Playing with Scales: Creating a Measurement Scale to Assess the Experience of Video Games

Mark James Parnell

Project report submitted in part fulfilment of the requirements for the degree of Master of Science (Human-Computer Interaction with Ergonomics) in the Faculty of Life Sciences, University College London, 2009

Acknowledgments

My heartfelt thanks to everyone that has helped me over the course of this project – from friends to family to participants, all of your help has been much appreciated. Many thanks to my supervisors, Dr. Nadia Berthouze and Dr. Duncan Brumby, for their kind guidance and patience, as well as to Dr. Eduardo Calvillo Gámez for his support.

Thanks also to Dr. Wijnand IJsselsteijn and Karolien Poels of Eindhoven University for the use of the GEQ questionnaire in the review, as well as to Laura Ermi and Dr. Frans Mäyrä for supplying a copy of their Immersion questionnaire for review.

Finally, special thanks to David Tisserand at SCEE for all of the insight, assistance and backing that he has provided me throughout this project. The project couldn't have happened without it.

- MJP

Abstract

A video game should be appealing to play. It should be usable, playable and provide enjoyable experiences. One tool for assessing the appeal of a game is to have gamers complete a questionnaire (or scale) after they have played the game. Of the current battery of scales that exist, none of them provides an integrated measure of a game's appeal. To address this gap a Gameplay Scale is presented that assesses gamers' attitudes towards a game's appeal and quality. The Gameplay Scale is validated across two studies. Study 1 had gamers (n = 98) respond to a web survey after playing the downloadable game *PixelJunk Eden* for 2 hours. Cluster analysis of responses found that the Gameplay Scale contained distinct subscales measuring different gameplay constructs: (1) Affective Experience, (2) Focus, (3) Playability Barriers, and (4) Usability Barriers. Overall, the Gameplay Scale accounted for 73% of the variance in a game's initial appeal. Study 2 validated the Gameplay Scale by showing how it generalizes to different genres of games (i.e. open-world) and is able to predict a game's appeal and quality (i.e. by review score) after a relatively short period of game play (1 hour). These findings suggest that the Gameplay Scale can predict the appeal and quality of a game. This information may be of value to game developers who wish to evaluate a game's likely appeal during the development process.

Contents

List of Illustrative Figures

List of Tables

1 Introduction

> "*Play is older than culture, for culture, however inadequately defined, always presupposes human society, and animals have not waited for man to teach them their playing*" (Huizinga, 1938/1998; p. 1.)

Academic video games researchers (both within and outside of the HCI community) now research play experiences extensively, yet few of their findings have any impact upon the video games industry (Hopson, 10 November 2006). A great challenge for researchers is to support industry practice, and one way to do this is to develop tools to improve the user experience of games. As with all software, the developers of video games are not the same as their users. Despite being gamers themselves, their attitudes towards their creations will inevitably differ to those of their audience. The result is that video games often have issues where the end user struggles to operate or understand the game. In productivity software development, the remedy for this has been to use usability principles and testing methods to detect and eliminate any such usability problems. However, such techniques have, until recently, been slow to catch on in video games development (Fulton, 2002), yet they are perhaps even more important here. A user who has to struggle with a poorly-designed word processor at their office may grumble (and have reduced efficiency) but in the end has to use the word processor. This is not the case with video games – playing video games is a choice, and the player can always put the controller down if the game is too hard, or clunky, or simply isn't any fun (Laitinen, 23 June 2005). This

means that removing any barriers to play – and to fun – is of the utmost importance if the video game is to be as appealing as it can be.

Methodologies to test the usability and user experience of video games (often called 'player testing' methods) do exist (e.g. Pagulayan *et al*, 2003; Kim *et al*, 2008) and have no doubt improved the play experience of many games. One important part of these methods involves the use of questionnaires to measure player attitudes after play, especially since 'think-aloud' protocols during the game can distort the player experience significantly (Pagulayan *et al*, 2003). No questionnaire exists in the literature that measures all aspects of play experiences; what Hassenzahl *et al* (2000) called the 'ergonomic' and 'hedonic' factors that make up user experience; there is a real need to measure all elements of the experience. Moreover, Hornbaek (2006) called the measurement of user experience via questionnaire 'in disarray' with little utilisation of existing research or methods; there is also a need to create a well-designed measurement tool. This thesis aims to determine how best to measure these 'hedonic' and 'ergonomic' factors using questionnaires, by developing a new questionnaire that is both valid and reliable. To prove that this is useful to the industry, whether review scores can be predicted by the scale will also be determined.

The first chapter reviews the literature, first to identify what elements the scale must measure, before examining best practice in questionnaire design. Previous similar questionnaires are then reviewed to determine what they got right (and wrong). The third chapter involves the first study, in which the questionnaire is initially developed and validated, whilst chapter four involves the further validation of that questionnaire in experimental conditions. Chapter five will discuss the successes, failures and implications of the study, whilst chapter six serves as a recapitulation and conclusion of the research.

2 Literature Review

2.0.1 Overview

There are numerous ways to measure the usability of productivity software that can be translated to video games; all have their strengths and weaknesses, but one that can be particularly useful in the context of user testing is the questionnaire. As will be argued, questionnaires play an important role in player testing yet there is no existing standardised questionnaire that measures all of the factors contributing to a game's appeal that need to be considered. Additionally, many existing questionnaires that do measure some of the factors are flawed. The first section of this review will define the key factors to be measured; these are usability, playability and player experience. The different sorts of player experience will then be examined, as will the application of the terms 'playability' and 'usability' to video games. Once we have considered what factors any novel questionnaire will need to include, best practice in questionnaire design will be examined, and these principles used to critique existing questionnaires. Such a review should provide principles with which a new player testing questionnaire can be developed.

2.0.2 Definitions of Terms Used

It is common in the field of Human-Computer Interaction (HCI) to divide the interaction between user and system into the overarching factors of usability and user experience. Usability is often described using the ISO 9241-11 definition.

> "*The extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use.*" (ISO 9241-11, 1998)

Such a definition provides us with the essence of usability – the quality of how well the software hinders or enables users' achievement of their goals when using software. We might also wish to add the factor of *learnability* to this definition, as a system could be effective, efficient and satisfying yet very difficult to learn (Abran *et al* 2003), which would inhibit system usability for novices.

Despite satisfaction being listed above as an element of usability it is rarely considered as such. Indeed, for a long time it was rarely considered at all, with the focus of research and evaluation purely being upon effectiveness and efficiency (i.e. Nielsen and Molich, 1990; Polson et al, 1994). Gradually, the HCI community realised that it needed to go 'beyond usability' and simple measures of satisfaction to examine the broader user experience, including issues such as self-efficacy, aesthetics, social factors and fun (Dillon, 2003). Applied to video games, this gives us the concept of player experience – the experience of the user playing the game.

It would, however, be erroneous to fully cleave the affective experience of an interaction from the usability of the software used. For example, it is now common knowledge that aesthetic properties of a system can influence the perceived usefulness of the system (Tractinsky *et al*, 2000). Moreover, both hedonic (i.e. experiential) and ergonomic (i.e. usability) qualities have been found to influence a system's appeal to users (Hassenzahl *et al*, 2000). This entails that these two factors interact and that both are important for systems. Whilst these factors vary in their importance for

different systems, it is likely that both influence the appeal of all software. This is the case with video games; games may not involve goals in the manner of productivity software but have the purpose of delivering a certain experience. Usability issues must be corrected so that such an experience can be delivered.

The final factor is *sui generis* to video games (or rather, to the domain of games in general): playability. The distinction between usability and playability is usually seen as usability relating to interface and control issues and playability relating to game mechanic issues (Korhonen and Koivisto, 2007; Febretti and Garzotto, 2009). A game menu being difficult to navigate would be a usability issue; ensuring that combat in a game has the correct pace would be a playability problem. Playability thus regards how the game itself operates; its rules and its level of challenge. Some playability problems, such as unfairly advantaged 'cheating-AIs' (Shelley, 15 August 2001) are clearly distinct from usability problems, yet others are not – when players feel they are not in control of their character, is that a playability issues relating to poor player empowerment or a usability issue relating to poor controls? Does a poor in-game camera impair playability or usability?

In short, whilst some playability concepts are clearly distinct from usability problems, many are not. Nevertheless, there is good reason to treat them as separate constructs in at least some respects; playability problems are more fundamental to the game design than usability problems, and these need to be prioritised, tested and caught sooner (Korhonen and Koivisto, 2006). For the evaluator then, playability is best treated as a domain-specific class of critical usability qualities. The next task is to examine what experiential, usability and playability factors relate to video games.

2.1 Factors Involved in the Player Experience

2.1.1 Player Experience and Engagement

Studies of player experience have focused upon a number of different constructs in an attempt to determine what makes video games so engaging. *Engagement* is a term used to characterise a state of involvement with a piece of software; a video game with an enjoyable player experience is thus said to be engaging. Whilst the term has been given various meanings in the literature (i.e. Lindley *et al* 2008; Douglas and Hardagon, 2000), Lazzaro (2004) models player engagement as resting upon the four 'keys' of Hard Fun (or challenge), Easy Fun (involving immersion, curiosity and delight), Altered States (emotion, relaxation) and The People Factor (social interaction). This model recognises that games do not need to be challenging to be engaging – the game world itself can engage players sufficiently. Nevertheless, whilst categorizing the aspects of engagement in such a way is useful, it doesn't examine the constituents of these factors in enough depth. Why should we include narrative? Can we have easy fun without narrative? What emotions are involved in 'Altered States'? Indeed, this is the problem with the entire notion of engagement; it simply restates player experience without unpacking it enough. Nevertheless, if we hold Lazzaro's concepts of Hard and Easy Fun to still be useful (as these are the most universal to all games) then the first element to examine in greater depth is challenge and the construct of flow.

2.1.2 Flow, Cognitive Absorption and Challenge

Mihalyi Csikszentmihalyi (1975; 1990) found that the experience of performing with a high level of skill at challenging tasks had a peculiar character that he called *flow*. Individuals who engage in tasks in order to experience flow are engaging in an *autotelic* (auto = self, telos = purpose) activity – their internal motivation replaces any external motivation. Indeed, people were found to be at their happiest when engaging in such an internally motivated task. Eight main elements of such flow experiences were identified by Csikszentmihalyi (1990):

1. A challenging but completable task
2. Attention is focused wholly upon the task
3. The task has clear, unambiguous goals
4. The task provides immediate feedback for actions.
5. The individual feels fully in control
6. Immersion in the task that removes awareness of everyday life
7. Sense of self diminishes, but is reinforced afterwards
8. Awareness of the passage of time is reduced.

However, these factors must occur during a task that balances the individual's skills with the challenges that they face; too little challenge vs. skill and the user can become bored; too much and they become anxious and lose their sense of control. In the narrow band between boredom and anxiety lies the flow channel; activities that elicit experiences in this band are so rewarding that individuals will go to great lengths to engage in them for the sake of the experience (see Figure 2.1 below).

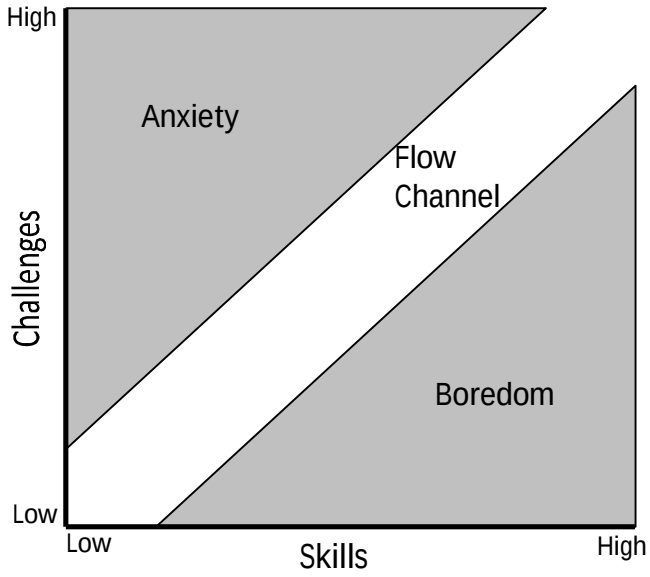

Figure 2.1. Graph showing the relationship between Challenge and Skills – the correct balance results in Flow experience. Redrawn from Csikszentmihalyi (1990)

This was later built upon by Massimini and Carli (1988) who noted that the individual's mean experience was neither optimal nor negative; rather it was neutral. Their experience fluctuation model (see Figure 2.2 below) better accounts the variety of human experience. For most activities with average challenges for which we possess average skills we do not experience flow; it is only when our skills and the corresponding challenge are high that flow is experienced. Indeed, the key to understanding flow is to recognise that it is an optimal experience, and certainly not a mundane one. To continue experiencing flow becomes a central goal for the individual, whatever the source of the optimal experience – including video games

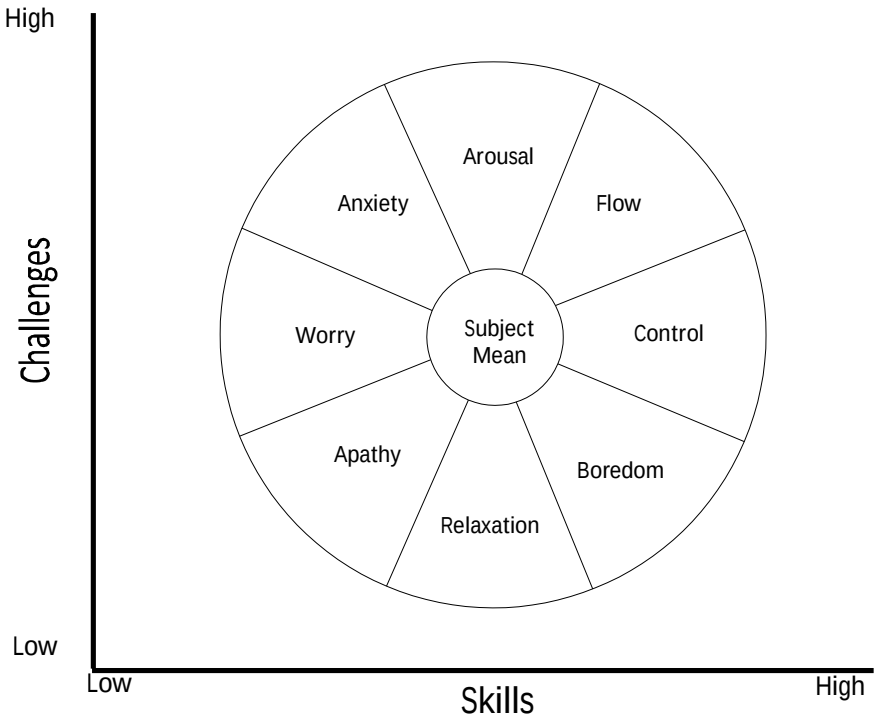

Figure 2.2. Graph showing how differing balances of challenge and skills result in different affective experiences. Adapted from Massimini and Carli (1988)

Chen (2007) held flow to be the *sine qua non* of an enjoyable game experience. Chen suggests that games must adapt to the different *flow zones* (i.e. difficulty level-tolerances) to ensure that as many users experience flow as possible. Another model that suggests ways to maximise player flow is the *GameFlow* model of Sweester and Wyeth (2005). This takes each element of flow (as listed above) and takes a game feature that must be present and/or optimised for flow to occur. These are concentration, challenge, player skills control, clear goals, feedback, immersion and social interaction. However, the faults of this, and similar models, are two-fold. First, flow is an optimal experience that gamers will only experience on occasions and perhaps only fleetingly (as noted by Jennett *et al*, 2008) whilst flow-like experiences are by no means a necessary component of an enjoyable gaming experience (Cowley

et al, 2008). Second, many of the descriptors used to signify flow (such as 'temporal dissociation', 'concentration' and 'control') can be explained by other constructs.

One such construct is Agarwal and Karahanna's (2000) concept of *cognitive absorption*, that describes a state of deep involvement with a piece of software through the factors of temporal dissociation, focused immersion or total engagement, heightened enjoyment, control and curiosity. The precedents of this deep involvement were stated as the perceived usefulness and perceived ease of use of the system (the 'usefulness' of games could be considered to be their enjoyability) – clearly less specific requirements than for an optimal flow experience. The key point is that cognitive absorption invokes many similar experiences to flow (though assumedly at a lesser intensity) without it being an optimal experience. Whilst gamers may occasionally experience flow, their mundane experience of involvement with a game is perhaps best explained by cognitive absorption. It is thus likely that when many investigators thought that they were examining flow they were in fact measuring milder, less optimal forms of experience.

However, we shouldn't as a result ignore flow; flow is still a measure of a successful game, but just as diamonds can measure one's wealth it is an exceptional measure and not the norm. When we come to measure cognitive absorption, very high reported levels of absorption (and one other factor – as below) should be taken to represent flow.

Challenge is the other factor that, when appropriate levels of it are reported, is likely to indicate flow. All games should provide an adequate challenge yet not be too difficult; this is both a basic factor underlying flow and a common heuristic suggested for game design (i.e. Federoff, 2002; Desurvire, Caplan and Toth 2004). By separating our measure of flow into measures of challenge and of cognitive absorption

it allows us to measure both suboptimal and optimal experience. Challenge and (the factors that underlie) cognitive absorption will describe flow if optimal; otherwise they will describe qualities of the average gaming experience.

2.1.3 Presence, Immersion and Fun

Having considered Lazzaro's (2004) 'Hard Fun' element of gaming fun, we must now consider its little brother, "Easy Fun". The aspect of altered states will be considered (under the banner of 'fun'), though 'the person factor' will not be (as though it is important, it is limited to multiplayer games).

The phenomenology of interacting with a game, especially if it involves avatars, is very idiosyncratic, and a number of constructs have been used to explain it. One such construct is the notion of presence. This concept arose from Virtual Reality (VR) research, where a peculiar feeling as though one is in the Virtual Environment (VE) was noticed by researchers. Floridi (2006) defines presence as:

> "*a* type of experience *of "being there", one loosely involving some technological mediation and often depending on virtual environments*" (Floridi, 2006)

Both Pinchbeck (2005) and Takatalo (2006) find that presence is a relevant concept for video game experience. This seems unlikely – not only was spatial presence (the *essence* of presence) the weakest extracted factor for Takatalo (2005) (behind such factors as role engagement and attention) but gamers do not speak of spatial presence in regards to their experiences (Jennett *et al*, 2009). Rather, their

'being there' is in a narrative, causal and social sense; it is categorically not a spatial presence.

If we discount the notion of presence, how then do we account for the sense of being "in" a game world? Brown and Cairns (2004) used grounded theory to examine what gamers meant when they spoke of being immersed in a game. They found three levels of immersion:

1.) *Engagement* – the player must invest time effort and attention to overcome barriers to the game – such as learning the controls or comprehending the setting.

2.) *Engrossment* – the game dominates player attention and players become emotionally invested, provided that the game mechanics and plot are well constructed.

3.) *Total Immersion* – players experience presence, empathy with characters and are totally absorbed with the game.

The first thing to note is that, as per Jennett *et al* (2009) it is unlikely that players experience presence when totally immersed, whilst immersion should be considered as a spectrum and not as a set of discrete stages. Secondly, this model of immersion is somewhat simplistic, and treats immersion as a single phenomenon that encompasses both challenge and diegesis. It is thus just another label for 'gameplay experience' if treated in this way. Considered in this way it still tells us something interesting; both that the gameplay experience is a spectrum and that barriers must be overcome to progress through this spectrum – including player experience, usability and playability problems.

Brown and Cairns (2004) failed to fully consider the narrative aspects of immersion, yet many have suggested that immersion is a narrative phenomenon, and that immersion involves deep engagement with a plot or setting. (McMahan, 2003; Douglas and Hargadon 2000). Ermi and Mäyrä (2005) account for this understanding of immersion in their SCI model of immersion, which identified three types of immersion:

Sensory immersion – the player becomes immersed in the sensory information – visual, auditory and tactile - that a game provides. Sega's *Rez* (http://www.thatgamecalledrez.com/ - see Figure 2.3 below) is probably the purest example of this.

Challenge immersion – immersion resulting from a balance of challenges and skills, requiring motor skills and/or strategy.

Imaginative immersion – immersion in the fantasy of the game, the plot, the game world and identification with the characters.

Figure 2.3. Sega's *Rez*, a classic example of how visual, auditory and haptic (a "trance vibrator" peripheral was released) interactions combine to induce sensory immersion in players. From www.ign.com.

A game can support all three of these types of immersion, whilst the elements of the SCI model are fully compatible with Brown and Cairn's (2004) levels of immersion giving us a more refined model of immersion. Two amendments are suggested. First, though challenge is an important element of any model of gameplay experience, calling it 'immersion' recalls the habit of redescribing gameplay experience in terms of one construct – be it immersion, flow or presence. It would be better to only consider sensory and narrative immersion in our questionnaire and just leave challenge as it is – as challenge. In addition, as per Arsenault (2005), imaginative immersion is best named fictional immersion to better capture its character.

Calvillo Gamez, Cairns and Cox (in preparation) created a grounded theory from score of press games reviews and articles that centred on the notion of puppetry. This involves factors of control (how to manipulate the game), ownership (the player comes to set personal goals and is provided with rewards) and facilitators (such aesthetics, that allow for control and ownership). Ownership is the most interesting

factor here, as it provides an explanation of *persistence* and *replayability* - why do players come back to a game? The model suggests that their own goals and sense of reward that keeps players going; they own the game in choosing which challenges to take. Some of these may be easy (and the player can thus feel rewards from showing their mastery) and some difficult. The idea of ownership must also be measured.

The final piece of the puzzle requires us to consider *fun*. If we consider fun in terms of dimensional affect, it involves high arousal and high positive valence. Not all games induce a sense of fun directly (e.g. survival horror games aim to scare; a well written roleplaying game may induce grief) but the net affective experience of playing a game should involve fun (Zagalo *et al*, 2005). In his classic research on fun, Malone (1981) found the key factors of fun are challenge, fantasy (akin to narrative) and curiosity. This provides us with two more factors – affect valance (which should be positive overall) and variety, as a lack of variety greatly inhibits curiosity and thus fun.

2.2 Usability and Playability in Games

As the earlier discussion of player experience noted, reaching deeper and more enjoyable levels of experience requires overcoming barriers; if these barriers are too great, the experience will be diminished. Some of these *barriers to play* relate to the player experience (such as the difficulty level). Others relate to the usability and playability aspects of the game (such as poor controls). This section will determine how questionnaires are an important tool for removing usability and playability barriers as well as player experience issues, and that there is a need for a new questionnaire with which to do this.

Gilleade and Dix (2004) distinguished between *at-game* frustration (resulting from poor controls and interfaces) and *in-game* frustration (resulting from unclear goals, navigation and similar). At-game frustration is always detrimental to the gameplay experience; in-game frustration (as IJsselsteijn *et al* (2007) note) is not always harmful. In-game frustration does not necessarily come from detrimental factors such as unclear goals, but also arises from the challenge of the game. If we remove all in-game frustration, the player has nothing to overcome and thus cannot experience *fiero*, or the experience of personal triumph over adversity (Lazzaro, 2004) – a critical emotion for gamers.

Standard usability evaluation techniques seek to remove all sources of frustration; we therefore need to tailor our evaluation methods to the video game domain. We should triangulate on usability (and playability) problems using a number of methods (Gray and Salzman, 1998). Kim *et al* (2008) describe the TRUE (Tracking Real-time User Experience) methodology. This involves recording user-initiated events; sets of data that describe what the user was doing when an event was initiated. So if a player crashes in a racing game, their speed, the track, the conditions, their location, etc are recorded. This is combined with observational (via video) and attitudinal (via questionnaire and interview) data to determine what a player was doing throughout a level or track that would lead them to enjoy or dislike it.

Questionnaires are thus an important part of this process, but they should not be understood as uncovering problems; interviews and observations are more effective for this. First, questionnaires can suggest to the evaluator where they must look in a huge dataset to uncover problems. If players found the difficulty too hard, this would suggest that they keep dying or losing a race, and the problem could be

uncovered by examining the relevant part of the dataset. Second, such questionnaires can act as a rubberstamp and quantify the severity of problems found in the other data sets

Which questionnaire to use to do this? An extensive sweep of the literature suggests that no validated usability and/or playability scale exists, compared to the numerous scales available for productivity software usability - e.g. Chin *et al* (1988). However, there are a number of studies that generated heuristics for evaluating games. Such heuristics could generate areas of interest or constructs that should be examined by any future questionnaire.

A number of existing sets of usability heuristics were examined (namely, Febretti and Garzotto, 2009; Desurvire and Wiberg, 2009; Pinelle, Wong and Stach, 2008; Federoff, 2002 and Korhonen and Koivisto, 2006). Desurvire and Wiberg (2008) was excluded from this analysis as their heuristics focused on game approachability (and thus focused on casual gamers – for whom a different set of factors are appropriate and are not the focus of this analysis) whilst Korhonen and Koivisto (2007) was excluded due to the focus on multiplayer games. The analysis is summarised in Table 2.1 below

Table 2.1. *Table Displaying Gameplay Heuristics Found Across the Literature*

Heuristic	*Previous Studies that Included It*	*Include in Current Study?*	*Why exclude from Study?*
Control(s)	All	Yes	N/A
Goals	All	Yes	N/A
Interface	All	Yes	N/A
Consistency	All but Febretti *et al* (2009)	Yes	N/A
Help	All but Febretti *et al* (2009)	Yes	N/A
Customisation	All but Korhonen *et al* (2006)	Yes	N/A
Variety	All but Pinelle *et al* (2008), Febretti *et al* (2009)	Yes	N/A
Navigation	Federoff *et al* (2002), Desurvire *et al* (2008)	Yes	N/A
Views	Pinelle *et al*, Febretti *et al* (2009)	Yes	N/A
Challenge	All but Pinelle *et* a*l* (2008)	No	Experience Factor
Immersion	All but Pinelle *et al* (2008)	No	Experience Factor
Feedback	All but Pinelle *et al* (2008*)*, Febretti *et al* (2009)	No	Covered by other Heuristic
Error Recovery	All but Pinelle *et al* (2008), Febretti *et al* (2009)	No	Covered by other Heuristic
Rewards	All but Pinelle *et al* (*2008*), Febretti *et al* (2009)	No	Experience Factor
Terminology	All but Pinelle *et al* (2008), Febretti *et al* (2009)	No	Covered by other Heuristic
AI	All but Korhonen *et al* (2006)	No	Genre-specific

In Table 2.1 above, 'Control(s)' refers to the quality of the game's controls and the player's feeling of control; 'Goals' to the need for clear player objectives and 'Customisation' to the need for customisable controls and settings. 'Consistency' means the consistency of input to output mappings; 'Views' to the quality of the in-game perspective; 'Interface' to the game's menus and (in-game) Heads-Up Display (HUD) and 'Help' to the need to provide help to the player. Finally, 'Navigation' entails that the player should not get lost in the game world (i.e. it has a slightly

different meaning to the concept of navigation in productivity software) and 'Variety' entails that the player should enjoy a range of gameplay elements.

The 'Challenge', 'Immersion' and 'Rewards' heuristics are already considered by experience and challenge items on the scale (reward being an element of challenge). 'Feedback', 'Error Recovery' and 'Terminology' are covered by other heuristics (i.e. 'Goals' and 'Consistency' cover feedback; 'Interface' largely exhausts terminology). Not all games have 'Artificial Intelligence' (AI) as not all have computer-controlled opponents (i.e. multiplayer games) so this was therefore excluded.

Overall, the heuristics have provided a foundation upon which a scale can be constructed. As was argued, both usability and playability should be evaluated as both are needed to improve a game. Given the distinction between usability and playability that was defined earlier, the above heuristics can be divided into playability and usability factors (e.g. 'Goals' involves a playability issues; 'Interface' is a usability issue, etc). Additionally, as no usability and playability scale exists creating a new one would clearly facilitate player testing. Both player experience and usability/playability factors to be included in any games evaluation have now been considered; the next task is to discuss the construction of the questionnaire.

2.3 Measuring the Video Game Experience

2.3.1 Evaluating Selected Constructs

Given the above review, the following constructs in Table 2.2 (below) were identified as needing to be evaluated for video games.

Table 2.2. *Four Main Factors of Video Game Experience and Sub-Constructs*

Factors Mediating Video Game Experience			
Experience	*Challenge*	*Playability*	*Usability*
Fictional Immersion	Challenge	Variety	Control
Sensory Immersion	Absorption	Clear Goals	Customisability
Affective Valence	Ownership	Navigation	Consistency
		Help/Training	Camera (Views)
			Game Interface

Table 2.2 shows that Experience, Challenge, Playability and Usability all need to be considered. By using a questionnaire involving closed questions we can quantify the degree to which users had a problem with, or enjoyed, an element of the game, just as we can quantify the nature of their overall experience. If well designed, such a scale would correlate with (and thus help us to predict) important measures of a game's success: review scores, sales, the game's appeal etc.

In the domain of productivity software, a questionnaire designed to do just that exists. Hassenzahl *et al's* (2000) *Attrakdiff* questionnaire. After examining one of seven prototypes that varied in terms of their ergonomic quality (i.e. usability) and hedonic quality (i.e. user experience), Hassenzahl *et al's* (ibid.) participants filled in scales that measured these two factors and the product's appeal. Both of these factors were found to correlate with the software's appeal.

There is little reason to suppose that this isn't the case for video games; the prior review has shown that both hedonic and ergonomic factors likely contribute to a game's quality and appeal. However, no such questionnaire currently exists for video games. If we are to aid player testing by creating such a questionnaire, we must first decide what type of questionnaire to design.

A commonly used type of questionnaire is the Likert scale (Likert, 1932, cited in Carifio and Perla, 2007). The Likert scale has a very particular process: a large number (80-100) of statements is generated that relate to a particular concept. Beneath

these is a response item where respondents mark how much they agree with the statement, usually rated 1-5 or 1-7 (see Figure 2.4). These are then rated by a number of judges in terms of how well they relate to the given concept. Inter-correlations between these items are then calculated, and the best 10-15 in terms of rating and inter-correlation are kept as the scale. By then summating the scores from each item to give a total scale score, we have a scale that measures the respondent's attitude towards the concept.

There are two important things to note here: first, that this method is rarely followed. Instead, the scale is usually administered to a large number of respondents (100+) which allows factor analysis to be performed (Oppenheim, 1992). This allows us to determine what the sub-scales that contribute to the scale are and ensure the reliability and validity of the scale far better than with Likert's original technique. Reliability and validity are the key metrics of scale success, with reliability referring to the degree to which the measurement if free of errors and validity referring to the usefulness and meaningfulness of a measure (Jensen, 2003).

Second, the scale is not (as is all too commonly believed) the response item beneath a statement, such as in Figure 2.4 below.

1	*2*	*3*	*4*	*5*
Disagree strongly	Disagree slightly	Neutral	Agree slightly	Agree strongly

Figure 2.4. Example Likert-type response item.

Figure 2.4 may have scalar properties but is not a scale. As Carifio (2008) contends, one should never call or treat single response items like a summated scale; only a summated scale can be considered as measuring at attitude. The whole

advantage of scaling is that the summated score increases reliability and validity when examining attitudes; testing single items massively increases the familywise error rate.

Using Likert scaling would allow us to quickly generate, pilot and validate a new questionnaire that can examine many constructs underlying the experiential and usability properties of a video game. It would do so in terms of participant's attitudes to the game that they just played. How to design such a scale is therefore explored next.

2.3.2 Scale Design and Validation

As per Hornbaek's (2006) call to improve the practice of usability measurement, the scale's content should be based upon research into scale design. Survey methodologies have progressed sufficiently over the past century or so for Schaeffer and Presser (2003) to confidently declare that there is no longer an art but rather a science of asking questions. Whilst the strength of this assertion is perhaps debateable, it is certainly true that a good deal of research has refined survey and scale design methods. The following section outlines what could be considered 'best practice' in scale design, providing a number of criteria that any new scale must meet.

Attitude judgements measured by scales reflect the information that was available at the time – this means that the context at the time the question is asked causes bias, and the most important such context is the scale construction (Tourangeau, 1999). The highest level sources of such error are order effects; these occur when responses to later questions are influenced by the content of earlier questions.

Order Effects

Effects regarding how questions appear to fit into a higher level category to a respondent are known as assimilation effects (Tourangeau, 1999). For simple Likert scales this is unavoidable – these effects act as a demand characteristic, and if respondents grasp the overall purpose of a scale it can bias their responses. However, if the scale is comprised of subscales the best solution is to separate out questions from each subscale. Asking questions in close proximity increases the likelihood of respondents altering their attitudes accordingly to increase consistency among responses (McGuire, 1960). This may increase correlations between items, but this greater correlation is illusory and a source of error. Mixing the order of subscales throughout a scale can reduce this bias, if not eliminate it.

Another major order effect is the *part-whole* effect (Krosnick, 1999; Lietz, 2008; Martin, 2006), whereby more general questions asked after specific questions can be misinterpreted – by respondents excluding the content of the specific question from the general one, for example. The cure for this is simple: ensure that more general questions are always asked before specific ones.

Question Wording

Moving now to the content of the questions themselves, all sources advise to keep questions as short as possible (Lietz, 2008; Foddy, 1993; Dillman, Tortora and Bowker, 1998) with the rule of thumb being a limit of around 20 words per question (Oppenheim, 1992). Overall scale length should also be minimised, especially when using web surveys (Ganassali, 2008).The wording of the question is advised to be kept as simple and unambiguous as possible, avoiding leading questions, ambiguity, double-barrelled questions (containing multiple clauses) or double negatives (Lietz,

2008; Martin, 2006; Foddy, 1993; Krosnick, 1999; Alwin and Krosnick; 1991). However, a fundamental element of Likert scaling involves including positive and negative statements about a potential viewpoint (Likert, 1932, cited in Carifio and Perla, 2007) - which can lead to double negatives. Whilst one should aim to use wording that will not lead to this – ugly' as opposed to 'not attractive' – this may be unavoidable. Nevertheless, negative statements also have the added benefit of helping to reduce the phenomenon of acquiescence – whereby many participants will simply agree with any statement provided (Hinz et *al*, 2007). Since they must express their level of agreement for both positive and negative positions, introducing negative statements should reduce the strength of this effect (Cox and Cairns, 2008), whilst there is evidence that both positively and negatively worded items do test the same construct (Bergstrom and Lunz, 1998), allaying any fears that they may not.

Response Item Design

In terms of the Likert-type scalar response item accompanying each question on the scale, there are a number of suggestions. Whilst some have suggested that response items with only three response options are adequate (Jacoby and Matell, 1971) the general consensus is that larger items of 5-7 options are required (Lietz 2008; Krosnick 1999; Preston and Colman 2000; Cox 1980; Lehmann and Hulbert 1972; Colman, Norris and Preston 1997) to ensure an adequate level of reliability and validity whilst reducing cognitive load on participants. Some suggest that larger items (of 11 items plus) are desirable (e.g. Dawes, 2001), yet other research has found indices for reliability and validity improve up to 7 response options (Masters, 1974) and decrease after 10 items (Preston and Colman, 2000). Indeed, Cox (1980) called it the 'lucky number 7, plus or minus two', in reference to Miller's (1956) dictum on

number span (though not suggesting that the two are linked). Seven response options are thus recommended.

All of the quantities suggested above are odd, which entails that a midpoint for the scale is endorsed. Inclusion or exclusion of this will change the data gathered (Garland, 1991), and whilst some have found that midpoints do not affect scale reliability (Alwin and Krosnick, 1991), the general consensus is that they do improve this measure (Lietz 2008; Oppenheim 1992). Good question clarity, meanwhile, reduces respondents adopting a satisficing strategy (reducing cognitive load by selecting the first acceptable response) and selecting the midpoint without further thought (Velez and Ashworth, 2007). The wording for the response option labels should be a balanced (i.e. 'like vs. dislike', not 'like vs. hate' etc) and the response item should be unipolar (i.e. run from '1-7' not '-3 to 3') (Lietz, 2008).

In terms of response item layout, the best layout is to run left to right, ascending numerically and from negative to positive responses (as in Figure 2.5 below). The opposite has been found to distort results (Hartley and Betts, 2009) and running left to right better matches the reading direction of Latin text.

1	*2*	*3*	*4*	*5*
Disagree strongly	Disagree slightly	Neutral	Agree slightly	Agree strongly

Figure 2.5. Response item runs from left to right, ascending numerically from negative to positive responses.

'Don't know' options are to be avoided, as many participants that fill them in do have genuine attitudes to provide, and only choose "don't know" options due to a satisficing strategy (Gilljam and Granberg, 1993). Labelling every response option

improves the quality of the data acquired (Krosnick, 1999), whilst ‘strongly’ and ‘slightly’ are good extreme point and mild qualifiers respectively (Lietz, 2008).

Web Surveys

As for web surveys (important, as the scale being designed may begin life as one) there is little evidence that paging or scrolling between web survey questions has an impact on response rate (Peytchev *et al*, 2006), although we may wish to follow good usability practice here and avoid scrolling. Finally, any demographics questions should come at the end of a survey, once respondents are committed to completing the scale (Lietz 2008; Oppenheim 1992).

2.3.3 Previous Gameplay Scales

Having reviewed good scale design practice, we must now consider the designs of existing scales. In terms of game experience, a number of scales are in use. The earliest was probably Witmer and Singer’s (1998) Presence Scale, originally designed for VR research and the experience of being in a VE. Whilst this is commonly used in VR studies and has been used in gaming studies (e.g. Eastin and Griffiths, 2006) it has several flaws. First, some of the questions are overly complex, making use of the dreaded conjunction ‘or’. For instance, “How much did the control devices interfere with the performance of assigned tasks or with other activities?” and “How responsive was the environment to actions that you initiated (or performed)?” have a great deal of complexity added by the conjunction. Second, some have argued (such as Slater, 2004) that presence should not be measured solely using scales; an

abstract concept such as presence needs more sources of evidence before we can label an inter-correlation between subscales 'presence'.

In his thesis, Kim (2006) makes use of a refined version of Chen's (2004) Game Engagement Scale (GEQ), the GEQ-R. This scale is rooted in the Presence Scale, but modifies it for video games. This scale fails to break the concept of engagement down; whilst there are questions measuring aspects such as control, graphics etc, no subscales are formed. Since we should avoid making judgements based upon single response items, this scale allows us to quantify how engaging a game was – and little else, something that is not useful when attempting to measure the qualities of a game in more detail.

Ermi and Mäyrä (2005) created a novel scale to measure their SCI model of immersion. Leaving aside "challenge immersion" (which, as aforementioned, is not being viewed as a form of immersion), one of the sensory immersion items is somewhat limited. "The game looked credible and real" would not apply to many non-realistic games that nevertheless have an immersive sensory experience (Sega's *Rez* again springs to mind). As for narrative immersion, some of the items seem poor due to the translation into English from the original Finnish (i.e. "I handled also my own emotions through the game") which is to be forgiven; the constant reference to 'characters' is less forgivable, as it limits the scope of the scale to those games that involve avatars (e.g. not Real Time Strategy (RTS) where there is no representation of the player, who instead controls whole armies).

Calvillo-Gamez (2009) developed the Core Elements of the Gaming Experience Scale (CEGEQ), whilst Jennett *et al* (2008) designed the Immersion Scale. Whilst these are not perfect, they are a considerable improvement on what came before. The former measures enjoyment and frustration in addition to factors relating

to Calvillo-Gamez's puppetry model – control, ownership, facilitators; the shift in focus from presence and flow is welcome, whilst the balance of positive and negative statements should reduce acquiescence. Nevertheless, there is still no focus on usability or playability factors as well as a lack of focus on narrative elements.

Finally, IJsselsteijn *et al's* (in preparation) Game Experience Scale (GEQ) has been used in a number of studies (e.g. Nacke and Lindley 2008; Lindley and Nacke, 2008) with some success. This scale divides flow and challenge (the former is probably better considered as some form of absorption) and includes subscales that don't necessarily aid evaluation (such as 'competence' – do high reported levels just suggest that a game is *too* easy?) or are genre-specific ('tension' is irrelevant for many games – i.e. life simulation games such as *Animal Crossing*). Despite this, the GEQ is well validated and has excellent question wording (using short, simple statements) and a useful, reduced form for administering between levels/missions etc.

In the productivity software literature, a number of scales measure system usability. The *Attrakdiff* (Hassenzahl, 2000) has already been discussed, but the most widespread is the System Usability Scale (SUS) developed by Brooke (1996). Though extremely short (10 items!) it has been found to be a remarkably robust measure of system usability (Bangor, Khortum and Miller, 2008), effective on a multitude of systems. However, it measures usability on a single scale; we might wish for a more fine-grained analysis. Lewis (1995) designed the Computer System Usability Scale (CSUQ) which does comprise of subscales – system usefulness, information quality and interface quality. Tullis and Stetson's (2004) review of usability scales found these two to be the most robust, with a sample of 12 users providing the correct findings (i.e. the same score as a larger sample) 90-100% of the time, and a sample of

10 users 75-80% of the time. In all, this suggests that such scales can be short, measure multiple factors yet still be reasonably robust when testing small samples.

Clearly, selection of the correct constructs is paramount, and the preceding review has shown what aspects should be measured. None of the existing battery of scales provides a truly integrated measure of a game's appeal or quality, although they should influence the content of any new scale. If a scale is to accurately assess a game, it must measure all of the pertinent factors. To address this need, the task now is to create a new validated scale, before we can determine if such a scale can reliably measure experience, challenge, playability and usability factors and predict video game review scores.

3 Study #1: Scale Construction, Validation and Refinement

3.1 Rationale

The literature review has shown that experience, challenge, playability and usability (broken down into the elements in Table 3.1 below) all need to be measured if we are to assess a video game. A scale that measures these factors now needs to be created, following the good design practice noted in section 2.3.2. As was discussed in section 2.3.1, the first step is to create a large pool of questions, test them on a large population and determine what factors emerge.

The questions that this first study seeks to answer are whether a scale developed along these principles can measure the appeal and quality of a video game reliably and accurately, and whether or not the scale has good validity. To determine this, participants will also complete a modified version of Hassenzahl *et al's* (2000) Appeal Scale, it being hypothesised that if the two scales correlate then the 'Gameplay Scale' is indeed measuring factors that influence a game's appeal – and therefore has good validity.

3.2 Questionnaire Construction

3.2.1 The Gameplay Scale

The main scale being devised, named the Gameplay Scale, aimed to measure 15 elements of video game experience divided into 4 factors as in Table 3.1 below.

Table 3.1. *Four Main Factors of Video Game Experience and Sub-Constructs*

Factors Mediating Video Game Experience			
Experience	*Challenge*	*Playability*	*Usability*
Fictional Immersion	Challenge	Variety	Control
Sensory Immersion	Absorption	Clear Goals	Customisability
Affective Valence	Ownership	Navigation	Consistency
		Help/Training	Camera (Views)
			Game Interface

(This is the same table as Table 2.2, redrawn here for clarity)

As Table 3.1 illustrates, on the basis of the literature review the factors of experience, challenge playability and usability are expected to emerge; there should thus be four subscales to the main scale after analysis. The player experience question content was drawn from existing questionnaires (notably, those of IJsselsteijn *et al*, in preparation; Ermi and Mäyrä; 2005, Calvillo-Gamez; 2009 and Jennett *et al*; 2008), modifying these questions as per the good questionnaire design review in section 2.3.2 and the review of how to measure player experience in section 2.1. This included questions such as "I thought that the game was fun" and "I felt the game was hard". The usability and playability questions were drawn from the gameplay heuristics reviewed in section 2.2 (i.e. those of Febretti and Garzotto, 2009; Desurvire and Wiberg, 2009; Pinelle, Wong and Stach, 2008; Federoff, 2002 and Korhonen and Koivisto, 2006), and were divided into playability and usability factors as in Table 3.1. These included questions such as "I found the controls to be difficult" and "The game provided me with an adequate tutorial".

This initial version of the scale had 49 questions (**see** Appendix A), 3-4 per element - meaning that any of the elements could still be measured if subsequent analysis showed that it formed a factor independent of anything else. These questions were refined via pilot testing and each question had a 7-point Likert response item

with labels on every point. The aim was for the following analysis to remove items with poor inter-correlations or reliability and thus leave a smaller, more accurate scale.

Initial pilot testing involved a cognitive interview, as described by Fowler (2002). This involves participants "thinking-aloud" as they work though the scale and was performed (after a short gaming session) by 4 participants, resulting in substantial amendments to the scale items and ordering.

3.2.2 The Appeal Scale

The Appeal Scale was a modified version of Hassenzahl *et al's* (2000) appeal scale; using 8 semantic differential items with a 7-point scale (see Appendix D). Following the pilot testing (as explained above) the original "sympathetic-unsympathetic" pair, was deemed unsuitable for testing videogames and was replaced with a "fun-boring" differential. This was viewed as a separate scale to the Gameplay Scale, as it used a different form of response item and was being used to assess the construct validity of the Gameplay Scale.

3.3 Methods

3.3.1 Participants

In all, there were 132 respondents to an online version of the scale that respondents accessed via a link to the survey posted (along with a substantive description of the survey) on both a private PlayStation Beta Testers forum and on the

public official (EU) English PlayStation Forums (http://community.eu.playstation.com/playstationeu/?category.id=55). This self-selecting sample was reduced to 98 respondents (*M*: 25 years old, *SD*: 6.96; 7% female) once partial responses had been filtered out. The gender bias is noted as a limitation for generalising the findings.

3.3.2 Materials

The study used an online version of the questionnaire that included both the Gameplay Scale and the Appeal Scale that was created using the SurveyMonkey survey software (http://www.surveymonkey.com).

3.3.3 Procedure

Participants were asked (via the forum posting) to play (the free downloadable demo of) Q Games's game *PixelJunk Eden* (http://pixeljunk.jp/), which was viewed as a reasonably simple yet aesthetic platformer/puzzle game. In *PixelJunk Eden*, players play as 'The Grimp', manoeuvring around a two-dimensional level (by swinging around or jumping) attempting to collect pollen that causes plants on the level to grow and allow other objectives to be met. This game was selected as it should raise interesting usability and playability issues (due to its novel premise and mechanics) whilst the demo is freely available to anyone on the PlayStation Network (PSN). Participants were requested to play the game for up to 2 hours (with a 15 minute break in the middle) and then complete both of the scales. Participants were informed that their participation was voluntary, that they were free to leave the study at any time

and that all data gathered was confidential. The study followed BPS ethical guidelines; ethics committee approval was sought and gained for the study, though no major ethical issues were foreseen

3.4 Results and Analysis

3.4.1 The Gameplay Scale

The first stage was to determine how the variables (in this case, each scale item) in the gameplay scale clustered. A hierarchical cluster analysis was performed on the 98 responses, using Ward's method. (This method was chosen as it is widely regarded as provided the most accurate and robust clustering solutions (Scheibler and Schneider, 1985) Selecting a clustering solution is as theory-driven as it is data driven (Thorndike, 1978), so a number of clustering solutions were outputted and the best selected on theoretical grounds. In this case, the clustering solution that fitted most closely with the expected 4 part structure to the scale was selected. This 4 cluster solution closely mirrored the expected experience/challenge/ usability/playability structure, as can be seen in Appendix B. The first cluster generally comprised of Affect, Sensory Immersion and Fictional Immersion questions; the second mostly consisted of Absorption and Challenge questions; the third included Navigation, Goals and Consistency and the fourth included Help, Controls and Menus. A few anomalies (such as some Controls and Menu items being clustered with the Affect and Immersion items) did occur however.

The scale as a whole had good reliability in terms of internal consistency, with a high Cronbach's alpha of 0.933. However, we have little certainty that it functions

as a unified scale measuring ‘game quality’ or similar, so cluster reliability is more important. Additionally, we need some way to validate the clustering solution selected; if the clusters possess good reliability it suggests that we can treat them as subscales.

Of the 4 clusters, cluster 1 was named ‘Affective Experience’, cluster 2 ‘Focus’, cluster 3 ‘Playability Barriers’ and cluster 4 ‘Usability Barriers’ – the reasons for naming the subscales this way are considered in depth in the discussion section. Affective Experience had an alpha of 0.933; Focus had an alpha of 0.757; Playability Barriers’ alpha was 0.857 and Usability Barriers’ was 0.783. In short, all of the clusters had good reliability, with alphas over 0.7. We thus have good reason to consider them subscales. The next task was to reduce the scale in size, removing questions that did not add (or indeed, detracted from) each subscale’s reliability. This was done in a stepwise process, with the impact on subscale reliability checked each time an item was removed. This process continued until the impact of removing any more items would reduce subscale reliability too much, lessen cluster integrity or leave constructs unaccounted for. The results of this are shown in Table 3.2 below. As the table illustrates, all the 26 remaining items had a significant correlation to both scale and subscale total scores.

This revised scale had a Cronbach’s alpha of .903, whilst the scales still had good reliability: Affective Experience had an alpha of 0.903; Focus had an alpha of 0.711; Playability Barriers’ alpha was 0.814 and usability Barriers’ was 0.760. It seems reasonable to therefore consider these clusters as subscales measuring specific constructs.

Table 3.2. *Revised Gameplay Scale Items and Correlations to Scale and Subscale Scores*

Question #	*Item*	*Construct*	*Subscale*	*Correlation to Subscale Total*	*Correlation to Scale Total*
1	I enjoyed the game.	Affect	AE	.821(**)	.704(**)
5	I thought that the game was fun.	Affect	AE	.778(**)	.713(**)
21	I found the appearance of the game world to be interesting.	Sensory	AE	.661(**)	.521(**)
43	The aesthetics of the game were unimpressive. ***	Sensory	AE	.695(**)	.560(**)
45	The game failed to motivate me to keep playing. ***	Ownership	AE	.882(**)	.764(**)
47	I wanted to explore the game world.	Fictional	AE	.831(**)	.740(**)
3	I was focused on the game.	Absorption	F	.558(**)	.405(**)
4	I could identify with the characters.	Fictional	F	.505(**)	.411(**)
20	I was unaware of the passage of time whilst playing.	Absorption	F	.547(**)	.405(**)
23	I forgot about my surroundings whilst playing.	Absorption	F	.580(**)	.287(**)
38	I found the game mechanics to be varied enough.	Variety	F	.524(**)	.462(**)
41	I thought about things other than the game whilst playing. ***	Absorption	F	.656(**)	.499(**)
42	My field of view made it difficult to see what was happening in the game. ***	Camera	F	.497(**)	.483(**)
44	I thought the camera angles in the game were appropriate.	Camera	F	.565(**)	.498(**)
48	I thought the level of difficulty was right for me.	Challenge	F	.487(**)	.474(**)
15	I always knew where to go in the game.	Navigation	PB	.777(**)	.442(**)
27	I knew how the game would respond to my actions.	Consistency	PB	.669(**)	.592(**)
28	I always knew how to achieve my aim in the game.	Goals	PB	.781(**)	.567(**)
30	My objectives in the game were unclear. ***	Goals	PB	.759(**)	.662(**)
37	I couldn't find my way in the game world. ***	Navigation	PB	.703(**)	.543(**)
8	The game trained me in all of the controls.	Help	UB	.579(**)	.414(**)
12	I knew how to use the controller with the game.	Controls	UB	.641(**)	.371(**)
14	I found the game's menus to be usable.	Menu	UB	.708(**)	.473(**)
16	I knew how to change the settings in the game.	Settings	UB	.613(**)	.241(*)
24	I found using the options screen to be difficult. ***	Settings	UB	.717(**)	.347(**)
36	I found the game's menus to be cumbersome. ***	Menu	UB	.650(**)	.503(**)

* Corr. Statistically significant to 0.05; ** Corr. Statistically significant to 0.01

*** Negative question – scoring reversed.

(AE= Affective Experience; F =Focus; UB= Usability Barriers; PB= Playability Barriers)

3.4.2 The Appeal Scale

As for the Appeal Scale, 94 participants completed it, and all items were significantly correlated to one another whilst the Cronbach's alpha was high at 0.939, suggesting good scale reliability. All scale items had significant Spearman's Correlations ($p < 0.01$), whilst Factor Analysis (which could be performed on this far smaller scale) found all of the items to fall into one factor with an Eigenvalue of 5.629 that accounted for 70% of the total variance. This allows us to view the Appeal Scale as measuring one underlying construct – the game's appeal.

3.4.3 Inter-Scale Correlations

Finally, relationships between the subscales were considered. Spearman's Correlations between the Gameplay Scale, the four subscales and the Appeal Scale revealed that they were all highly correlated with both each other and with the total score for the scale (See Table 3.3 below). As Table 3.3 illustrates, all of these correlations were highly significant. This suggests that there may well be an overall construct of "game quality" to which each of the constructs measured by the relevant subscale contributes. It also implies that the four subscales measure appeal.

Table 3.3. *Spearman's Correlations between Subscales, the Gameplay Scale and the Appeal Scale for Study 1*

	Appeal	*Affective Experience*	*Focus*	*Playability Barriers*	*Usability Barriers*	*Gameplay Scale*
Appeal Scale	-	.757	.576	.531	.358	.746
Affective Experience	.757	-	.587	.513	.366	.839
Focus	.576	.587	-	.351	.321	.775
Playability Barriers	.531	.513	.351	-	.480	.737
Usability Barriers	.358	.366	.321	.480	-	.606
Gameplay Scale	.746	.839	.775	.737	.606	-

(All correlations statistically significant to $p < 0.01$)
(For Appeal Scale correlations, $N = 94$; for all others, $N = 98$)

Construct validity was investigated by multiple regression of each of the Gameplay Scale subscales against the Appeal Scale. This would allow us to determine which of the subscales best predicted the appeal rating - if the subscales correlated with the Appeal Scale we have good reason to infer that they are measuring elements of the game's appeal and thus that the scale had good construct validity. This found an R^2 of 0.731, suggesting that the constructs measured by the four subscales collectively account for 73% of the variance in a game's appeal. The overall effect was significant ($F(4,93)=7027.668$, $p = 0.000$); moreover, the Affective Experience subscale had the highest contribution, with a beta coefficient of 0.599, followed by Focus (0.209), Playability Barriers (0.164) and Usability Barriers (-0.003). Only the contribution of the Usability Barriers scale was non-significant ($p>0.05$). All in all, these results imply that each of the subscales measure constructs that are related to a game's appeal, but that in this case the Usability Barriers measure did not make a significant contribution. This either entails that Usability Barriers do

not contribute to a game's initial appeal at all or that they did not influence this game's appeal.

3.5 Interim Discussion

The aim of this first study was to construct and refine the questionnaire, ensuring that it and its subscales had good reliability and, by its correlating with the Appeal Scale, good construct validity.

Overall, this interim stage in the validation of the scale has refined the number of items in the scale, based on how well the items fit into clusters as revealed by cluster analysis. However, most scale construction involves using the statistical method of Exploratory Factor Analysis to determine what underlying variables are represented by each item in the scale and thus ensure good content validity. This is a large sample technique, and the 98 responses collected may not be enough for Exploratory Factor Analysis to be used successfully. It is generally maintained that a larger sample is needed - a minimum of 250 participants or more (e.g. Guilford, 1954) or a ratio of 5:1 or 10:1 participants to items or more (e.g. Everitt, 1975) being recommended. As such, Cluster Analysis is often recommended for scale development involving smaller samples (i.e. Thorndike, 1978). Whilst the technique lacked the statistical rigour of factor analysis it does assign all of an item's variance to a particular cluster – making it ideal for dividing questions into subscales. Indeed, this method is what Witmer and Singer (1998) used when developing the presence scale.

If we accept the cluster analysis, the four cluster solution broadly matched the expected structure (given the literature review) of Experience, Challenge, Usability

and Playability subscales with a few key differences, whilst the stated aim of using the analysis to select the most effective questions and reduce the size of the scale was achieved. The first cluster was very similar to the hypothesised 'Experience' factor, but with a slight shift towards the affective aspects of such experience, hence it now being called the 'Affective Experience' subscale. One anomaly was the inclusion of 'Controls' and 'Menu' construct items (Q25 and Q34 respectively in the original Gameplay Scale) in this cluster. However, both of these items included the word 'intuitively', thus it is likely that the use of this word biased responses by adding a more experiential element to the judgement being requested.

The 'Challenge' subscale that was initially hypothesised ended up being less about challenge and more about cognitive absorption (and included the Camera items), hence it now being called the 'Focus' subscale. It still covers challenge, but absorption is taken to be the key measurement here; the level of absorption or focus influences the player's ability to engage in a challenge and is itself a result of acceptable challenges. This scale involves absorption, challenge and variety, as well as two interesting results. Items for the 'Fictional Immersion' construct correlated to both the Affective Experience and Focus subscales (see Appendix A). One's experience of the game world (via Q47 in the original Gameplay Scale) was part of the Affective Experience cluster; one's ability to empathise with characters (Q4 in the original scale) was part of the Focus cluster. Previous studies have found empathy to correlate with absorption (Wickramasekera, 2007); it seems quite plausible that absorption is required for one to empathise with another, so this relationship is accepted.

Less understandable is the high correlation of the 'Camera' construct items with the Focus cluster. Whilst it is plausible that one's perspective of a game and the

ease at which this can be altered could mediate one's ability to focus on the characters and challenges, this is still a supposition, as we have no definite evidence of this relationship. As such, this is precisely the sort of cluster analysis result that we should be wary of. The subscale will be kept as it is for now, but this must be considered again after the second part of the scale validation. Indeed, removing these items now was found to negatively impact the subscale's reliability, which is why the Focus scale is (at 9 items) much larger than the other three.

The items in the other Playability Barriers and Usability Barriers were much as expected and offered few surprises. A plus point for these subscales (along with the other two) is that all of the expected constructs are still covered by at least one item, meaning that all of the various elements of game usability and playability can be measured. We may not expect these various elements to form a consistent scale (e.g. because consistency may be poor in a game but navigation and goals good) but as Sauro and Lewis (2009) found, there is a general construct of usability, which suggests that users form an overall impression of a system's usability which informs their judgements for the usability of each element of the system; the same is likely true for playability. Thus the players likely formed an overall impression of playability and usability which resulted in the correlation between their judgments about each element of the game.

It could be argued that this is problematic, as it lessens our trust in user's judgements. – How do we know that the navigation in the game is poor if the user's overall 'playability' judgement for the game influences their attitude towards this construct? However, if true, this would also be an issue for many usability evaluation methods (such as' think-aloud' user testing, interviews and focus groups) and simply underlies both the fact that individual Likert items should not be analysed (at least not

formally) and that triangulation using numerous methods is required to identify the source of poor usability and playability; methods such as user testing and interviews may be more appropriate here.

The Usability Barriers subscale is so called because it measures the severity of barriers to player engagement that are rooted in usability issues. This did not significantly contribute to the game's appeal this time, but this suggests that no major usability issues arose. Such barriers are likely to significantly detract from a game's appeal when they are severe, but fade into the background when usability and playability are good – and thus contributed little to a game's appeal (though a broader survey of games would be required to establish this)

Finally, the strong inter-correlations between subscales are suggestive that an overall measure of 'game quality' or similar can be measured by the scale; whilst it would be useful for the scale to function like this, it is too early to claim that such a measure exists. Either way, the strong relationship between the Gameplay Scale and the Appeal Scale suggests that the Gameplay Scale is indeed measuring something of interest, and that it will be worth continuing to validate the scale. It must be noted (Hassenzahl *et al* (2000) failed to) that the Appeal Scale is really a measure of *initial* appeal. How a user views any system (games included) will change over time (Grodal, 2000) – something especially true of a large game world that is to be explored. Thus in the ~2 hours of play that respondents had, they could only form their initial attitudes towards the game.

These attitudes are still important, however. In an age when most major games will have a free playable demo available before their release, ensuring that this initial appeal is high is very important for a game to sell well. Indeed, it is likely this initial appeal that motivates players to continue playing a game. However, further calibration

of the scale is required before we can state that a game scored well or scored badly on any of the qualities that it measures.

This study has thus established that the four elements of a video game's quality and appeal– Affective Experience, Focus, Usability and Playability – are relevant when assessing video games; that the Gameplay Scale designed is reliable and that it does measure a video game's initial appeal. However, further analysis is required to determine if these results generalise to different game genres and to determine if the scale can predict review scores. A further study was therefore performed to investigate this.

4 Study #2: Further Exploration of Scale Validity

4.1 Rationale

The previous study established the constituent elements of the Gameplay Scale, reduced it to 26 items and suggested that it has good construct validity. To further refine the gameplay scale and ensure that the revised version had good construct validity a further study was undertaken. This study aims to determine if the findings of the previous study can be generalised to other game genres, and to determine if the Gameplay Scale correlates with review scores, which would demonstrate the Gameplay Scale's usefulness to industry. To do this, a player testing study in which participants would play the game in a laboratory before completing the Gameplay and Appeal Scales was performed. This would determine if the Gameplay Scale would still correlate to initial appeal for open-world games.

The 'open-world' or 'sandbox' genre involves providing players with a large environment and allowing them to choose which tasks they perform in a highly non-linear way. The open-world genre was selected for two reasons – to ensure that the Gameplay Scale had good generalisability and could be used to determine the appeal of games in many genres (since these non-linear games are very different to the linear *PixelJunk Eden* used in study 1) and because there should be interesting playability and usability issues arising in this genre, especially regarding navigation, camera and controls as path-finding in open-worlds can be difficult. Controlling the game genre in this way lessens the ecological validity of the study and our ability to generalise these findings to all genres of game; however, it does increase internal validity by

controlling for each player's genre preference and allows for more reliable between-group comparisons

The two games selected, Radical Entertainment's *Prototype* (http://www.prototypegame.com/) and Sega's *The Incredible Hulk* (http://incrediblehulkthegame.com/) - see Figure 4.1 - were selected as they were both very similar in terms of genre (open-world action games involving a super-powered protagonist, with a linear first level) and setting (modern-day New York City) but had been reviewed very differently on the review compilation site Metacritic – with aggregate 'metascores' of 79 and 55 (out of 100) respectively - see http://www.metacritic.com. *Prototype* was thus fairly well reviewed (though by no means perfect) whilst *Hulk* was quite poorly reviewed (Whilst 55% would usually seem like an average score, video game review scores are usually shifted higher, with 70% or so being an average review score; see http://www.joystiq.com/2006/08/07/ign-gamespot-review-score-inflation-revealed/ for more).

Figure 4.1. Screenshots from the games *The Incredible Hulk* (left) and *Prototype* (right). Both images from www.ign.com.

Therefore, the further aim of this study was to determine if the Gameplay Scale would give significantly higher ratings to games given higher metascores than games given lower metascores. It is debatable whether review scores are a perfect metric to which the scale can be compared (especially for initial appeal), but given that the difference between the metascores (compiled from 51 (for *Prototype*) or 26 (for *Hulk*) magazine and website reviews) is significant ($p < 0.01$) we should nevertheless expect our scale to detect this apparent difference in game quality if it measures factors that influence a game's perceived quality.

The experimental hypotheses were that a well reviewed game would be rated significantly more highly on all or at least some of the Gameplay subscales and the (initial) Appeal Scale than the non-well reviewed game and that all (or at least some) of the Gameplay subscales would correlate with the (initial) Appeal Scale.

4.2 Methods

4.2.2 Participants

Seventeen subjects participated in the experiment (*M*: 23.6 years old, *SD*: 2.0; 4 females). The participants were recruited in an opportunistic sample around the university campus and paid £6. The participants were recruited using the following criteria to ensure that they were in the target demographic for the games being tested:

- They were 18-27 years old
- They had not played the test game
- They regularly played games (a median of 5-10 hours per week)

- They had played open-world games for at least 6hrs+; most had played the genre for 30hrs+
- They owned an average of 2 gaming platforms (including PC).

These participants were then randomly assigned to one of the two groups (n = 9 for the game *Prototype*; n = 8 for the game *Hulk*); of the four females, two were assigned to each game to counter-balance the genders.

Figure 4.2. The *DualShock 3* Controller. Like most modern controllers, this has vibrating tactile feedback or 'rumble'. From www.gizmodo.com

4.2.3 Materials

A PlayStation 3 console was connected to an LCD colour projector in a laboratory. One of two games was projected onto a wall in the lab. Players sat in front of the wall and played the game using a standard DualShock 3 controller (see Figure 4.2 above) Lights in the lab were dimmed to enhance player concentration on the game. A digital video camera (filming in night mode) was positioned to the front of

the player to capture their reactions to the game; whilst a researcher sat to the rear of the lab in order to take notes on their activities in the game world. Printed copies of the revised Gameplay Scale (see Appendix C) and Appeal Scales (see Appendix D) were used in this study.

4.2.1 Design

The experiment had a one-way between-subjects design – one group played *Hulk* and the other group played *Prototype*. The independent variable was the game that each group played (and the review score given to that game); the dependent variable being the difference in subjective ratings (on both the Gameplay and Appeal Scales) that participants gave to the games, which should give us a measure of the game's initial appeal.

4.2.4 Procedure

Participants played one of the games for one hour (including any non-playable cut scenes) in the laboratory whilst being videoed. Participants then completed the Gameplay and Appeal Scales by hand before being given a short, semi-structured interview. It is maintained that the 1 hour play session gives enough time to form attitudes about most of the game's aspects. The video and interview data were gathered to resolve any intractable issues that may arise from the scale data; as such, they were not the focus of the investigation. The study followed BPS ethical guidelines; ethics committee approval was sought and gained for the study, though no major ethical issues were foreseen.

4.3 Results

4.3.1 Significance Testing Between Groups

The overall scale scores and mean subscale scores were first calculated for each participant; these are summarised in Table 4.1 below. As Table 4.1 shows, *Prototype* was given higher ratings than *Hulk* on every single scale. Shapiro-Wilk tests found all groups of data to be normally distributed (all $p > 0.05$) with the exception of the *Hulk* Appeal Scale data ($p = 0.02$), whilst all of the data sets passed Levene's test for equality of variance (all $p > 0.05$) with the exception of the Playability Barriers (PB) data ($p = 0.012$); as parametric tests are normally robust to such minor violations of their assumptions a MANOVA test was performed on the data.

Table 4.1. *For Each Game, the Mean Score for Each Subscale and the Mean Scale Scores, with Standard Deviation*

Scale	***Prototype***		***Hulk***	
	M	*SD*	*M*	*SD*
Affective Experience Subscale	34.111	4.935	29.750	4.920
Focus Subscale	44.444	5.174	41.125	5.436
Playability Barriers Subscale	24.555	6.267	24.250	3.105
Usability Barriers Subscale	31.444	4.034	25.375	3.204
Summed Gameplay Scale	134.555	11.673	120.500	9.242
Appeal Scale	42.777	7.446	38.750	5.849

However, including the summed Gameplay Scale data in the analysis resulted in singularity, preventing Box's M (a measure of covariance) from being calculated.

This is a violation of one of the key assumptions of MANOVA, so the summed Gameplay Scale data (seen in Figure 4.3 below, which shows that *Prototype* was given higher scores than *Hulk*) were thus analysed in a separate one-tailed independent samples t-test. The Bonferroni correction was applied to the *p* value to account for the additional test (the correction is *new p = p/n* where *n* is the number of analyses being performed) and prevent the familywise error rate from increasing. Including all of the analyses to be performed, *n*= 6 and this entailed that the new *p* = 0.008. The t-test revealed that for the summed Gameplay Scale, *Prototype* was scored significantly higher than *Hulk* (*t* (15) = 2.73, p = 0.000) by participants, as hypothesised. The $\eta^2 = 0.33$, a large effect size that suggests 33% of the total variance in the GS scores is a result of varying the game (generally, .01 is a small effect size, .06 and above is moderate effect size and .14 and above is a large effect size (Cohen, 1988, cited in Pallant, 2001).

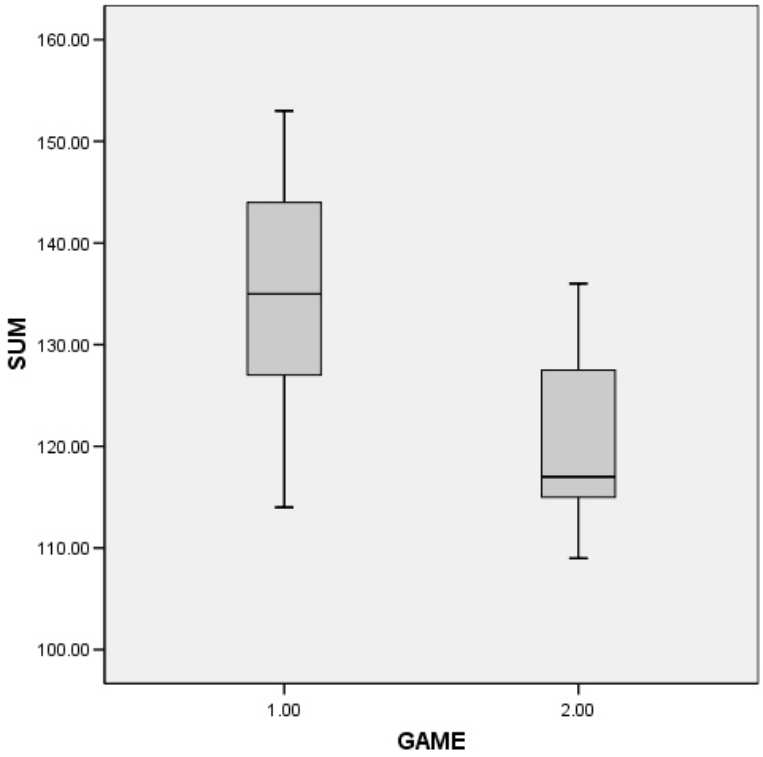

Figure 4.3. Mean summed Gameplay Scale score and *SD* for each game. ('1' = *Prototype*; '2'= *Hulk*)

The MANOVA analysis of the average subscale scores, sans the summed Gameplay Scale data, could then be performed as Box's $M = 12.302$, $p = 0.924$, suggesting that the data had homogeneity of covariance matrices - a key assumption for MANOVA analysis. Only the Usability Barriers subscale showed a significant difference between games (F (1, 15) = 11.580, p = 0.004); for all other measures $p > 0.008$. The partial $\eta^2 = 0.44$ for the Usability Barriers subscale, meaning that 44% of the subscale variance was accounted for by the game manipulation. Interestingly, for the Affective Experience subscale the partial $\eta^2 = 0.18$, for the Appeal Scale the partial $\eta^2 = 0.09$, whilst for the Focus subscale the partial $\eta^2 = 0.1$ – a large and two moderate effect size respectively, despite the difference being non-significant.

4.3.2 Inter-Scale Correlations

The other important question involved determining which (if any) scales correlated with the Appeal scale. The inter-scale correlations for the scale sums are shown in Table 4.2 below. Note that both groups were added together to ensure a large enough sample for correlational procedures; any correlations that exist should hold for both games given their high level of similarity.

Table 4.2. *Spearman's Correlations between Subscales, the Gameplay Scale and the Appeal Scale for Study 2. For each correlation value, its statistical significance is also reported.*

	Affective Experience	*Appeal Scale*	*Focus*	*Playability Barriers*	*Usability Barriers*	*Gameplay Scale*
Affective Experience	-	0.779**	0.711**	0.034	0.129	0.735**
Appeal Scale	0.779**	-	0.737**	0.095	0.143	0.708**
Focus	0.711**	0.737**	-	0.125	-0.061	0.760**
Playability Barriers	0.034	0.095	0.125	-	0.082	0.521*
Usability Barriers	0.129	0.143	-0.061	0.082	-	0.394
Gameplay Scale	0.735**	0.708**	0.760**	0.521*	0.394	-

** Correlation is significant at the 0.01 level (1-tailed).
* Correlation is significant at the 0.05 level (1-tailed).

As Table 4.2 and Figure 4.4 (below) illustrate, significant correlations exist between the Affective Experience, Focus, Appeal and overall Gameplay Scales ($p < 0.01$); the only other significant relationship is between the Playability Barriers and Gameplay Scales ($p = 0.015$). To further ensure construct validity, simple linear regression was performed, this time only between the Appeal and summed Gameplay Scales with both groups again added together as linear regression requires a minimum of at least 15 participants per dependent variable to be accurate (Stevens, 2002). This found an R^2 of 0.577, suggesting that 58% of the variance in the Appeal rating given by participants of both groups can be considered due to factors measured by the Gameplay Scale, the subsequent ANOVA finding this relationship to be significant (F (1,15) = 22.786, $p = 0.000$).

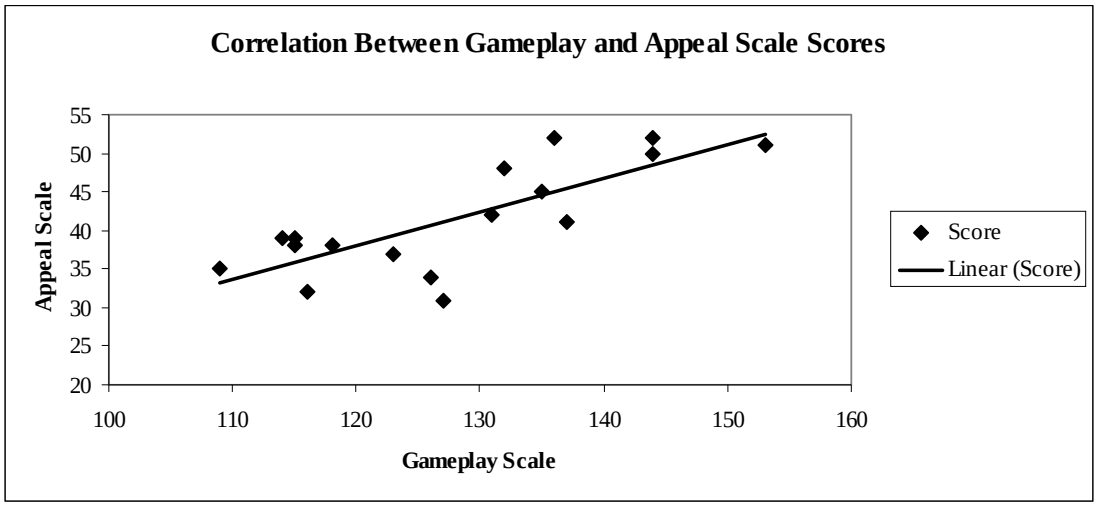

Figure 4.4. Correlation between Gameplay and Appeal Scale scores, with line of best fit.

4.3.3 Comparison with Study 1 Data

Finally, the data gathered in study 2 was compared to the data gathered in study 1. This involved performing a one-way ANOVA between the *Prototype*, *Hulk* and (study 1 game) *PixelJunk Eden* Gameplay Scale mean scores (seen below in Figure 4.5), followed by two planned comparisons: *PixelJunk Eden* vs. *Prototype* and *PixelJunk Eden* vs. *Hulk*. *PixelJunk Eden* was given a metascore of 80; we should expect its Gameplay Scale score to be significantly different to *Hulk* but not to *Prototype*. Five outliers (as identified by the SPSS boxplot) with Gameplay Scale scores of 80 or below were removed from the *PixelJunk Eden* data before analysis. The Levene's statistic was narrowly significant ($p = 0.046$) meaning that the planned comparison statistic in which equal variances are not assumed was used The ANOVA found no significant difference between the scores for the games (F (2,107) = 1.81867, $p = 0.16$); however the planned comparison shows the *PixelJunk Eden* vs. *Prototype* score difference to be non-significant (F (1, 11) = 0.5329, $p = 0.09$) and the *PixelJunk Eden* vs. *Hulk* comparison to be significant (F (1, 11) = 7.29, $p = 0.02$).

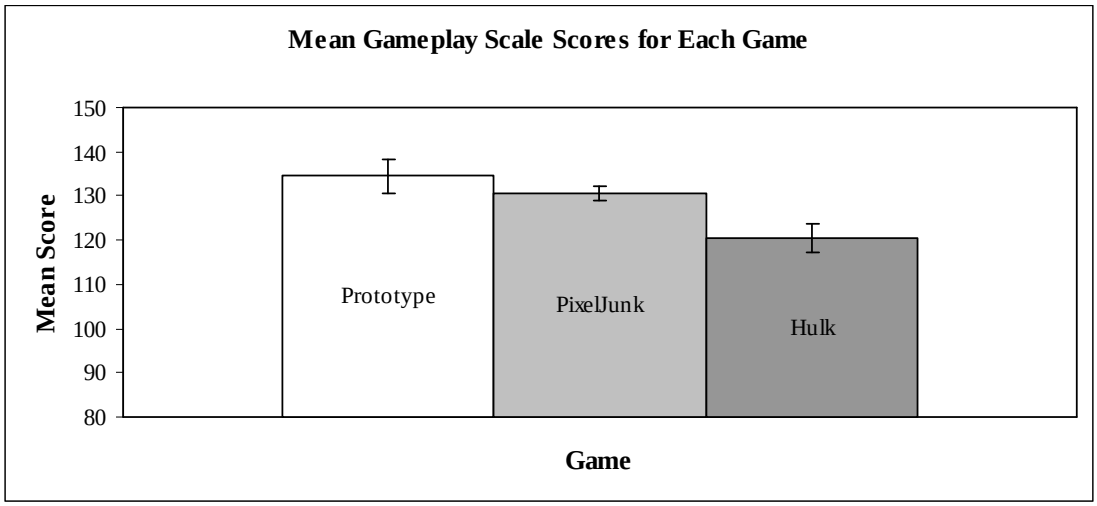

Figure 4.5. Mean Gameplay Scale scores for each game. Error bars show standard error.

4.4 Interim Discussion

The second study aimed to examine the generalisability of the Gameplay Scale and the results from study 1 by having participants play open-world action games. It also aimed to determine if the Gameplay Scale could distinguish between games based on their review scores. The study found that, using the Gameplay Scale, players rated *Prototype* significantly higher than *The Incredible Hulk,* allowing us to reject the null hypothesis. As *Prototype* was significantly better reviewed than *Hulk* by the gaming press, this suggests that the scale has good construct validity. This is further confirmed by the scale correctly ranking the *PixelJunk Eden* data. As *PixelJunk Eden* received almost the same metascore as *Prototype,* that the scale found no-significant difference between those two games but did find a significant difference between *PixelJunk Eden* and *Hulk* further suggests that the scale is measuring game quality in some way – thus making it useful to player testing. Furthermore, the Gameplay Scale correlated with the Appeal Scale, suggesting that the scale does measure the initial appeal of a game; it accounted for 58% of the variance in player appeal ratings which

gives the scale good construct validity. Finally, the large overall Gameplay Scale effect size showed that 33% of the variance in the scores came from varying the game played (and not just individual differences or error), again demonstrating that the scale can detect differences in game appeal and quality.

Moving to the specific subscales, the picture provided by the analysis is less clear. There being no significant difference in *Prototype's* or *Hulk's* Appeal Scale scores as well as the Affective Experience and Focus subscales correlating most strongly with the Appeal Scale both undermine the construct validity of the scale as there was no significant difference between each game's scores on these subscales. Moreover, the only subscale with a significant difference between each game's scores was the Usability Barriers subscale; a large effect was found to exist, with 44% of the variance being explained by the manipulation. However, reviews (e.g. Goldstein, 09 June 2008; McGarvey, 17 June 2008) generally focused on the high level of repetition in *Hulk* as its major failing – a Focus and not a Usability Barriers issue, according to the scale at least.

How then do we account for these apparently anomalous results? The first two are most likely due to the small sample size; despite its flaws, Hulk still provided an enjoyable experience for the hour that participants played it (no participant stated that they didn't enjoy playing it) , and thus the difference in initial appeal is probably rather small. The medium effect size for the Appeal Scale ($\eta^2 = 0.09$) suggests that with even a few more players, a significant difference is likely to be found.

As for the Affective Experience and Focus subscales, the same is probably true. With effect sizes of $\eta^2 = 0.18$ and $\eta^2 = 0.1$ respectively, it is clear that manipulating the game has a considerable influence of these scores, even if the

difference is not significant. With a larger sample (n = 15+) it is hypothesised that a significant difference will be found.

Given the above considerations, why should the Usability Barriers scale show a significant difference? It is likely that during initial play experiences, usability barriers are especially pronounced; players have to grapple with new control systems, menus and ranges of settings with which they are not familiar, so a larger effect size is perhaps to be expected. From observing the play sessions, these indeed appeared far worse for the *Hulk*, with many players having a lot of (initial) trouble with the healing, climbing and targeting controls, as well with the menu screens. That isn't to say that there were no issues with *Prototype*; the 'disguise' mechanic in this game was very poorly explained and thus resulted in a lot of player confusion.

However, these usability problems did not correlate with the Appeal Scale, suggesting that Usability Barriers factors do not determine a game's initial appeal (the same was found in the study 1 regression analysis – see section 3.4.3). Referring to the interview transcripts, most players did not view the game as having poor usability even if they struggled with a game mechanic unnecessarily – e.g. from the interviews:

> "Researcher: *Do you think the game could have made that [the picking up objects mechanic] clearer?*
> Hulk Participant 8: *Maybe. But maybe I could have concentrated more*"

> "Hulk Participant 4 [On having difficulty with the jumping controls] …*I think it's a good thing, because I've never played a game where you jump like that before, so it obviously takes a bit of getting used to…*"

This implies that ‘core gamers’ (i.e. those that play regularly) expect some degree of struggle when it comes to learning a new game’s control - unless the usability problems are very severe, it is only when these problems persist through long-term play experiences that they inhibit play. The opposite is likely true for casual gamers (i.e. those that do not play very often).

For the same reason, the reviews did not mention these usability issues, largely because the longer play session had resulted in the reviewers transcending these initial usability problems, and instead encountering Focus issues (such as a lack of variety) that were not present in the initial play session. The interesting result that still requires explanation, then, is how the Gameplay Scale was able to arrange the games in order of their metascores despite the most significant difference being in a subscale that determined neither the review scores nor the game’s initial appeal to players (indeed, the Usability Barriers subscale didn’t actually correlate with the scale as a whole for study 2). The first thing to note is that, despite having non-significant differences with *Hulk*, both *PixelJunk Eden* and *Prototype* scored higher than *Hulk* for almost all subscales; this would have contributed to the overall significant difference, meaning that these factors were still important in ranking the games. Second, the poor Usability Barriers ratings for *Hulk* likely counteracted the Affective Experience and Focus scores that the game received, which were probably boosted by the short play session that prevented much repetition from occurring. Thirdly, it is possible that games that suffer from such initial usability problems are more likely to be poorly designed in a way that can also inhibit the long-term appeal of the game and thus review scores.

Finally, the Playability Barriers subscale showed no indication of either a meaningful effect size or of a significant difference between games. Whilst a larger

sample could again remedy this, another cure would be to broaden its scope and include some of the playability issues that were excluded from the initial scale (such as rewards, feedback, etc) for the sake of brevity; this would make the instrument more sensitive to differences in playability between games.

Despite some anomalous results with the individual subscales, study 2 has shown the Gameplay Scale as a whole to have good construct validity by measuring a video game's initial appeal and good generalisability by measuring differing genres. In addition, both construct validity and the utility of the Gameplay Scale to industry has been demonstrated by the scale ranking games according to their review scores. The final task is to discuss the wider implications of these findings.

5 General Discussion

Through study 1 an instrument was developed to measure two major aspects of player experience, as well as the usability and playability factors of the game, revealing that all except the usability factors contributed to a game's initial appeal. Study 2 has shown that the Gameplay Scale instrument should have predictive power when it comes to anticipating the quality and appeal of video games across different genres, although reference to the individual subscales reveals that it may not have behaved in this way for the expected reasons. In short, the scale developed thus far should be considered a proof of principle; that, as Hassenzahl *et al* (2000) found for productivity software, both hedonic and ergonomic factors influence a video game's appeal and that it is appropriate to measure these factors in the manner described by this study. Whilst the ergonomic usability and playability factors generally were less important in determining initial appeal, there are good reasons to believe that this was only the case for initial appeal, an that longer gaming sessions increase the importance of these factors.

It perhaps goes against the heuristics examined (in Febretti and Garzotto, 2009; Desurvire and Wiberg, 2009; Pinelle, Wong and Stach, 2008; Federoff, 2002 and Korhonen and Koivisto, 2006) to find that usability and playability factors were less important in determining a game's initial appeal than expected. However, these factors were still important to some degree, whilst it is maintained that longer testing sessions will highlight their importance. In terms of the player experience factors, considering challenge as associated with cognitive absorption instead of immersion (*contra* Ermi and Mäyrä, 2005) seems to have been the correct choice given the clustering, even if cognitive absorption (included as per Jennett *et al*, 2008) was the

dominant factor in the Focus subscale. Nevertheless, the Focus subscale should act as a measure of flow that can also measure suboptimal experiences, improving on how many existing scales (e.g. IJsselsteijn *et al*, in preparation) chose to measure flow.

The Focus subscale being considerably longer (9 items as opposed to 5 or 6 for the other subscales) was an artifact of the process used to whittle the scale down from 49 questions, but has important implications. If we accept the success of the entire scale in ranking the games according to review score as a reason to accept the Focus subscale's larger size, it suggests that the Focus factors are of greater importance in determining a game's reviews. Whilst Affective Experience factors generally correlated more strongly with initial appeal, the *Hulk* reviews (as an anecdotal example) did note Focus issues as determining the game's quality and appeal, so long term play might value Focus factors more highly. It is interesting that none of the models in the previous literature (save perhaps for Chen, 2007) noted Focus constructs as being more important than other player experience factors. It is possible that not all constituents of the gameplay experience are equally important, but it is likely that this varies from game to game and genre to genre.

Finally, the Affective Experience scale measures especially important aspects of the game experience (judging by its correlation to the Appeal Scale) and so the models of Ermi and Mäyrä (2005), Brown and Cairns (2005) and Malone (1981) were highly successful in capturing this rich phenomenology through the Gameplay Scale.

There are also other, quite different approaches to the video game experience that were not covered in the literature review. Whilst the current study summarised much of the HCI literature, Ryan *et al* (2006) and Rigby and Ryan (2007) examined player experience factors in a very different way. Their Player Experience of Need Satisfactions (PENS) model was founded upon motivational psychology, and included

factors such as the need for player feelings of autonomy and competence in addition to more familiar constructs such as presence and intuitive controls. This model was also found to discriminate between games with high and low metascores (although it could be argued that almost any model could distinguish between a game rated 56.6% and a game rated 97.8%, as Ryan *et al* (2006) were able to) using a scale, and so future work should also consider including such motivational factors.

As noted though, this should be considered an exploratory study. The sample size was small (n = 98 for study 1, n = 17 for study 2) preventing a full factor analysis from being performed in study 1 and limiting the strength of the conclusions in study 2. Indeed, the sample size of study 2 fell below Tullis and Stetson's (2004) recommended minimum of 10 per group when using such scales. Moreover, further analysis is required to establish what a high score on the scale is and what is a low score – and thus establish benchmarking. Not only that, but some of the clustering of the questions is perhaps suspect (especially the empathy and camera questions being in the Focus subscale, which was quite unexpected), whilst – against the received wisdom on scale design – moving the general questions to the end may encourage players to base their judgements on a all of the elements referred to in the preceding specific questions.

The generalisability from these results can also be questioned. Only three games were studied, whereas previous scales were built upon the experiences of players across many games (e.g. Jennett *et al*, 2008). Yet throughout the two studies, such generalisability was sacrificed to increase the internal validity of the results. By only having players (recruited from a console manufacturer's official forum, a reasonably reliable source for a web survey) play one game in study 1 the error introduced by recruiting any players from a score of third party websites was reduced.

Moreover, controlling the sort of games (and sort of player) in study 2 allowed for direct comparison between two games even if each player only played one. Of course, repeated measures would have been desirable here, but ensuring that players played the game for long enough (i.e. at least an hour; even longer would have been preferable) prevented this – and resources were not available to study longer sessions involving multiple games or use greater triangulation.

The choice of game is then perhaps suspect – ones that could have been better rated by the players in an hour may have been a better choice, although the remarkable similarity between the two titles did help to ensure a reasonable level of control. Finally, it is recognised that metascores, and the opinions of reviewers, are not an ideal external measure of game quality, especially given repeated allegations of games publishers tampering with review scores (e.g. Plunkett, 10 July 2009). Measures of player arousal - such as galvanic skin response or facial electromyography (Nacke and Lindley, 2008) may have been preferable. Nevertheless, metascores are held to be highly important in the industry, with anecdotal evidence suggesting that many publishers can correlate the metascores of games to the sales of said games (Stuart, 17 January 2008,). This is why metascores were chosen – by being be able to predict metascores when assessing in-development games, the Gameplay Scale may predict the sales of a game and suggests areas for improvement that should increase the appeal of the game and thus its sales. By doing this, the Gameplay Scale (or at least, the rationale behind it) is potentially very useful to the video game industry.

In addition to overcoming the weaknesses already mentioned, future studies will need to expand such scales to cover multiple genres; even the core constructs that this scale represents will not cover every game genre; e.g. Massive-Multiplayer

Roleplaying Games (MMORPG) have social factors that it is very important to measure. One solution to this is to add 'modules' to the scale; extra (validated) subscales that cover elements that are important to each genre. For instance, we might add a 'social factors' subscale when assessing MMORPGs. We also need to study more types of players, the current study assessing only dedicated or 'core' gamers; if we want to cover more casual gamers, we might add an 'approachability subscale', as per Desurvire and Wiberg (2008). In addition, whilst the androcentric gender bias mirrors that of core gamers, more female participants will be required to study casual gamers.

Nevertheless, despite the limitations noted the study has achieved its ultimate aim – to establish that player experience, usability and playability factors are all important in player perception of video game quality, and then to create a scale to measure these characteristics during player testing. Moreover, by drawing on previous questionnaires and the questionnaire design literature, the issues documented by Hornbaek (2006) were largely avoided by increasing the rigour involved in the scale's design.

The Gameplay Scale, and the reasoning behind it, therefore has utility to the games industry due its predictive strength. Whilst the review scores of the games tested were already known, there is no reason (once benchmarking is established) that the Gameplay Scale couldn't be used to predict the likely review score of a game. Combined with the usual uses of such scales in player testing (i.e. to add attitudinal data or locate important areas in a dataset), a scale that measures both hedonic and ergonomic factors is very useful indeed.

6 Conclusions

Many studies on video game experience end with a complex diagram, showing the interrelations between the analysed constructs. This one does not, as it is recognised that player experiences are so varied that such formulations of how 'experience *x* leads to experience *y* or is a species of experience *z*' fail to capture this variety. Instead, the key message from this study is that the appeal of video games does have many components, but that the relations between these constructs can vary depending on the game. What is more certain, however, is that whatever the enjoyable experience of a particular game entails this enjoyment cannot be fully realised if barriers are in place that prevent the user from engaging in play. Some of these barriers are deeply embedded in the experience (such as the quality of the characters or fundamental game mechanics) and may be difficult to improve; other however, are not.

These are the playability and usability factors measured by this study, which were found to be important predictors of review scores and perceived appeal. Those involved in player testing are thus recommended to measure both hedonic and ergonomic factors. Whilst the current study focused only on initial appeal, this is still of importance in player testing. Both Pagulayan *et al* (2003) and Fulton (2002) note the need to optimise the player experience from 1 minute to 10 minutes to 1 hour to 10 hours, whilst the proliferation of easily downloadable demos means that initial gameplay experiences are much more closely tied to purchase choices than before.

This thesis has thus shown that player experience, usability and playability all contribute to a game's initial appeal and devised a scale to measure all three of these factors, neither of which had been done before. Moreover, scores derived using the

scale correlate to the review scores of games, suggesting that the scale may have practical applications in industry player testing. Although the scale designed in this study may only make a modest contribution to player testing protocols, this could nevertheless be a useful contribution, allowing us to quantify the degree to which the game presents barriers to player enjoyment. Again, players choose to play games, and unless we remove or reduce such barriers they can always put the controller down and do something else.

References

Abran, A., Khelifi, A., Suryn, W., and Seffah, A. (2003). Usability meanings and interpretations in ISO standards. *Software Quality Journal*, 11(4):325–338.

Agarwal, R. and Karahanna, E. (2000). Time flies when you're having fun: Cognitive absorption and beliefs about information technology usage. *MIS Quarterly*, 24(4):665–694.

Alwin, D. F. and Krosnick, J. A. (1991). The reliability of survey attitude measurement: The influence of question and respondent attributes. *Sociological Methods Research*, 20(1):139–181.

Arsenault, D. (2005). Dark waters: Spotlight on immersion. In *Game On North America 2005 Conference Proceedings*, pages 50–52.

Bangor, A., Kortum, P. T., and Miller, J. T. (2008). An empirical evaluation of the system usability scale. *International Journal of Human-Computer Interaction*, 24(6):574–594.

Bergstrom, B. A. and Lunz, M. E. (1998). Rating scale analysis: Gauging the impact of positively and negatively worded items. In *Annual Meeting of the American Educational Research Association*. April 13-17 1998; San Diego, CA

Brooke, J. (1996). SUS: A quick and dirty usability scale. In Jordan, P. W., Thomas, B., Weerdmeester, B. A., and McClelland, I. L., editors, *Usability Evaluation in Industry*. Taylor & Francis., London.

Brown, E. and Cairns, P. (2004). A grounded investigation of game immersion. In *CHI '04: CHI '04 Extended Abstracts On Human Factors In Computing Systems*, pages 1297–1300, New York, NY, USA. ACM Press.

Cairns, P. and Cox, A. L. (2008). *Research Methods for Human-Computer Interaction*. Cambridge University Press, New York, NY, USA.

Carifio, J. and Perla, R. J. (2007). Ten common misunderstandings, misconceptions, persistent myths and urban legends about likert scales and likert response formats and their antidotes. *Journal of Social Sciences*, 3(3):106-116.

Carifio, J. and Perla, R. (2008). Resolving the 50-year debate around using and misusing likert scales. *Medical Education*, 42(12):1150–1152.

Chen, J. (2007). Flow in games (and everything else). *Commun. ACM*, 50(4):31–34.

Chen, M. and Johnson, S. (2004). *Measuring flow in a computer game simulating a foreign language environment*. Retrieved 17 June 2009 from http://markdangerchen.net/papers/

Chin, J. P., Diehl, V. A., and Norman, K. L. (1988). Development of an instrument measuring user satisfaction of the human-computer interface. In *CHI '88: Proceedings of the SIGCHI Conference On Human Factors In Computing Systems*, pages 213-218, New York, NY, USA. ACM Press.

Colman, A. M., Norris, C. E., and Preston, C. C. (1997). Comparing rating scales of different lengths: Equivalence of scores from 5-point and 7-point scales. *Psychological Reports*, (80):355–362.

Cowley, B., Charles, D., Black, M., and Hickey, R. (2008). Toward an understanding of flow in video games. *Comput. Entertain.*, 6(2):1–27.

Csikszentmihalyi, M. (1975) *Flow: The Psychology of Optimal Experience* Harper Perennial; New York.

Csikszentmihalyi, M. (2000). *Beyond Boredom and Anxiety: Experiencing Flow in Work and Play*. Jossey-Bass; New York.

Dawes, J. (2001). Comparing data gathered using 5 point versus 11 point scales. In *Australian & New Zealand Marketing Academy Conference*. Massey University.

Desurvire, H., Caplan, M., and Toth, J. A. (2004). Using heuristics to evaluate the playability of games. In *CHI '04: CHI '04 extended abstracts on Human factors in computing systems*, pages 1509–1512, New York, NY, USA. ACM Press.

Desurvire, H. and Wiberg, C. (2008). Master of the game: assessing approachability in future game design. In *CHI '08: CHI '08 Extended Abstracts On Human Factors In Computing Systems*, pages 3177–3182, New York, NY, USA. ACM.

Desurvire, H. and Wiberg, C. (2009). Game usability heuristics (play) for evaluating and designing better games: The next iteration. *Lecture Notes in Computer Science*, 56(21); 557-556

Dillman, D. A., Tortora, R. D., and Bowker, D. (1998). *Principles for Constructing Web Surveys.* Technical report, SESRC, Washington.

Dillon, A. (2001). Beyond usability: process, outcome and affect in human computer interactions. *Canadian Journal of Information Science*, 26(4):57-69.

Douglas, Y. and Hargadon, A. (2000). The pleasure principle: immersion, engagement, flow. In *HYPERTEXT '00: Proceedings of the Eleventh ACM on Hypertext and Hypermedia*, pages 153–160. ACM Press.

Eastin, M. S. and Griffiths, R. P. (2006). Beyond the shooter game: Examining presence and hostile outcomes among male game players. *Communication Research*, 33(6):448–466.

Ermi, L. and Mäyrä, F. (2005). Fundamental components of the gameplay experience. In de Castell, S. and Jenson, J., editors, *Changing Views: Worlds in Play. Selected papers of the 2005 Digital Games Research Association's (DiGRA) Second International Conference*, pages 15–27.

Everitt, B. S. (1975). Multivariate analysis: the need for data, and other problems. *The British Journal of Psychiatry*, 126(3):237–240.

Febretti, A. and Garzotto, F. (2009). Usability, playability, and long-term engagement in computer games. In *CHI EA '09: Proceedings Of The 27th International Conference Extended Abstracts On Human Factors In Computing Systems*, pages 4063–4068, New York, NY, USA. ACM.

Federoff, M. A. (2002). *Heuristics And Usability Guidelines For The Creation And Evaluation Of Fun In Video Games.* Master's thesis, Indiana University

Foddy, W. (1993). *Constructing Questions for Interviews and Questionnaires: Theory and Practice in Social Research*. Cambridge University Press, Cambridge.

Fowler, F. J. (2001). *Survey Research Methods (Applied Social Research Methods)*. Sage Publications, Inc.

Fulton, B. (2002). Beyond psychological theory: getting data that improves games, in *Proceedings of the Game Developer Conference*. (San Jose, CA, March 2002).

Ganassali, S. (2008). The influence of the design of web survey questionnaires on the quality of responses. *Survey Research Methods*, 2(1):21–32.

Garland, R. (1991). The mid-point on a rating scale: Is it desirable? *Marketing Bulletin*, (2):66–70.

Gilleade, K. M. and Dix, A. (2004). Using frustration in the design of adaptive videogames. In *ACE '04: Proceedings of the 2004 ACM SIGCHI International Conference on Advances In Computer Entertainment Technology*, pages 228–232, New York, NY, USA. ACM.

Gilljam, M. and Granberg, D. (1993). Should we take don't know for an answer? *Public Opin Q*, 57(3):348–357.

Goldstein, H. (09 June 2008). The incredible hulk review. In *IGN*. Retrieved 08 August 2009 from http://uk.ps3.ign.com/articles/880/880381p1.html

Gray, W. D. and Salzman, M. C. (1998). Damaged merchandise? a review of experiments that compare usability evaluation methods. *Human-Computer Interaction*, 13(3):203–261.

Grodal, T. (2000) Video games and the pleasures of control. In D. Zillmann and P . Vorderer (eds) *Media Entertainment*, pp. 197–212. Mahwah, NJ: Erlbaum

Guilford, J. P. (1954). *Psychometric Methods*. McGraw-Hill, New York.

Hartley, J. and Betts, L. R. (2009). Four layouts and a finding: the effects of changes in the order of the verbal labels and numerical values on likert-type scales. *International Journal of Social Research Methodology*, 99(1):1–11.

Hassenzahl, M., Platz, A., Burmester, M., and Lehner, K. (2000). Hedonic and ergonomic quality aspects determine a software's appeal. In *CHI '00: Proceedings of the SIGCHI Conference On Human Factors In Computing Systems*, pages 201–208, New York, NY, USA. ACM.

Hinz, A., Michalski, D., Schwarz, R., and Herzberg, P. Y. (2007). The acquiescence effect in responding to a questionnaire. *GMS Psycho-Social-Medicine*, 4.

Hopson, J. (10 November, 2006) We're Not Listening: An Open Letter to Academic Game Researchers. In Gamasutra, Retrieved 21/August/2009 from http://www.gamasutra.com/features/20061110/hopson_01.shtml

Hornbaek, K. (2006). Current practice in measuring usability: Challenges to usability studies and research. *International Journal of Human-Computer Studies*, 64(2):79–102.

Huizinga, J. (1998) *Homo Ludens*. (R.F.C. Hull, Trans.) Routledge; New York (Original work published 1938)

IJsselsteijn, W., de Kort, Y., Poels, K., Jurgelionis, A., and Bellotti, F. (2007). Characterising and measuring user experiences in digital games. In *International Conference on Advances in Computer Entertainment*.

IJsselsteijn, W. A., de Kort, Y. A. W., and Poels, K. (in preparation). *The Game Experience Questionnaire: Development of a Self-Report Measure to Assess the Psychological Impact of Digital Games*. Manuscript in preparation.

ISO 9241-11 (1998) *Ergonomic Requirements for Office Work With Visual Display Terminals (VDTs) – Part 11: Guidance on Usability*

Jacoby, J. and Matell, M. S. (1971). Three-point likert scales are good enough. *Journal of Marketing Research*, 8(4):495–500.

Jennett, C., Cox, A., and Cairns, P. (2009). Being 'in the game'. In Gunzel, S., Liebe, M., and Mersch, D., editors, *Proc. of the Philosophy of Computer Games 2008*,, pages 210–227. Potsdam University Press,.

Jennett, C., Cox, A., Cairns, P., Dhoparee, S., Epps, A., Tijs, T., and Walton, A. (2008). Measuring and defining the experience of immersion in games. *International Journal of Human-Computer Studies*, 66(9):641–661.

Jensen, M. P. (2003). Questionnaire validation: a brief guide for readers of the research literature. *The Clinical Journal Of Pain*, 19(6):345–352.

Kim, J. H., Gunn, D. V., Schuh, E., Phillips, B., Pagulayan, R. J., and Wixon, D. (2008). Tracking real-time user experience (TRUE): a comprehensive instrumentation solution for complex systems. In *CHI '08: Proceeding Of The Twenty-Sixth Annual SIGCHI Conference On Human Factors In Computing Systems*, pages 443–452, New York, NY, USA. ACM.

Kim, W. W. (2006). *Engagement, Body Movement and Emotions in Games: Relationships And Measurements.* Master's thesis, University College London.

Kline, P. (1998). *New Psychometrics: Science, Psychology and Measurement.* Routledge.

Korhonen, H. and Koivisto, E. M. I. (2006). Playability heuristics for mobile games. In *MobileHCI '06: Proceedings Of The 8th Conference On Human-Computer Interaction With Mobile Devices And Services*, pages 9–16, New York, NY, USA. ACM.

Korhonen, H. and Koivisto, E. M. I. (2007). Playability heuristics for mobile multi-player games. In *DIMEA '07: Proceedings Of The 2nd International Conference On Digital Interactive Media In Entertainment And Arts*, pages 28–35, New York, NY, USA. ACM.

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology*, 50(1):537–567.

Laitinen, S (June 23 2005) Better games through usability evaluation and testing. In *Gamasutra.* Retrieved 10 June 2009 from http://www.gamasutra.com/features/20050623/laitinen_01.shtml

Lazzaro, N. (2004). Why we play games: Four keys to more emotion in player experiences. In *Proceedings of the Game Developers Conference.*

Lehmann, D. R. and Hulbert, J. Are three-point scales always good enough? *Journal of Marketing Research*, 9(4):444–446.

Lewis, J. R. (1995). IBM computer usability satisfaction questionnaires: psychometric evaluation and instructions for use. *Int. J. Hum.-Comput. Interact.*, 7(1):57–78.

Lietz, P. (2008) *Questionnaire design in attitude and opinion research: Current state of an art.* Technical report number: FOR 655, Jacobs University Bremen.

Lindley, C. and Nacke, L. (2008). Boredom, immersion, flow - a pilot study investigating player experience. In *IADIS Gaming 2008: Design for Engaging Experience and Social Interaction*,. IADIS.

Lindley, S. E., Le Couteur, J., and Berthouze, N. L. (2008). Stirring up experience through movement in game play: effects on engagement and social behaviour. In *CHI '08: Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, pages 511–514, New York, NY, USA. ACM.

Malone, T. W. (1981). Toward a theory of intrinsically motivating instruction. *Cognitive Science*, 5(4):333–369.

Martin, E. (2006). *Survey questionnaire construction*. Technical report, US Census Bureau.

Massimini, F. and Massimo, C. (1988). The systematic assessment of flow in daily experience. In Csikszentmihalyi, M. and Csikszentmihalyi, I., editors, *Optimal Experience*, pages 288–306. Cambridge University Press, New York.

Masters, J. R. (1974). The relationship between number of response categories and reliability of likert-type questionnaires. *Journal of Educational Measurement*, 11(1):49–53.

McGarvey, S. (17 June 2008). Reviews: The incredible hulk. In *GameSpy*. Retrieved 08 August 2009 from http://uk.ps3.gamespy.com/playstation-3/the-incredible-hulk-the-movie/882516p1.html

McGuire, W. J. (1960). Cognitive consistency and attitude change. *Journal of Abnormal and Social Psychology*, 60:345–353.

McMahan, A. (2003). Immersion, engagement, and presence: A method for analyzing 3d videogames. In Wolf, M. J. P. and Perron, B., editors, *The Video Game Theory Reader*, pages 67–86. Routledge, New York.

Miller, G. A. (1956). The magical number seven plus or minus two: some limits on our capacity for processing information. *Psychol Rev*, 63(2):81–97.

Nacke, L. and Lindley, C. A. (2008). Flow and immersion in first-person shooters: measuring the player's gameplay experience. In *Future Play '08: Proceedings of the 2008 Conference on Future Play*, pages 81–88, New York, NY, USA. ACM.

Nielsen, J. and Molich, R. (1990). Heuristic evaluation of user interfaces. In *CHI '90: Proceedings Of The SIGCHI Conference On Human Factors In Computing Systems*, pages 249–256, New York, NY, USA. ACM Press.

Oppenheim (1992). *Questionnaire Design, Interviewing & Attitude Measurement.* Pinter, London.

Pallant, J. (2001*) SPSS Survival Manual*. Open University Press, Maidenhead, UK.

Pagulayan, R. J., Keeker, K., Wixon, D., Romero, R. L., and Fuller, T. (2003). User-centered design in games. In J.A. Jacko and A. Sears (Eds.) *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications*. Lawrence Erlbaum Associates; New York, pages 883–906.

Peytchev, A., Couper, M. P., McCabe, E. S., Crawford, S.D. (2006). Web survey design. *Public Opinion Quarterly*, 70(4):596–607.

Pinchbeck, D. (2005). Is presence a relevant or useful construct in designing game environments? In *Proceedings of the 2nd Annual International Workshop in Computer Game Design and Technology*. Liverpool John Moores University.

Pinelle, D. and Wong, N. (2008). Heuristic evaluation for games: usability principles for video game design. In *Chi '08: Proceeding Of The Twenty-Sixth Annual SIGCHI Conference On Human Factors In Computing Systems*, pages 1453–1462, New York, NY, USA. ACM.

Plunkett, L. (10 July 2009). Eidos once again attempting to mess with review scores? In *Kotaku*. Retrieved 08 August 2009, from http://kotaku.com/5311606/%5Bupdate%5D-eidos-once-again-attempting-to-mess-with-review-scores.

Polson, P. G., Lewis, C., Rieman, J., and Wharton, C. (1992). Cognitive walkthroughs: a method for theory-based evaluation of user interfaces. *Int. J. Man-Mach. Stud.*, 36(5):741–773.

Preston, C. and Colman, A. M. (2000). Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica*, 104(1):1-15.

Rigby, S. and Ryan, R. (2007). *The Player Experience Of Need Satisfaction (Pens):An Applied Model And Methodology For Understanding Key Compone nts Of The Player Experience.* Technical report, Immersyve, Celebration, FL.

Ryan, R. M., Rigby, C. and Przybylski, A. (2006). The motivational pull of video games: A self-determination theory approach. *Motivation and Emotion*, 30(4):344–360.

Sauro, J. and Lewis, J. R. (2009). Correlations among prototypical usability metrics: evidence for the construct of usability. In *Chi '09: Proceedings Of The 27th International Conference On Human Factors In Computing Systems*, pages 1609–1618, New York, NY, USA. ACM.

Schaeffer, N. C. and Presser, S. (2003). The science of asking questions. *Annual Review of Sociology*, 29:65–88.

Scheibler, D. and Schneider, W. (1985). Monte Carlo tests of the accuracy of cluster analysis algorithms: A comparison of hierarchical and nonhierarchical methods. *Multivariate Behavioral Research*, 20(3):283–304.

Shelley, B. (15 August 2001). Guidelines for developing successful games. *Gamasutra*. Retrieved 15 June 2009, from http://www.gamasutra.com/features/20010815/shelley_01.htm

Slater, M. (2004). How colorful was your day? why questionnaires cannot assess presence in virtual environments. *Presence: Teleoper. Virtual Environ.*, 13(4):484–493.

Stevens, J. (2002). *Applied Multivariate Statistics For The Social Sciences*. Lawrence Erlbaum Associates, Philadelphia.

Stuart, K. (17 January 2008). Interview: the science and art of metacritic. In *The Guardian Gamesblog*. Retrieved 08 August 2009 from http://www.guardian.co.uk/technology/gamesblog/2008/jan/17/interviewtheartofmetacriti

Sweetser, P. and Wyeth, P. (2005). GameFlow: a model for evaluating player enjoyment in games. *Comput. Entertain.*, 3(3):3.

Takatalo, J., Hokkinen, J., Komulainen, J., Sarkela;, H., and Nyman, G. (2006). Involvement and presence in digital gaming. In *NordiCHI '06: Proceedings of the 4th Nordic conference on Human-computer interaction*, pages 393–396, New York, NY, USA. ACM Press.

Thorndike, R. M. (1978). *Correlational Procedures for Research*. Gardner Press, New York.

Tourangeau, R. (1999). Context effects on answers to attitude questions. In Sirken, M. G., Herrmann, D. J., Schechter, S., Schwarz, N., Tanur, J. M., and Tourangeau, R., editors, *Cognition and Survey Research*, pages 111-132. Wiley, New York.

Tractinsky, N., Katz, A., and Ikar, D. (2000). What is beautiful is usable. *Interacting with Computers*, 13(2):127–145.

Tullis, T. S. and Stetson, J. N. (2004). A comparison of questionnaires for assessing website usability. In *Proceedings of the Usability Professionals Association Conference*, Minneapolis, MN: UPA.

Velez, P. and Ashworth, S. D. (2007). The impact of item readability on the endorsement of the midpoint response in surveys. *Survey Research Methods*, 1(2):69–74.

Wickramasekera, I. E. (2007). Empathic features of absorption and incongruence. *The American Journal of Clinical Hypnosis*, 50(1):59–69.

Witmer, B. G. and Singer, M. J. (1998). Measuring presence in virtual environments: A presence questionnaire. *Presence: Teleoper. Virtual Environ.*, 7(3):225–240.

Zagalo, N., Torres, A., and Branco, V. (2005). Emotional spectrum developed by virtual storytelling. In *Virtual Storytelling,* pages 105-114.

Appendix A

The Initial Gameplay Scale

1.) Information provided to participants before beginning questionnaire:

Participant Information - Read This

We would like to invite you to participate in this research project. Before you decide whether you want to take part, it is important for you to read the following information carefully and discuss it with others if you wish.

- The purpose of this study is to investigate player experience of video games using a questionnaire. You will have been asked to play a game for up to 2 hours.
- You are now asked to complete the following survey, answering each of the statements as truthfully as possible. The survey should take around 10 minutes to complete.
- This study is being performed by University College London (UCL) with the cooperation of Sony Computer Entertainment Europe (SCEE); please note that whilst SCEE will have to access to the results of the study they will not have access to any private details for use in marketing, etc. All data will be collected and stored in accordance with the Data Protection Act 1998.
- If you decide to take part you are still free to withdraw at any time and without giving a reason.
- Ask us if there is anything that is not clear or you would like more information. If you do have any questions, please contact the researcher for this study, Mark Parnell, at mjparnell@gmail.com or the supervisor for the research, Dr. Nadia Bianchi-Berthouze, at n.berthouze@ucl.ac.uk

HEALTH WARNING

Always play in a well lit environment. Take regular breaks, 15 minutes every hour. Discontinue playing if you experience dizziness, nausea, fatigue or have a headache. Some individuals are sensitive to flashing or flickering lights or geometric shapes and patterns, may have an undetected epileptic condition and may experience epileptic seizures when watching television or playing videogames. Consult your doctor before playing videogames if you have an epileptic condition and immediately should you experience any of the following symptoms whilst playing: altered vision, muscle twitching, other involuntary movement, loss of awareness, confusion and/or convulsions.

2.) Example of appearance of question with response item in both the initial and revised Gameplay Scales:

1.)

	Strongly Disagree	Disagree	Slightly Disagree	Neither Agree Nor Disagree	Slightly Agree	Agree	Strongly Agree
I enjoyed the game.	◯	◯	◯	◯	◯	◯	◯

Figure A.1 Example response item from the initial Gameplay Scale.

3.) Full list of questions in the initial Gameplay Scale, in correct order:

Q1 I enjoyed the game.
Q2 I felt the game was hard.
Q3 I was focused on the game.
Q4 I could identify with the characters.
Q5 I thought that the game was fun.
Q6 I found the game boring.
Q7 I liked how the game looked.
Q8 The game trained me in all of the controls.
Q9 I thought that the game was repetitive.
Q10 I found the game to be easy.
Q11 Playing the game made me happy.
Q12 I knew how to use the controller with the game.
Q13 I felt like events in the game were happening to me.
Q14 I found the game's menus to be usable.
Q15 I always knew where to go in the game.
Q16 I knew how to change the settings in the game.
Q17 It felt like I was responsible for what happened in the game.
Q18 Moving my point of view in the game was easy.
Q19 The game would provide help at appropriate moments.
Q20 I was unaware of the passage of time whilst playing.
Q21 I found the appearance of the game world to be interesting.
Q22 I found the controls to be difficult.
Q23 I forgot about my surroundings whilst playing.
Q24 I found using the options screen to be difficult.
Q25 I thought the controls were intuitive.
Q26 I concentrated on sounds in the game.
Q27 I knew how the game would respond to my actions.
Q28 I always knew how to achieve my aim in the game.
Q29 I thought the game mechanics were consistent.
Q30 My objectives in the game were unclear.
Q31 The game provided me with an adequate tutorial.
Q32 I lost my direction through the game.
Q33 I knew when my goal in the game had changed.

Q34 I thought that the game's menus were intuitive.
Q35 I felt that the game provided enough variety.
Q36 I found the game's menus to be cumbersome.
Q37 I couldn't find my way in the game world.
Q38 I found the game mechanics to be varied enough.
Q39 I found the game's story to be dull.
Q40 I played by my own rules in the game.
Q41 I thought about things other than the game whilst playing.
Q42 My field of view made it difficult to see what was happening in the game.
Q43 The aesthetics of the game were unimpressive.
Q44 I thought the camera angles in the game were appropriate.
Q45 The game failed to motivate me to keep playing.
Q46 The game responded to my inputs in an inconsistent way.
Q47 I wanted to explore the game world.
Q48 I thought the level of difficulty was right for me.
Q49 I knew how to customize the way that the game was set up.

Demographic Questions (asked at the very end):

- How old are you?

- What is your gender?

- Is English your first language?

- Thank you for completing the questionnaire. If you have any further comments about the content of the questionnaire - or what it lacked - please provide them below.

Appendix B

Results of Cluster Analysis on the Initial Gameplay Questionnaire

Table B.1. *Cluster Membership of Each Gameplay Scale Item.*

Construct	*Question*	*Question Number*	*Cluster*
Affect	I enjoyed the game.	Q1	AE
Affect	Playing the game made me happy.	Q11	AE
Sensory	I found the appearance of the game world to be interesting.	Q21	AE
Controls	I thought the controls were intuitive.	Q25	AE
Sensory	I concentrated on sounds in the game.	Q26	AE
Menu	I thought that the game's menus were intuitive.	Q34	AE
Variety	I felt that the game provided enough variety.	Q35	AE
Fictional	I found the game's story to be dull.	Q39	AE
Sensory	The aesthetics of the game were unimpressive.	Q43	AE
Ownership	The game failed to motivate me to keep playing.	Q45	AE
Fictional	I wanted to explore the game world.	Q47	AE
Affect	I thought that the game was fun.	Q5	AE
Affect	I found the game boring.	Q6	AE
Sensory	I liked how the game looked.	Q7	AE
Challenge	I found the game to be easy.	Q10	F
Fictional	I felt like events in the game were happening to me.	Q13	F
Ownership	It felt like I was responsible for what happened in the game.	Q17	F
Camera	Moving my point of view in the game was easy.	Q18	F
Absorption	I was unaware of the passage of time whilst playing.	Q20	F
Absorption	I forgot about my surroundings whilst playing.	Q23	F
Absorption	I was focused on the game.	Q3	F
Variety	I found the game mechanics to be varied enough.	Q38	F
Fictional	I could identify with the characters.	Q4	F
Ownership	I played by my own rules in the game.	Q40	F
Absorption	I thought about things other than the game whilst playing.	Q41	F
Camera	My field of view made it difficult to see what was happening in the game.	Q42	F
Camera	I thought the camera angles in the game were appropriate.	Q44	F
Challenge	I thought the level of difficulty was right for me.	Q48	F
Navigation	I always knew where to go in the game.	Q15	PB
Consistency	I knew how the game would respond to my actions.	Q27	PB
Goals	I always knew how to achieve my aim in the game.	Q28	PB
Consistency	I thought the game mechanics were consistent.	Q29	PB
Goals	I always knew how to achieve my aim in the game.	Q30	PB
Navigation	I lost my direction through the game.	Q32	PB
Goals	I knew when my goal in the game had changed.	Q33	PB
Navigation	I couldn't find my way in the game world.	Q37	PB
Controls	I knew how to use the controller with the game.	Q12	UB
Menus	I found the game's menus to be usable.	Q14	UB
Settings	I knew how to change the settings in the game.	Q16	UB
Help	The game would provide help at appropriate moments.	Q19	UB
Challenge	I felt the game was hard.	Q2	UB
Controls	I found the controls to be difficult.	Q22	UB

Table B.1. *Cluster Membership of Each Gameplay Scale Item* (Cont'd)

Construct	*Question*	*Question Number*	*Cluster*
Settings	I found using the options screen to be difficult.	Q24	UB
Help	The game provided me with an adequate tutorial.	Q31	UB
Menus	I found the game's menus to be cumbersome.	Q36	UB
Consistency	The game responded to my inputs in an inconsistent way.	Q46	UB
Settings	I knew how to customize the way that the game was set up.	Q49	UB
Help	The game trained me in all of the controls.	Q8	UB
Variety	I thought that the game was repetitive.	Q9	UB

("AE" = Affective Experience; "F" = Focus; "PB" = Playability Barriers; "UB" = Usability Barriers)

Appendix C

The Revised Gameplay Scale

Full List of Questions in the Revised Gameplay Scale, in Correct Order

Q1 I enjoyed the game.
Q2 I was focused on the game.
Q3 I could identify with the characters.
Q4 I thought that the game was fun.
Q5 The game trained me in all of the controls.
Q6 I thought the level of difficulty was right for me.
Q7 I found the game's menus to be usable.
Q8 I knew how to use the controller with the game.
Q9 I was unaware of the passage of time whilst playing.
Q10 I found the appearance of the game world to be interesting.
Q11 I knew how to change the settings in the game.
Q12 My objectives in the game were unclear.
Q13 I thought about things other than the game whilst playing.
Q14 I knew how the game would respond to my actions.
Q15 I couldn't find my way in the game world.
Q16 I always knew how to achieve my aim in the game.
Q17 I found the game's menus to be cumbersome.
Q18 I found the game mechanics to be varied enough.
Q19 I forgot about my surroundings whilst playing.
Q20 My field of view made it difficult to see what was happening in the game.
Q21 I found using the options screen to be difficult.
Q22 The aesthetics of the game were unimpressive.
Q23 I thought the camera angles in the game were appropriate.
Q24 The game failed to motivate me to keep playing.
Q25 I always knew where to go in the game.
Q26 I wanted to explore the game world.

Demographic Questions (asked at the very end):

- How old are you?

- What is your gender?

- Is English your first language?

- What is your favourite video game genre?

Appendix D

The Appeal Scale

1.) Example of appearance of response items in the Appeal Scale:

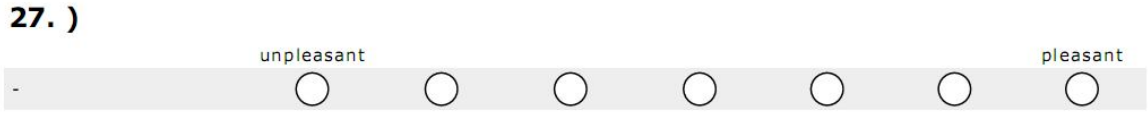

Figure D.1 Example response item from the Appeal Scale.

2.) Full list of word-pairs in the Appeal Scale, in correct order:

1. unpleasant - pleasant
2. bad - good
3. unaesthetic - aesthetic
4. rejecting - inviting
5. unattractive - attractive
6. discouraging - motivating
7. undesirable - desirable
8. boring – fun

(Appeal Scale always administered after the Gameplay Scale)

Appendix E

Information Sheet and Consent Form used in Study #2

Information Sheet for Participants in Research Studies

You will be given a copy of this information sheet.

Title of Project: **Playing with Scales: Creating a Measurement Scale to Assess Player Experience When Testing Video Games**

This study has been approved by the UCL Research Ethics Committee [Project ID Number]: **XXXXX**

Name, Address and Contact Details of Investigators:

Nadia Bianchi-Berthouze (Lecturer)
UCL Interaction Centre (UCLIC)
University College London
Malet Place Engineering Building, 8th floor
Gover Street
London WC1E 6BT, UK
Email: n.berthouze@ucl.ac.uk
Tel: +44 (0)20 7679 0690 (internal: X30690)

Mark Parnell (MSc Student)
Oakdene
Farthings Hill
Horsham
West Sussex
RH12 1TS
Email: mjparnell@gmail.com
Tel: 07792812843

We would like to invite you to participate in this research project. You should only participate if you want to; choosing not to take part will not disadvantage you in any way. Before you decide whether you want to take part, it is important for you to read the following information carefully and discuss it with others if you wish. Ask us if there is anything that is not clear or you would like more information.

The purpose of this study is to investigate player experience of videogames, using both a questionnaire and an interview. You will be asked to play a game for one hour whilst being videotaped playing. You will then be asked to complete a questionnaire, after which you will be given a short interview by the researcher, during which the video will be viewed and discussed. The experiment will then end. This study is being performed in conjunction with Sony Computer Entertainment Europe; please note that whilst they will have to access to the results of the study they will not have access to any private details for use in marketing etc.

It is up to you to decide whether or not to take part. If you choose not to participate it will involve no penalty or loss of benefits to which you are otherwise entitled. If you decide to take part you will be given this information sheet to keep and be asked to sign a consent form. If you decide to take part you are still free to withdraw at any time and without giving a reason.

All data will be collected and stored in accordance with the Data Protection Act 1998.

Informed Consent Form for Participants in Research Studies

Title of Project: **Playing with Scales: Creating a Measurement Scale to Assess Player Experience When Testing Video Games**

This study has been approved by the UCL Research Ethics Committee
[Project ID Number]: **XXXXX**

Participant's Statement

I .....

agree that I have

- read the information sheet and/or the project has been explained to me orally;
- had the opportunity to ask questions and discuss the study;
- received satisfactory answers to all my questions or have been advised of an individual to contact for answers to pertinent questions about the research and my rights as a participant and whom to contact in the event of a research-related injury.
- understood that my participation will be taped/video recorded and I am aware of and consent to, any use you intend to make of the recordings after the end of the project.
- read and understood the involvement of Sony Computer Entertainment Europe in this study.

I understand that I am free to withdraw from the study without penalty if I so wish and I consent to the processing of my personal information for the purposes of this study only and that it will not be used for any other purpose. I understand that such information will be treated as strictly confidential and handled in accordance with the provisions of the Data Protection Act 1998.

Signed: Date:

I .....

agree to the publishing of frames from my videos (in which my face will be blanked out) in academic publications

Yes ☐ No ☐

Investigator's Statement

I ……..

confirm that I have carefully explained the purpose of the study to the participant and outlined any reasonably foreseeable risks or benefits (where applicable).

Signed: Date: