Research Article

How Hard Is It Really? Assessing Game-Task Difficulty Through Real-Time Measures of Performance and Cognitive Load

Simulation & Gaming
2023, Vol. 54(3) 294–321

DOI: 10.1177/10468781231169910
journals.sagepub.com/home/sag


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

Abstract

Background. Serious and entertainment game designers strive to create engaging, immersive, and often, challenging games. This task involves modifying game mechanics or environments to create experiences with differing levels of challenge to meet player skill. The balance between different game mechanics or environments, and the differing levels of challenge they pose, is typically understood through iterative testing. Balance and challenge becomes increasingly important in serious games and simulation training as these games commonly need to be engaging and impart learning content. Overburdening players' cognitive capacity with either too much gameplay challenge or learning content may reduce the educational effectiveness of the game.

Aim. In this research, we develop a game-based driving simulation with different gameplay tasks to explore the impact of different types of challenges and game aesthetics on real-time cognitive load and task performance, which may inform serious game design. We also test the validity of a game-embedded real-time cognitive load measuring method.

Method. A total of 31 participants undertook the driving simulation experiment under three different aesthetic conditions using a within-subject experimental

[1]School of Information and Physical Sciences, University of Newcastle, Callaghan, NSW, Australia

Corresponding Author:
Andrew J.A. Seyderhelm, School of Information and Physical Sciences, University of Newcastle, University Drive, Callaghan, NSW 2308, Australia.
Email: andrew.seyderhelm@uon.edu.au

design. Cognitive load was measured using three different methods, and performance was measured via in-game metrics. Additionally, demographic and engagement surveys were also completed.

Results. Player performance and cognitive load respond differently to different types of challenge, and an appropriate level of game challenge can lower cognitive load. The embedded cognitive load measure was validated as an effective method for evaluating real-time cognitive load during gameplay.

Conclusion. The results demonstrate the validity of a dual measure approach for future adaptive serious games and simulation training environments combining performance and cognitive load. An easy to implement, and robust, in-game measure for cognitive load has been validated in real-world conditions. From these results, a system for dynamic difficulty adjustment is proposed tailored towards serious games and simulation.

Keywords
dynamic difficulty adjustment, serious games, adaption, simulation training, cognitive load

1 Introduction

Serious games and simulation training applications are used in a broad range of education and training settings (Chemikova et al., 2020; Csikszentmihalyi et al., 2014; Wilkinson, 2016). Achieving the optimum balance of gameplay difficulty and learning content is challenging due to player proficiency, prior knowledge, and learning aptitude (Ravyse et al., 2016). Dynamic difficulty adjustment (DDA) is a form of adaption in games that helps address differing and evolving player skill in real-time. Entertainment games seek to achieve the optimum level of challenge for players, either through predetermined difficulty levels or via the application of DDA systems. A similar goal is desired in serious games, however, this incorporates the additional requirement of ensuring effective learning content delivery. DDA is explored in a serious games context to adapt the serious game to the needs of the learner and ultimately achieve better learning outcomes (Landsberg et al., 2010). Striking the right balance of challenge is critical to maximizing the chance of positive learning outcomes by both engaging the learner and helping achieve a Flow state (Hamari et al., 2016), or similar positive motivational or affective states.

The success of serious games is greatly influenced by participant prior knowledge, pace of learning, and learning design implementation using methods such as scaffolding (Chemikova et al., 2020). How much the player knows on a subject, or how readily they grasp the learning material, can greatly influence the level of engagement they have for the activity. One way to address this issue is via DDA, where both the gameplay and learning content are varied to adjust the challenge in a game. This dual

adjustment is critical, as it is feasible a player may understand the learning content but have less developed gameplay skills and vice versa. For example, when using a driving training game, someone may know the rules of the road very well but have never played a driving game before and struggle with the gameplay controls. Therefore, adjusting the gameplay to be easier, but increasing the learning challenges as a separate factor, may be beneficial. A recent literature review on DDA identified that in complex 3D game environments, a multiple measure combined with a multiple adaption strategy is more likely to be effective (Seyderhelm & Blackmore, 2021). However, very few serious games systems adopted this approach, with the same review showing only 10.2% of studies incorporating DDA systems incorporate dual measures (Seyderhelm & Blackmore, 2021).

2 Literature Review

The following section provides a review of literature on the key concepts of Flow Theory (Section 2.1) and Cognitive Load Theory (Section 2.2) relevant to DDA in serious games and simulation training.

2.1 Flow Theory

Serious games are often considered in respect of how they improve training or learning through their impact on motivational, behavioral and/or cognitive processes; recent indications are that for best results these three aspects need to be combined (Krath et al., 2021). Reflecting this, there are many educational, behavioral and psychological theories applied to serious game design with overlapping concepts and goals (Krath et al., 2021). Krath et al., identified 10 core theoretical principles that can be applied to serious game design, including Flow theory and Cognitive Load theory. Within DDA system research, Flow theory is the most referred to concept (Seyderhelm & Blackmore, 2021). Similarly, the concepts of engagement, immersion and presence are often referred to in serious game design and have many overlaps with the concept of Flow (Hookham & Nesbitt, 2019).

In some instances, a state of flow may not be attainable, or desirable, however it can be considered as a related state to engagement (Hookham & Nesbitt, 2019). Either flow or engagement remain desirable states depending on the purpose of the training (Kiili, 2006; Mills et al., 2013). Further research has indicated that motivation has a strong impact on learning success and "[F]low was an especially strong predictor of motivation" (Özhan & Kocadere, 2020). In this context, using the concept of flow as a basis for developing a serious game DDA system is valid: even if a flow state is not achieved, flow is interlinked with other motivational theories and it follows that a general increase in motivation is therefore likely, improving learning outcomes.Lastly, a key aspect of the concept of flow is to do with the challenge of the activity being consummate with the capability of the participant. A DDA system strives to adapt the challenge of a game

(or serious game in this context) to match the participants ability. Achieving the correct challenge is a strong indicator of learning success (Hamari et al., 2016).

Therefore, an ideal goal for a DDA system is for users to achieve a 'flow state' (Alves et al., 2018). Flow is a mental state akin to *being in the zone*, or achieving absorbed and focused engagement (Csikszentmihalyi et al., 2014). DDA research often describes flow as an appropriate level of challenge, whereby the difficulty of the activity is neither too hard (leading to frustration) nor too easy (leading to boredom). Csikszentmihalyi et al. (2014) define the conditions for attaining flow in further detail, which include three primary criteria that need to be met:

1. A "clear set of goals" that add "direction and purpose to behavior" (p.232); this also is influenced by, and influences, player motivation with both intrinsic and extrinsic motivation being impacted.
2. "A balance between perceived challenges and perceived skill" (p.232); this is the aspect of flow most commonly discussed in DDA research for games, and helps lead to mastery of a task or mechanic.
3. Flow requires "clear and immediate feedback" (p.232); this informs the participant how they are performing and can identify where the participant needs to focus, as long as this area of focus is not perceived as too difficult.

Understanding these three principles of flow may help lead to better serious game experiences (Hamari et al., 2016) and is a valid serious games design approach (Kiili, 2006).

2.2 Cognitive Load Theory (CLT)

In a serious game, the complexity of the environment, game aesthetics, and different mechanics may impact on player performance in a variety of ways. This impact is best described through the prism of cognitive load theory (CLT) which explains how the brain processes information in working memory, which is limited, and consolidates it into long-term memory schema (Paas et al., 2004). CLT refers to three main cognitive load structures: Intrinsic, extraneous, and one termed germane processing. Intrinsic cognitive load refers to the essential complexity of the material, or how much mental effort is required in order to grasp the information contained (Sweller, 2011). Extraneous cognitive load considers the way material is presented, or actions required of the learner; that is, how much mental effort is needed to receive and perceive the information (Sweller, 2011). This can also be impacted by environmental conditions (Choi et al., 2014).

Finally, germane processing refers to the amount of mental effort used to store information in working memory and into long-term memory schema (Sweller et al., 2019).

Good design reduces extraneous cognitive load, manages intrinsic load, and fosters positive germane load. A balance must be struck between the information that relates to

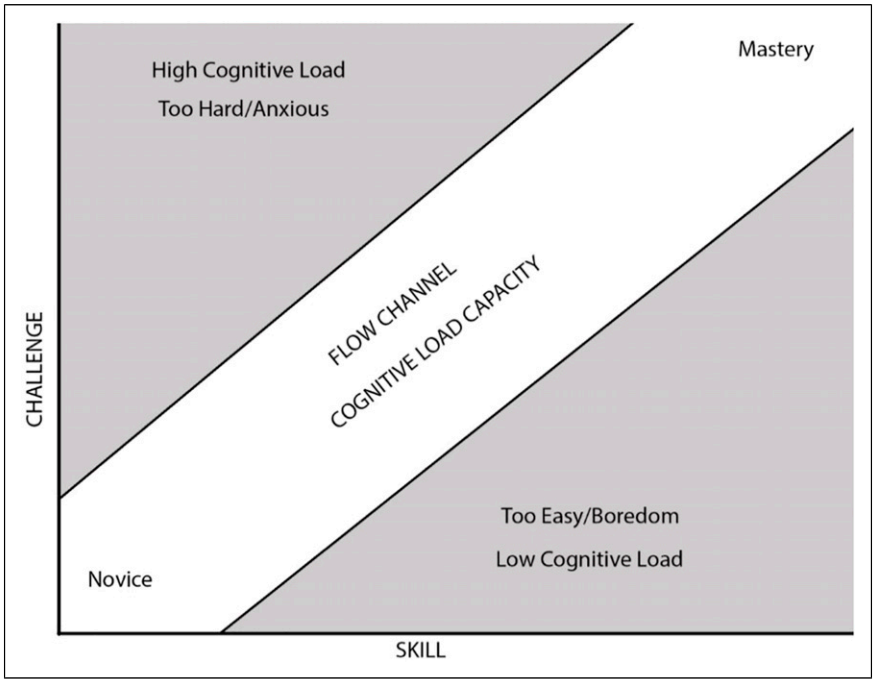

Figure 1. Flow channel combined with cognitive load capacity leading from novice to mastery.

gameplay versus the learning content. Distracting gameplay may help foster flow through challenge and excitement but serve as extraneous cognitive load impacting absorption of learning content. Ideally, a balance between the learning and non-learning content to promote flow through engagement and fun, balanced against cognitive load to achieve the best learning outcomes, is desired (Chang et al., 2017). Player performance in this context may be simple to measure, for example, through correct answers, time taken, or a plethora of other metrics. What is more challenging to measure in real-world settings is cognitive load, in a cheap, easy to implement, robust and reliable manner.

This dichotomy presents two opportunities to challenge the player, and by extension, potentially regain interest and focus. One is via the learning content and the second is to add complexity, variation, or challenge into the gameplay elements. These opportunities can be realized through DDA, which can adapt the gameplay and learning content separately driven by real-time measures of player performance and cognitive load.

A DDA system developed with a better understanding of the challenge impacts of game design, aesthetics, and pedagogical theories, may improve the likelihood of achieving a state of flow. A flow state may enhance the value of the training material (Hamari et al., 2016) and a flow state may be sustained when an appropriate DDA system is applied, potentially leading to mastery (Figure 1).

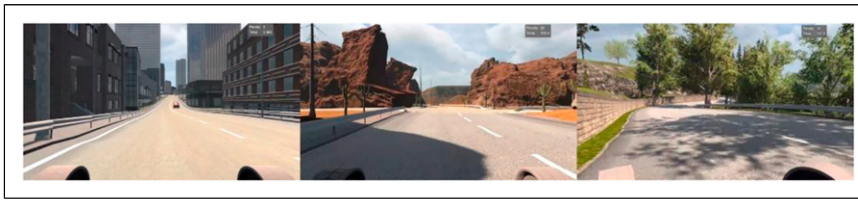

Figure 2. The three biomes of the CEDG: city (left), desert (middle), and forest (right).

The concepts of flow and CLT present challenges in serious games in knowing what to change and what impact that may have on player performance, learning, engagement, flow, and cognitive load. The research presented in this paper seeks further understanding of these factors to inform the design and implementation of appropriate serious game DDA systems that can potentially approach the effectiveness of one-on-one training.

3 Experiment: Cognitive-Effect Driving Game

The COGNITIVE-EFFECT DRIVING GAME (CEDG) is a first-person 3D driving game developed for this research with specific goals to test cognitive load and performance. The CEDG incorporates a range of player performance measures, and tests an embedded cognitive load measure termed the *virtual detection response task* (virDRT) (see Section 3.2). Full specification of the CEDG is beyond the scope of this paper and is fully defined in (Seyderhelm & Blackmore, 2021b). The CEDG has been designed to assess the impact of different aesthetic conditions on player cognitive load and performance across a wide range of tasks in a complex virtual environment. The CEDG also tests the layering of primary and secondary tasks to allow a robust measurement of cognitive load using the virDRT approach.

The CEDG design consists of three levels, each having the same track layout with different surrounding biomes consisting of a city, forest, and desert environment (Figure 2). Each zone has a different challenge or task (Table 1). For each level, the player completed two circuits - one with the virDRT active and one without. The order of each level is randomized, and within each level, the order in which the virDRT was active is also randomized.

4 Method

This research explores how different gameplay tasks, difficulties, and aesthetics impact player performance and cognitive load in a complex driving video game. Additionally, we seek to validate an embedded game controller-based version of the detection response task (DRT) as a cost effective and easy to implement measure of cognitive load. Specifically, we seek to answer the following research questions:

Table 1. Zone and their challenges.

Zone	Challenge or task
Zone 1	Count specified colored vehicles parked by the road
Zone 2	The player must listen to, recall and follow a series of verbal driving directions and apply them through a suburban street area, with traffic and turns
Zone 3	Short zone
Zone 4	Longer driving section in which absolute difficulty (Adams, 2014) of the primary task is increased
Zone 5	Directed to follow and maintain a distance behind another vehicle, the player also needs to recall the number of cars counted in zone 1
Zone 6	Continue to follow vehicle from zone 5, but now with rain and thunder
Zone 7	Drive through a narrow entrance and proceed in a moderately long zone with rain and thunder
Zone 8	Drive through a tunnel with only car headlights causing limited visibility.
Zone 9	Short zone
Zone 10	The player has to cross a bridge with roadworks, one lane is blocked and the player must judge the distances and timing of oncoming cars to proceed to waiting bays or continue across the bridge safely

RQ1. Does an embedded in-game cognitive load measure (virDRT) effectively capture cognitive load without impacting task performance?

RQ2. How do different tasks and challenges affect cognitive load and performance in a complex driving game-based environment?

To address these questions, data was collected using a within-subjects experimental design, approved by the Human Research Ethics Committee [details omitted for double-anonymized peer review].

4.1 Demographics

The CEDG experiment was conducted across two weeks with staff or students at the University of Newcastle, Australia in October 2020. Prior to commencing, all participants ($n = 33$) completed a demographic and game preferences survey. Participants ranged from 19 to 50 years old (mean=28.03, SD=8.97). Of the 33 participants, 23 identified as male and 10 identified as female; 30 of the participants were right-handed, 2 left-handed, and one ambidextrous. Each participant session took approximately 90 minutes, and included obtaining informed consent, providing a task and intervention briefing, pre-game demographic and game preference surveys, CEDG tutorial, playing the CEDG, and post-test completion of the NASA-TLX and Game Engagement survey. Two participants experienced technical issues during the experiment leading to their data being rejected leaving 31 participants (22 males and

9 females). Of these, 22 (70.97%) were 19-30 years old, five (16.13%) were 31-40, and four (12.90%) were over 40 years old. Accordingly, 24 (77.42%) of the participants were 34 or younger, which is the average age of gamers in Australia (Brand et al., 2019).

4.2 Cognitive Load and the Virtual Detection Response Task (virDRT)

We propose that measuring cognitive load alone is insufficient to fully inform a robust DDA system. Cognitive load implies the cognitive burden but doesn't necessarily indicate where this burden originates; it doesn't identify if cognitive load is high due to task complexity or the extraneous load is too high due to poor task design and presentation, necessitating the use of performance measures, differing from game-to-game dependent on task type. A comprehensive categorization of performance measures and types is detailed in (Seyderhelm & Blackmore, 2021).

During the CEDG experiment, cognitive load was measured using the virDRT employing the standards detailed in International Organization for Standardization (2016).. In addition, the NASA Task Load Index (NASA-TLX) was used to provide an overall assessment of the cognitive load from the entire gameplay period. The design and use of the virDRT is described below.

The virDRT is delivered as a secondary task by measuring the reaction time relating to a stimulus while a person performs a primary task or function. A decrease in the response time to the secondary task stimulus indicates an increased cognitive burden from the primary task (Paas et al., 2003). A DRT requires minimal impact on learning tasks and therefore minimal cognitive impact in itself (Brunken et al., 2003) It should also be easy to grasp, recognize, and respond to. A wide range of secondary task methods have been developed in different experiment settings (Brunken et al., 2003; Chandler & Sweller, 1996; Haji et al., 2015; Park & Brünken, 2015; Sweller, 2011), and the DRT has been shown to be particularly successful and relatively simple to implement. The virDRT used in the experiment is very similar to the remote DRT described by Harbluk et al. (2013), however it is integrated into the game controller via shoulder button presses (Figure 3a) rather than via a separate finger switch typically used (Harbluk et al., 2013). By integrating the virDRT into the UI layer of the serious game and using a standard game-controller, ease of use and deployment opportunities for a DRT-based cognitive measure are greater and cheaper, as the need for specialized equipment is removed. The virDRT is designed to be noticeable without being obtrusive, and the player was tasked to respond to the virDRT as the least important element of their current activity.

The virDRT records the reaction time (RT) to the stimulus, in this case a red dot to the lower left of the screen (Figure 3b), 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; anything outside of this is captured in the HR as 0 (miss) and 1 (successful hit). The standard requires a minimum of five data points (hits or misses) to provide valid cognitive load measurement.

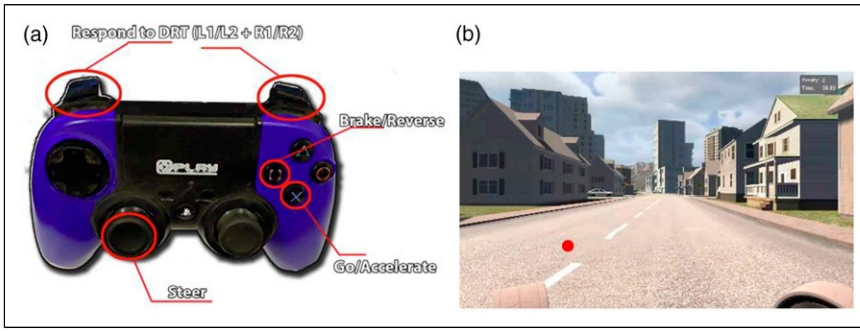

Figure 3. (a) (Left) The game controller used in the CEDG, with the shoulder buttons used as virDRT response triggers. (b) (Right) The virDRT is active in the city level (Zone 2), the red dot towards the bottom left.

Therefore, for each zone in the CEDG each group of five to nine virDRT responses are averaged to provide a measure of current cognitive load. These measures are grouped by zone as this experiment seeks to understand the impact of different challenges and aesthetic conditions on a zone-by-zone basis.

RQ1 assesses the efficacy of the virDRT for measuring cognitive load. This was compared to the ISO 17488:2016 (Standardization, 2016) that details the validated DRT methods and also includes a method for checking data quality, providing a frequency histogram for reaction time (RT) as a guide. Each participant's virDRT RT (Mean=0.7 sec; SD = 0.29 sec) was plotted for visual comparison to the ISO histogram. This comparison confirms that the data collected follows the expected RT response distribution pattern specified in ISO 17488:2016, validating the virDRT data (Figure 4).

4.3 Performance Measures

We propose that combining measures of cognitive load and performance will be more reflective of task difficulty. For the primary task (driving), performance was measured by three elements: time taken, lane deviations, and crashes into other vehicles or objects. The vehicle had a maximum speed of 80 kilometers per hour (kph) to create a standard baseline where driving skill would differentiate performance. The participants were briefed to drive as swiftly and accurately as possible, avoiding accidents and lane deviations. Lane deviation has been used extensively in driver studies as a measure of performance (Beede & Kass, 2006; Irwin et al., 2015; Shinar et al., 2005).

Additionally, secondary tasks were designed to test different cognitive processes (Table 2).

A key element for defining the difficulty of challenges in the CEDG for comparisons is to identify a meaningful method for combining and assessing time taken and driving

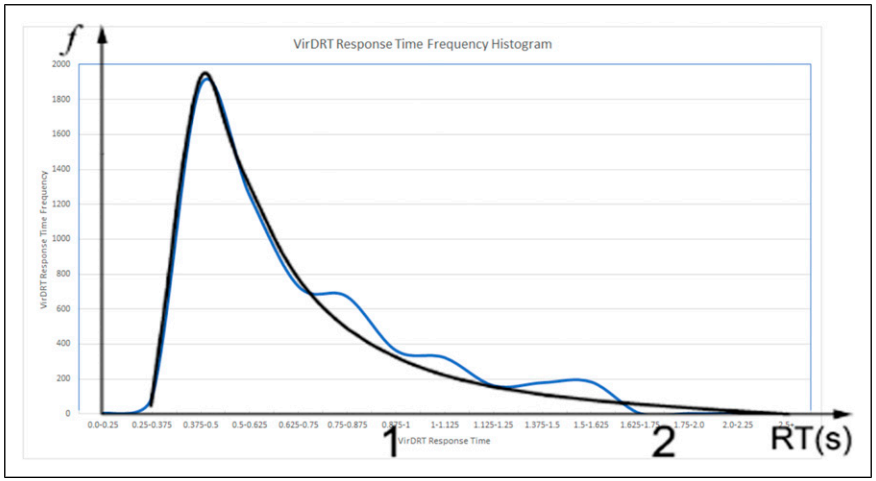

Figure 4. Frequency histogram from the CEDG, the blue line, overlayed with the ideal histogram example from ISO 17488:2016 (Section 6.10, page 13), black line.

Table 2. Task description and performance measure used.

Task Description	Measure
Count specific-colored vehicles parked beside a section of the road. Four zones later the player is asked to recall and input the number counted.	This is marked as right (1) or wrong (0).
The player is given a series of directions to follow in a residential zone and must remember them.	If the player successfully followed the directions, yes (1) or no (0).
The player is asked to follow a car maintaining a distance range.	One penalty is awarded per second for failing to stay within the range.

accuracy. Here we rate each zone to enable a meaningful comparison of zone difficulty irrespective of length using a single score for each zone. Zones were analyzed and plotted reflecting the total range of time taken, as well as the combined total of crashes and lane deviations (driving errors). This produced a range for each zone that was then divided by 10 to provide discrete, evenly distributed groupings (Table 3). The groupings were given a score ranging from 1 – 10, with one being the quickest time and ten being the slowest. Driving accuracy (derived from crashes and lane deviations) was assessed similarly, however a result of zero (0) crashes or deviations was given a score of 1 (as it was always the best result) and the remainder grouped according to variance (Table 3).

Time score and accuracy score were then multiplied together to provide a final combined score with a possible range of 1 - 100. This produces a normalized zone difficulty score accounting for the different lengths.

Table 3. Presents an example (Zone 1) of the way time and accuracy were grouped and scored from 1 – 10; the count columns show how many times across all levels and players that range was achieved (31 players x 3 levels x 2 laps = 186 instances).

Time taken (range)	Time count	Time score	Driving Accuracy (range)	Accuracy count	Accuracy score
20.608-27.108	149	1	0	49	1
27.108-33.608	18	2	1	42	2
33.608-40.108	5	3	2-3	51	3
40.108-46.608	7	4	4-5	23	4
46.608-53.108	3	5	6-7	9	5
53.108-59.608	2	6	8-9	6	6
59.608-66.108	1	7	10-12	3	7
66.108-72.608	0	8	13-15	2	8
72.608-79.108	0	9	16-18	0	9
79.108-85.608	1	10	19-21	1	10

5 virDRT Results

The following sections present the results of the statistical analysis of the experimental data (descriptive statistics and two-tailed *T*-Tests) to assess if the virDRT impacted overall player performance, validating the usefulness of the virDRT approach.

The virDRT was active 52 times in the first circuit of each level versus 41 in the second circuit. This equated to 55.9% of the time in the first loop compared to 44.1% in the second. The CEDG was developed using Unity and C#, with the inbuilt random integer function used to derive the order of the virDRT. There is no reported bias in the Unity random integer function, and it is likely a higher number of instances may have evened out the percentages.

5.1 virDRT Impact on Performance and Cognitive Load

This section details the results used to determine if there was a statistically significant impact from the virDRT on player performance (defined as time taken and driving accuracy). Each zone is played twice, both with and without the virDRT active, so the impact on performance of the virDRT can be assessed. Performance was measured in three ways: time, accuracy, and secondary task accuracy. These results were then used to create a total combined score (Section 4.3).

The total combined scores were collated, and a paired two tailed *T*-Test was used to identify any differences for each zone (Table 4). The results indicate that there is no statistically significant difference in performance when the virDRT is active or not active for all but Zone 9, which is marginally significant at the 0.05 alpha level, Zone 9 results are discussed in further detail below.

Table 4. Mean, standard deviation and test results (*p-value*) of the total combined scores for each individual zone.

Zone	1	2	3	4	5	6	7	8	9	10
DRT - M	5.12	14.32	4.14	13.54	7.41	6.85	6.97	3.83	3.10	9.02
DRT - SD	10.05	13.46	5.57	16.86	11.46	5.92	11.89	12.11	3.09	13.49
No DRT - M	3.96	14.42	4.49	11.27	5.30	6.81	5.90	2.75	4.19	6.69
No DRT - SD	6.79	16.96	10.73	13.39	5.85	9.06	14.43	6.21	4.39	8.82
t(92)	1.04	−0.06	−0.29	1.73	1.77	0.04	0.60	0.80	−2.01	1.55
p-value	.30	.95	.77	.09	.08	.97	.55	.42	.05[a]	.13

[a]Marginally statistically significant result.

5.1.1 Zone 9 Analysis. The only zone with a marginally statistically significant difference is Zone 9, so a more detailed analysis was undertaken for time and accuracy individually, to determine where the near impact originated. The following table (Table 5) lists the results of a two tailed *T*-Test for time and accuracy, for each level, of Zone 9.

As detailed in Table 5, time taken to complete the level for Level 3 is the only statistically significant result. Reviewing the data reveals that in Zone 9, players were quicker overall when the virDRT was active. Table 6 provides further analysis of time taken for Zone 9.

Zone 9 was also the only zone that took less time in all three levels when the virDRT was active. The virDRT had no programmed impact on vehicle speed, and no other zones demonstrated this anomaly indicating that other factors were the likely cause of the difference, such as the number of NPC vehicles, a level design anomaly, and/or practice effect (Duff et al., 2007).

5.2 Summary of virDRT Validation Results

The virDRT was shown to collect input results commensurate with the DRT standard data frequency histogram indicating the data collection was valid. The use of the virDRT had no statistically significant impact on player performance in a complex driving game. Therefore, the game console controller-based implementation of the virDRT was effective and presents as a suitable and easy to implement method for measuring real-time cognitive load during game-play.

6 Impact of cognitive load and performance by zone and challenge in the CEDG

The cognitive load and performance measures were used to rate the difficulty of each type of challenge in the CEDG. Each zone consists of different challenges to the primary task (driving) as well as additional tasks such as counting vehicles, following directions, or judging distance. To assess difficulty, each player received a total score

Table 5. *T*-Test results, *p-value*, for zone 9 for time and driving accuracy across all three levels.

Statistic Category	Level 1	Level 2	Level 3
Time	0.79	0.12	0.02[a]
Accuracy	0.26	0.80	0.54

[a]indicates a statistically significant result.

Table 6. Time taken analysis for Zone 9.

Category	virDRT active	virDRT inactive
Total Time Taken	390.76	410.28
M - Time per level	12.605	13.235
SD - Time per level	1.662	1.743

per zone via the combined measure system (Section 4.3). These scores were then added together for all three levels providing an indication of zone difficulty (Figure 5).

Table 7 lists the zones in order of difficulty and identifies if a secondary task was involved.

6.1 Impact of challenges on cognitive load

When assessing the data collected from a DRT, it is important to consider both the response time and the misses (non-responses) as both data provide information on current cognitive burden (Standardization, 2016; Vandierendonck, 2017). The inverse efficiency score (IES) (Bruyer & Brysbaert, 2011; Vandierendonck, 2017) method was selected to derive a single cognitive load value. 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:

$$IES = RT/1-PE.$$

To understand the cognitive impact of each zone, the IES was applied on a per zone basis. However, there were 18 total instances across all levels, zones and players in which there was no response to the virDRT recorded. This null response is taken to mean a high cognitive load is in effect, and a DDA system can respond accordingly. However, for this analysis where the zone cognitive difficulty is being sought, instances of a null result are replaced with the maximum score that the player could have achieved if they had made the minimal possible response. For example, player 104923 failed to record any responses in Zone 7 of the first level where there were six virDRT stimuli. To derive a maximal score, using the details outlined in Section 4.2, we take the maximum response time of 2.5 seconds and the minimum response value of 1 response and divide by 1- proportion of errors: 2.5/(1-0.833) = 14.997. The methodology explained here, and in Section 4.2, was applied to each of the 18 instances of a null result to derive a maximal score proxy so that cognitive challenge could be determined for each zone.

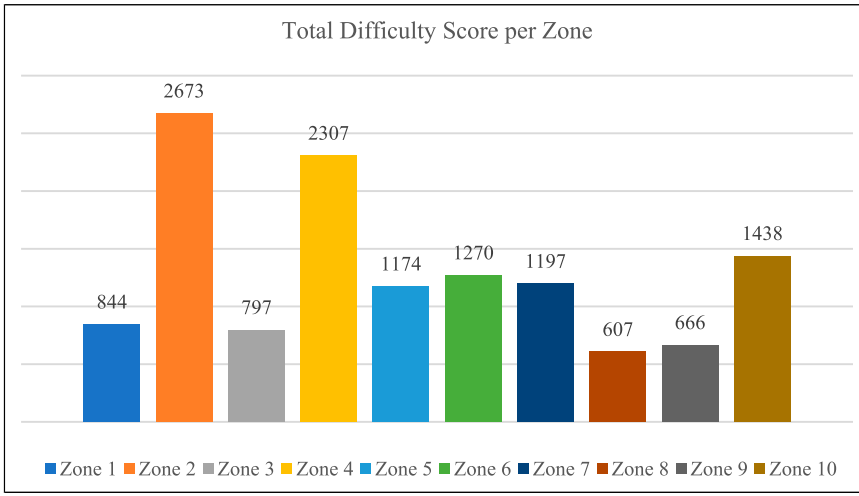

Figure 5. The total difficulty scores for each zone.

Table 7. Zones listed in order of overall difficulty.

Zone	Total Score	Secondary Task	Task
Zone 2	2673	Yes	Recall and follow directions
Zone 4	2307	No	No secondary task, more difficult primary task
Zone 10	1438	No	Judge distance and space
Zone 6	1270	Yes	Follow vehicle and maintain distance with weather
Zone 7	1197	No	Narrow entrance and weather
Zone 5	1174	Yes	Follow vehicle and maintain distance, recall count task (Zone 1)
Zone 1	844	Yes	Count cars
Zone 3	797	No	Short zone
Zone 9	666	No	Short zone
Zone 8	607	No	Drive through tunnel with low light

The sum of the IES scores (Figure 6) for each zone provides a score for how challenging each zone is from a cognitive load perspective, allowing zone difficulty based on performance (Section 5.1) to be compared to cognitive load for a more holistic view.

Zones 5 and 10 have the greatest impact on cognitive load. Zones 1, 2, 5 and 6 all include additional tasks, while Zone 10 includes the requirement for judgement and spatial assessment. In comparing task difficulty with cognitive load (Figure 7), there

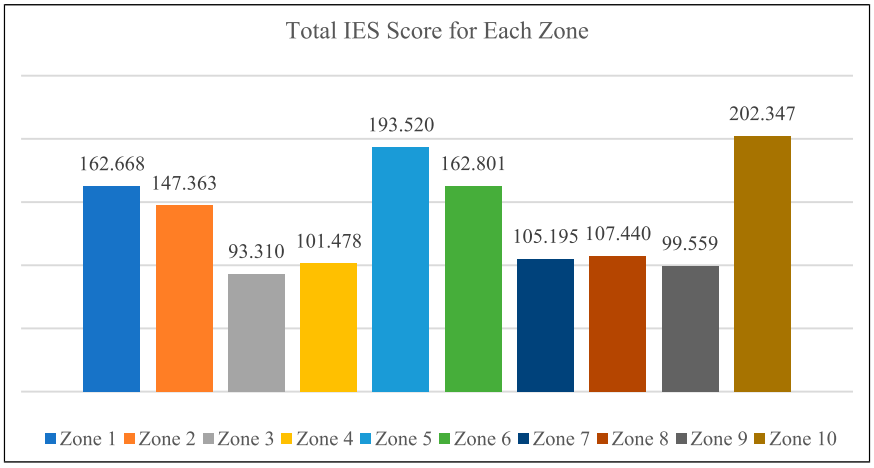

Figure 6. IES-based combined cognitive load score for each zone for all three levels.

are some areas of interest. For example, Zone 4 is ranked as the second most challenging zone from a performance perspective, and yet has a relatively low cognitive load score (third lowest overall). This may be explained by Flow theory (Csikszentmihalyi et al., 2014) where the correct level of challenge engenders a state of high engagement and focus. This high level of focus indicates that cognitive resources are martialed to a single task and other distractions have minimal impact leading to a reduction of cognitive load. Zone 1 was rated as quite easy from a performance perspective, but the player had the additional task of counting specific color vehicles as they were driving, and as a result, cognitive load in this zone was the 4th highest overall.

6.2 Per Zone and Level Comparison of Cognitive Load and Performance

Figure 7 combines the results for cognitive load and performance. This makes the relationship between cognitive load and performance clear, although there are some notable differences, particularly Zone 5 where the graph lines head in opposite directions, and Zone 2 for different reasons.

Zone 2 has difficult driving conditions; it was ranked as the hardest overall (Figure 5) and had the 5th highest overall score for cognitive load (Figure 6). However, there were very large differences between the first level and subsequent in both performance and cognitive load (Figure 7). This demonstrates a strong practice effect (Duff et al., 2007), and possibly some form of strategy, for example cue utilization (Brouwers et al., 2016).

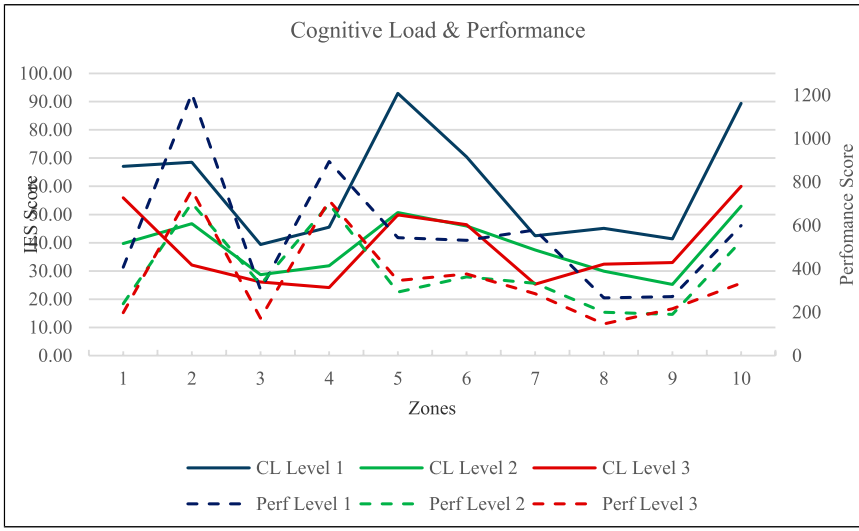

Figure 7. Cognitive Load (CL) and performance (Perf) overlaid for comparison.

6.3 Discussion on Challenge

The CEDG involved different challenges or variations to the primary task of driving, including changes to lighting (Zone 8), adjustment of the absolute difficulty (Zone 4), and addition of complexity by incorporating distance and safety judgements (Zone 10). Zones 1, 2, 5 and 6 are discussed in Section 6.4 below as they contain secondary tasks with specific separate measures and results.

Zones 3 and 9 were designed to be relatively short with easy driving difficulty, which is reflected in both performance and cognitive load measures. These zones serve as a good baseline for measuring the impact of other driving challenges.

Zone 4 was mentioned in Section 6.1, but it is worth reiterating that the 'right' level of challenge can impact cognitive load positively and potentially help engender a state of flow and enhance learning (Csikszentmihalyi et al., 2014; Hamari et al., 2016). The driving task (ie. sharp corners and twists) in this zone was hard but had low cognitive load, indicating heightened engagement or a sense of flow.

Zone 7 is one of the longer sections and included a narrow entry into a rain and thunder affected zone, allowing exploration of the impact of aesthetic only weather effects on performance and cognitive load. Zone 7 is equivalent to Zone 5 and 6 in terms of primary task difficulty but resulted in lower cognitive load. This suggests that purely aesthetic weather conditions have minimal impact on either cognitive load or performance. However, navigating the narrow entry to the zone, and the type of road made it the fifth hardest zone.

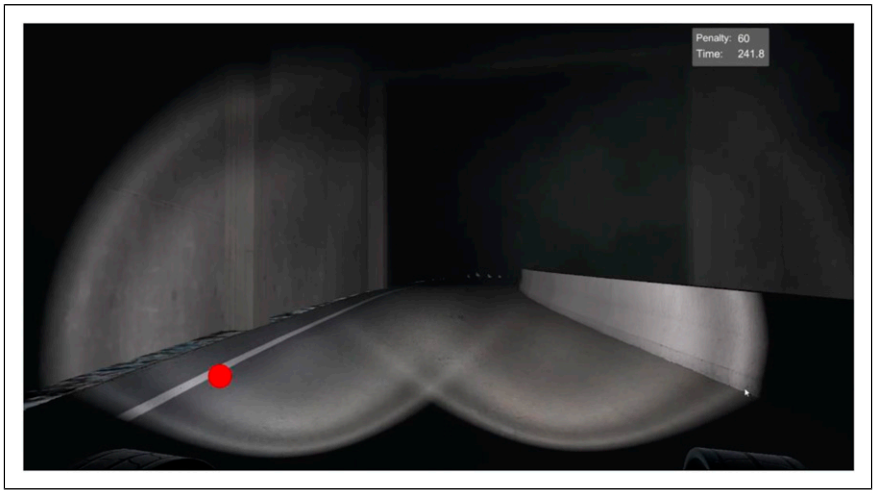

Figure 8. Inside zone eight, the low light tunnel section.

Zone 8 explores the impact of lighting conditions, a reduction in view distance, and visually restricted environment (Figure 8). It required the player to drive through a tunnel with only vehicle headlights. The tunnel appears more cramped due to the addition of walls and a center crash barrier; however, these do not restrict road dimensions. The challenges in this zone had no negative impact on the difficulty; this zone was the easiest overall in terms of performance, and 6th lowest for cognitive load. This suggests narrowing the field of view and providing clear visual guidance, may be an effective method to make tasks easier. In short, this zone may reduce cognitive load, and enhance performance, by the subtraction of extraneous aspects of the environment (Choi et al., 2014).

Zone 10 includes a long bridge to cross with contraflow barriers; the driving lane is blocked off and there are two waiting bays for the player to reach (Figure 9). There is oncoming traffic, and the player must judge the correct time to progress or wait until the entire bridge is clear of traffic. This zone explored the cognitive burden of judging distance, speed, and timing. Spatial judgement tasks like this have been explored in other driving environments that have indicated this type of task is complex (Alexander et al., 2002; Feldstein, 2019). Zone 10 was ranked 3rd hardest overall for performance and produced the highest overall cognitive load burden. In principle, the primary driving task was easy with a straight road and barriers that reduced the road access. Yet adding oncoming traffic and the need to judge when to drive, or when to wait, added to the challenge and cognitive load greatly. Adding a risk factor element based on spatial judgement, speed, and timing can increase the level of difficulty and cognitive load.

Figure 9. Zone 10 bridge contraflow with oncoming traffic.

6.4 Analysis of Secondary Task Effects on Performance

There were four zones with specific secondary tasks in addition to driving, the following section explores the results from those secondary tasks and seeks to compare secondary task performance with primary task results by considering zone scores and penalties (Table 8).

6.4.1 Zone 1 – Observation and Recall as a Secondary Task. In Zone 1 players were tasked to count how many specific-colored cars were parked beside the road (Figure 10). Later, in Zone 5, players were asked to recall the number counted with the result recorded as a binary pass (1) or fail (0) in each circuit completed. The players completed three levels, with two circuits of each level, requiring them to complete this task six times. From the overall results, this task was completed correctly in 73.12% of instances (Table 8).

The total for the first level was much lower (worse) than the following two. Interestingly, the third level had worse performance than the second level, which may have been caused by external factors; we suspect fatigue. Overall, this indicates practice effect (Duff et al., 2007) as a factor in the results.

The counting task appears to have had little impact on performance (Figure 5), however the recall element was clearly a challenge for the first circuit with a 56.45% success rate (Table 8). It is likely that after a relatively poor performance at this task, and being provided with visual and auditory feedback, the need for greater focus or applying a strategy was realized. This is relevant to serious games designers who may include similar counting and recall tasks; the first attempt at the task is unlikely to be

Table 8. Scores for each secondary task zone.

Zone	Result Category	1st Level	2nd Level	3rd Level
Zone 1	Score	35	53	48
Zone 1	Percent Correct	56.45%	85.48%	77.42%
Zone 2	Score	35	45	53
Zone 2	Percent Correct	56.45%	72.58%	85.48%
Zone 5	Penalties[a]	626	359	302
Zone 6	Penalties[a]	368	357	316

[a]Six player results were removed due to a design error in Zone 5 (Section 5.5.3).

Figure 10. Zone one (1) – player tasked to count specific color cars parked on the side of the road.

particularly successful from a learning context until players develop a strategy and realize the importance of the task through in-game feedback.

6.4.2 Zone 2 – Follow & Remember Driving Directions. In Zone 2 players were recorded as either correctly (1) or incorrectly (0) following driving directions. The directions consisted of either three or four verbal instructions given at the start of the zone, that the player needed to remember and follow for the remainder of the zone. For example: "*Take the next right, take the first left, then the next right and at the end turn left*."

This task was completed correctly in 71.51% of instances (Table 8). This suggests that following and remembering directions may be marginally more difficult than the counting (Zone 1) and recalling (Zone 5) task.

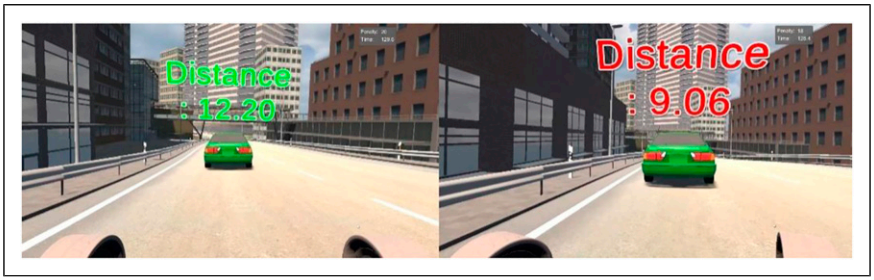

Figure 11. Zone 5 showing the correct distance range (left) and getting too close (right).

Following and remembering directions was clearly challenging, particularly the first time encountered, and the players improved substantially in a relatively short time (Table 8). Using audio directions that require memory and recall in different locations and with differing complexities may have a strong impact on cognitive load and performance, known as a transient information effect (Sweller et al., 2019). In the CEDG, the zone layout and style of directions were the same, and players clearly demonstrate practice effect (Duff et al., 2007). It would be valuable to further explore the use of audio directions with a recall component in different settings to understand the impact of this type of challenge more fully.

6.4.3 Zone five and six – following a vehicle maintaining a distance. Zones 5 and 6 consist of the same basic secondary challenge with a few minor differences and are thus addressed together. In these zones players follow a specified vehicle maintaining a distance between 10 and 18 meters, with in-game text and visual elements to aid in judgement (Figure 11). For every second spent either too close or too far away, the player suffers a penalty point that accumulates over the duration of the task. In the first circuit of each level, the car being followed drives slower than the one in the second circuit. Unfortunately, Zone 5 has a collision error near the start of the zone, that could cause players to crash unexpectedly leading to higher penalty scores. This had a knock-on effect to Zone 6 for those players as they sought to catch up.

Approximately halfway through Zone 5 the game pauses and the player must recall the number of cars they counted in Zone 1, they receive audio and visual feedback as to whether they were correct or not.

The following task from Zone 5 continues into Zone 6 and introduces a change in weather and lighting effects with rain, thunder, and lightning (Figure 12).

Six player results were removed from Zone 5 due to the impact of the collision error detailed above. Performance improves markedly across the levels (Table 8). Zone 5 was the 5th hardest zone in respect of performance (Figure 5), but the second hardest in respect of cognitive load (Figure 6). Zone 6 was the 3rd hardest in respect of cognitive load, indicating that the addition of the recall of cars counted from Zone 1 has an impact

Figure 12. Zone 6 with grey skies, rain drops and the distance display.

on cognitive load in Zone 5, suggesting designers need to be mindful of how tasks are stacked in simulations and serious games.

There was a slight impact on performance from the addition of weather effects from Zone 5 to 6. However, there was no significant difference between the performance of Zone 5 ($M = 6.35$, $SD = 9.13$) and Zone 6 ($M = 6.83$, $SD = 7.63$) indicating that adding aesthetic-only weather effects has no significant impact on performance ($t(185) = -0.65$, $p = .517$).Weather may have a positive impact on player engagement and enjoyment, and is worth utilizing in order to enhance realism in simulations irrespective of the impact on performance or cognitive load (Roberts & Patterson, 2017).

7 Understanding the tasks and challenges for future game design and DDA implementation

Section 6 detailed the impact of different challenges and secondary tasks on both player performance and cognitive load. Specifically, the section determined which challenges were most impactful in terms of performance and cognitive load and ranked them in order of this impact. There are a few key lessons from this analysis that can be used to inform DDA systems and serious game design more broadly, namely:

- Adjusting the absolute difficulty of a task, as seen in Zone 4, to the ideal challenge level may engender a flow state. Flow states have been shown to reduce extraneous cognitive load and positively impact germane cognitive load (Chang et al., 2017; Chang et al., 2018).

- Audio directions and recall are difficult when combined with other tasks (Knight & Tlauka, 2017), in this case driving or following, and have an impact on both cognitive load (Klatzky et al., 2006) and performance.
- Weather and lighting effects, not impacting gameplay mechanics (Zone 6 to 8 respectively), have little effect on either cognitive load or performance, but may help increase engagement (Roberts & Patterson, 2017).
- Tasks that require distance judgment and constant in-game responses, for example maintaining a distance behind a vehicle (Zones 5 and 6), navigating a narrow space (Zone 7), or judging when it is clear to drive (Zone 10), have a strong impact on cognitive load and performance.
- Reducing extraneous details and variables (Zone 8) by limiting interactions or field of view can make a task easier and help reduce cognitive load, likely due to reduced extraneous cognitive load (Skulmowski & Xu, 2021; Sweller et al., 2019).

From these tasks and mechanics, an understanding of how different types of challenges can be used to impact cognitive load and performance emerges. Extrapolating the type of task into different games and scenarios to validate the findings in other game genres and mechanics is warranted. However, results indicate that mechanics can be combined and varied in different ways to impact challenge, making a serious game either easier or harder in deterministic ways.

Our findings show that while there may be a connection between cognitive load and performance, this relationship is not always clear or obvious. For example, Zone 4 was one of the hardest rated zones for performance, but one of the lowest for cognitive load. Conversely the challenges in Zone 10 were rated high for both cognitive load and performance, highlighting the need for careful consideration of performance or cognitive load in design choices. A decrease in cognitive load, paired with an increase in performance provides an indication of mastery of the challenges and game play elements. While also considered as training effect, this occurs over the three levels as players' performance and cognitive load improve. Continued gameplay at this level of challenge would likely lead to boredom or frustration as a player needs to be continually challenged in the appropriate way (Hamari et al., 2016). Ideally, measuring cognitive load and performance will help inform a highly effective DDA system – if cognitive load is too high game challenge can be altered, or additional tasks tweaked to modify the cognitive burden. Similarly, if the intrinsic game difficulty is too hard, or easy, it can be modified to help improve flow or engagement to enhance the learning outcomes (Skulmowski & Xu, 2021; Sweller et al., 2019). Performance of primary tasks and additional tasks can be measured and compared with cognitive load to help strike the correct balance between game difficulty and learning content.

Previous research (Seyderhelm & Blackmore, 2021) suggests that dual adaptive measures using cognitive load and performance have a greater chance of success than single measures. Also, it is suggested that adapting multiple elements has a correlation with greater efficacy for DDA systems. The findings of this research highlight the value

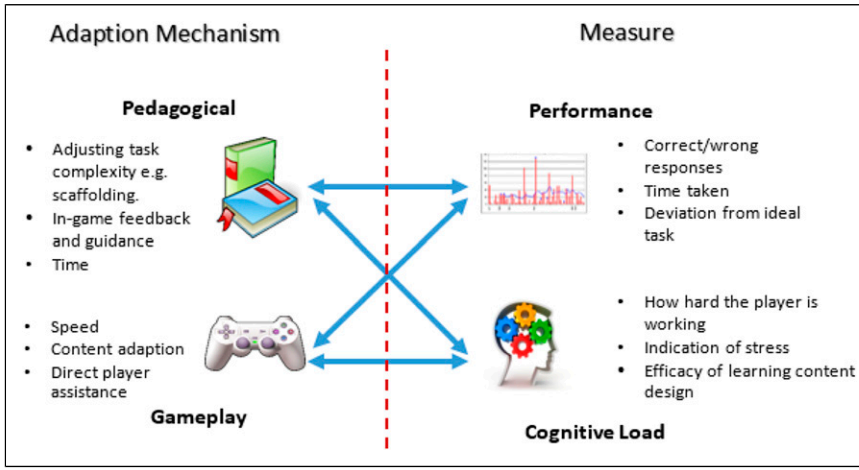

Figure 13. The dual measure and multi-adaption concept.

in measuring both performance and cognitive load, supporting the premise for dual adaption systems in learning contexts (Figure 13).

Another valuable aspect of measuring both performance and cognitive load is that this data may further enhance and empower after action review (AAR). Research has shown that a quality AAR can improve training efficacy significantly (Keiser & Arthur Jr, 2021). Being able to report on a player's cognitive load over the duration of the intervention, or training exercise, can reveal a great deal about how that person is learning and journeying towards mastery. Cognitive load theory outlines that as information is fully understood it is stored in memory schema leading to a lowering of cognitive load (Sweller, 2011). Similarly, as a person understands a topic it is expected that their performance improves. Therefore, as performance improves and cognitive load lowers, we can surmise that the learner has achieved mastery of that content. This is important in some circumstances and may be useful as an AAR report to inform training development and future deployment or enhance the direction of future training.

8 Conclusion

This research explored how performance and cognitive load are impacted in simulated driving tasks and validated a novel implementation of the detection response task - the virDRT. We developed an easy to implement, robust and cost-effective cognitive load measure that can be applied to almost any game or simulation environment through an in-game user interface (UI) element and the use of a standard video game controller. Importantly, the results of this research confirm that this implementation does not impact on task performance. In doing this, we contribute an approach to make real-time

cognitive load measures more accessible for use in a wider range of games and simulations suitable for real-world applications. By detailing the challenge level of different tasks, and the cognitive burden thereof, the identified tasks and challenges represent adaption measures that can be used to modify difficulty more accurately and confidently in other game-based environments.

Some limitations were present in this research, principally pertaining to sample size and gender balance. With 31 participants and only 9 female participants, a larger sample size and greater parity between genders would be preferable. As discussed in Section 5, the limited sample size led to an uneven balance in the Unity randomization function. Future research should use a stratified randomization method to ensure greater balance in random functions, tasks, levels, or assignments. The implementation of the virDRT used a specific controller type, game genre, and interface implementation. To ensure the robustness of the virDRT, further research should explore different control mechanisms for the virDRT for use on a wide range of hardware and game-play types. It will also be worth exploring the efficacy of the DDA mechanism when applied to complex learning material (e.g. chemistry or physics), particularly as this learning content may not be assimilated easily by all players. The value of measuring cognitive load and performance to inform a DDA mechanism presents as a more effective method of DDA for serious games and simulations than a single measure approach. While performance and cognitive load are linked, there are clear instances where having a measure of each is beneficial. This work has shown the value in measuring both, and the work serves as a useful launching point for future research. Specifically, the results of this paper form a foundation for future research exploring the implementation of a dual measure and multiple-adaption DDA system that considers the impact of this on both training performance and trainee experience. In doing so, the research presented contributes important insights for the development of effective serious games and simulation training platforms.

Authors' contributions

Andrew Seyderhelm and Karen Blackmore contributed to the design of the study, and the development of the study protocol. Andrew Seyderhelm developed the game-based experiment task, completed the data collection, and analysis of the results. Karen Blackmore provided project supervision. Both authors discussed the results and contributed to the final manuscript.

Declaration of conflicting interests

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

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Defence Innovation Network NSW PhD Grant, and the University of Newcastle Postgraduate Research Scholarship.

Ethics Approval Statement

This research was approved by our institution's human research ethics committee, application #H-2020-0069.

ORCID iDs

Andrew J.A. Seyderhelm https://orcid.org/0000-0003-0191-0367
Karen L Blackmore https://orcid.org/0000-0002-9111-0293

References

Adams, E. (2014). *Fundamentals of game design* (Third Edition ed.). Pearson Education.

Alexander, J., Barham, P., & Black, I. (2002). Factors influencing the probability of an incident at a junction: results from an interactive driving simulator. *Accident Analysis & Prevention*, *34*(6), 779–792, https://doi.org/10.1016/s0001-4575(01)00078-1

Alves, T., Gama, S., & Melo, F. S. (2018). Flow adaptation in serious games for health. In 2018 IEEE 6th International Conference on Serious Games and Applications for Health (SeGAH).

Beede, K. E., & Kass, S. J. (2006). Engrossed in conversation: The impact of cell phones on simulated driving performance. *Accident Analysis & Prevention*, *38*(2), 415–421, https://doi.org/10.1016/j.aap.2005.10.015

Brand, J. E., Jervis, J., Huggins, P. M., & Wilson, T. W. (2019). *Digital Australia 2020*. Retrieved January 29, 2020, from

Brouwers, S., Wiggins, M. W., Helton, W., O'Hare, D., & Griffin, B. (2016). Cue utilization and cognitive load in novel task performance. *Frontiers in psychology*, *7*, 435. https://doi.org/10.3389/fpsyg.2016.00435

Brunken, R., Plass, J. L., & Leutner, D. (2003). Direct measurement of cognitive load in multimedia learning. *Educational psychologist*, *38*(1), 53–61, https://doi.org/10.1207/s15326985ep3801_7

Bruyer, Raymond, & Brysbaert, Marc (2011) (In press). 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*, *51*(1), 5–13. https://doi.org/10.5334/pb-51-1-5

Chandler, P., & Sweller, J. (1996). Cognitive load while learning to use a computer program. *Applied Cognitive Psychology*, *10*(2), 151–170, https://doi.org/10.1002/(sici)1099-0720(199604)10:2<151::aid-acp380>3.0.co;2-u

Chang, C.-C., Liang, C., Chou, P.-N., & Lin, G.-Y. (2017). 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*, *71*, 218–227, https://doi.org/10.1016/j.chb.2017.01.031

Chang, C.-C., Warden, C. A., Liang, C., & Lin, G.-Y. (2018). Effects of digital game-based learning on achievement, flow and overall cognitive load. *Australasian Journal of Educational Technology*, *34*(4). https://doi.org/10.14742/ajet.2961

Chemikova, O., Heitzmann, N., Stadler, M., Holzberger, D., Seidel, T., & Fischer, F. (2020). Simulation-based learning in higher education. *Review of Educational Review*, (4).

Choi, H.-H., Van Merriënboer, J. J., & Paas, F. (2014). Effects of the physical environment on cognitive load and learning: towards a new model of cognitive load. *Educational psychology review*, *26*(2), 225–244, https://doi.org/10.1007/s10648-014-9262-6

Csikszentmihalyi, M., Abuhamdeh, S., & Nakamura, J. (2014). Flow. In *Flow and the foundations of positive psychology* (pp. 227–238). Springer.

Duff, K., Beglinger, L. J., Schultz, S. K., Moser, D. J., McCaffrey, R. J., Haase, R. F., Westervelt, H. J., Langbehn, D. R., Paulsen, J. S., & Group, H. s. S. (2007). Practice effects in the prediction of long-term cognitive outcome in three patient samples: A novel prognostic index. *Archives of Clinical Neuropsychology*, *22*(1), 15–24, https://doi.org/10.1016/j.acn.2006.08.013

Feldstein, I. T. (2019). Impending collision judgment from an egocentric perspective in real and virtual environments: A review. *Perception*, *48*(9), 769–795, https://doi.org/10.1177/0301006619861892

Haji, F. A., Rojas, D., Childs, R., de Ribaupierre, S., & Dubrowski, A. (2015). Measuring cognitive load: performance, mental effort and simulation task complexity. *Medical education*, *49*(8), 815–827, https://doi.org/10.1111/medu.12773

Hamari, J., Shernoff, D. J., Rowe, E., Coller, B., Asbell-Clarke, J., & Edwards, T. (2016). Challenging games help students learn: An empirical study on engagement, flow and immersion in game-based learning. *Computers in Human Behavior*, *54*, 170–179. https://doi.org/10.1016/j.chb.2015.07.045

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

Hookham, G., & Nesbitt, K. (2019). A Systematic Review of the Definition and Measurement of Engagement in Serious Games. Proceedings of the Australasian Computer Science Week Multiconference - ACSW 2019, Sydney, NSW, Australia,29-31 January, 2019.

Irwin, C., Monement, S., & Desbrow, B. (2015). The influence of drinking, texting, and eating on simulated driving performance. *Traffic injury prevention*, *16*(2), 116–123, https://doi.org/10.1080/15389588.2014.920953

Keiser, N. L., & Arthur, W. Jr (2021). A meta-analysis of the effectiveness of the after-action review (or debrief) and factors that influence its effectiveness. *Journal of Applied Psychology*, *106*(7), 1007, https://doi.org/10.1037/apl0000821

Kiili, K. (2006). Evaluations of an experiential gaming model. *Human Technology: An Interdisciplinary Journal on Humans in ICT Environments*.

Klatzky, R. L., Marston, J. R., Giudice, N. A., Golledge, R. G., & Loomis, J. M. (2006). Cognitive load of navigating without vision when guided by virtual sound versus spatial language. *Journal of Experimental Psychology: Applied*, *12*(4), 223, https://doi.org/10.1037/1076-898X.12.4.223

Knight, M. J., & Tlauka, M. (2017). Interactivity in map learning: The effect of cognitive load. *Spatial Cognition & Computation*, *17*(3), 185–198, https://doi.org/10.1080/13875868.2016.1211661

Krath, J., Schürmann, L., & Von Korflesch, H. F. (2021). Revealing the theoretical basis of gamification: A systematic review and analysis of theory in research on gamification, serious games and game-based learning. *Computers in Human Behavior*, *125*, 106963, https://doi.org/10.1016/j.chb.2021.106963

Landsberg, C. R., Van Buskirk, W. L., Astwood, R. S. Jr, Mercado, A. D., & Aakre, A. J. (2010). *Adaptive training considerations for use in simulation-based systems*.

Mills, C., D'Mello, S., Lehman, B., Bosch, N., Strain, A., & Graesser, A. (2013). What makes learning fun? exploring the influence of choice and difficulty on mind wandering and engagement during learning. *International Conference on Artificial Intelligence in Education*.

Özhan, Ş. Ç., & Kocadere, S. A. (2020). The effects of flow, emotional engagement, and motivation on success in a gamified online learning environment. *Journal of Educational Computing Research*, *57*(8), 2006–2031, https://doi.org/10.1177/0735633118823159

Paas, F., Renkl, A., & Sweller, J. (2004). Cognitive load theory: Instructional implications of the interaction between information structures and cognitive architecture. *Instructional science*, *32*(1/2), 1–8, https://doi.org/10.1023/b:truc.0000021806.17516.d0

Paas, F., Tuovinen, J. E., Tabbers, H., & Van Gerven, P. W. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational psychologist*, *38*(1), 63–71, https://doi.org/10.1207/s15326985ep3801_8

Park, B., & Brünken, R. (2015). The rhythm method: A new method for measuring cognitive load—An experimental dual-task study. *Applied Cognitive Psychology*, *29*(2), 232–243, https://doi.org/10.1002/acp.3100

Ravyse, W. S., Seugnet Blignaut, A., Leendertz, V., & Woolner, A. (2016). Success factors for serious games to enhance learning: a systematic review. *Virtual Reality*, *21*(1), 31–58. https://doi.org/10.1007/s10055-016-0298-4

Roberts, S., & Patterson, D. (2017). Virtual weather systems: measuring impact within videogame environments. In Proceedings of the Australasian Computer Science Week Multiconference.

Seyderhelm, A. J., & Blackmore, K. (2021). Systematic Review of Dynamic Difficulty Adaption for Serious Games: The Importance of Diverse Approaches. *Available at SSRN 3982971*.

Seyderhelm, A. J., & Blackmore, K. L. (2021). Quantifying In-Game Task Difficulty Using Real-Time Cognitive Load. *i3 Lab Working Paper Series*. Retrieved from https://nova.newcastle.edu.au/vital/access/manager/Repository/uon:38525?view=null&f0=sm_identifier%3A%22http%3A%2F%2Fhdl.handle.net%2F1959.13%2F1427274%22&sort=sort_ss_title%2F

Shinar, D., Tractinsky, N., & Compton, R. (2005). Effects of practice, age, and task demands, on interference from a phone task while driving. *Accident Analysis & Prevention*, *37*(2), 315–326, https://doi.org/10.1016/j.aap.2004.09.007

Skulmowski, A., & Xu, K. M. (2021). Understanding Cognitive Load in Digital and Online Learning: a New Perspective on Extraneous Cognitive Load. *Educational psychology review*, *34*(1), 171–196. https://doi.org/10.1007/s10648-021-09624-7

Standardization, I. O. f. (2016). *Road vehicles — Transport information and control systems — Detection-response task (DRT) for assessing attentional effects of cognitive load in driving*.

Sweller, J. (2011). Cognitive load theory. In *Psychology of learning and motivation* (Vol. *55*, pp. 37–76). Elsevier.

Sweller, J., van Merriënboer, J. J. G., & Paas, F. (2019). Cognitive Architecture and Instructional Design: 20 Years Later. *Educational psychology review*, *31*(2), 261–292. https://doi.org/10.1007/s10648-019-09465-5

Vandierendonck, André (2017) (In press). A comparison of methods to combine speed and accuracy measures of performance: A rejoinder on the binning procedure. *Behavior research methods*, *49*(2), 653–673. https://doi.org/10.3758/s13428-016-0721-5

Wilkinson, P. (2016). A brief history of serious games. In *Entertainment computing and serious games* (pp. 17–41). Springer.

Author Biographies

Andrew J. A. Seyderhelm is a PhD candidate at the University of Newcastle, Australia. He has several years commercial games experience, has worked for over ten years in law enforcement and is applying a real-world approach to his research in enhancing the efficacy of simulation training and serious games.

Karen Blackmore, Associate Professor Karen Blackmore is a leading researcher in human-computer interaction and simulation-based training. She has expertise in the modelling and simulation of complex social and environmental systems. Her research interests cover the use of agent-based models for simulation of socio-spatial interactions, and the use of virtual environments and serious games for learning. She is the Deputy Head of School (Industry Engagement) for the School of Information and Physical at the University of Newcastle, Australia and an IEEE Senior Member.