International Journal of Serious Games

ISSN: 2384-8766
https://journal.seriousgamessociety.org/

Article

Dynamic Adaptative Surveillance Training in a Virtual Environment Using Real-Time Cognitive Load and Performance

Andrew J.A. Seyderhelm[1], and Karen L. Blackmore[1]

[1]*School of Information and Physical Sciences, University of Newcastle, Callaghan, NSW Australia*
Andrew.Seyderhelm@uon.edu.au ; Karen.Blackmore@Newcastle.edu.au

Keywords:

Dynamic difficulty adjustment
serious games
simulation training
immersive environments
cognitive load
training performance

Received: January 2024
Accepted: August 2024
Published: August 2024
DOI: 10.17083/ijsg.v11i3.733

Abstract

Dynamic difficulty adjustment of serious games is a field of research seeking to assist participants attain an ideal learning state. To achieve this, the difficulty of a game is adjusted in real-time to approach the capabilities of the player. Many dynamic difficulty adjustment systems only measure a few variables, often solely player performance, and adjust a limited number of in-game aspects. Little research has sought to ascertain if measuring a combination of cognitive load with measures of performance in real-time leads to a more effective dynamic difficulty adjustment system. Building on research which defined 'mental efficiency', we conducted a novel experiment to address this gap. This is achieved by comparing two versions of a surveillance training serious game; one a linear approach and the other with our unique cognitive load and performance-based dynamic difficulty adjustment system. Our experiment included n=52 participants (26 per treatment group). The experiment demonstrates that our approach achieved similar performance outcomes with lower cognitive load, in less time than the linear difficulty approach. These results indicate that our system enhances learning capacity and may prove beneficial for future serious games.

1. Introduction

Serious games are increasingly used in a wide range of fields to meet education and training needs, and to foster greater engagement, interest, and motivation amongst the training participants [1, 2]. Serious games serve a dual purpose of having a serious aim (e.g. learning, training, recruitment, etc.) as well as entertainment value [3, 4]. This study explores a new dynamic difficulty adjustment (DDA) approach combining cogntive load (CL) and performance into a more effective system.

Within serious games research, the application of DDA is a method to adjust gameplay and learning content to meet learner needs [5]. DDA systems adapt in real-time and are informed by various player measures, they include the adjustment of various elements to alter the

difficulty of either, or both, gameplay and learning content. Numerous approaches have been explored and are discussed in Section 2.1 and examined further in a systematic literature review [6]. DDA systems applied to serious games deliver cost-effective training that can approximate one-on-one tutoring, increasing overall effectiveness [7]. However, the application of CL measures, combined with performance scores, is an under-explored area of research in this field, which we seek to address.

CL represents the mental capacity of a person in processing new tasks, learning, or knowledge. Cognitive load theory (CLT) describes the way new information is processed and subsequently stored in long-term memory [8]. This is critical in designing complex learning experiences, where moving outside the limits of CL capacity can have negative impacts on learning outcomes [9]. Integrating CL measures into a DDA system may help inform the system to adjust the game-play or learning content based on real-time cognitive capacity. This paper describes the implementation of a new DDA system combining CL with performance measures, termed the CL and performance DDA (CLP-DDA). The CLP-DDA system uses a matrix to define current player states and applies multiple adaption strategies to adjust variables in-game. The goal is to achieve a serious game experience balanced to the needs of participants. The serious game developed for this experiment is a surveillance training serious game, including two levels as the primary experiment focus (the Park and City). This is described in detail in Sections 2.3, 3.2, and 3.3.

This paper is structured as follows: the first sections present a background and core related concepts of the CLP-DDA and serious games; then introduces the experiment concept, demographic information, and method; next extensively reviews and presents the experiment results; following that discusses the results and what they mean; and finally outlines some limitations and future research ideas.

2. Related Work

Within serious games, a one-size-fits all approach may not be optimal when considering learner prior knowledge, pace of learning, interest, natural talent, and so forth. DDA systems offer personalised experiences when compared to serious games with no adaption [10-16].

A previous systematic literature review demonstrated that multiple adaption strategies were more effective in 3D games than adapting a single element, showing an 89% success rate versus 73% respectively [6]. Thus, a suitably complex serious game was created to test the novel CLP-DDA implementation. This literature review identified only one experiment that explored working memory as part of a DDA system [17], and none that adapted via combined measures of CL and performance; highlighting the novelty of this research.

Previous research combining mental effort and performance measures to understand the relative efficiency of instructional conditions, undertaken in 1993, serves as a forerunner to this work [18]. The research by Paas et al., (1993) measured mental effort with a post hoc questionnaire. They created a formula based on the scores from that questionnaire, combined with test performance, to derive relative efficiency, resulting in an efficiency rating that was determined after the test and mental effort questionnaire had been completed.

The combination of performance and mental effort was explored further in 2004 by Salden, Paas, Broers and Van Merrienboer [19]. This study considered dynamic task selection, where tasks in a training program were automatically selected between levels based on a combined measure of mental effort and performance. Mental effort was scored based on what the players indicated after completing a task. While confirming the usefulness of combined measures in training contexts, these approaches do not allow for real-time adjustments of task difficulty. In contrast, this research presents CL and performance measured during the task, with no interruption to game-play or experience, allowing the serious game to be adapted in real-time (Section 4.1).

Measuring CL, in real-time, via various forms of secondary task has been undertaken in previous studies [20, 21]. There are several methods used to measure CL [22], however to be effective in a DDA system these methods need to work during gameplay, thus excluding post-game surveys. Several physiological methods have been used in serious games to record CL, including EEG systems, eye and pupil movement, or heart rate [22]. Sections 2.2 and 4.1 describe the novel approach taken in this research.

The following sections provide background relevant to this research, including literature on DDA and CLT, and context for the serious game developed here: the Surveillance Training Serious Game (STSG) (Section 3).

2.1 Dynamic Difficulty Adjustment (DDA)

There is a growing interest in adaptive serious games that adjust various elements of the gameplay and/or learning content to help meet the specific needs of the learner [6, 23-25]. These adaptive systems alter various aspects of gameplay and/or learning content, in real-time, to make the experience either more challenging or easier to help foster greater learning success, engagement and flow [6, 26, 27]. Unlike entertainment games, it is necessary to adjust both gameplay and learning content in serious game to achieve optimal outcomes.

Flow theory is prevalent in DDA research [6] recommending difficulty should be balanced to achieve the optimal level of challenge. This is usually described in terms of a flow channel whereby difficulty is increased or decreased to maintain an optimal state (Figure 1) [28] leading to an ideal learning state [27] where participants direct more mental resources to the task at hand, rejecting extraneous distractions, resulting in a more positive learning experience.

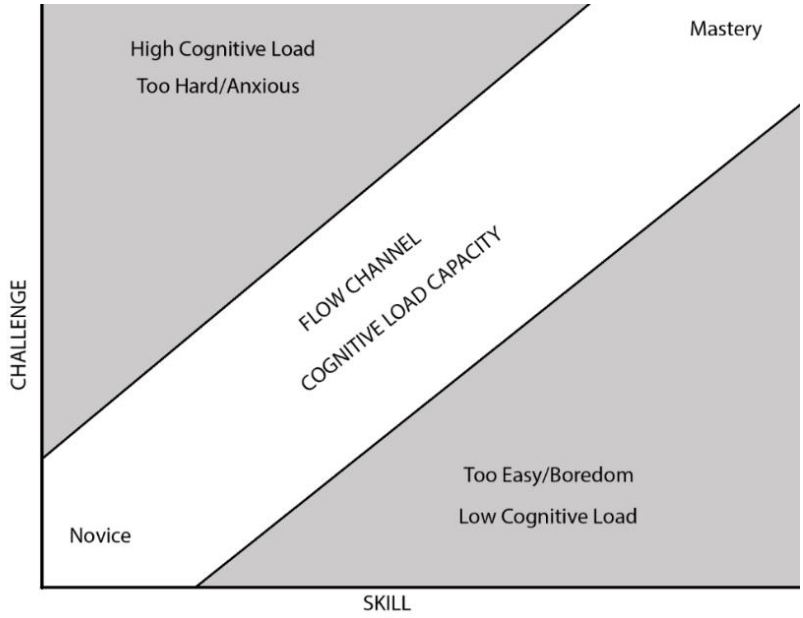

Figure 1. The flow channel overlaid with cognitive load considerations.

DDA systems often measure performance or player affect to trigger the adjustments, however very few consider CL (see Section 2.2) [6]. In an existing systematic literature review of DDA systems in serious games [6] the highest success rate (89%) was for 3D games that employed multiple adaptions, that is when more than one game or learning mechanic were made easier or harder. Additionally, the review identified that there are few applications that use multiple metrics to drive DDA systems, with only 10.2% of experiments using multiple measures to effect changes in the gameplay experience.

Of the 59 games reviewed in the same literature review,there were 23 serious games and 36 entertainment games. 18 of the serious games were successful (78.3%) versus 22 of the entertainment games (61.1%). Most of the entertainment games were 2D (24) versus nearly half of the serious games (11). Therefore, it is likely the greater proportion of success with serious games is due to the greater incidence of multiple adaptions being used in 3D games.

There is a fundamental difference in the purpose of DDA for serious games versus entertainment games because they have a non-entertainment 'serious' purpose (e.g. training, education, or therapy) as opposed to enjoyment or recreation [4]. Therefore, DDA systems for serious games need to adjust both the learning content as well as entertainment aspects. Thus, a multiple measure DDA approach may prove beneficial for serious games as more than one aspect can be measured: performance relating to pedagogical content, performance or response to gameplay, and/or the level of engagement or CL [29] (Figure 2). The serious game designed for this research is sufficiently complex to enable multiple adaptation and performance metrics to be used, including a measure of CL.

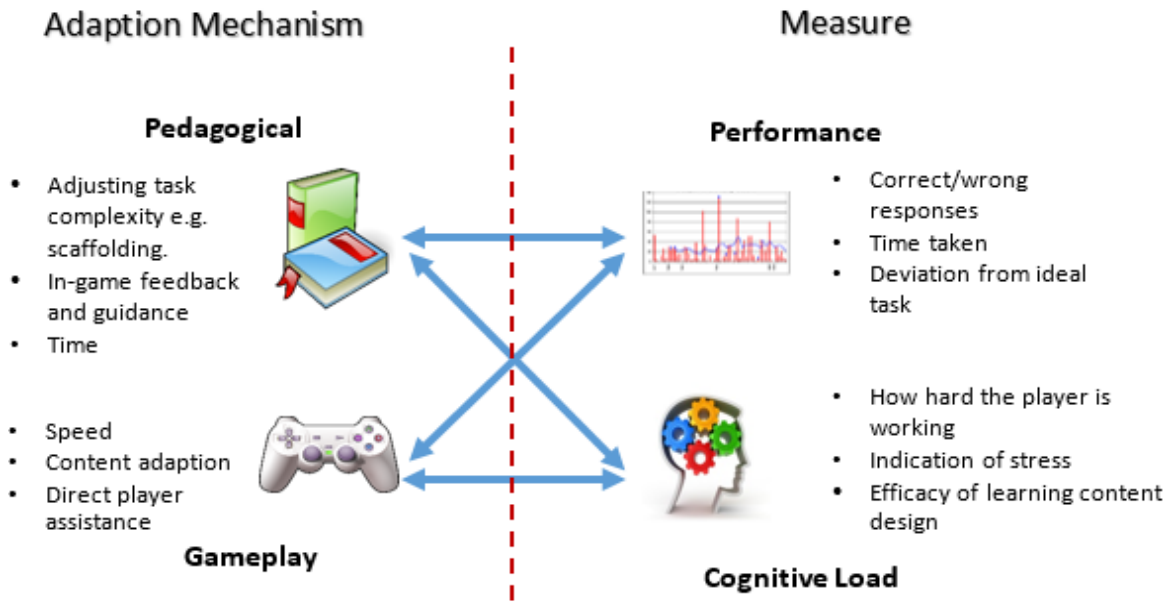

Figure 2. The concept of multiple measures effecting multiple adaptions in the CLP-DDA [28].

2.2 Cognitive load theory and associated measures

CLT is concerned with managing working memory, categorized by three cognitive constructs: intrinsic, extraneous, and germane CL [9]. CLT asserts that the brain has finite resources to deal with new information, when this information becomes familiar it is stored in near unlimited long-term memory resources (termed schema) [9]. When presenting learners with new material, it is important that learning content is presented to avoid overloading the limited CL resources and foster efficient conversion of the learning content into memory schema. CL in serious games and training has been shown to be an indicator of success [30, 31], and therefore we apply a measure of CL in the CLP-DDA system as a key aspect of this experiment [32]. While numerous objective and subjective approaches to the measurement of CL exist [8], we apply a real-time measure validated in complex 3D game environments through a novel variation on a detection response task (DRT), having previously been demonstrated as effective [28]. This modified method is based on the DRT ISO standard [33], and termed the Virtual Detection Response Task (virDRT). It adapts the DRT ISO standard to function with a typical video game controller for an affordable, robust, and effective CL measuring system. This existing virDRT is further adapted here to have five times the resolution of the original versions (Section 4.2).

2.3 Surveillance Training Serious Game

There has been little published research on the effectiveness of serious games in law enforcement training [34, 35], although increasing adoption by law enforcement agencies whas arisen in pursuit of more efficient training more [35, 36]. Thus, developing a serious game that serves a dual purpose of assessing the CLP-DDA and presenting opportunities for future research in the application of serious games in a law enforcement context is of interest.

The foot-based surveillance task implemented here includes the same core controls from a previous virDRT implementation [28], although they are expanded to include some additional surveillance specific tasks (Section 3.2). The STSG goes a step beyond earlier VR surveillance research [37] and provides the player with the opportunity to undertake a surveillance training task within large virtual environments as described in Section 3.3.

In addition to the first author's 12-year background in law enforcement, two key documents helped inform the design of the STSG: Behind the private eye – surveillance tales and techniques [38] and the Perform foot surveillance training package [39]. The former provided context and ideas for the game design, and the latter outlined key areas for performance measures. Two versions of the STSG were developed and evaluated; one version included the CLP-DDA system while the other was non-adaptive, using linear difficulty progression.

2.4 Study Objectives

We combine multiple performance measures with a robust measure of CL, in real-time, to drive the CLP-DDA system with several adaption outputs. The experiment aims to test if adapting multiple game aspects using CL and performance measures is successful. Few studies have compared non-adaptive versus adaptive serious games that incorporate multiple measures and multiple adaptions [40]. The current study directly addresses this by comparing the results of two versions of the STSG; one with linear difficulty and one with the CLP-DDA. The following research questions are posed:

RQ1: Does the combination of cognitive load and performance measures for DDA deliver more equal performance results across levels with different difficulty combined with lower cognitive load?

RQ2: Are there any statistical differences between linear difficulty versus the CLP-DDA approach, across various performance and cognitive measures?

3. Participant demographics & STSG level descriptions

3.1 Demographics

The experiment was run across two weeks in September 2023 at the University of Newcastle Callaghan campus in New South Wales, Australia with participants including staff and students of the University (Figure 3). Prior to commencing the STSG, all participants (*n=57)* completed a demographic and game preferences survey. Participants ranged in age from 18 to 36 years (mean = 22.67, SD = 3.89). Of the 57 participants, 48 identified as male and 9 identified as female. Four participants suffered technical issues and one participant withdrew during the experiment, leading to their data being rejected, leaving 52 participants (43 male and 9 female).

Figure 3. A participant playing the STSG.

3.2 The STSG tasks and levels

Participants were allocated to one of the two versions of the game (linear versus CLP-DDA) using stratified random assignment to ensure equal group sizes (n=26), and they were not informed which version they were playing. In both versions of the game, the virDT was active to capture real-time CL and ensure the game interface was the same for all participants.

The STSG consists of four levels: a tutorial level, a counting task at a train station, a surveillance task in a large park environment, and a similar surveillance task in a city environment. The CLP-DDA was implemented into the Park and City levels only. Table 1 describes the tasks involved in the game.

Table 1. The details of the tasks involved in the Park and City levels.

Task	Description
Follow target	This is the core task within the game and has varying levels of difficulty associated. For this task, the player must follow the target, paying attention to a range of factors while simultaneously maintaining a good distance and avoiding detection.
In-game quizzes	Throughout the game the player is asked questions that relate to environment or target actions. For example, what was next to the person the target stopped by, what is the name of the street, what was the name of a café the target stopped at, etc.
Take Photos	Use a virtual camera to take photos of the target. The camera system is a simplified representation of camera functionality and includes a screen overlay and a single button click to take the photo.
Stay Safe	Avoid traffic in the City level when required to cross the roads

3.3 Level Descriptions

The tutorial level is set in a warehouse environment and a 3D non-player character (NPC) provides instructions to the player involving following a target and practicing their skills with a camera (Figure 4).

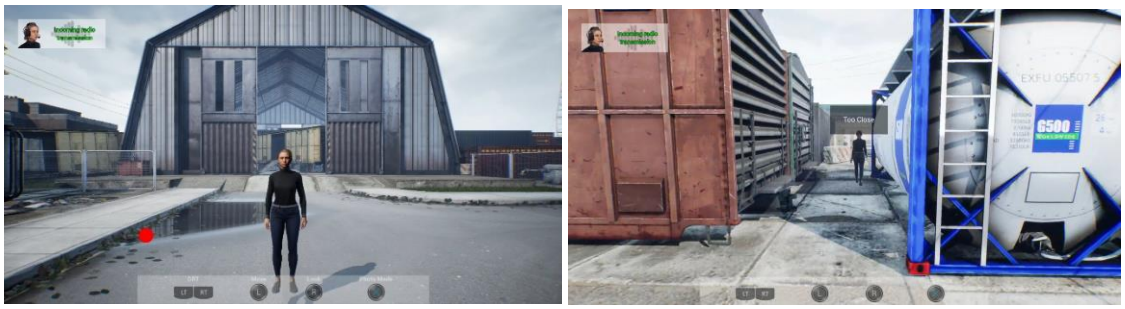

Figure 4. The start of the tutorial level (Left) and learning follow distances (right).

The first level after the tutorial is based in a train station where players must count target groups within waves of disembarking passengers. Due to time limitations, a lack of level balancing led to the results of this level being discarded.

The Park level is the first of the two levels in which the CLP-DDA system was implemented. This is a large level, commencing in a café courtyard, and then progressing on a long winding path through various park locations (Figure 5).

Figure 5. Various dynamic game scenes from the Park level; note the red dot from the virDRT in the bottom right.

The final level is the City level, a relatively busy European style city with roads, pedestrian crossings, cafes, and cars (Figure 6), and containing the same tasks as the Park.

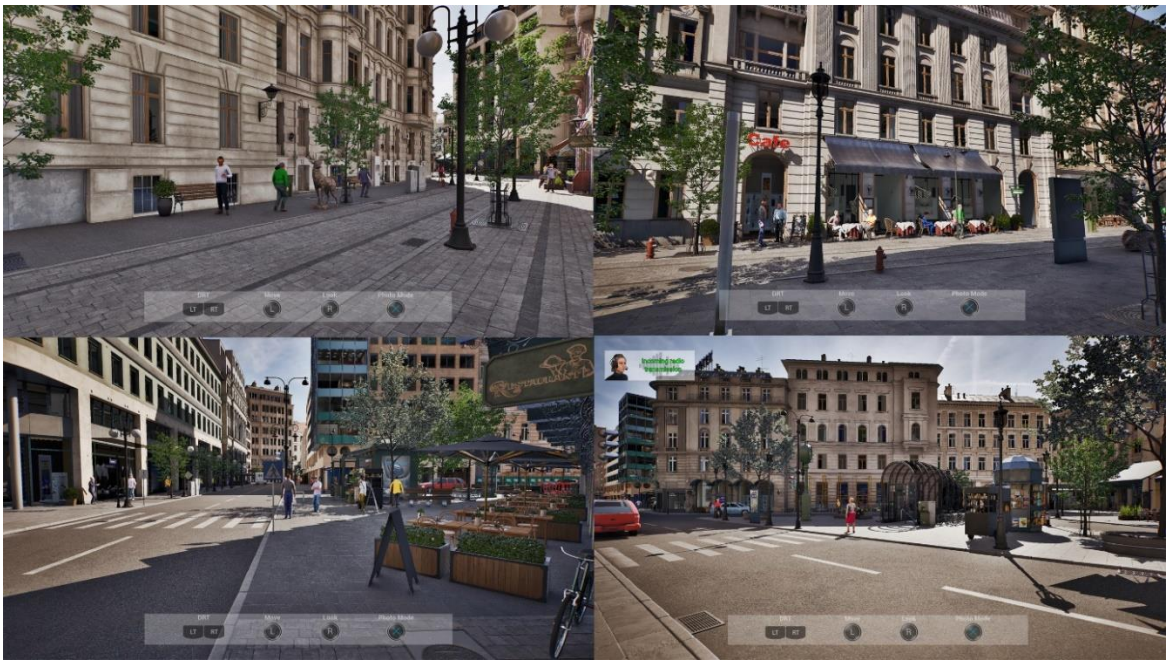

Figure 6. Various dynamic game scenes from the City level.

4. Methods

For this experiment, the results are expected to show less variation in performance scores and CL in the CLP-DDA version of the game compared to the linear adaption variant. This is because the CLP-DDA version aims to adjust to the individual needs of each player and make it more challenging for those doing well and vice versa. Experiment design and implementation was approved by the Human Research Ethics Committee (HREC) of the University of Newcastle, Australia (Approval Number: #H-2020-0069).

Before commencing the STSG, each player received a briefing that outlined the length of the experiment session, introduction to the task and game controls, and explained the priority of tasks in-game (also reiterated in the tutorial). This included: following the target maintaining a good distance and not being seen; taking photos and answering awareness quizzes equally important; and responding to the virDRT as the least important of the tasks.

4.1 Cognitive load and performance dynamic difficulty adjustment system

The CLP-DDA system combines CL with performance metrics consisting of follow distance, target awareness, quiz scores, photography, and player safety. (see Section 4.3). These measures are assessed to determine if performance and CL are high, medium, or low. These ratings were compared to a matrix to adapt difficulty (Figure 7). This matrix has similarities to the approach taken by Camp, Paas, Rikers and van Merrienboer [41], with the primary difference being this matrix is used in real-time and informed by the virDRT, whereas their approach was used between tasks and is based on subjective post task questionnaires.

	Low Cognitive Load	Medium Cognitive Load	High Cognitive Load
High Performance	Player is finding the task easy, so increase the challenge. **+2 Difficulty**	Player is close to mastering content, slightly harder. **+1 Difficulty**	Player is doing well, however they have not yet mastered content - more time needed. **No Change**
Medium Performance	Player is finding the task easy – increase difficulty. **+ 1 Difficulty**	Player is competent and using average CL. No changes. **No Change**	Player is finding it challenging and only scoring ok – make easier. **-1 Difficulty**
Low Performance	Likely player has disengaged – this could be for many reasons. Consider other intervention. **No Change**	Player has average CL, but performing poorly, decrease difficulty. **-1 Difficulty**	Player struggling, decrease difficulty. **-2 Difficulty**

Figure 7. Adaption Matrix for the Surveillance Training Serious Game (STSG)

4.1.1 Rules for the CLP-DDA matrix

Every 20 seconds, the current primary task performance score (a combination of follow distance score added with the target's awareness of the player) is converted to low, medium, or high that informs the difficulty adjustment. There are five potential changes of difficulty per 20 second cycle: 0 (no change), +1 (slightly harder), +2 (much harder), -1 (slightly easier), and -2 (much easier). A limit of three adjustments for any specific CLP-DDA mechanism is implemented to provide a ceiling difficulty level. For example, the target's walk speed increases and decreases depending on the requirement of the CLP-DDA; this speed is set to a maximum of 95% of the player maximum walk speed to prevent issues where the player cannot catch-up to the target if they fall behind.

4.2 Cognitive Load measure – the Virtual Detection Response Task (virDRT)

The experiment implements the virDRT approach to real-time measurement of CL following the design detailed in previous work [28]. The virDRT is a secondary task that measures reaction time relating to a stimulus while performing a primary task. A decrease in the response time to the stimulus indicates an increased cognitive burden from the primary task [42]. When appropriately implemented, DRT systems have minimal impact on CL or on primary task performance [43]. A wide range of secondary task methods have been developed in different experiment settings [20, 21, 43-45]. The virDRT used in the STSG experiment is similar to the remote DRT described by Harbluk et., al. [46], however, it is integrated into the game controller via shoulder button presses rather than via a separate finger switch [46]. The virDRT is designed to be noticeable without being obtrusive, with the player tasked to respond to the virDRT as the least important element of the STSG.

The virDRT records the reaction time (RT) to the stimulus, in this case a red dot to the lower left of the screen, as well as the hit rate (HR), which is the number of times the player successfully responded to the stimulus within the allotted time. The RT records the result of a hit from 100ms to 2500ms; successful responses are recorded as HR = 1 and misses as HR = 0. The ISO standard requires a minimum of five data points (hits or misses) to provide valid CL measurement [33].

It is important to consider both the response time and the misses as both provide information on current cognitive burden [33, 47]. The inverse efficiency score (IES) [47, 48] was selected to derive a single CL value as it incorporates both hits and misses in its calculation. The IES includes the virDRT reaction time (RT) and instances where the virDRT signal is not responded to, termed the proportion of errors (PE). The IES is expressed as:

$$IES = RT/1\text{-}PE.$$

The virDRT system records response time as well as PE then applies an IES rating every five activations. The aim of the CLP-DDA system is to return a low, medium, or high CL value. To achieve this rating, the data from [28] was arranged into thirds (Table 2) and used as the thresholds for the STSG.

Table 2. IES percentile score ranges to inform the rating system, High, Medium and Low.

Value	Percent in quantile	CL Rating
0.1 – 0.670	32.99%	low
0.671 – 1.089	34.03%	medium
1.09+	32.99%	high

The DRT method detailed in the ISO standard results in a valid measure every 5 DRT activations. However, this results in a CL measure of once approximately every 25 seconds.

This lacks detail for the CLP-DDA system and a multi-channel version of the virDRT was devised to provide a valid measure approximately every 5 seconds (Figure 8).

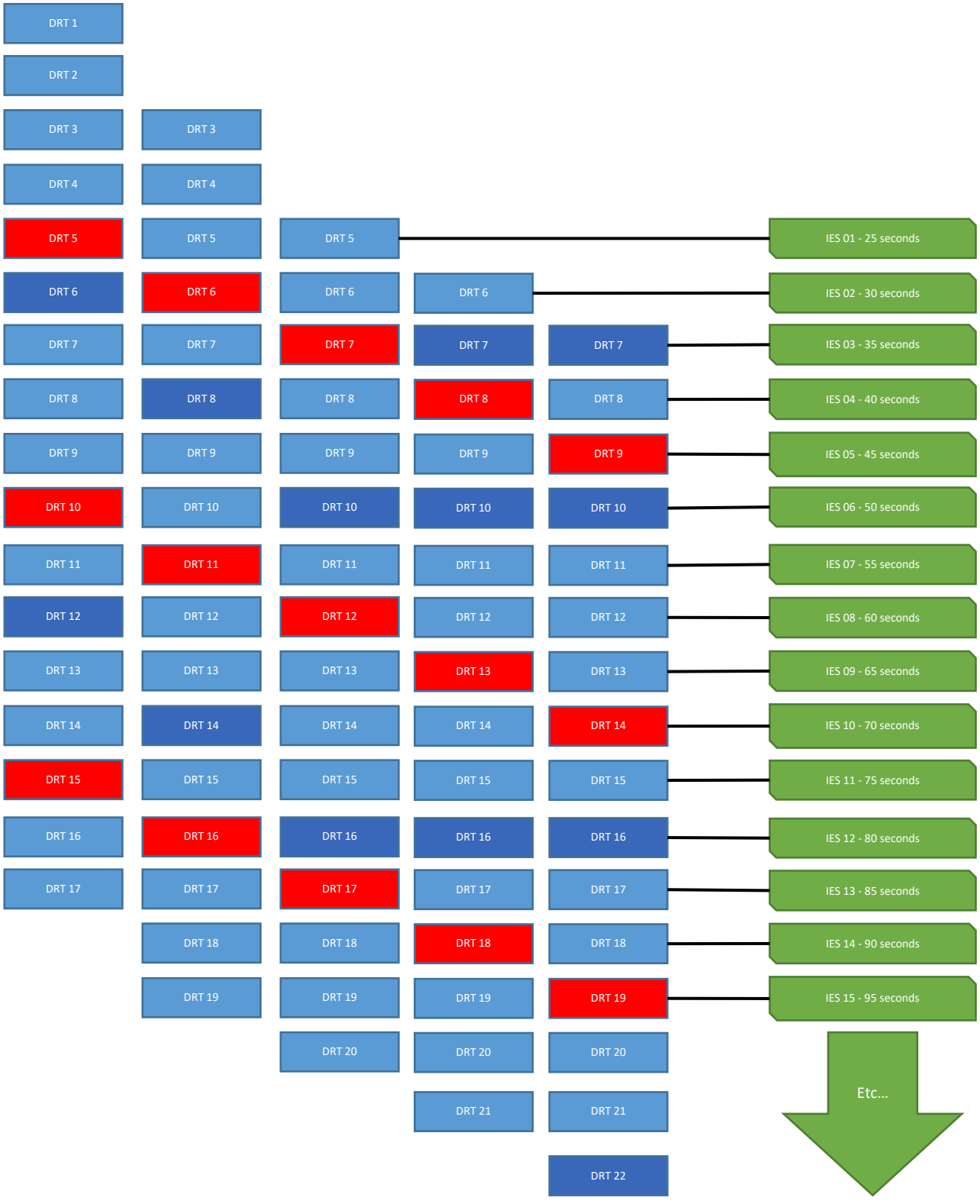

Figure 8. The multiple channel virDRT system developed for the CLP-DDA

4.3 Performance measures

The core elements of foot-based surveillance were identified and refined to be achievable in the STSG. These tasks were developed into systems and became the primary and secondary performance tasks (Table 3).

Table 3. The key performance metrics within the STSG.

Task Priority	Task summary	Measure	Description
Primary Task	Follow target	Distance from player to target	The player was tasked to follow a target maintaining a distance range – neither too close nor too far (10-22.5m).
Primary task	Target awareness	Peripheral vision and clear vision system	A target vision system was developed which produced a score if the target could "see" the player with either peripheral vision or in plain sight.
Secondary task	Take photographs	Framing of the target	The player was tasked to take photos, the first photo was rated based on the framing of the target in the camera view.
Secondary task	Environment awareness	In-game quizzes	At points in each level the game paused and the player was asked questions about aspects in the environment and given a correct/incorrect result.
Secondary Task (only in the City)	Safety	A minus score for being hit by a vehicle	Due to the nature of the City with the target crossing multiple roads and vehicles driving on streets. The player received a penalty if they we hit by a car.

4.4 Dynamic difficulty adjustment system

All elements detailed in the previous sections were combined to inform the CLP-DDA system. This system adapted four elements in the STSG to increase or decrease the challenge level (Figure 9 and Table 4). The categories of adjustment chosen were identified from an earlier systematic literature review [6], with time, speed, in-game feedback, and challenge adjustment (Table 4) adopted for this research.

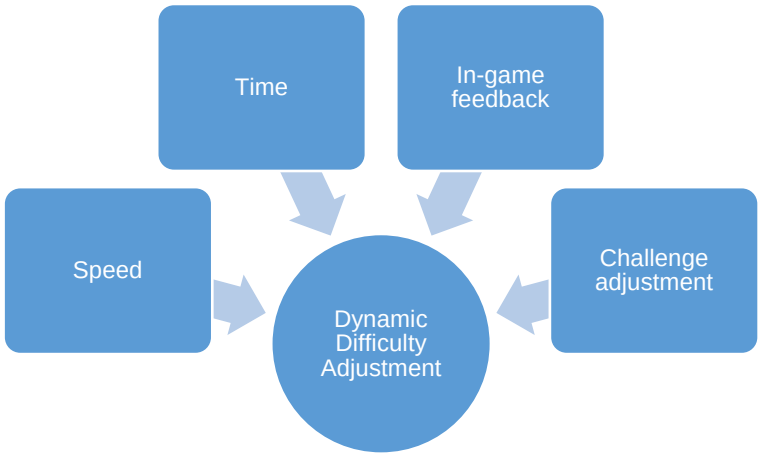

Figure 9. A representation of the elements that inform the CLP-DDA

Table 4. DDA adjustment techniques applied in the STSG.

Adjustment Technique	Specific Elements Adjusted
Speed	The walking speed of the target was made slower or faster. The maximum speed of the target was set at no more than 95% of the player speed; this was to enable the player to catch-up if they fell behind. This also ensured the player was required to periodically adjust their speed and distance.
Time	Time is adjusted to give the player more, or less, opportunity to achieve success. In the STSG this was achieved by altering the path length the target followed. There are three intersecting path lengths, short, medium, and long. If the player is performing well, or poorly, the target may switch paths to make the game longer or shorter.
In-game feedback	In-game feedback was provided via different hints, e.g. the player was getting too close or too far away from the target, and an admonition to pay more attention to their surrounds if they answered an environment awareness question incorrectly.
Challenge adjustment	Challenge adjustment was achieved by increasing or decreasing the number of pedestrians in the levels. More people makes it harder to follow the target and increases the chance of losing the target, with fewer people the follow task is easier.

5. Results

In this section the results of the STSG experiment are explored in detail. First an overview of the scoring system is presented (Section 5.1), followed by a breakdown of each of the core measures with statistical analysis applied to the results.

5.1 Overview of the scoring system

Each of the individual task performance scores were converted into decimals and then divided by the maximum score possible. Decimals were used to account for the different game lengths, due to the CLP-DDA system and player actions, to achieve normalized scores (0 to 1).

The two primary tasks, following the target and staying out of the target line of sight, were grouped together to form a single score as they are both critical in successfully completing the primary task. Following this, the combined primary task score was multiplied by three to weight its higher importance.

5.1.1 Distance to the target - following.

The player received one point for every second they maintained a good distance from the target (between 10m and 22.5 m) and lost a point for each second either too close (< 10m) or too far away (>22.5m). This distance was arrived at by gameplay testing and feedback rather than a distance defined by surveillance methodology.

5.1.2 Staying out of the target's view

The player received 1 point for each second outside of view from the target's perspective. Conversely, they lost a point if they strayed into the target's peripheral vision and two points if they were in clear view.

Being out of view was defined using a two-step process: first, a check is conducted to see if the player is within the targets field of view (defined as clear vision and peripheral vision) and then another one is conducted to see if there is an unobstructed line of sight.

Clear vision is defined as a 90° arc from the targets front, with peripheral vision extending from 90° - 200° (Figure 10). If the player fell in either of these arcs, a test is run to check if the player is in view, or if intervening obstructions provide cover.

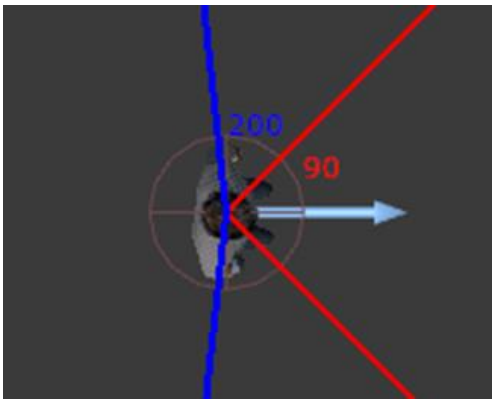

Figure 10. A top-down view of the target demonstrating the fields of view, the red represents clear view; the blue is peripheral vision, the arrow is the front facing direction of the target.

The visibility state of the player is calculated by a series of primary and secondary points. Primary points are red and secondary are blue (Figure 11 left). If the player is within clear vision, then all the red and blue points are considered in the calculation. If the player is in the target's periphery, then only the red points are considered in the visibility calculation.

Red points are given a value of 0.5 and blue points 0.3, if the player returns a visibility value of greater than 1.0 then they are deemed to be visible. The more obstructed the target's view, the lower the score, and the less visible the player (Figure 11 right). This system was designed to allow the player to hide in bushes or behind other obstructions to take photos or observe target actions without being "seen".

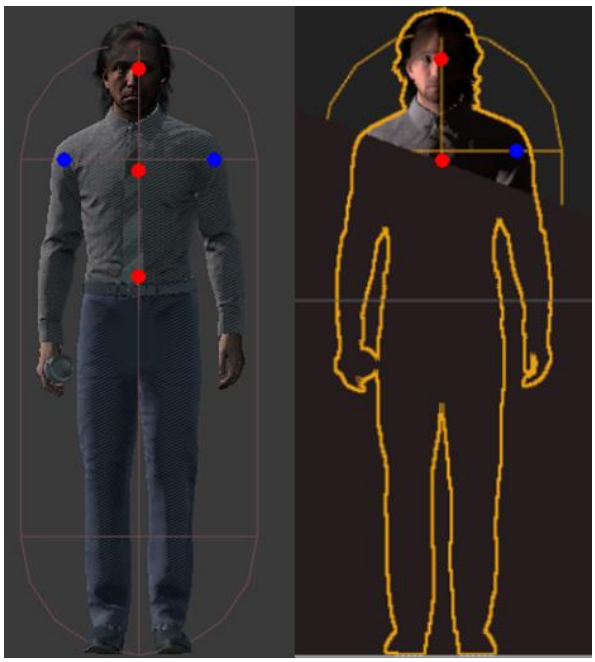

Figure 11. (Left) the red and blue dots representing visibility calculation points. (Right) an example of how occlusion will eliminate points from the visibility calculation.

5.1.3 Primary tasks score

At the end of the level points from following were totaled and divided by twice the number of data points (which represent the maximum score the player could have achieved if they had performed flawlessly).

5.1.4 *Taking Photos of the target*

Photography is a secondary task involving taking a virtual photo of the target when they stop. The score was derived from the quality of the first photo taken per stop. The quality of the photo was defined using the visibility system in reverse, where the target has the primary (red) and secondary (blue) visibility points (Figure 11). The more points visible to the camera when the photo is taken, the higher the score. The total score the player achieved for each level was divided by the maximum possible score, normalising results to a value between 0 – 1.

5.1.5 *Environmental awareness and traffic safety*

In real-world foot-based surveillance, an officer is required to follow a specified target. Throughout this activity they would be required to report on locations and target actions, for example, street names and direction the target takes, descriptions of persons-of-interest the target interacts with, and so forth. This type of information can be described as environmental awareness, whereby the surveillance officer needs to be cognizant of a range of environmental information. To simulate this players were periodically presented with multiple-choice questions testing their observaiton skills. As with the photography scores, the player's score was divided by the maximum possible score to return a result between 0 and 1. In the City level, an additional 'safety' metric was introduced which assesses avoiding being hit by vehicles. Each time a player was struck by a vehicle, 0.25 points were deducted from their score.

5.1.6 *Total combined scores with cognitive load*

The scores for each of the tasks were then combined to produce a total performance score (TPS), where TPS = $(F \times 3) + P + Q$, and F=follow score, P=photo score, and Q= quiz score. For an overall final score that draws together both CL and performance, the TPS for each player is divided by their CL IES average.

5.2 Cognitive load results

CL IES scores were calculated for the Park and the City levels. Figure 12 presents the mean CL for each level and standard deviation (SD).

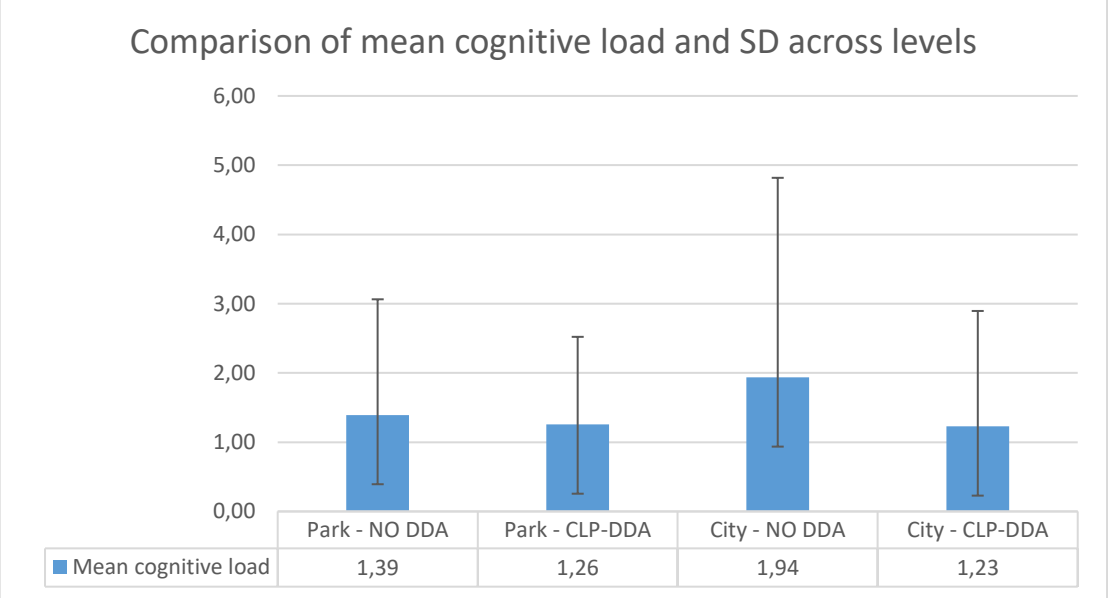

Figure 12. Mean cognitive load and SD results per level.

The City level resulted in higher mean CL compared to the Park level in the No-DDA trial, indicating that the City level is cognitively harder than the Park. Similar CL was recorded for both levels of the CLP-DDA trial, indicating the CLP-DDA was working effectively. CL was lower for the City CLP-DDA level versus the No-DDA version.

By adjusting the challenge to the needs of the player, the optimal level of difficulty is achieved in relation to their cognitive capacity. That is, if the STSG is too hard and their CL was high, the difficulty was reduced, as per the matrix in Section 4.1 (Figure 7) and vice versa, aligning to the CL ratings described in Section 4.2 (Table 2). The results (Figure 12) indicate this was successful, as each CLP-DDA level had lower CL than its NO-DDA counterpart and with reduced SD. The SD results for the No-DDA levels were City = 2.88, Park = 1.67, a difference of 1.21. The CLP-DDA levels were CLP-DDA City = 1.67, CLP-DDA Park = 1.27, a difference of 0.4. The two CLP-DDA levels were balanced sufficiently to have very similar levels of CL to each other even though the levels were inherently different in cognitive complexity (as demonstrated by the linear difficulty version results), demonstrating that the CLP-DDA was working effectively to homogenize results.

A series of two tailed *t*-tests were conducted with the null hypothesis that there is no difference in CL between the levels. H^0*:* Level1 = Level 2, with an *alpha value* of 0.05 (Table 6):

Table 5. Results of the two tailed *t*-test conducted across levels to determine if cognitive load measurement via the virDRT impacts on task performance.

Null hypothesis Question (*H*)	Result	Significance
Park No-DDA = Park CLP-DDA	$t(51) = 0.32$, $p = .75$	no significant difference
City No-DDA = City CLP-DDA	$t(51) = 1.05$, $p = .31$	no significant difference
Park No-DDA = City No-DDA	$t(51) = -2.09$, $p = 0.046$	Statistically significant
Park CLP-DDA = City CLP-DDA	$t(51) = 0.13$, $p = 0.90$	no significant difference

The results indicate the only statistically significant difference is between the two levels with No-DDA. The mean scores of all the results indicate that the City level with No-DDA was the most cognitively challenging level.

No significant difference is recorded in the comparison between the two City levels, yet the difference in CL scores is large, with a mean of 1.94 for the NO-DDA City level and 1.23 for the CLP-DDA level. The SD between the City levels is also large, with the NO-DDA City having an SD of 2.88 versus 1.67 in the CLP-DDA City.

Typically, any value more than three times the SD over the mean is regarded as an outlier [49]. In this case, the No-DDA version had an outlier threshold of: mean (1.9359) + 3 x SD (2.8814) = 10.5801. The same formula is applied to the CLP-DDA version: mean (1.2270) + 3 x SD (1.667) = 6.228. In each set of data, one outlier was found based on this method: No-DDA = 11.8906 and CLP-DDA = 9.2420.

A 90% Winsorization pass [50] was undertaken, ignoring the bottom 5% as there were no outliers in this range; it is common to replace the outliers with the top 5% percentile value [51]. The Winsorization values are: No-DDA = 8.36, and the CLP-DDA = 1.67. As such the two outliers were replaced with these values respectively.

A new two tailed *t*-test was undertaken with the null hypothesis that there is no difference in CL between the levels. H^0*:* Level1 = Level 2, with an *alpha value* of 0.05. The result is: $t(51) = 2.05$, $p = 0.05$. This result is marginal but suggests the differences are more significant than with the outliers unaltered. A trend in this experiment shows that the No-DDA versions result in higher CL, therefore a one tailed *t*-test was also conducted using the same Winsorized values as the two tailed *t*-test. This resulted in: $t(51) = 2.05$, $p = 0.03$, indicating these differences are significant and that the CLP-DDA system was effective at lowering CL scores compared to the NO-DDA levels, meeting a core goal of the experiment.

5.3 Follow task (primary task) performance results.

The primary task in the STSG is to follow the target and maintain a suitable distance and simultaneously stay out of the target's view (detailed in Sections 5.1.1 and 5.1.2). Figure 13 shows the mean scores and SD for each level.

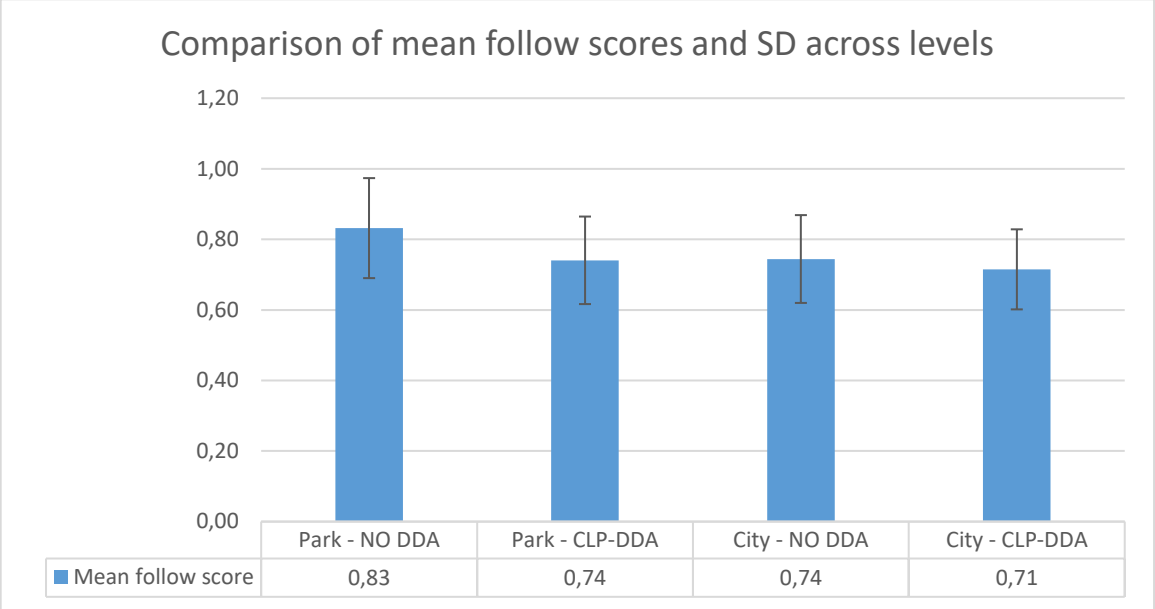

Figure 13. A comparison of the mean follow scores per level, and the SD.

Like the measure of CL, the No-DDA version in the City versus the Park level demonstrate similar difficulty. Yet, in the CLP-DDA version, there is some change in performance. A series of two tailed *t*-tests were conducted with the null hypothesis that there is no difference in follow scores between the levels. H^0*:* Level1 = Level 2, with an *alpha value* of 0.05 (Table 6).

Table 6. Results of the two tailed *t*-tests conducted across levels for follow scores.

Null hypothesis Question (*H*)	Result	Significance
Park No-DDA = Park CLP-DDA	$t(51) = 3.07$, $p = .005$	Statistically significant
City No-DDA = City CLP-DDA	$t(51) = 0.87$, $p = .39$	no significant difference
Park No-DDA = City No-DDA	$t(51) = 3.68$, $p = .001$	Statistically significant
Park CLP-DDA = City CLP-DDA	$t(51) = 0.91$, $p = .37$	no significant difference

Two results showed clear statistical differences; Park No-DDA versus Park CLP-DDA and Park No-DDA versus City No-DDA, indicating that the No-DDA City level is significantly more difficult than the No-DDA Park level, and the Park with CLP-DDA is more difficult than the No-DDA Park. This is important as it demonstrates the CLP-DDA system working as designed, that is, the Park level is significantly less difficult than the City level, thus the CLP-DDA system made the Park level more challenging. This brings it closer in-line with the difficulty of the City level, balancing the difficulty of the two levels. This is evidenced in the mean scores detailed in Figure 13, where the Park (CLP-DDA), City (NO-DDA), and City (CLP-DDA) resulted in very similar scores. Interestingly, the CLP-DDA system reduced the SD for both the Park and City levels. Exploring this further a series of equality of variance *f*-tests were conducted with the null hypothesis that there is no difference in follow scores between the levels. H^0*:* Level1 = Level 2, with an *alpha value* of 0.05. None of these results were statistically significant, however the results do show a slight reduction in variance for the CLP-DDA versions of the levels versus the No-DDA versions: No-DDA Park (0.021), CLP-DDA Park (0.016), No-DDA City (0.016), and CLP-DDA City (0.013).

5.4 Photography performance scores

In addition to following the target, the player was given two secondary tasks; the first of which is to take photographs of the target when they stop. Figure 14 details the mean and SD results.

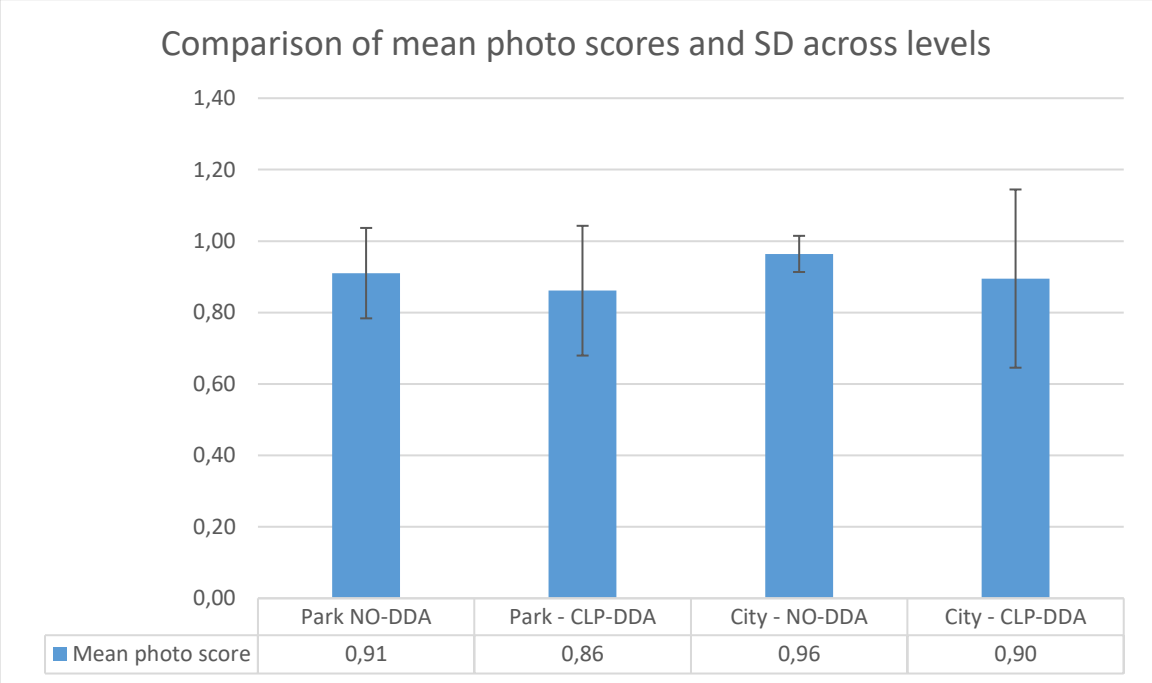

Figure 14. A comparison of the mean photography scores per level, and the SD.

The photography performance scores are similar, reflecting the simple mechanic implemented in the game. However, as with previous levels, there is a larger difference between the two No-DDA levels than the CLP-DDA version. A series of two tailed *t*-tests were conducted with the null hypothesis that there is no difference in photography performance scores between the levels. H^0*:* Level1 = Level 2, with an *alpha value* of 0.05 (Table 7). Again, the No-DDA version resulted in a significant difference whereas there was no difference in the CLP-DDA. This demonstrates that the CLP-DDA system was successful in moderating the challenge level to help players achieve consistent results.

Table 7. Results of the two tailed *t*-test conducted across levels for photography performance scores.

Null hypothesis Question (*H*)	Result	Significance
Park No-DDA = Park CLP-DDA	$t(51) = 1.00$, $p = .33$	no significant difference
City No-DDA = City CLP-DDA	$t(51) = 1.37$, $p = .18$	no significant difference
Park No-DDA = City No-DDA	$t(51) = -2.08$, $p = .048$	Statistically significant
Park CLP-DDA = City CLP-DDA	$t(51) = -1.15$, $p = .26$	no significant difference

5.5 Environmental Awareness Quiz performance scores

Figure 15 details the mean quiz scores and SD for each level.

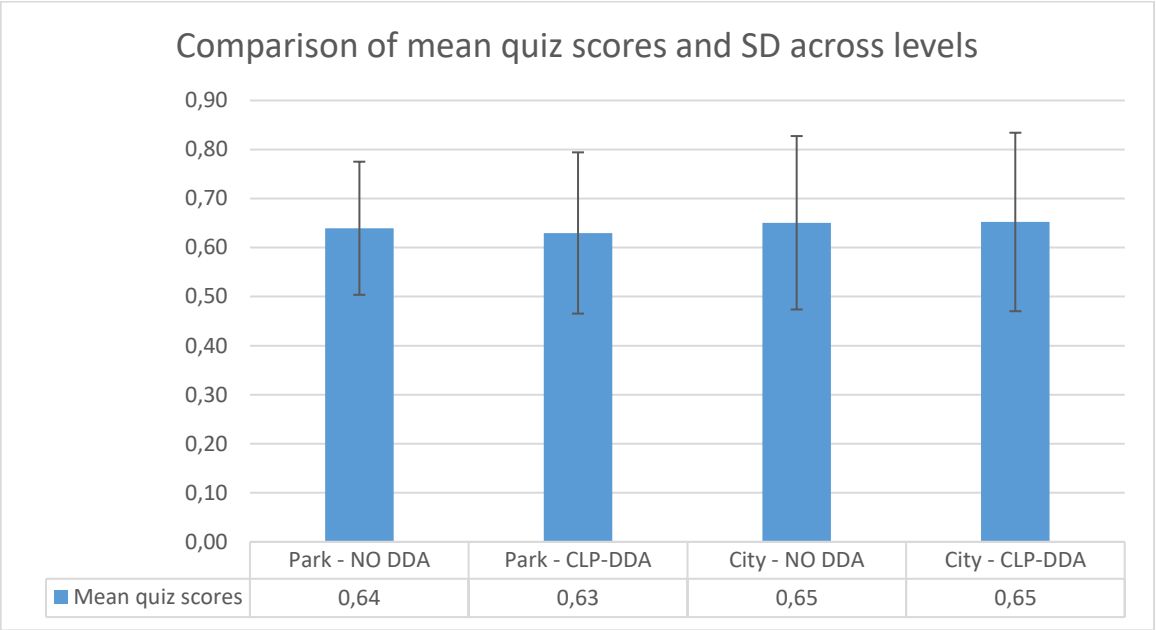

Figure 15. Comparison of the mean quiz scores and standard deviation for each level.

The CLP-DDA system had little impact on performance in the environmental awareness quizzes. A series of two tailed *t*-tests were conducted with the null hypothesis that there is no difference in quiz scores between the levels. H^0: Level1 = Level 2, with an *alpha value* of 0.05 (Table 8):

Table 8. Results of the two tailed *t*-test conducted across levels for quiz scores.

Null hypothesis Question (*H*)	Result	Significance
Park No-DDA = Park CLP-DDA	$t(51) = 0.22$, $p = .83$	no significant difference
City No-DDA = City CLP-DDA	$t(51) = -0.04$, $p = .97$	no significant difference
Park No-DDA = City No-DDA	$t(51) = -0.24$, $p = .81$	no significant difference
Park CLP-DDA = City CLP-DDA	$t(51) = -0.59$, $p = .56$	no significant difference

There were no statistically significant differences between the levels regarding environmental awareness quiz scores.

5.6 Combined cognitive load and performance score results.

The final analysis of results considers the combined score, with weighting for the primary task (Section 5.1.6). Figure 16 details the combined score averages.

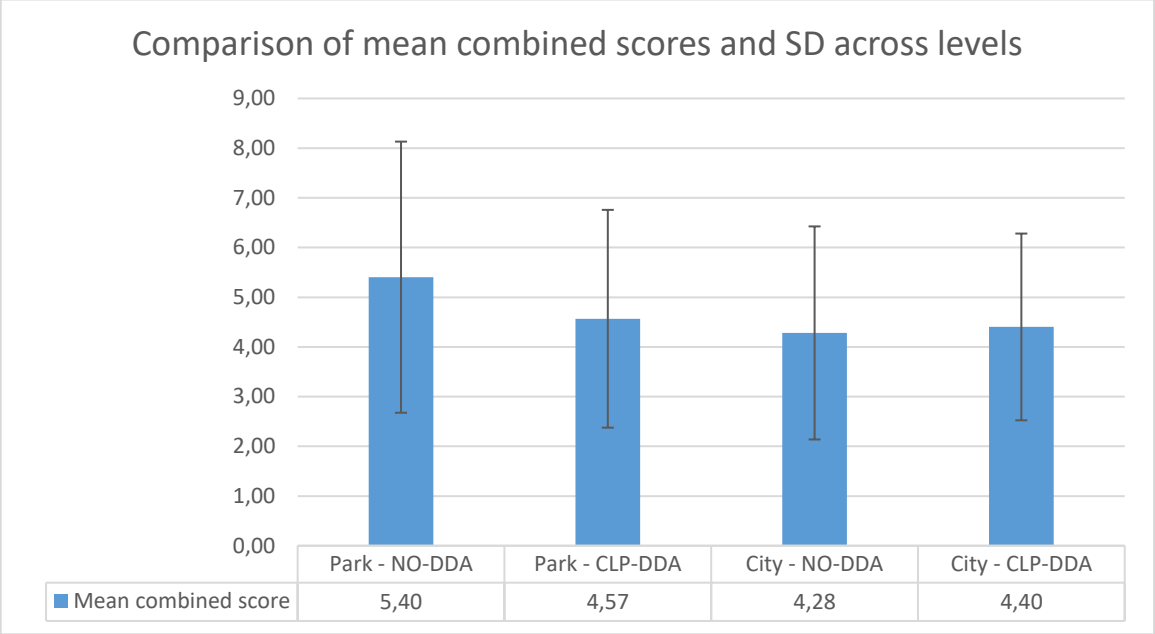

Figure 16. Comparison of the final combined and standard deviation scores for each level .

Interestingly, the NO-DDA City level is harder than the NO-DDA Park level, with the NO-DDA City average being significantly worse than the NO-DDA Park level (Table 9). Differences in scores between the CLP-DDA levels are not significant, score results are subjected to further two tailed *t*-tests (Table 9).

Table 9. Results of the two tailed *t*-test conducted across levels for combined total scores.

Null hypothesis Question (*H*)	Result	Significance
Park No-DDA = Park CLP-DDA	$t(51) = 1.21$, $p = .24$	no significant difference
City No-DDA = City CLP-DDA	$t(51) = -0.22$, $p = .83$	no significant difference
Park No-DDA = City No-DDA	$t(51) = 4.19$, $p = .0003$	Statistically significant
Park CLP-DDA = City CLP-DDA	$t(51) = 0.41$, $p = 0.69$	no significant difference

Crucially, there is a significant difference between the NO-DDA levels. The combined scores show that there is a large difference between these levels with the Park being significantly easier than the City. Furthermore, the SD for the NO-DDA levels is noticeably different, with a greater spread of results in the Park level (City = 2.14, Park = 2.73, a difference of 0.58). However, these differences are less in the CLP-DDA levels (CLP-DDA City = 1.88, CLP-DDA Park = 2.19, a difference of 0.31), an analysis of variance was conducted on these results. The variance scores for each level were found: No-DDA Park (7.74), CLP-DDA Park (4.99), No-DDA City (4.78), and CLP-DDA City (3.67). These variance results show that the CLP-DDA version had lower variance than the No-DDA versions, however a series of *f*-tests were conducted that revealed these variances were not statistically significant.

5.7 Time taken to complete each level

The time taken to complete each level provides useful performance insights. The CLP-DDA system adjusted two key mechanisms that led to varied game times. One was the path length, where a player doing well would move onto a shorter path reducing game time, and vice versa. Additionally, target walking speed was increased or decreased dynamically, also impacting game duration, with results shown in Table 10.

Table 10. Average time and standard deviation (SD) in seconds for each level.

	No-DDA Park	No–DDA City	CLP-DDA Park	CLP-DDA City
Average time	874 seconds	705 seconds	723 seconds	453 seconds
SD	20.97	14.62	37.32	81.07

Of note, the time taken was less for the CLP-DDA version compared to its respective No-DDA counterpart. This may have ramifications in real-world training as time can have an impact on costs, engagement, and performance. Reducing the time and achieving suitable performance with lower CL, may indicate a more effective serious game solution than alternatives that do not employ the CLP-DDA system. A factor of effectiveness arises from working memory resource depletion where working memory capacity reduces over time from sustained mental effort [52]. This means that long periods of high mental effort without breaks can reduce cognitive capacity, therefore achieving success in shorter periods of time may be more effective in complex learning tasks.

Standard deviation in the CLP-DDA instances was greater for both levels compared to the NO-DDA versions, indicating the CLP-DDA system was accounting for both lower and higher performance and CL, and adapting the time required commensurate to player needs, this means if correctly working the SD for time should have greater variance. Thus, players that needed more, or less, time were provided what they required, demonstrating the CLP-DDA time adaption working effectively.

While the CLP-DDA system still needs further development, it shows promise as core results demonstrated more equal player scores across levels of different difficulty with lower CL and provided more detailed CL metrics to help inform debriefing.

6. Discussion

The aims of this study were to assess if the combination of CL and performance measures for DDA deliver more equal performance results across levels with differing difficulty, and with lower CL. Also, if there were any statistical differences between the linear NO-DDA version versus the CLP-DDA version. Finally, we detailed an updated and improved version of the virDRT that has higher resolution than an earlier version [28].

The first research question asked if combining CL and performance measures deliver an effective DDA system in terms of raising or lowering the difficulty of levels to achieve more equal scores and simultaneously lower CL. Overall results confirmed that the CLP-DDA system was successful, with several key findings extracted from the results. This included lower CL in the CLP-DDA versions of the STSG and adjustments to the level of challenge that equalized the difficulty, resulting in very similar combined scores, performance scores, and CL respectively in the CLP-DDA version. The No-DDA levels were significantly different from each other in combined scores, as well as the individual performance and CL results.

The CLP-DDA system increased difficulty for the Park level to a similar level as the No-DDA City level; the Park level performance was medium to high and CL was relatively low, causing the CLP-DDA system to increase difficulty as intended. In the City CLP-DDA level, there was minimal performance difference when compared to the No-DDA City. However, there was a significant reduction in CL suggesting that participants achieved a similar level of performance with less cognitive effort, a desired outcome from a learning perspective. CL was marginally lower in the CLP-DDA Park level than its No-DDA counterpart, but this was not statistically significant. The CLP-DDA Park level was made harder, as reflected in the performance scores, yet CL remained stable indicating the CLP-DDA was successful in improving the cognitive performance of the participants. The adaptions lead to these outcomes in significantly less time than the No-DDA version. All these factors indicate that the CLP-

DDA system operated effectively, and the combination of CL and performance measures used to adapt multiple serious game elements was successful.

The second research question considers if there are any statistical differences between the linear difficulty and the CLP-DDA approach, in the STSG, across various performance and cognitive measures. The experiment identified some positive differences, the key elements were a reduction in CL, and more equal performance scores in the CLP-DDA version. The results express that the City level was more challenging than the Park level in the No-DDA group, in terms of CL ($p = 0.046$), primary task performance ($p = .001$), and combined scores ($p = .0003$). The CLP-DDA version had the same challenges within the same levels, yet there was little variation in scores between the two CLP-DDA levels. This indicates that the CLP-DDA system normalized the difficulty for the levels, matching the task challenge to player proficiency and CL.

Achieving the correct balance of challenge for players leads to a lowering of CL which aligns with the concepts of flow theory [53, 54]. Flow states are associated with total focus or absorption in a task; by achieving an optimum challenge, this state is attained whereby the player marshals all their mental resources on the task at hand, lowering CL. This may be because extraneous thoughts, distractions, and other impacts are reduced by being in a flow state, freeing up mental resources [55]. CL was lower in the CLP-DDA City level versus the No-DDA City level; analysis of these results using Winsorization show that this difference was significant. Equally important was that in the comparison, the means and SD for the CLP-DDA versions were lower than the No-DDA versions, particularly in the City (harder) levels. Lower CL, in challenging learning experiences, is valuable as high CL may lead to overload or indicate that the participant is struggling to master the material [56]. Aiding participants via approaches such as the CLP-DDA, that lower CL while attaining similar performance, may free up cognitive resources to better absorb and master the content.

Finally, the virDRT system was enhanced in this research from a previous version providing a valid CL measure approximately every 25 seconds, to one in which a valid CL measure was provided approximately every five seconds. This was achieved by recording five separate streams of overlapping responses. This system was implemented and provided greater granularity, which is necessary for meaningful real-time adaption, debriefing, or other analysis.

6.1 Limitations and future directions

Some limitations in the research are noted that could be addressed for future experiments using the CLP-DDA system. The first is the relatively low number of participants. There were 52 participants in this research, however, to achieve more robust statistical analysis, it would be advantageous to increase this to over 60. The demographics show little diversity within the participants, with similar ages, similar backgrounds (as all were recruited from a single university campus, and the majority were from the same undergraduate courses) and were predominantly male (83%). In future, drawing upon a more diverse group, both in age and gender, may help explore the CLP-DDA more robustly across a more representative sample of the community.

From a game development perspective, this is an early implementation of the CLP-DDA, and as a result includes a relatively simple adaption strategy that could be refined and improved in future. For example, instead of fixed paths with set lengths for the time adjustment, the paths would be better if defined by a waypoint system that provides opportunity for more changes and greater replay-ability.

Finally, the development of the STSG encountered some development hurdles, particularly with regards to the Train Station level. Due to these issues, this level was disregarded from analysis. Some aspects of the STSG would improve with further development, for example voiceover content, reducing repetitive feedback, NPC animation bugs and so forth. Nothing that inherently affected the experiment but would create a more polished experience.

6.2 Future Research

While the results are promising, there are opportunities for future research and improvements. The overarching sense is that some form of target or limit needs to be more clearly defined. The concept of a defined learning target was demonstrated and proved effective in [16]. In this case, instead of simply making the serious game easier or harder in response to CL or performance, it should also consider a learning goal and adapt in a weighted manner towards that outcome. This may be particularly relevant where a certain pass mark is required, and by implementing a target minimum result, the CLP-DDA may help to achieve improved performance in specific knowledge domains. Experimenting with different delivery and control systems, such as virtual reality, may also be beneficial for broader serious games opportunities. Finally, an assessment of how performance and CL can be analysed to better inform debriefing would be beneficial, particularly in law enforcement contexts relevant to this implementation.

7. Conclusions

There is a lack of research comparing complex 3D serious games that use linear difficulty approaches to DDA variants. We have addressed this gap by creating a serious game, the STSG, that includes a new DDA approach that combines cognitive and performance-based measures to trigger adjustments and then compared this with a variant using linear difficulty. In doing this, we sought to understand if the CLP-DDA system would be successful in improving serious game performance outcomes for participants, while also concurrently minimizing cognitive burden. This experiment demonstrated successful adaption strategies with easier challenges made more difficult for those performing well, and more difficult challenges made less challenging for those under-performing, leading to a more balanced outcome when compared to a control version of the game using linear difficulty. The CLP-DDA approach reduced CL while also reducing the time taken in-game. This may have important implications to future serious game development by identifying a more optimized approach to learning, potentially reducing costs and reducing the risks of working memory resource depletion. In summary, the specific contributions of the paper are as follows:

- Demonstrated a new DDA system (CLP-DDA) and determine the efficacy of this combined CL and performance-based approach.
- Detailed an updated and improved version of the virDRT.
- Provided a serious game revolving around surveillance that may be used in future law enforcement research.

We presented a new version of the virDRT where multiple streams of input were recorded to increase the resolution of participant responses, and therefore a higher resolution view of participant CL. This is important as the number of responses outlined in the ISO standard DRT were insufficient for use in the CLP-DDA as the temporal span was too great for meaningful adjustment. This leads to a new version of a DRT that can be applied to future serious game development.

Acknowledgments

This work was supported by the Defence Innovation Network NSW PhD Grant, and the University of Newcastle Postgraduate Research Scholarship.

The authors would like to acknowledge the hard work for the development, programming, testing and design of the STSG by Justin Cragg and Rhys Adams, as well as further development effort and support provided by staff of Kite Shield Interactive Pty Ltd.

Conflicts of interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

[1] O. Chemikova, N. Heitzmann, M. Stadler, D. Holzberger, T. Seidel, and F. Fischer, "Simulation-based learning in higher education," *Review of Educational Review*, no. 4, 2020, doi: 10.3102/0034654320933544.

[2] D. Charsky, "From edutainment to serious games: A change in the use of game characteristics," *Games and culture*, vol. 5, no. 2, pp. 177-198, 2010, doi: 10.1177/1555412009354727.

[3] P. Caserman *et al.*, "Quality criteria for serious games: serious part, game part, and balance," *JMIR serious games*, vol. 8, no. 3, p. e19037, 2020, doi: 10.2196/19037.

[4] J. Breuer and G. Bente, "Why so serious? On the relation of serious games and learning," *Journal for computer game culture*, vol. 4, pp. 7-24, 2010.

[5] M. Aydin, H. Karal, and V. Nabiyev, "Examination of adaptation components in serious games: a systematic review study," *Education and Information Technologies*, vol. 28, no. 6, pp. 6541-6562, 2023, doi: 10.1007/s10639-022-11462-1.

[6] A. Seyderhelm and K. L. Blackmore, "Systematic Review of Dynamic Difficulty Adaption for Serious Games: The Importance of Diverse Approaches," *Available at SSRN 3982971*, 2021. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3982971.

[7] C. R. Landsberg, R. S. Astwood, W. L. Van Buskirk, L. N. Townsend, N. B. Steinhauser, and A. D. Mercado, "Review of Adaptive Training System Techniques," *Military Psychology*, vol. 24, no. 2, pp. 96-113, 2012, doi: 10.1080/08995605.2012.672903.

[8] J. Sweller, J. J. G. van Merriënboer, and F. Paas, "Cognitive Architecture and Instructional Design: 20 Years Later," *Educational Psychology Review*, vol. 31, no. 2, pp. 261-292, 2019, doi: 10.1007/s10648-019-09465-5.

[9] F. Paas, A. Renkl, and J. Sweller, "Cognitive load theory: Instructional implications of the interaction between information structures and cognitive architecture," *Instructional science*, vol. 32, no. 1/2, pp. 1-8, 2004, doi: https://www.jstor.org/stable/41953634.

[10] B. S. Avi Shena, B. Sitohang, and S. A. Rukmono, "Application of Dynamic Difficulty Adjustment on Evidence-centered Design Framework for Game Based Learning," presented at the 2019 International Conference on Data and Software Engineering (ICoDSE), 2019. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85085546149&doi=10.1109%2fICoDSE48700.2019.9092725&partnerID=40&md5=e1d8e0830a0455179d440a7820ed0dc9.

[11] D. Hooshyar, L. Malva, Y. Yang, M. Pedaste, M. Wang, and H. Lim, "An adaptive educational computer game: Effects on students' knowledge and learning attitude in computational thinking," *Computers in Human Behavior*, vol. 114, 2021, doi: 10.1016/j.chb.2020.106575.

[12] M. Ninaus, K. Tsarava, and K. Moeller, "A pilot study on the feasibility of dynamic difficulty adjustment in game-based learning using heart-rate," presented at the International Conference on Games and Learning Alliance, 2019. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85082481045&doi=10.1007%2f978-3-030-34350-7_12&partnerID=40&md5=bd6c9a542ebb7656fc53eda1ed234405.

[13] S. Papadimitriou, K. Chrysafiadi, and M. Virvou, "FuzzEG: Fuzzy logic for adaptive scenarios in an educational adventure game," *Multimedia Tools and Applications*, vol. 78, no. 22, pp. 32023-32053, 2019, doi: 10.1007/s11042-019-07955-w.

[14] L. Tahai, J. R. Wallace, C. Eckhardt, and K. Pietroszek, "Scalebridge: Design and evaluation of adaptive difficulty proportional reasoning game for children," presented at the 2019 11th International Conference on Virtual Worlds and Games for Serious Applications (VS-Games), September, 2019. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85074289631&doi=10.1109%2fVS-Games.2019.8864526&partnerID=40&md5=e779bf4aa1987e7232be9577009d50ae.

[15] H. van Oostendorp, E. D. van der Spek, and J. Linssen, "Adapting the Complexity Level of a Serious Game to the Proficiency of Players," *EAI Endorsed Transactions on Game-Based Learning*, vol. 1, no. 2, 2014, doi: 10.4108/sg.1.2.e5.

[16] R. Hare, N. Patel, Y. Tang, and P. Patel, "A Graph-based Approach for Adaptive Serious Games," presented at the 2022 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl

Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech), 2022.
[17] D. Afergan *et al.*, "Dynamic difficulty using brain metrics of workload," presented at the Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI '14, Toronto, ON, CA, 26 April 2014 - 1 May 2014, 2014.
[18] F. G. Paas and J. J. Van Merriënboer, "The efficiency of instructional conditions: An approach to combine mental effort and performance measures," *Human factors*, vol. 35, no. 4, pp. 737-743, 1993, doi: 10.1177/001872089303500412.
[19] R. J. Salden, F. Paas, N. J. Broers, and J. J. Van Merrienboer, "Mental effort and performance as determinants for the dynamic selection of learning tasks in air traffic control training," *Instructional science*, vol. 32, pp. 153-172, 2004, doi: 10.1023/B:TRUC.0000021814.03996.ff.
[20] B. Park and R. Brünken, "The rhythm method: A new method for measuring cognitive load—An experimental dual-task study," *Applied Cognitive Psychology*, vol. 29, no. 2, pp. 232-243, 2015, doi: 10.1002/acp.3100.
[21] F. A. Haji, D. Rojas, R. Childs, S. de Ribaupierre, and A. Dubrowski, "Measuring cognitive load: performance, mental effort and simulation task complexity," *Medical education*, vol. 49, no. 8, pp. 815-827, 2015, doi: 10.1111/medu.12773.
[22] L. M. Naismith and R. B. Cavalcanti, "Validity of cognitive load measures in simulation-based training: a systematic review," *Academic Medicine*, vol. 90, no. 11, pp. S24-S35, 2015, doi: 10.1097/ACM.0000000000000893.
[23] S. Bakkes, C. T. Tan, and Y. Pisan, "Personalised gaming: a motivation and overview of literature," presented at the Proceedings of The 8th Australasian Conference on Interactive Entertainment: Playing the System, 2012.
[24] B. Bontchev, "Adaptation in Affective Video Games: A Literature Review," *Cybernetics and Information Technologies*, vol. 16, no. 3, pp. 3-34, 2016, doi: 10.1515/cait-2016-0032.
[25] M. Zohaib, "Dynamic Difficulty Adjustment (DDA) in Computer Games: A Review," *Advances in Human-Computer Interaction*, vol. 2018, pp. 1-12, 2018, doi: 10.1155/2018/5681652.
[26] Z. Buzady, "Flow, leadership and serious games–a pedagogical perspective," *World Journal of Science, Technology and Sustainable Development*, 2017, doi: 10.1108/WJSTSD-05-2016-0035.
[27] J. Hamari, D. J. Shernoff, E. Rowe, B. Coller, J. Asbell-Clarke, and T. Edwards, "Challenging games help students learn: An empirical study on engagement, flow and immersion in game-based learning," *Computers in Human Behavior*, vol. 54, pp. 170-179, 2016, doi: 10.1016/j.chb.2015.07.045.
[28] A. J. Seyderhelm and K. L. Blackmore, "How Hard Is It Really? Assessing Game-Task Difficulty Through Real-Time Measures of Performance and Cognitive Load," *Simulation & Gaming*, vol. 54, no. 3, pp. 294-321, 2023, doi: 10.1177/10468781231169910.
[29] D. Sharek and E. Wiebe, "Measuring video game engagement through the cognitive and affective dimensions," *Simulation & Gaming*, vol. 45, no. 4-5, pp. 569-592, 2014, doi: 10.1177/1046878114554176.
[30] S. Aldekhyl, R. B. Cavalcanti, and L. M. Naismith, "Cognitive load predicts point-of-care ultrasound simulator performance," *Perspectives on medical education*, vol. 7, no. 1, pp. 23-32, 2018, doi: 10.1007/s40037-017-0392-7
[31] J.-C. Woo, "Digital game-based learning supports student motivation, cognitive success, and performance outcomes," *Journal of Educational Technology & Society*, vol. 17, no. 3, pp. 291-307, 2014, doi: http://www.jstor.org/stable/jeductechsoci.17.3.291.
[32] A. Dideriksen, C. Reuter, T. Patry, T. Schnell, J. Hoke, and J. Faubert, "Define Expert - Characterizing Proficiency for Physiological Measures of Cognitive Workload," presented at the Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC) 2018, I/ITSEC Knowledge Repository, January 30, 2020, 2018. [Online]. Available: http://www.iitsecdocs.com/.
[33] I. O. f. Standardization, *Road vehicles — Transport information and control systems — Detection-response task (DRT) for assessing attentional effects of cognitive load in driving*. 2016.
[34] M. DeGeurin. "Police VR Training: Empathy Machine or Expensive Distraction?" Gizmodo. https://gizmodo.com.au/2022/05/police-vr-training-empathy-machine-or-expensive-distraction/ (accessed 01/01/2024, 2024).
[35] A. T. Hussain, E. Halford, and F. AlKaabi, "The Abu Dhabi Police Virtual Training Centre: A case study for building a virtual reality development capacity and capability," *Policing: A Journal of Policy and Practice*, vol. 17, p. paad028, 2023, doi: 10.1093/police/paad028.
[36] L. Kleygrewe, R. V. Hutter, M. Koedijk, and R. R. Oudejans, "Virtual reality training for police officers: a comparison of training responses in VR and real-life training," *Police Practice and Research*, vol. 25, no. 1, pp. 18-37, 2024, doi: 10.1080/15614263.2023.2176307.

[37] O. Binsch *et al.*, "The effect of virtual reality simulation on police officers' performance and recovery from a real-life surveillance task," *Multimedia Tools and Applications,* vol. 82, no. 11, pp. 17471-17492, 2023, doi: 10.1007/s11042-022-14110-5.

[38] C. Cooper, *Behind the Private Eye : the private investigator's secrets : surveillance tales and techniques*. Port Macquarie, NSW:: Chris Cooper, 2005.

[39] training.gov.au. "DEFIN011A - Perform foot surveillance." https://training.gov.au/Training/Details/DEFIN011A (accessed 15/10/2023, 2023).

[40] S. Vanbecelaere, K. Van den Berghe, F. Cornillie, D. Sasanguie, B. Reynvoet, and F. Depaepe, "The effectiveness of adaptive versus non-adaptive learning with digital educational games," *Journal of Computer Assisted Learning,* vol. 36, no. 4, pp. 502-513, 2020, doi: 10.1111/jcal.12416.

[41] G. Camp, F. Paas, R. Rikers, and J. van Merrienboer, "Dynamic problem selection in air traffic control training: A comparison between performance, mental effort and mental efficiency," *Computers in Human Behavior,* vol. 17, no. 5-6, pp. 575-595, 2001.

[42] F. Paas, J. E. Tuovinen, H. Tabbers, and P. W. Van Gerven, "Cognitive load measurement as a means to advance cognitive load theory," *Educational psychologist,* vol. 38, no. 1, pp. 63-71, 2003.

[43] R. Brunken, J. L. Plass, and D. Leutner, "Direct measurement of cognitive load in multimedia learning," *Educational psychologist,* vol. 38, no. 1, pp. 53-61, 2003, doi: 10.1207/S15326985EP3801_7.

[44] J. Sweller, "Cognitive load theory," in *Psychology of learning and motivation*, vol. 55: Elsevier, 2011, pp. 37-76.

[45] P. Chandler and J. Sweller, "Cognitive load while learning to use a computer program," *Applied cognitive psychology,* vol. 10, no. 2, pp. 151-170, 1996, doi: 10.1002/(SICI)1099-0720(199604)10:2<151::AID-ACP380>3.0.CO;2-U.

[46] J. L. Harbluk, P. C. Burns, J. Tam, and V. Glazduri, "Detection response tasks: Using remote, headmounted and Tactile signals to assess cognitive demand while driving," presented at the PROCEEDINGS of the Seventh International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design, Bolton Landing, New York, USA, 17-20 June, 2013, 2013.

[47] A. Vandierendonck, "A comparison of methods to combine speed and accuracy measures of performance: A rejoinder on the binning procedure," *Behavior research methods,* vol. 49, no. 2, pp. 653-673, 2017, doi: 10.3758/s13428-016-0721-5.

[48] R. Bruyer and M. Brysbaert, "Combining speed and accuracy in cognitive psychology: Is the inverse efficiency score (IES) a better dependent variable than the mean reaction time (RT) and the percentage of errors (PE)?," *Psychologica Belgica,* vol. 51, no. 1, pp. 5-13, 2011.

[49] S. K. Kwak and J. H. Kim, "Statistical data preparation: management of missing values and outliers," *Korean journal of anesthesiology,* vol. 70, no. 4, pp. 407-411, 2017, doi: 10.4097/kjae.2017.70.4.407.

[50] D. Ruppert, "Trimming and winsorization," *Wiley StatsRef: Statistics Reference Online,* 2014.

[51] D. Ghosh and A. Vogt, "Outliers: An evaluation of methodologies," *Joint statistical meetings,* vol. 12, no. 1, pp. 3455-3460, 2012.

[52] O. Chen, J. C. Castro-Alonso, F. Paas, and J. Sweller, "Extending cognitive load theory to incorporate working memory resource depletion: evidence from the spacing effect," *Educational Psychology Review,* vol. 30, pp. 483-501, 2018, doi: 10.1007/s10648-017-9426-2.

[53] C.-C. Chang, C. A. Warden, C. Liang, and G.-Y. Lin, "Effects of digital game-based learning on achievement, flow and overall cognitive load," *Australasian Journal of Educational Technology,* vol. 34, no. 4, 2018, doi: 10.14742/ajet.2961.

[54] M. Csikszentmihalyi, S. Abuhamdeh, and J. Nakamura, "Flow," in *Flow and the foundations of positive psychology*: Springer, 2014, pp. 227-238.

[55] C.-C. Chang, C. Liang, P.-N. Chou, and G.-Y. Lin, "Is game-based learning better in flow experience and various types of cognitive load than non-game-based learning? Perspective from multimedia and media richness," *Computers in Human Behavior,* vol. 71, pp. 218-227, 2017, doi: 10.1016/j.chb.2017.01.031.

[56] F. A. Haji, J. J. Cheung, N. Woods, G. Regehr, S. de Ribaupierre, and A. Dubrowski, "Thrive or overload? The effect of task complexity on novices' simulation-based learning," *Medical education,* vol. 50, no. 9, pp. 955-968, 2016, doi: 10.1111/medu.13086.