in managing CL levels in educational settings. Further studies are needed to clarify the effectiveness of CTML in other learning scenarios and different target populations.

CRediT authorship contribution statement

Vito Candido: Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Alberto Cattaneo:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used Chat GPT 4 o and Claude Sonnet 3.5 exclusively in order to improve the language correctness, fluency and readability of the self-written manuscript. After using this tool/service, the authors reviewed and edited the content again as needed. The authors take full responsibility for the content of the publication.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This research was funded by SERI (State Secretariat for Education, Research and Innovation). (Contract number 1315002129). MARHVL project (Mixing Augmented Reality and Hyper-video for Learning).

We would like to thank Prof. Dr. Lorenzo Sommaruga, Dr. Nadia Catenazzi, and Chiara Locatelli, collaborators at the University of Applied Sciences and Arts of Southern Switzerland, for the development of the various versions of the AR application, and in particular Chiara Locatelli for her collaboration during the data collection.

Special thanks go to Prof. Dr. Dominik Petko, who provided valuable suggestions on the analyses to be conducted and contributed to the review of the manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chbr.2025.100678.

Data availability

Data will be made available on request.

References

Akçayir, M., & Akçayir, G. (2017). Advantages and challenges associated with augmented reality for education: A systematic review of the literature. *Educational Research Review, 20*, 1–11. https://doi.org/10.1016/j.edurev.2016.11.002

Alpizar, D., Adesope, O. O., & Wong, R. M. (2020). A meta-analysis of signaling principle in multimedia learning environments. *Educational Technology Research & Development, 68*, 2095–2119. https://doi.org/10.1007/s11423-020-09748-7

Altmeyer, K., Kapp, S., Thees, M., Malone, S., Kuhn, J., & Brünken, R. (2020). The use of augmented reality to foster conceptual knowledge acquisition in STEM laboratory courses—theoretical background and empirical results. *British Journal of Educational Technology, 51*(3), 611–628. https://doi.org/10.1111/bjet.12900

Alves, J., Marques, B., Oliveira, M., Araújo, T., Dias, P., & Santos, B. S. (2019). Comparing spatial and mobile augmented reality for guiding assembling procedures with task validation. In *2019 IEEE international conference on autonomous robot systems and competitions (ICARSC)* (pp. 1–6). IEEE.

Ashman, G., Kalyuga, S., & Sweller, J. (2020). Problem-solving or explicit instruction: Which should go first when element interactivity is high? *Educational Psychology Review, 32*, 229–247. https://doi.org/10.1007/s10648-019-09500-5

Avila-Garzon, C., Bacca-Acosta, J., Kinshuk, Duarte, J., & Betancourt, J. (2021). Augmented reality in education: An overview of twenty-five years of research. *Contemporary Educational Technology, 13*(3), ep302. https://doi.org/10.30935/cedtech/10865

Azuma, R. (1997). A survey of augmented reality. *Presence: Teleoperators and Virtual Environments, 6*(4), 355–385. https://doi.org/10.1162/pres.1997.6.4.355

Baceviciute, S., Lucas, G., Terkildsen, T., & Makransky, G. (2022). Investigating the redundancy principle in immersive virtual reality environments: An eye-tracking and EEG study. *Journal of Computer Assisted Learning, 38*(1), 120–136. https://doi.org/10.1111/jcal.12595

Baddeley, A. (2012). Working memory: Theories, models, and controversies. *Annual Review of Psychology, 63*, 1–29. https://doi.org/10.1146/annurev-psych-120710-100422

Baran, B., Dogusoy, B., & Cagiltay, K. (2007). How do adults solve digital tangram problems? Analyzing cognitive strategies through eye tracking approach. In J. A. Jacko (Ed.), *Lecture notes in computer science: Vol. 4552. Human-computer interaction. HCI intelligent multimodal interaction environments. HCI 2007*. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-540-73110-8_60.

Batty, M. (2018). Digital twins. *Environment and Planning B: Urban Analytics and City Science, 45*(5), 817–820. https://doi.org/10.1177/2399808318796416

Baumeister, J., Ssin, S. Y., ElSayed, N. A., Dorrian, J., Webb, D. P., Walsh, J. A., … Thomas, B. H. (2017). Cognitive cost of using augmented reality displays. *IEEE Transactions on Visualization and Computer Graphics, 23*(11), 2378–2388. https://doi.org/10.1109/TVCG.2017.2735098

Bressler, D. M., & Bodzin, A. M. (2013). A mixed methods assessment of students' flow experiences during a mobile augmented reality science game. *Journal of Computer Assisted Learning, 29*(6), 505–517. https://doi.org/10.1111/jcal.12008

Buchner, J., Buntins, K., & Kerres, M. (2021a). The impact of augmented reality on cognitive load and performance: A systematic review. *Journal of Computer Assisted Learning, 38*(1), 285–303. https://doi.org/10.1111/jcal.12617

Candido, V., Raemy, P., Amenduni, F., & Cattaneo, A. (2023). Could vocational education benefit from augmented reality and hypervideo technologies? An exploratory interview study. *International Journal for Research in Vocational Education and Training, 10*(2), 138–167. https://hdl.handle.net/10419/273532.

Carmigniani, J., Furht, B., Anisetti, M., et al. (2011). Augmented reality technologies, systems and applications. *Multimedia Tools and Applications, 51*, 341–377. https://doi.org/10.1007/s11042-010-0660-6

Castro-Alonso, J., & Jansen, P. (2019). Sex differences in visuospatial processing. In *Visuospatial processing for education in health and natural Sciences* (pp. 1–22). Springer. https://doi.org/10.1007/978-3-030-20969-8_4.

Çeken, B., & Taşkın, N. (2022). Multimedia learning principles in different learning environments: A systematic review. *Smart Learn. Environ., 9*, 19. https://doi.org/10.1186/s40561-022-00200-2

Cheng, K., & Tsai, C. C. (2013). Affordances of augmented reality in science learning: Suggestions for future research. *Journal of Science Education and Technology, 22*(4), 449–462. https://doi.org/10.1007/s10956-012-9405-9

Chin, K. Y., & Wang, C. S. (2021). Effects of augmented reality technology in a mobile touring system on university students' learning performance and interest. *Australasian Journal of Educational Technology, 37*(1), 27–42. https://doi.org/10.14742/ajet.5841

Clark, R. E. (1994). Media will never influence learning. *Educational Technology Research & Development, 42*(2), 21–29. https://doi.org/10.1007/BF02299088

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24*(1), 87–114. https://doi.org/10.1017/s0140525x01003922

Cowan, N. (2010). The magical mystery four: How is working memory capacity limited, and why? *Current Directions in Psychological Science, 19*(1), 51–57. https://doi.org/10.1177/0963721409359277

Dabor, O., Longford, E., & Walker, S. (2019). Design guidelines for augmented reality user interface: A case study of simultaneous interpretation. *2019 11th Computer Science and Electronic Engineering (CEEC)*, 164–166. https://doi.org/10.1109/CEEC47804.2019.8974331

Delgado, C. Y., & Mayer, R. E. (2024). Implementing pretraining to optimise learning in immersive virtual reality. *Journal of Computer Assisted Learning, 41*(1). https://doi.org/10.1111/jcal.13099

Dunleavy, M., Dede, C., & Mitchell, R. (2009). Affordances and limitations of immersive participatory augmented reality simulations for teaching and learning. *Journal of Science Education and Technology, 18*(1), 7–22. https://doi.org/10.1007/s10956-008-9119-1

Elford, D., Lancaster, S. J., & Jones, G. A. (2022). Exploring the effect of augmented reality on cognitive load, attitude, spatial ability, and stereochemical perception. *Journal of Science Education and Technology, 31*, 322–339. https://doi.org/10.1007/s10956-022-09957-0

Garzón, J., & Acevedo, J. (2019). Meta-analysis of the impact of Augmented Reality on students' learning gains. *Educational Research Review, 27*, 244–260. https://doi.org/10.1016/j.edurev.2019.04.001

Gecu-Parmaksiz, Z., & Delialioğlu, Ö. (2020). The effect of augmented reality activities on improving preschool children's spatial skills. *Interactive Learning Environments, 28*(7), 876–889. https://doi.org/10.1080/10494820.2018.1546747

Ginns, P. (2005). Meta-analysis of the modality effect. *Learning and Instruction, 15*(4), 313–331. https://doi.org/10.1016/j.learninstruc.2005.07.001

Ginns, P. (2006). Integrating information: A meta-analysis of the spatial contiguity and temporal contiguity effects. *Learning and Instruction, 16*(6), 511–525.

Goad, K., Mangold, K., & Furman, S. (2023). *It's not always What the eye can see – Challenges in the Evaluation of augmented reality* (NIST IR 8489). *National Institute of Standards and Technology*. https://doi.org/10.6028/NIST.IR.8489

Gonzalez, A. A., Lizana, P. A., Pino, S., Miller, B. G., & Merino, C. (2020). Augmented reality-based learning for the comprehension of cardiac physiology in undergraduate biomedical students. *Advances in Physiology Education, 44*(3), 314–322. https://doi.org/10.1152/advan.00137.2019

Goodboy, A. K., & Martin, M. M. (2020). Omega over alpha for reliability estimation of unidimensional communication measures. *Annals of the International Communication Association, 44*(4), 422–439. https://doi.org/10.1080/23808985.2020.1846135

Goss-Sampson, M. A. (2020). Bayesian inference in JASP: A guide for students. https://doi.org/10.17605/OSF.IO/CKNXM.

Grieves, M. W. (2019). Virtually intelligent product systems: Digital and physical twins. *Progress in Astronautics and Aeronautics*. https://doi.org/10.2514/4.105654

Harp, S. F., & Mayer, R. E. (1997). The role of interest in learning from scientific text and illustrations: On the distinction between emotional interest and cognitive interest. *Journal of Educational Psychology, 89*(1), 92–102. https://doi.org/10.1037/0022-0663.89.1.92

Hochreiter, J., Daher, S., Bruder, G., & Welch, G. (2018). Cognitive and touch performance effects of mismatched 3D physical and visual perceptions. In *2018 IEEE conference on virtual reality and 3D user interfaces (VR)* (pp. 1–386). IEEE. https://doi.org/10.1109/VR.2018.8446574.

Ibáñez, M. B., Di Serio, Á., Villarán, D., & Delgado Kloos, C. (2014). Experimenting with electromagnetism using augmented reality: Impact on flow student experience and educational effectiveness. *Computers & Education, 71*, 1–13. https://doi.org/10.1016/j.compedu.2013.09.004
Jeffreys, H. (1939). *Theory of probability*. Oxford: Clarendon Press.
Kalyuga, S. (2005). Prior knowledge principle in multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 325–337). Cambridge University Press. https://doi.org/10.1017/CBO9780511816819.022.
Kalyuga, S. (2011). Cognitive load theory: How many types of load does it really need? *Educational Psychology Review, 23*, 1–19. https://doi.org/10.1007/s10648-010-9150-7
Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*(430), 773–795.
Klepsch, M., Schmitz, F., & Seufert, T. (2017). Development andvvalidation of two instruments measuring intrinsic, extraneous, and germane cognitive load. *Frontiers in Psychology, 8*, 1997. https://doi.org/10.3389/fpsyg.2017.01997
Korbach, A., Brünken, R., & Park, B. (2018). Differentiating different types of cognitive load: A comparison of different measures. *Educational Psychology Review, 30*, 503–529. https://doi.org/10.1007/s10648-017-9404-8
Krieglstein, F., Beege, M., Rey, G. D., Ginns, P., Krell, M., & Schneider, S. (2022). A systematic meta-analysis of the reliability and validity of subjective cognitive load questionnaires in experimental multimedia learning research. *Educational Psychology Review, 34*(4), 2485–2541. https://doi.org/10.1007/s10648-022-09683-4
Küçük, S., Kapakin, S., & Göktaş, Y. (2016). Learning anatomy via mobile augmented reality: Effects on achievement and cognitive load. *Anatomical Sciences Education, 9*(5), 411–421. https://doi.org/10.1002/ase.1603
Lee, I. (2020). Using augmented reality to train students to visualize three-dimensional drawings of mortise–tenon joints in furniture carpentry. *Interactive Learning Environments, 28*(7), 930–944. https://doi.org/10.1080/10494820.2019.1572629
Makransky, G. (2021). The immersion principle in multimedia learning. In R. E. Mayer, & M. Fiorella (Eds.), *The Cambridge handbook of multimedia learning* (3rd ed., pp. 296–303). Cambridge University Press.
Makransky, G., Borre-Gude, S., & Mayer, R. E. (2019). Motivational and cognitive benefits of training in immersive virtual reality based on multiple assessments. *Journal of Computer Assisted Learning, 35*(6), 691–707. https://doi.org/10.1111/jcal.12375
Makransky, G., Terkildsen, T., & Mayer, R. E. (2017b). Adding immersive virtual reality to a science lab simulation causes more presence but less learning. *Learning and Instruction, 60*, 225–236. https://doi.org/10.1016/j.learninstruc.2017.12.007
Mann, S., Do, P. V., Furness, T., Yuan, Y., Iorio, J., & Wang, Z. (2023). Fundamentals of all the realities: Virtual, augmented, mediated, multimediated, and beyond. In A. Y. C. Nee, & S. K. Ong (Eds.), *Springer handbook of augmented reality*. Cham: Springer Handbooks. Springer. https://doi.org/10.1007/978-3-030-67822-7_1.
Marques, B., Ferreira, C., Silva, S., Dias, P., & Sousa Santos, B. (2023). Is social presence (alone) a general predictor for good remote collaboration? Comparing video and augmented reality guidance in maintenance procedures. *Virtual Reality, 27*(1), 1783–1796. https://doi.org/10.1007/s10055-023-00770-7
Mayer, R. E. (1989). Systematic thinking fostered by illustrations in scientific text. *Journal of Educational Psychology, 81*(2), 240–246. https://doi.org/10.1037/0022-0663.81.2.240
Mayer, R. E. (1999). Multimedia aids to problem-solving transfer. *International Journal of Educational Research, 31*(7), 611–623. https://doi.org/10.1016/S0883-0355(99)00027-0
Mayer, R. E. (2002). Multimedia learning. *Psychology of Learning and Motivation, 41*, 85–139. https://doi.org/10.1016/S0079-7421(02)80005-6. Academic Press.
Mayer, R. E. (2005). In *The Cambridge handbook of multimedia learning*. Cambridge university press.
Mayer, R. E. (Ed.). *The Cambridge handbook of multimedia learning* (Cambridge University Press).
Mayer, R. E. (2020). *Multimedia learning* (3rd ed.). Cambridge: Cambridge University Press.
Mayer, R. E. (2021). *Multimedia learning* (3rd ed.). Cambridge University Press. https://doi.org/10.1017/9781316941355
Mayer, R. E. (2021b). Evidence-based principles for how to design effective instructional videos. *Journal of Applied Research in Memory and Cognition, 10*(2), 229–240. https://doi.org/10.1016/j.jarmac.2021.03.007
Mayer, R. E., & Anderson, R. B. (1991). Animations need narrations: An experimental test of a dual-coding hypothesis. *Journal of Educational Psychology, 83*(4), 484–490. https://doi.org/10.1037/0022-0663.83.4.484
Mayer, R. E., & Chandler, P. (2001). When learning is just a click away: Does simple user interaction foster deeper understanding of multimedia messages? *Journal of Educational Psychology, 93*(2), 390–397. https://doi.org/10.1037/0022-0663.93.2.390
Mayer, R. E., Hegarty, M., Mayer, S., & Campbell, J. (2005). When static media promote active learning: Annotated illustrations versus narrated animations in multimedia instruction. *Journal of Experimental Psychology: Applied, 11*(4), 256–265. https://doi.org/10.1037/1076-898X.11.4.256
Meyer, O. A., Omdahl, M. K., & Makransky, G. (2019). Investigating the effect of pre-training when learning through immersive virtual reality and video: A media and methods experiment. *Computers & Education, 140*, 1–17. https://doi.org/10.1016/j.compedu.2019.103603
Miller, G. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*(2), 81–97. https://doi.org/10.1037/h0043158
Moreno, R., & Mayer, R. E. (2002). Learning science in virtual reality multimedia environments: Role of methods and media. *Journal of Educational Psychology, 94*(3), 598–610. https://doi.org/10.1037/0022-0663.94.3.598
Morton, J., Zheleva, A., Van Acker, B. B., Durnez, W., Vanneste, P., Larmuseau, C., De Bruyne, J., Raes, A., Cornillie, F., Saldien, J., De Marez, L., & Bombeke, K. (2022). Danger, high voltage! Using EEG and EOG measurements for cognitive overload detection in a simulated industrial context. *Applied Ergonomics, 102*, Article 103763. https://doi.org/10.1016/j.apergo.2022.103763
Mousavi, S. Y., Low, R., & Sweller, J. (1995). Reducing cognitive load by mixing auditory and visual presentation modes. *Journal of Educational Psychology, 87*(2), 319–334. https://doi.org/10.1037/0022-0663.87.2.319
Nakano, Y. (2017). Cognitive and attentional process in insight problem solving of the puzzle game "Tangram". In G. Gunzelmann, A. Howes, T. Tenbrink, & E. J. Davelaar (Eds.), *Proceedings of the 39th annual meeting of the cognitive science society* (pp. 2778–2783). Cognitive Science Society. Link https://escholarship.org/uc/item/74v7p35h.
Oberauer, K. (2022). The importance of random slopes in mixed models for Bayesian hypothesis testing. *Psychological Science, 33*(4), 648–665. https://doi.org/10.1177/09567976211046884
Osborne, J. (2010). Improving your data transformations: Applying the Box-Cox transformation. *Practical Assessment, Research and Evaluation, 15*(12), 1–9.
Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist, 38*(1), 1–4. https://doi.org/10.1207/s15326985ep3801_1
Paas, F., Van Gog, T., & Sweller, J. (2010). Cognitive load theory: New conceptualizations, specifications, and integrated research perspectives. *Educational Psychology Review, 22*, 115–121.
Peterson, L. R., & Peterson, M. G. E. (1959). Short-term retention of individual verbal items. *Journal of Experimental Psychology, 58*(3), 193–198. https://doi.org/10.1037/h0049234
Rey, G. D. (2012). A review of research and a meta-analysis of the seductive detail effect. *Educational Research Review, 7*(3), 216–237. https://doi.org/10.1016/j.edurev.2012.05.003
Rey, G. D., Beege, M., Nebel, S., Wirzberger, M., Schmitt, T. H., & Schneider, S. (2019). A meta-analysis of the segmenting effect. *Educational Psychology Review, 31*, 389–419. https://doi.org/10.1007/s10648-018-9456-4
Roca-González, C., Martín Gutiérrez, J., García-Dominguez, M., & Mato Carrodeguas, M. D. C. (2017). Virtual technologies to develop visual-spatial ability in engineering students. *Eurasia Journal of Mathematics, Science and Technology Education*. https://doi.org/10.12973/eurasia.2017.00625a
Rodenburg, D., Hungler, P., Etemad, S. A., Howes, D., Szulewski, A., & Mclellan, J. (2018). Dynamically adaptive simulation based on expertise and cognitive load. *2018 IEEE Games, Entertainment, Media Conference (GEM)*, 1–6. https://doi.org/10.1109/GEM.2018.8587618
Santos, M. E. C., Chen, A., Taketomi, T., Yamamoto, G., Miyazaki, J., & Kato, H. (2014). Augmented reality learning experiences: Survey of prototype design and evaluation. *IEEE Transactions on Learning Technologies, 7*(1), 38–56. https://doi.org/10.1109/TLT.2013.37
Schroeder, N. L., & Cenkci, A. T. (2020). Do measures of cognitive load explain the spatial split-attention principle in multimedia learning environments? A systematic review. *Journal of Educational Psychology, 112*(2), 254–270. https://doi.org/10.1037/edu0000372
Seeliger, A., Netland, T., & Feuerriegel, S. (2022). Augmented reality for machine setups: Task performance and usability evaluation in a field test. *Procedia CIRP, 107*, 570–575. https://doi.org/10.1016/j.procir.2022.05.027
Sinha, T., & Kapur, M. (2021). When problem solving followed by instruction works: Evidence for productive failure. *Review of Educational Research, 91*(5), 761–798. https://doi.org/10.3102/00346543211019105
Skulmowski, A. (2023). Guidelines for choosing cognitive load measures in perceptually rich environments. *Mind, Brain, and Education, 17*, 20–28. https://doi.org/10.1111/mbe.12342
Skulmowski, A., & Xu, K. M. (2022). Understanding cognitive load in digital and online learning: A new perspective on extraneous cognitive load. *Educational Psychology Review, 34*, 171–196. https://doi.org/10.1007/s10648-021-09624-7
Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*, 257–285.
Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction, 4*(4), 295–312. https://doi.org/10.1016/0959-4752(94)90003-5
Sweller, J. (2011). Cognitive load theory. In J. P. Mestre, & B. H. Ross (Eds.), *Psychology of learning and motivation* (Vol. 55, pp. 37–76). Elsevier.
Sweller, J. (2022). The role of evolutionary psychology in our understanding of human cognition: Consequences for cognitive load theory and instructional procedures. *Educational Psychology Review, 34*, 2229–2241. https://doi.org/10.1007/s10648-021-09647-0
Sweller, J., Van Merriënboer, J., & Paas, F. G. W. C. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*(3), 251–296.
Sweller, J., van Merriënboer, J., & Paas, F. G. W. C. (2019a). Cognitive architecture and instructional design: 20 Years later. *Educational Psychology Review, 31*(2), 261–292. https://doi.org/10.1007/s10648-019-09465-5
Sweller, J., van Merriënboer, J. J. G., & Paas, F. (2019b). Cognitive architecture and instructional design: 20 Years later. *Educational Psychology Review, 31*, 261–292. https://doi.org/10.1007/s10648-019-09465-5
Tabachnick, B. G., & Fidell, L. S. (2013). In *Using multivariate statistics* (6th ed.). Pearson.
Thees, M., Kapp, S., Strzys, M. P., Beil, F., Lukowicz, P., & Kuhn, J. (2020). Effects of augmented reality on learning and cognitive load in university physics laboratory

courses. *Computers in Human Behavior, 108*, Article 106316. https://doi.org/10.1016/j.chb.2020.106316

Vanneste, P., Raes, A., Morton, J., Bombeke, K., Van Acker, B. B., Larmuseau, C., Depaepe, F., & Van Den Noortgate, W. (2021). Towards measuring cognitive load through multimodal physiological data. *Cognition, Technology & Work, 23*(3), 567–585. https://doi.org/10.1007/s10111-020-00641-0

Venkatesh, V., & Davis, F. D. (2000). A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management Science, 46*(2), 186–204. https://doi.org/10.1287/mnsc.46.2.186.11926

Wagenmakers, E.-J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Selker, R., Gronau, Q. F., Dropmann, D., Boutin, B., Meerhoff, F., Knight, P., Raj, A., van Kesteren, E.-J., van Doorn, J., Smira, M., Epskamp, S., Etz, A., Matzke, D., … Morey, R. D. (2018). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review, 25*(1), 58–76. https://doi.org/10.3758/s13423-017-1323-7

Wilson, A. D., & Golonka, S. (2013). Embodied cognition is not what you think it is. *Frontiers in Psychology, 4*, 58. https://doi.org/10.3389/fpsyg.2013.00058

Wu, H. K., Lee, S. W. Y., Chang, H. Y., & Liang, J. C. (2013). Current status, opportunities and challenges of augmented reality in education. *Computers & education, 62*, 41–49. https://doi.org/10.1016/j.compedu.2012.10.024

Yarbus, A. L., & Yarbus, A. L. (1967). Eye movements during perception of complex objects. *Eye movements and vision*, 171–211.

Yu, Z. (2023). Meta-analyses of effects of augmented reality on educational outcomes over a decade. *Interactive Learning Environments*, 1–15. https://doi.org/10.1080/10494820.2023.2205899

Zhou, F., Duh, H. B. L., & Billinghurst, M. (2008). Trends in augmented reality tracking, interaction and display: A review of ten years of ISMAR. In *2008 7th IEEE/ACM international symposium on mixed and augmented reality* (pp. 193–202). IEEE. https://doi.org/10.1109/ISMAR.2008.4637362.