Title

Empirical evaluation of player experience using a machine-learning approach to dynamic difficulty adjustment in video games.

Author names and affiliations

Nigel Robb[a], Bo Zhang[b]

[a] University of Tokyo.

[b] East China Normal University.

Corresponding author

Nigel Robb

Center for Global Communication Strategies, University of Tokyo, Japan

nigelrobb@g.ecc.u-tokyo.ac.jp

Empirical evaluation of player experience using a machine-learning approach to dynamic difficulty adjustment in video games.

Abstract

Dynamic difficulty adjustment (DDA) in video games involves altering the level of challenge provided based on real-time feedback from the player. Some approaches to DDA use measurements of player performance, such as success rate or score. Such performance-based DDA systems aim to provide a bespoke level of challenge to each player, so that the game is neither too hard nor too easy. Previous research on performance-based DDA shows that it is linked to better player performance, but finds mixed results in terms of player experience (e.g., enjoyment). Also, while the concept of flow is regarded as an important aspect of video game experience, little research has considered the effects of performance-based DDA on flow. We conducted an experiment on the effects of performance-based DDA on player performance, enjoyment, and experience of flow in a video game. DDA was achieved using a generalised algorithm. 221 participants played either the DDA version of the game, a control version (difficulty remained constant), or an incremental version (difficulty increased regardless of performance). Results show that the DDA group performed significantly better. However, there were no significant differences in terms of enjoyment or experience of flow.

Keywords

video games; dynamic difficulty adjustment; game balancing; flow; performance; adaptive software

1 Introduction

Most video games involve some challenge for the player. Some games are generally easy, some are generally hard, and almost all games feature some change in the difficulty level over time; typically, games get harder the further the player progresses (Chang, 2013). In this paper, we refer to this traditional approach as “incremental difficulty adjustment”. Furthermore, a diverse range of people (e.g., in terms of age, gender, disability, motivations, and preferences) play games (Entertainment Software Association, 2017; Williams et al., 2008). Taken together, these points begin to illustrate the complex challenges involved for the game designer when determining the difficulty of a game, to enhance the experience for a range of players. If the game is too easy, more skilled players may be bored; but if it’s too hard, less skilled players may be frustrated (Leiker et al., 2016). It is likely that this applies regardless of the genre or intended purpose of the game. Whether it is a fast-paced action game intended to entertain, a puzzle game for mathematics education, a cognitive training game for children with cognitive impairments, or even a language-learning application with game-like features, it is obviously essential that players engage optimally with the software to ensure the desired outcome. As such, the level of challenge provided by a game is an important consideration.

Within this broad issue, one interesting potential solution to some of these challenges lies in dynamic difficulty adjustment (DDA). DDA in video games refers to any technique in which

the difficulty of the game is altered during the game (or perhaps between games) in response to some feedback about the player's experience (Xue, et al., 2017). The aim is to continuously tailor the difficulty of the game to each individual player. DDA has featured in video games since at least 1981 (Adams, 2008). The promise of DDA lies in the fact that the game designer does not need to pre-determine one specific difficulty curve (or even a range of pre-determined curves, as in games which let the player select, e.g., Easy, Medium, or Hard mode). Instead, the designer can effectively provide a range of possible difficulties which are dynamically selected for each individual player based on their experience of the game. Essentially, this provides a bespoke difficulty curve for each player.

1.1 Approaches to dynamic difficulty adjustment

One important distinction when using DDA is based on the metric used to provide feedback for game adaptation. In this paper, we have so far (intentionally) characterised this feedback very broadly, as "player experience". In practice, player experience can be determined in a variety of ways. We can broadly categorise approaches to DDA in two ways, depending on whether they use players' in-game performance to provide the feedback, or use some information about the players affective state. A combination of these approaches would of course also be possible.

Performance-based DDA involves measuring the player's performance in the game, and adjusting the level of challenge provided accordingly. For example, the survival horror game Left 4 Dead generates a unique experience for individual players by tailoring – among other things – the enemies encountered by the player, based upon measurements of player performance (Booth, 2009). In other words, as the player performs better, the game gets harder, and vice versa. Previous research on performance-based approaches to DDA demonstrates the wide range of choice available in the design of such systems, both in terms of the indicator of player skill (i.e., the feedback), and the game features that are subsequently adjusted. Regarding the measurement of player skill, previous approaches have used, for example, the time taken to complete a task (Sharek & Wiebe, 2015), players' scores (Bateman et al., 2011), or, in more complex systems, multiple measurements may be combined and evaluated to determine the current state of the player (Hunicke, 2005). Regarding the game features that are adjusted, these range from simple adjustments such as changing the speed of the game (Alexander et al., 2013) or changing the number of objects a player must interact with (Nagle et al., 2015; Robb et al., 2019), to more complex systems which alter the behaviour or characteristics of computer-controlled enemy characters (Hunicke, 2005) or dynamically generate the layout of the game environment (Shaker et al., 2010).

Affective DDA refers to any approach which aims to use some feedback about the player's emotional state as the basis for the adaptivity. A growing body of research has investigated the feasibility of using psychophysiological measurement to provide an index of players' emotions during gameplay, and adapt the game experience based on these measurements. Measures used include cardiovascular (e.g., heart rate), electro-dermal activity (i.e., galvanic skin response, which is directly dependent on sweat-gland activity), electromyography (which measures muscle movements), and neuroimaging techniques. The latter category includes techniques such as electroencephalography and functional near-infrared spectroscopy; both of which use sensors attached to the head to provide real-time measurements of brain activity with a relatively high temporal resolution (Thibault et al.,

2016). These techniques have been used to adapt game difficulty, for example, by increasing the speed of the game or decreasing the size of targets. In addition, some studies have used affective feedback to alter other features of a game not related to difficulty, such as lighting and audio effects. For a comprehensive review of research in this area and references for all examples discussed in this paragraph, see Bontchev (2016). One obvious disadvantage of affective-based DDA is the requirement for additional equipment (which is often large and expensive) to obtain the psychophysiological feedback. Therefore, affective DDA is most likely not yet suitable for widespread use in video games. However, technological advances will undoubtedly address this issue. For example, preliminary work shows the potential of using machine vision techniques to infer players' emotional states based on real-time data obtained from the front-facing camera in smartphones and tablets (Bevilacqua et al., 2015; Bevilacqua et al., 2016). However, due to this current limitation of affective DDA, we will focus on performance-based DDA in the remainder of this paper. We will, however, return to the issue of affective DDA in Section 4, and show how the system used in the present study may also be used with affective feedback.

1.2 Effects of performance-based dynamic difficulty adjustment

Previous research has investigated the effects of using performance-based DDA on various aspects of the game playing experience. Several studies in this area have considered the relationship between DDA and player enjoyment. Alexander et al. (2013) used an experimental game to investigate how DDA compared with incremental difficulty adjustment. They found that, while casual gamers reported enjoying a simple 2D game more when the difficulty was dynamically adjusted according to their performance, experienced gamers (who made up most of the sample) enjoyed the game more when the difficulty was adjusted incrementally. This study also showed that players enjoyed the game more when the difficulty was tailored to their gaming experience (casual vs experienced) rather than their performance. However, in a sample of 90 players, only 19 were categorised as casual players. Furthermore, this classification was determined by the players' response to the question "are you a casual or experienced gamer?"; it is therefore difficult to determine how to understand the distinction between casual and experienced gamers as the classification criteria are not explicit.

Nagle et al. (2015) also found that performance-based DDA led to lower player enjoyment than an alternative system in which players could control the level of difficulty throughout the game themselves. They created a 3D game in which players had to memorize a list of objects and locations, then find the objects and place them in the correct location. The difficulty of the game was determined by the number of objects (more is harder) and the number of times they could view the list of objects and location numbers (fewer is harder). However, while player enjoyment was lower with DDA, DDA was associated with better player performance. That is, when they allowed players to control the number of objects and number of times the list could be consulted, performance was significantly lower than when these values were automatically adjusted based on player performance.

Sharek & Wiebe (2015) also showed that DDA was associated with better player performance. Using an isometric puzzle game, they created over 100 different levels which were tested and ordered by difficulty. They had three difficulty conditions: DDA, in which more difficult levels were provided to players based primarily on measures of performance; incremental; and a choice condition, in which players were given the option to select a

harder or easier level after each completed level. Players in the DDA condition showed significantly higher performance (indexed as reaching more difficult levels more quickly) than players in the other two conditions (see also Baldwin et al., 2014).

However, in a study by Orvis et al. (2008), no significant differences in performance or motivation were found between 4 groups, playing versions of a game with either no difficulty adjustment, incremental difficulty adjustment, or adaptive adjustment (two versions, distinguished in terms of the starting difficulty, which was either easy or hard). Other research on motivation finds similar negative results, with DDA not associated with significantly different levels of player motivation than incremental difficulty adjustment in a Spanish language education game (Sampayo-Vargas et al., 2013). However, in line with previous results showing increased performance, the authors showed that DDA led to significantly higher learning outcomes, which they attribute to a scaffolding effect wherein the reductions in difficulty (the "scaffold") are provided when students need support, then removed when students were ready to progress.

While Sharek & Wiebe (2015) claim that DDA increases performance with no decrease in engagement, Altimira et al. (2017) found that DDA increased player engagement in a digitally augmented game of table tennis. By projecting images onto a surface, they could increase or decrease the difficulty of the game (e.g., by altering the size of the virtual table projected onto the surface). They found that adjusting the difficulty in response to the score differentials between the two players (e.g., by making one player's half of the table smaller, thus making the game harder for the opposing player), was associated with significantly higher player engagement (self-report questionnaire) than no adjustment. Preliminary results from by research by Xue et al. (2017) using DDA in a mobile game distributed via the Google Play Store and Apple App Store show that DDA increases player engagement over a longer period (4 months). This study is notable as it measures player engagement objectively, in terms of total time spent playing the game. The authors also note that using DDA had no effect on the amount of revenue generated from in-game transactions.

Altimira et al. (2017) also highlight another potential application of DDA technology, in that they showed that using DDA significantly reduced the score differences between players, thus allowing less skilled players to be more competitive against more skilled players. Other studies have successfully used DDA to reduce skill differentials between players of different abilities (e.g., Jensen & Grønbæk, 2016; Hwang et al., 2017). Gerling et al. (2014) created a rhythm game (i.e., in which players must perform steps in time with music) which could either be controlled by a dance mat (i.e., the player inputs the rhythm with their feet), buttons on a standard video game controller, or a via a wheelchair input (wheelchair movements are captured by a motion sensor camera). Using various techniques, they produced adaptive versions of the game which allowed less-skilled able-bodied players to compete with more-skilled able-bodied players, and players with and without mobility disabilities to compete together. Bateman et al. (2011) also decreased performance differentials between players by providing adaptive targeting assistance in a simple shooting game. DDA may therefore be important for enabling people with disabilities to play multiplayer games with those without disabilities (Hernandez et al., 2013; Hwang et al., 2017). DDA may also be one factor which can increase the effectiveness of rehabilitation games for people with disabilities, both in terms of making such games accessible and in terms of adapting the difficulty of the games to suit players of a wide range of abilities and

provide an optimum level of challenge, which is shown to increase the effectiveness of such games (Barrett et al., 2016; Hocine et al., 2015).

To summarise, previous research generally supports the idea that performance-based DDA can increase player performance, and that it can be used to reduce performance differentials between players of different abilities. Proposed applications of this include making games more accessible or inclusive for people with disabilities, and, related to this, increasing the effectiveness of games for rehabilitation. However, in terms of player experience – i.e., engagement, enjoyment, and motivation – previous research provides mixed results on the effects of DDA.

1.3 Flow

Several issues relevant to DDA are captured in the concept of flow, which was first introduced by Csikszentmihalyi (1975). During a flow state, an individual is completely focused on an activity; it is an enjoyable and fulfilling experience (an "optimal" experience), often described colloquially as being in "the Zone" (Chen, 2007). Csikszentmihalyi identifies several characteristics of the flow state, including having clear goals, concentrating on the task at hand, feeling in control, and being engaged in a challenging activity requiring skill. Flow was originally modelled by Csikszentmihalyi as an optimal balance between anxiety and boredom; the zone in which one is challenged enough to not be bored, but skilled enough to not be anxious about one's performance (Csikszentmihalyi, 1988). Since Csikszentmihalyi's original work, a large body of research has further investigated and characterised flow, and several instruments have been developed for measuring the subjective experience of flow (Weibel et al., 2008). Some of this work has explicitly focused on video games, which have been claimed to "possess ideal characteristics to create and maintain flow experiences in that the flow experience of video games is brought on when the skills of the player match the difficulty of the game" (Sherry, 2004). The importance of the flow experience is shown by empirical research suggesting that flow is one of the reasons why people play video games (Perttula et al., 2017). In educational games, flow has been used as a measure of game quality, and a small amount of research suggests that the experience of flow is associated with the effectiveness of game-based learning (Perttula et al., 2017).

The notion of a balance between player skill and game difficulty shows the direct relevance of flow to performance-based DDA. If the aim of these approaches to DDA is to match the difficulty of a game to each unique player's skill level, then it seems likely that DDA could be used to achieve a balance between these two factors, and therefore encourage flow experiences. However, although theoretical discussions of flow feature in much research on DDA, we are aware of only one previous study investigating the relationship between difficulty adjustment and flow empirically. Using a modified version of *Tetris* and a within-subjects design with 30 participants, Ang & Mitchell (2017) showed that playing with incremental difficulty adjustment and with a version of DDA in which the player could control the difficulty were both associated with significantly different scores (compared to no DDA) on several constructs of the Flow State Scale (Jackson & Marsh, 1996). This included challenge-skill balance, which was greater when participants played a game with DDA. Note, however, that the DDA used in this study is not performance-based as discussed in this paper, in that players manually (and voluntarily) increased or decreased the difficulty

during the game themselves, rather than have the difficulty automatically adjusted based on a measurement of player performance.

1.4 Machine learning for a generalised adaption mechanism

One of the limitations of many previous DDA systems is their specificity. Much of the research discussed in this paper makes use of context-specific algorithms; that is, the algorithm is designed and implemented for the game at hand, based on "domain-specific requirements" (Chang, 2013). In other words, many previous systems rely on a heuristic that only applies to either a specific game, or a specific genre of games. This limits the extent to which these systems can be readily applied to other games, and limits the applicability of research findings using these systems.

The machine learning approach to DDA developed by Missura & Gärtner (2011) is intended to address this issue. Their partially ordered set master (POSM) algorithm is a generalised DDA mechanism which does not rely on a domain-specific heuristic. A detailed description of the algorithm is beyond the scope of this paper, but the procedure can be summarised as follows. Given a set of possible difficulty settings, some will be too hard for the player and some will be too easy. Furthermore, because difficulty settings can be ordered (i.e., they can be sorted from least to most difficult), it follows that if a setting is judged to be too easy (or too hard), then all settings easier (or harder) than it will also be too easy (or too hard). Based on this, the POSM algorithm operates by allowing the player to play one round of the game (this can be a level, or a period of any duration); next, based on feedback from the game (expressed simply as "too hard" or "too easy"), the algorithm updates its "belief" (a numerical value) about the suitability of that setting, and all settings easier (or harder) than that setting. The algorithm then selects the setting which currently has the highest belief value as the most appropriate for the player in the next round (see Missura & Gärtner, 2011, for a detailed description of the POSM algorithm). The advantage of this approach is that the algorithm is blind to the both what the settings are (they could be game speed, number of enemies, or anything that can be quantified) and the heuristic that determines the feedback; in other words, the message "too hard"/"too easy" is sent from the game to the POSM algorithm, and deciding which message to send is the responsibility of the game, not the POSM algorithm. This shows how POSM is generalised: it selects settings based on feedback, but it does not know what those settings are or how the feedback is determined. As such, it should be possible to easily apply POSM in almost any game, without modification. However, as far as we are aware, POSM has only been implemented and evaluated (with human players, as opposed to simulations) in turn-based strategy games (Ilici et al., 2012). Therefore, the feasibility of such an approach to DDA in other kinds of video game remains to be investigated.

1.5 The present study

Our aim in this study was to investigate how a performance-based approach to DDA based on POSM affects player performance, enjoyment, and experience of flow, using an experimental game created for this study. We conducted a controlled experiment with 3 groups, with each group playing a different version of the game. The independent variable was the way in which the difficulty of the game was adjusted, with 3 levels: (1) DDA, (2) incremental difficulty adjustment (in which the game gets progressively harder irrespective of player experience), and (3) no difficulty adjustment (control group). In the DDA version of the game, we used a generalised machine learning algorithm (based on POSM) to adjust the

difficulty, and a measure of player performance provided the feedback upon which the adjustments were based.

1.5.1 Hypotheses

Our hypotheses were (1) DDA will produce greater player performance than either incremental or no difficulty adjustment; (2) DDA will lead to greater experience of flow than either incremental or no difficulty adjustment; and (3) DDA will lead to greater enjoyment than either incremental or no difficulty adjustment.

2 Materials and methods

2.1 The game

To test our hypotheses, we designed and implemented a simple video game called “Meteor Shower” (Figure 1). The object of the game is to avoid the meteors which continually fall from the top of the screen while catching the pink falling stars. The player controls the yellow character by moving left or right along the bottom of the screen (using the left and right directional arrows on the keyboard). The velocity of the meteors determines the difficulty of the game, with higher velocities making the game more difficult (i.e., the meteors are harder to avoid). The game consists of 20 levels, each lasting 45 seconds, with a short pause between each level. Players are awarded one point for each star they catch, and lose a point each time a meteor hits the character. The score is reset to 0 at the end of each level. There are 8 falling stars to catch in each level (each falling 5 seconds apart), and so the maximum score available on any level is 8 (i.e., the player catches all stars and avoids all meteors). While the meteors and stars appear to the player to originate from random locations, the game in fact uses a seeded random number generator to ensure that the pattern of locations at which stars and meteors appear is the same for each player. When generated, stars fall in a straight line. Meteors move in a straight line from their point of origin, to the location of the player-controlled character at the time the meteor was generated. This ensures that players always must move the character to avoid every meteor. Three versions of the game were created. The sole difference between each version is the way in which the difficulty (i.e., the velocity of the meteors) was adapted during gameplay.

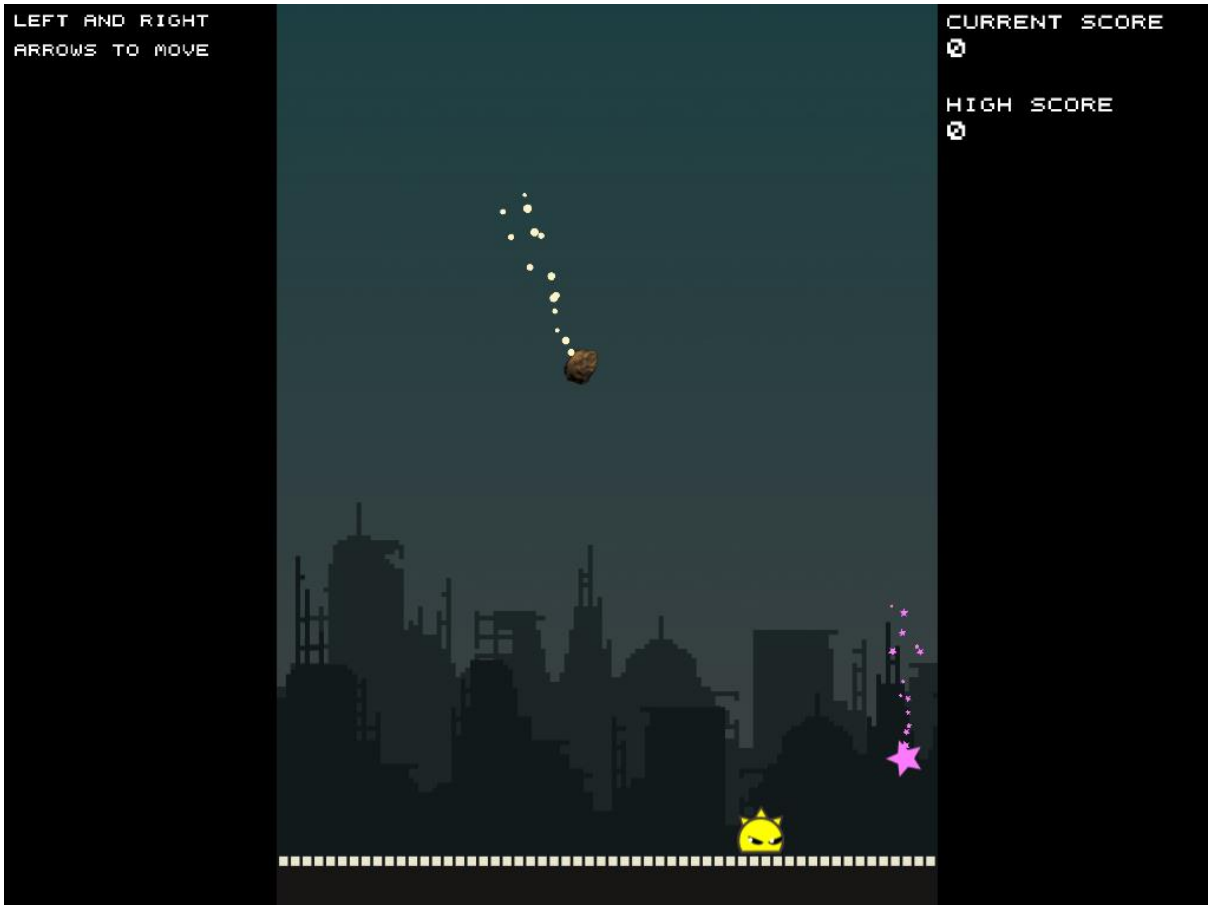

Figure 1. The experimental game, Meteor Shower. The player moves the yellow character left and right using the directional arrows on a keyboard. The meteor, which is moving towards the yellow character, should be avoided. The pink star, which is falling in a straight line, should be collected. The velocity at which the meteor falls dictates the difficulty of the game (faster is more difficult).

2.1.1 Control version of the game

In this version of the game, the velocity of the meteors remained constant at 800 throughout.

2.1.2 Incremental version of the game

In this version of the game, the velocity of the meteors increased by 50 at the start of each new level. The velocity on level 1 was 200; the velocity on level 20 was 1150.

2.1.3 DDA version of the game

This version of the game used a machine learning algorithm, based on the POSM algorithm, to adapt the velocity of the meteors in response to measurements of player performance. The starting value on level 1 was 900. The possible settings ranged from 200 to 1700, increasing in increments of 50. Due to an error in the programming, value 1200 was not used.

As described in Section 1.4, the POSM algorithm requires a message to be regularly sent from the game describing if the current setting is too hard or too easy for the player. In this study, this message was sent every 5 seconds. The decision was based on two measures of player success rate. The first of these measured the player's success at avoiding meteors over the previous 5 seconds of play (*success1* = (*nM* – *nH*)/*nM*, where *nM* is the number of meteors in the previous 5 seconds and *nM* is the number of times the player was hit by a

meteor in the previous 5 seconds). The second measurement was the player's overall success rate based on their score and the potential maximum score possible at that point in the level (*success2* = *score*/*nS*, where *nS* is the number of stars that have fallen so far in the level).

Determining which value to use for the target success rate (i.e., the value above which the game was judged to be too easy), proved to be one of the most challenging design decisions in the development of the game, and we were unable to find previous research to guide this decision. Therefore, during the development process, we played versions of the game using success rates ranging from 0.75 (i.e., 75%) to 1 (i.e., 100%). We determined that the most satisfying experience was provided when we used a target success rate of 1 for both measures.

2.2 Questionnaire

Flow was assessed using the Flow Short Scale (Rheinberg et al., 2003), which is shown to be a reliable measure of flow experience in video games (Wiebel et al., 2008). We used the online version of the scale (http://www.psych.uni-potsdam.de/people/rheinberg/messverfahren/fks1-e.html) which has 14 items. Items 1 – 10 measure flow, items 11 – 13 measure anxiety, and item 14 measures perceived skill demands (challenge) (see Appendix in Engeser, 2012). In addition, demographic information (age, gender, frequency of video game play) was collected.

2.3 Performance data

During gameplay, each participant's score was recorded for levels 1 – 19. Due to a bug in the software (which we identified after running the experiment), the score for level 20 was not recorded. For the DDA group only, we also recorded the velocity of the meteors at 5 second intervals (i.e., each time the velocity was updated). This provided a measure of the difficulty of the game (higher velocity is more difficult), and a measure of the player's skill level (better players will reach higher velocities). Scores were recorded for all participants in the DDA group for 855 seconds of gameplay (i.e., not the full 15 minutes, due to a technical problem or bug, currently unidentified). We did not record velocity for the control group, as this remained constant throughout the game (800) or for the incremental group, as this increased on a pre-determined scale with each level.

2.4 Data collection method: Amazon Mechanical Turk

We conducted the experiment online using Amazon Mechanical Turk (MTurk). MTurk has been described as a "marketplace for work that requires human intelligence" (Rouse, 2015). Users of the site are classified as "requesters" (who post tasks to the site) and "workers" (who complete these tasks in return for payment). Example tasks include completing surveys to provide feedback about a website, classifying images based on their content, or providing translations of short pieces of text. Typically, MTurk is appropriate for tasks which can be completed quickly, and for which many instances of the task must be completed. MTurk is now frequently used to conduct research in psychology (Paolacci & Chandler, 2014; Mason & Suri, 2012), and it has been used in at least one previous study on the effects of DDA in video games (Sharek & Wiebe, 2015).

However, several issues have been identified which may threaten the validity of data obtained from MTurk (Cheung et al., 2017). Some of these issues are not unique to MTurk. For example, the issue of selection bias, in the sense that MTurk workers choose the tasks

they wish to complete, applies in any research wherein participants voluntarily opt to take part after viewing, e.g., a poster or advertisement on social media. However, all participants in research conducted via MTurk will (obviously) have self-selected to become MTurk workers. Related to this, Cheung et al. (2017) note that samples obtained from MTurk may not be representative of the population of interest (e.g., all MTurk participants are internet users, which not be appropriate for some studies). On this issue, we make two observations. Firstly, we note that, in some respects, samples obtained from MTurk may be more diverse than samples obtained by traditional means. Casler et al. (2013) point out that most participants in psychological research are American college students; they showed that a sample of MTurk workers was significantly, desirably more diverse in terms of ethnicity, economic status, and age, than a sample of undergraduate students. Secondly, we point out that the nature of our study – in which participants are required to play an online video game remotely – dictates that participants would be required to have internet access and be reasonably computer literate whether recruited through MTurk or not.

Perhaps the most important concern with MTurk data highlighted by Cheung et al. (2017) is the possibility of participants answering questions without paying attention to the content (either fully, or at all, i.e., selecting random answers). However, steps can be taken to mitigate this risk (see also Fleischer et al., 2015). Firstly, MTurk incorporates a rating system for workers, so that requesters can specify that only workers of a suitable quality can access their tasks. We will discuss this further in Section 2.6, where we describe how we used this system to specify that only workers of a certain quality could access our experiment. Secondly, items can be included in questionnaires to check for attentiveness (e.g., a multiple-choice item that states which option the respondent should select). Finally, it may also be possible to detect inattentive responses by analysing data gathered, although this would presumably depend on the nature of the data. The screening process we used to identify inattentive participants in the present study is described in Section 2.7.

As should be the case in all data collection processes, we recommend that consideration is given to potential limitations of data collected using MTurk, and that these limitations are addressed as fully as possible.

2.5 Procedure

The task was made available using MTurk's Survey Link template. This provides a link to an external website, where participants complete a task (typically a questionnaire) and receive a completion code. They then enter the completion code in MTurk to receive credit for completing the task. The default configuration of the Survey Link template allows each worker to only complete the task once. In our case, the survey link took participants to a site hosting the game. Which version of the game was loaded was determined randomly by the software. Participants pressed a button to start the game when they were ready. After 15 minutes of play, the game automatically ended, and participants were presented with a link to the webpage containing the questionnaire. When they submitted the questionnaire with all questions completed, the data were stored on a server, and a unique completion code was generated on the server and returned to participants. The completion code, a record of which version of the game they had played, and performance data automatically recorded during gameplay, were also stored on the server. Participants then entered the completion code in MTurk, and were paid $0.99. All participants were paid, whether their data were included in the analysis or not.

2.6 Pilot

Initially, we ran a pilot with 10 participants. By considering participants' performance data, it appeared that some participants did not actually play the game. It would be possible to merely let the game run for 15 minutes, then select random answers to the questions. To address this, we included two attention check items in the questionnaire. These items stated that the participant should select option 7 ("very much") in the Likert-style scale. We also screened the data collected during the full experiment and removed data which appeared to show no engagement with the game (see Section 2.7). In addition, we decided to use MTurk's qualifications feature to ensure that the task was only available to workers who (1) had completed over 10000 tasks on MTurk; and (2) and an approval rate of 97% or higher.

2.7 Participants

Data were collected from 300 adults via MTurk, each randomly assigned to play either the control, DDA, or incremental versions of the game. Entries in which answers to either of the attention check questions were incorrect were removed. We also removed entries in which a score of zero was recorded for every level, as this suggested that the participants did not actually play the game, but merely let it run for the required time. Finally, we considered the velocity data recorded for participants in the DDA group, and removed any entries in which the pattern of velocities across the levels suggested that participants had not actually played the game (i.e., when the velocity quickly decreased to 200 and did not rise above 250 for the remainder of the game).

This left 221 participants (93 female) whose data were retained for analysis. Ages ranged from 20 years to 65 years, with a mean age of 36.51 years (std. deviation 9.73). There were 82 participants in the control group, 68 in the DDA group, and 71 in the incremental group.

3 Results

The 14-item online version of the Flow Short Scale was shown to have acceptable reliability (Cronbach's a = .834).

Chi-square tests of homogeneity showed that the three groups did not significantly differ in terms of gender ($\chi^2(3) = 1.4$, $p = .497$), number of days per week spent playing video games ($\chi^2(3) = 23.348$, $p = .055$), or age ($\chi^2(3) = 69.525$, $p = .902$).

3.1 Significant results: performance

To analyse group differences in terms of overall mean score (i.e., participants' mean score over 19 levels of play), we ran a Kruskal-Wallis H test. The distributions of overall mean score were not similar for all groups. The DDA group had a smaller range (2.89) and smaller interquartile range (.83) than both the control group (range = 7.32, interquartile range = 3.01) and incremental group (range = 7.74, interquartile range = 2.89) (see Figure 2). The median values increased from the incremental group (3.74), to the control group (4.61) to the DDA group (5.05). These differences in median values were statistically significantly different between the groups, $\chi^2(3) = 16.148$, $p < .0005$. Pairwise comparisons were then performed using Dunn's (1964) procedure with a Bonferroni correction for multiple comparisons. This analysis revealed statistically significant differences (adjusted p-values presented) in median values between the DDA group (mean rank = 135.31) and control

group (mean rank = 106.91) (p = .02), and between the DDA group and incremental group (mean rank = 92.44) (p < .0005) groups, but not between the control group and the incremental group (p = .488).

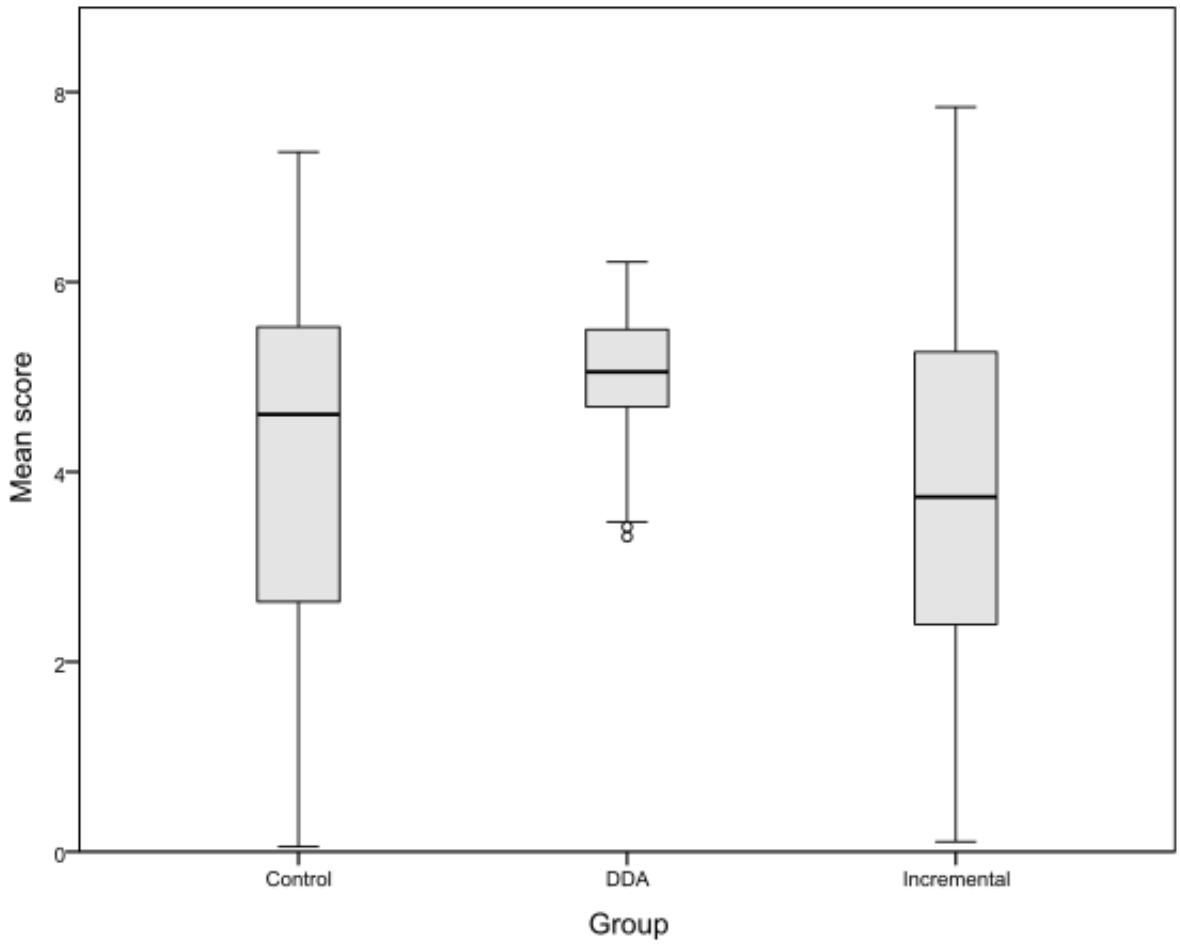

Figure 2. Boxplots of overall score for each of three groups, Control, Dynamic Difficulty Adjustment (DDA), and Incremental.

We analysed the differences in scores between the three groups further by conducting one-way Welch ANOVAs on mean score per level (i.e., mean score for each group for each of 19 levels of play). A small number of outliers in the DDA group and Incremental group were not removed. Levene's test for equality of variances showed that the assumption of homogeneity of variances was violated in levels 4 – 19 (p-values ranging from .012 to <.0005). Scores were significantly different between the groups during levels 1, 3, 5, and levels 9 – 19. Games-Howell post hoc analyses revealed that the DDA group had higher mean scores than the control group in levels 3 – 19; these differences were significant in levels 3, 5, and 9 – 19 (p-values ranging from .019 to <.0005). The DDA group also had significantly higher mean scores than the incremental group in levels 10 – 19 (p-values ranging from .013 to <.0005). The control group had significantly higher mean scores than the incremental group in levels 14 – 19 (p-values ranging from .014 to <.0005). These results are shown in Figures 3.1, 3.2, and 3.3.

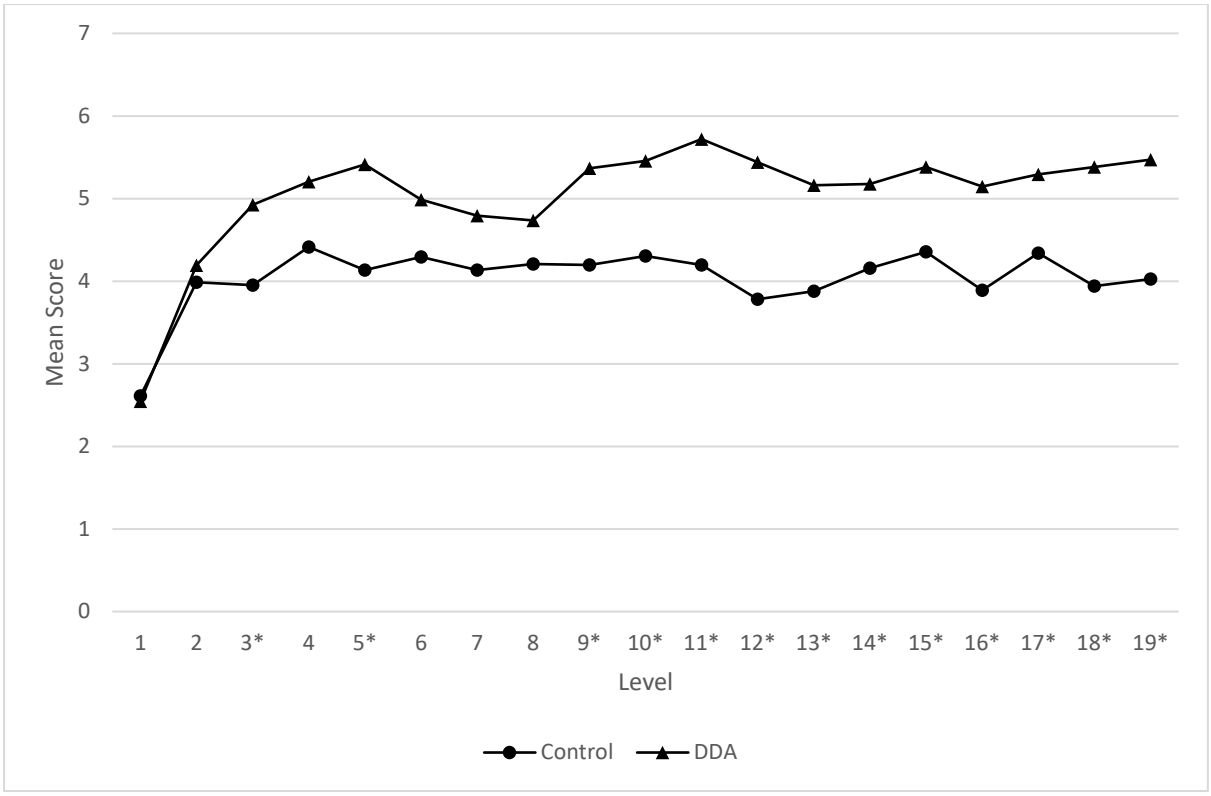

Figure 3.1. A comparison of mean score per level for the Control and Dynamic Difficulty Adjustment (DDA) groups. Significant differences are marked with a *.

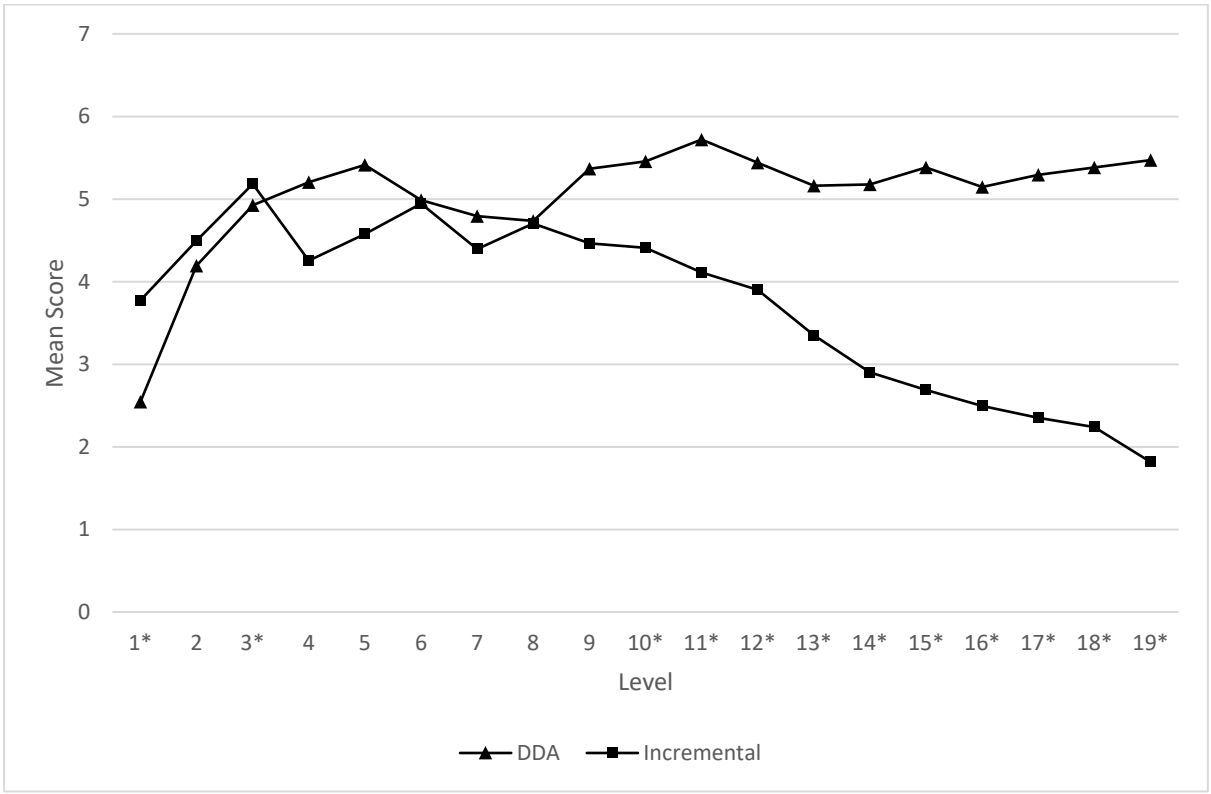

Figure 3.2. A comparison of mean score per level for the Dynamic Difficulty Adjustment (DDA) and Incremental groups. Significant differences are marked with a *.

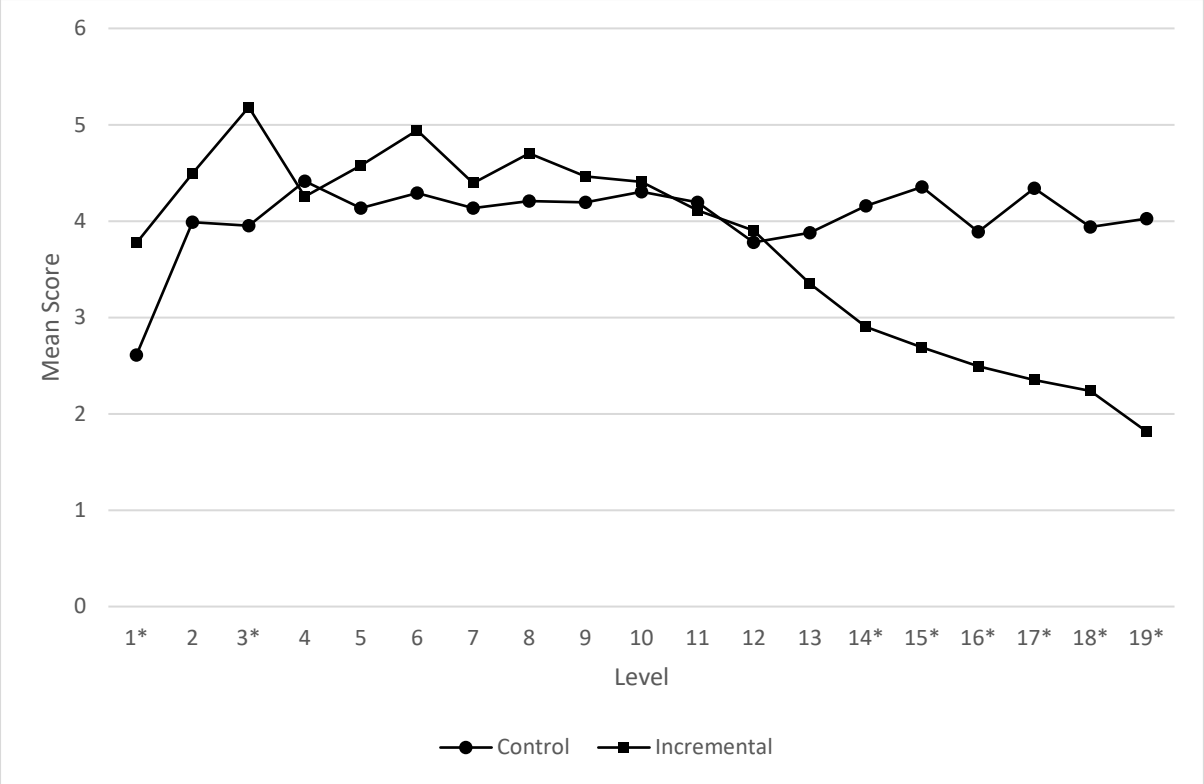

Figure 3.3. A comparison of mean score per level for the Control and Incremental groups. Significant differences are marked with a *.

For the DDA group, we analysed the velocity values, which indicates the difficulty of the game (higher velocity is more difficult). First, we considered the mean velocity across participants at each measurement point (5 seconds between each measurement, 171 measurements considered). Figure 4 shows how velocity ranged across participants over time. We considered the relationship between each participant's (DDA group only) overall mean velocity over 855 seconds of gameplay, and each participant's overall mean score across 19 levels, using Pearson's correlation test. The data were linear, overall mean score was normally distributed ($p > .05$), while overall mean velocity was not normally distributed ($p < .0005$). There was a strong positive correlation between overall mean score and overall velocity, $r(68) = .554$, $p < .0005$, with overall mean velocity explaining 31% of the variation in overall mean score

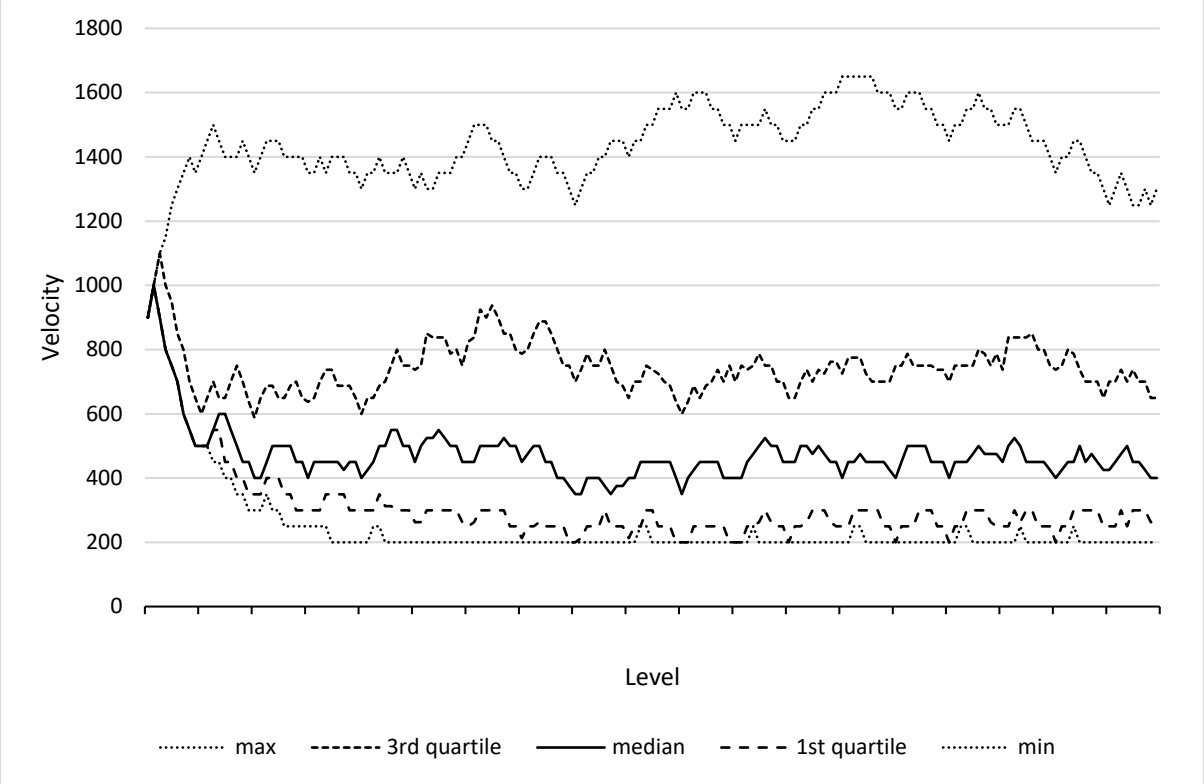

Figure 4. Range of meteor velocity per level for the Dynamic Difficulty Adjustment (DDA) group over 19 levels. As a higher velocity makes the game more difficult, and higher velocities are only achieved by players who perform better, velocity here can be used as an index of player performance.

3.2 Non-significant results: flow, anxiety, challenge, and enjoyment

We also ran Kruskal-Wallis H tests to analyse differences between the groups in terms of the 3 factors of the Flow Short Scale, and the single-item for Enjoyment.

In the case of Flow, the distributions were not similar for all groups (assessed by visual inspection of a box plot). The median values increased from the incremental group (4.9), to the DDA group (4.95), to the control group (5.3). These differences were not statistically significantly different between the groups, $\chi^2(3) = 4.087$, $p = .130$.

In the case of Anxiety, the distributions were similar for all groups. The median values increased from the incremental group (3.67), to the control group (4.0) and DDA group (4.0) These differences were not statistically significantly different, $\chi^2(3) = 3.336$, $p = .189$.

In the case of Challenge, the distributions were not similar for all groups and the median values were the same for all three groups (4.0).

In the case of Enjoyment, the distributions were not similar for all groups. The median values increased from the incremental group (5.0), to the control group (6.0) and the DDA group (6.0). These differences were not statistically significantly different, $\chi^2(3) = 3.628$, $p = .163$.

3.3 Correlations

We also considered relationships between flow, anxiety, challenge, enjoyment, and overall mean score across all three groups. Flow was positively correlated with Anxiety ($r(221) = .462$, $p < .0005$), Challenge ($r(221) = .476$, $p < .0005$), and Enjoyment ($r(221) = .675$, $p < .0005$), Anxiety was positively correlated with Challenge ($r(221) = .531$, $p < .0005$) and

Enjoyment ($r(221) = .511$, $p < .0005$), and Challenge was positively correlated with Enjoyment ($r(221) = .501$, $p < .0005$). Overall mean score was significantly negatively correlated with Challenge ($r(221) = -.181$, $p = .007$).

3.4 Hypotheses

In the case of hypothesis (1) – that DDA will produce greater player performance than either incremental or no difficulty adjustment – we are able to reject the null hypothesis.

However, for hypotheses (2) and (3), we are unable to reject the nulls. That is, we cannot reject the hypotheses that there is no difference between DDA, incremental difficulty adjustment, and the control group in terms of either player experience of flow or player enjoyment.

4 Discussion, limitations, and future work

The research presented here investigated dynamic difficulty adjustment in a video game, with the adjustment based on feedback about the player's performance, and using a machine learning algorithm to select the appropriate setting for a single variable which directly affected game difficulty. We found that this type of difficulty adjustment led to greater overall player performance over approximately 15 minutes of play, than either incremental difficulty adjustment or a no-adjustment control group. In addition, the range of performance in the DDA group was smaller than the other groups, and overall performance in the DDA group correlated with overall mean difficulty across 15 minutes of play.

In line with previous research, our results suggest that performance-based DDA is suitable for reducing skill differentials between players. This provides further evidence of the suitability of this relatively simple approach to DDA in facilitating competitive and or collaborative play between players with different skill levels. We believe that this approach could therefore be used to increase the accessibility of video games for people with disabilities.

It is also interesting to note that the players in the DDA group showed a wide range of abilities, as indexed by the range of difficulty settings recorded during the experiment (Figure 4). However, the overall mean performance of this group increased more than both the control and incremental groups, with significant differences in mean performance in most levels of the game between the DDA and other groups. It is therefore possible, in line with the results of Sampayo-Vargas et al. (2013), that DDA provides a scaffold to players of a range of abilities, by making the game easier when their performance drops; this scaffold is then removed when their performance increases. However, there may be detrimental effects associated with DDA if players are aware that it is operating (Baldwin et al., 2014; Baldwin et al., 2016), and further research should investigate how players' awareness of DDA is related to their experience of playing a video game.

Our results show no significant differences between the three conditions on all the self-reported measures of player experience (flow, enjoyment, challenge, and anxiety). Previous research on performance-based DDA has found mixed results on self-report measures of player experience such as enjoyment, engagement, motivation, and challenge. There are several possible explanations for this. Firstly, it may be the case that performance-based approaches to DDA have less or no effect on self-reported player experience than affective

approaches. This is feasible, as several studies have found that player perceptions of gameplay experience are related to factors other than player skill or performance, such as gameplay experience (Alexander et al., 2013), motivations, personality, or preferences (Karpinskyj et al., 2014). If the aim of DDA is to increase players' positive perception of the gameplay experience, then affective approaches to DDA may be more useful. Secondly, within performance-based DDA approaches, there is a large range of factors which could influence player experience. In this study, we used simple measures of player performance to alter a single variable affecting game difficulty. There are many other factors we could have chosen. In addition, we could have provided a different range of difficulty settings (e.g., with larger or smaller increments), used a different target success rate, adapted the difficulty more or less frequently, and so on. These adjustments could lead to different results, and future research should consider, not just the difference between adaptive and non-adaptive difficulty adjustment, but also differences between alternative approaches to DDA. Thirdly, as discussed in Section 1.4, self-report data obtained from MTurk may be less reliable than data obtained from traditional sources. While we included attention-check items to identify participants who were potentially selecting random answers to the questions, found high reliability for our questionnaire, and found expected correlations between flow, anxiety, challenge, and enjoyment, we did not use any other techniques to identify participants who were not engaged with the questionnaire. For example, it has been suggested that MTurk data can be made more reliable by explicitly asking participants if they answered the questions genuinely and assuring them that they will still be paid if they admit they did not (Rouse, 2015). Note that this limitation is somewhat mitigated in this study as we removed responses from participants whose performance data indicated that they had not engaged with the game. However, it is still feasible that some participants engaged with the game and read the questions but still provided unreliable data simply by not providing considered answers.

While this study focused on a performance-based approach to DDA, it is important to recall that the POSM algorithm used is not specific to this context or to the game used in our experiment. This means that our approach, in principle, could be matched with any other feedback that can be quantified. It could therefore be used, not only with other performance measures, but, crucially, with affective measures, as any real-time quantitative data could be used as the feedback upon which difficulty is updated. This could include, for example, heart rate, galvanic skin response, or neurological activity. In future work, it would be informative to test the feasibility of the POSM algorithm in affective DDA systems.

5 Conclusion

This study makes several contributions to research on dynamic difficulty adjustment in video games. Our results show that performance-based dynamic difficulty adjustment, based on the Partially ordered set master (POSM) algorithm (Missura & Gärtner, 2011) can be used to increase player performance and reduce performance differentials in a 2D video game. We also demonstrate the feasibility of conducting video games research using an experimental game via Amazon Mechanical Turk. We make recommendations for future research to further investigate the effects of dynamic difficulty adjustment on enjoyment and experience of flow (for which we found non-significant results), such as using different feedback measures (including affective feedback), adjusting different game variables, and

implementing additional steps to ensure the reliability of data gathered by player self-report.

References

Adams, E. (2008, May 14). The designer's notebook: Difficulty modes and dynamic difficulty adjustment. Retrieved from *https://www.gamasutra.com/view/feature/132061/the_designers_notebook_.php*

Alexander, J. T., Sear, J., & Oikonomou, A. (2013). An investigation of the effects of game difficulty on player enjoyment. *Entertainment Computing*, *4*(1), 53-62.

Altimira, D., Clarke, J., Lee, G., Billinghurst, M., & Bartneck, C. (2017). Enhancing player engagement through game balancing in digitally augmented physical games. *International Journal of Human-Computer Studies*, *103*, 35-47.

Ang, D. & Mitchell, A. (2017). Comparing effects of dynamic difficulty adjustment systems on video game experience. In Proceedings of the Annual Symposium on Computer-Human Interaction in Play, 317-327. ACM.

Baldwin, A., Johnson, D., & Wyeth, P. A. (2014). The effect of multiplayer dynamic difficulty adjustment on the player experience of video games. In *Proceedings of the extended abstracts of the 32nd annual ACM conference on Human factors in computing systems* (pp. 1489-1494). ACM.

Baldwin, A., Johnson, D., & Wyeth, P. (2016). Crowd-pleaser: Player perspectives of multiplayer dynamic difficulty adjustment in video games. In *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play* (pp. 326-337). ACM.

Barrett, N., Swain, I., Gatzidis, C., & Mecheraoui, C. (2016). The use and effect of video game design theory in the creation of game-based systems for upper limb stroke rehabilitation. *Journal of Rehabilitation and Assistive Technologies Engineering*, *3*, 2055668316643644.

Bateman, S., Mandryk, R. L., Stach, T., & Gutwin, C. (2011). Target assistance for subtly balancing competitive play. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2355-2364). ACM.

Bevilacqua, F., Backlund, P., & Engstrom, H. (2015). Proposal for non-contact analysis of multimodal inputs to measure stress level in serious games. In *Games and Virtual Worlds for Serious Applications (VS-Games), 2015 7th International Conference on*. IEEE.

Bevilacqua, F., Backlund, P., & Engstrom, H. (2016). Variations of Facial Actions While Playing Games with Inducing Boredom and Stress. In *Games and Virtual Worlds for Serious Applications (VS-Games), 2016 8th International Conference on*. IEEE.

Booth, M. (2009). The AI systems of Left 4 Dead. Keynote presented at the 5th Artificial Intelligence and Interactive Digital Entertainment Conference, Stanford, CA.

Bontchev, B. (2016). Adaptation in Affective Video Games: A Literature Review. *Cybernetics and Information Technologies*, *16*(3), 3-34.

Casler, K., Bickel, L., & Hackett, E. (2013). Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Computers in Human Behavior*, *29*(6), 2156-2160.

Chang, D. M. J., (2013). Dynamic difficulty adjustment in computer games. Retrieved from http://studylib.net/doc/8266212/dynamic-difficulty-adjustment-in-computer-games

Chen, J. (2007). Flow in games (and everything else). *Communications of the ACM*, *50*(4), 31-34.

Cheung, J. H., Burns, D. K., Sinclair, R. R., & Sliter, M. (2017). Amazon Mechanical Turk in organizational psychology: An evaluation and practical recommendations. *Journal of Business and Psychology*, *32*(4), 347-361.

Csikszentmihalyi, M. (1975). Beyond boredom and anxiety: Experiencing flow in work and play. San Fransisco: Jossey-Bass.

Csikszentmihalyi, M. (1988). The flow experience and its significance for human psychology. In M. Csikszentmihalyi & I. Csikszentmihalyi (Eds.), Optimal experience: Psychological studies of flow in consciousness (pp. 15–35). Cambridge: Cambridge University Press.

Dunn, O. J. (1964). Multiple comparisons using rank sums. *Technometrics*, *6*(3), 241-252.

Engeser, S. (Ed.). (2012). *Advances in flow research*. Springer Science & Business Media.

Entertainment Software Association (2017). Essential facts about the computer and video game industry. http://www.theesa.com/wp-content/uploads/2017/09/EF2017_Design_FinalDigital.pdf

Fleischer, A., Mead, A. D., & Huang, J. (2015). Inattentive responding in MTurk and other online samples. *Industrial and Organizational Psychology*, *8*(2), 196-202.

Gerling, K. M., Miller, M., Mandryk, R. L., Birk, M. V., & Smeddinck, J. D. (2014). Effects of balancing for physical abilities on player performance, experience and self-esteem in exergames. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems* (pp. 2201-2210). ACM.

Hernandez, H. A., Ye, Z., Graham, T. C., Fehlings, D., & Switzer, L. (2013). Designing action-based exergames for children with cerebral palsy. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1261-1270). ACM.

Hocine, N., Gouaïch, A., Cerri, S. A., Mottet, D., Froger, J., & Laffont, I. (2015). Adaptation in serious games for upper-limb rehabilitation: an approach to improve training outcomes. *User Modeling and User-Adapted Interaction*, *25*(1), 65-98.

Hunicke, R. (2005). The case for dynamic difficulty adjustment in games. In *Proceedings of the 2005 ACM SIGCHI International Conference on Advances in computer entertainment technology* (pp. 429-433). ACM.

Hwang, S., Schneider, A. L. J., Clarke, D., Macintosh, A., Switzer, L., Fehlings, D., & Graham, T. C. (2017). How Game Balancing Affects Play: Player Adaptation in an Exergame for Children with Cerebral Palsy. In *Proceedings of the 2017 Conference on Designing Interactive Systems* (pp. 699-710). ACM.

Ilici, L., Wang, J., Missura, O., & Gärtner, T. (2012). Dynamic difficulty for checkers and Chinese chess. In *Computational Intelligence and Games (CIG), 2012 IEEE Conference on* (pp. 55-62). IEEE.

Jackson, S. A., & Marsh, H. W. (1996). Development and validation of a scale to measure optimal experience: The Flow State Scale. *Journal of sport and exercise psychology*, *18*(1), 17-35.

Jensen, M. M., & Grønbæk, K. (2016). Design strategies for balancing exertion games: A study of three approaches. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems* (pp. 936-946). ACM.

Karpinskyj, S., Zambetta, F., & Cavedon, L. (2014). Video game personalisation techniques: A comprehensive survey. *Entertainment Computing*, *5*(4), 211-218.

Leiker, A. M., Bruzi, A. T., Miller, M. W., Nelson, M., Wegman, R., & Lohse, K. R. (2016). The effects of autonomous difficulty selection on engagement, motivation, and learning in a motion-controlled video game task. *Human movement science*, *49*, 326-335.

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior research methods*, *44*(1), 1-23.

Missura, O., & Gärtner, T. (2011). Predicting dynamic difficulty. In *Advances in Neural Information Processing Systems* (pp. 2007-2015).

Nagle, A., Novak, D., Wolf, P., & Riener, R. (2014). The effect of different difficulty adaptation strategies on enjoyment and performance in a serious game for memory training. In *Serious Games and Applications for Health (SeGAH), 2014 IEEE 3rd International Conference on* (pp. 1-8). IEEE.

Orvis, K. A., Horn, D. B., & Belanich, J. (2008). The roles of task difficulty and prior videogame experience on performance and motivation in instructional videogames. *Computers in Human behavior*, *24*(5), 2415-2433.

Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science*, *23*(3), 184-188.

Perttula, A., Kiili, K., Lindstedt, A., & Tuomi, P. (2017). Flow experience in game based learning–a systematic literature review. *International Journal of Serious Games*, *4*(1).

Rheinberg, F. Vollmeyer, R. Engeser, S. (2003). Die Erfassung des Flow-Erlebens [Measuring flow experiences]. In J. Stiensmeier-Pelster, F. Rheinberg (Eds.), *Diagnostik von Motivation und Selbstkonzept. Tests und Trends, Vol. 2*, Hogrefe, Göttingen (2003), pp. 261-279

Robb, N., Waller, A., & Woodcock, K. A. (2019). Developing a task switching training game for children with a rare genetic syndrome linked to intellectual disability. *Simulation & Gaming*, *50*(2), 160-179.

Rouse, S. V. (2015). A reliability analysis of Mechanical Turk data. *Computers in Human Behavior*, *43*, 304-307.

Sampayo-Vargas, S., Cope, C. J., He, Z., & Byrne, G. J. (2013). The effectiveness of adaptive difficulty adjustments on students' motivation and learning in an educational computer game. *Computers & Education*, *69*, 452-462. doi: 10.1016/j.compedu.2013.07.004

Shaker, N., Yannakakis, G., & Togelius, J. (2010). Towards automatic personalized content generation for platform games. In 6th AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment, AIIDE 2010.

Sharek, D., & Wiebe, E. (2015). Investigating Real-time Predictors of Engagement: Implications for Adaptive Videogames and Online Training. *International Journal of Gaming and Computer-Mediated Simulations (IJGCMS), 7*(1), 20-37. doi:10.4018/IJGCMS.2015010102

Sherry, J. L. (2004). Flow and media enjoyment. Communication Theory, 14(4), 328–347.

Thibault, R. T., Lifshitz, M., & Raz, A. (2016). The self-regulating brain and neurofeedback: experimental science and clinical promise. *Cortex*, *74*, 247-261.

Weibel, D., Wissmath, B., Habegger, S., Steiner, Y., & Groner, R. (2008). Playing online games against computer-vs. human-controlled opponents: Effects on presence, flow, and enjoyment. *Computers in Human Behavior*, *24*(5), 2274-2291. doi: 10.1016/j.chb.2007.11.002

Williams, D., Yee, N., & Caplan, S. E. (2008). Who plays, how much, and why? Debunking the stereotypical gamer profile. *Journal of Computer-Mediated Communication*, *13*(4), 993-1018.

Xue, S., Wu, M., Kolen, J., Aghdaie, N., & Zaman, K. A. (2017). Dynamic Difficulty Adjustment for Maximized Engagement in Digital Games. In *Proceedings of the 26th International Conference on World Wide Web Companion* (pp. 465-471). International World Wide Web Conferences Steering Committee.