The Edge of Glory: The Relationship between Metacritic Scores and Player Experience

Daniel Johnson
Queensland University of Technology
Brisbane, Queensland
dm.johnson@qut.edu.au

Christopher N. Watling
Queensland University of Technology
Brisbane, Queensland
christopher.watling@qut.edu.au

John Gardner
CSIRO Ecosystem Sciences
Brisbane, Queensland
john.gardner@csiro.au

Lennart E. Nacke
University of Ontario Institute of Technology, Canada
lennart.nacke@acm.org

ABSTRACT

This study sought to examine how measures of player experience used in videogame research relate to Metacritic Professional and User scores. In total, 573 participants completed an online survey, where they responded the Player Experience of Need Satisfaction (PENS) and the Game Experience Questionnaire (GEQ) in relation to their current favourite videogame. Correlations among the data indicate an overlap between the player experience constructs and the factors informing Metacritic scores. Additionally, differences emerged in the ways professionals and users appear to allocate game ratings. However, the data also provide clear evidence that Metacritic scores do not reflect the full complexity of player experience and may be misleading in some cases.

Author Keywords

Videogames; player experience; Metacritic; Psychology; Player Experience of Need Satisfaction; Game Experience Questionnaire

ACM Classification Keywords

J.4 [**Computer Applications**]: Sociology, Psychology – Social and Behavioural Sciences; K.8.0 [**Personal Computing**]: General – games;

INTRODUCTION

Review scores sell videogames. However, little is known about how much aggregated reviewer and user scores reflect the actual player experience. Videogame reviews are, for some consumers, an important source of information that can guide purchasing of videogame titles [6, 8]. In fact, research has shown that reviews (when read before playing a game for the first time) can actually influence player experience [20, 21]. In essence, videogame reviewers summarise and quantify their own subjective player experience. They provide a benchmark for consumers. This benchmark has a critical impact on video game sales and publisher deals in an increasingly competitive environment of videogame development. Score aggregators like the Metacritic website[1] provide an aggregation of a number of different reviews for major videogame releases. Videogame publishers have used Metacritic scores to adjust developers' monetary compensation tied to specific video game titles[2]. The relationship between sales and Metacritic scores is still debated. However, an important research question that arises from this debate is how representative Metacritic scores are of different facets of player experience.

Scientific studies increasingly recognize the potential impact (both positive and negative) of videogames on personal wellbeing [2, 10, 18, 37]. To study this impact, a growing number of psychometric scales or questionnaires for measuring player experience have been developed. These measures represent attempts to better understand the engagement and enjoyment associated with player experience as well as the factors that motivate people to play videogames.

A noteworthy distinction between psychometric measures of player experience and game reviews is that the former generally assume player experience is a multidimensional construct. This multidimensionality is important, because it allows for the possibility that a videogame might do some things well and other things poorly. Different players might enjoy different aspects of the same game. It is counterintuitive to reduce a videogame to a single score as a measure of quality and success. However, this is common practice when a videogame is reviewed and it is also a result of Metacritic's review aggregation process. This inevitably hides some of the complexities of player experience.

We can assume that the quality of a reviewer's player experience directly informs review scores. However, to date, little is known about which facets of player experience influence how a game is judged in a review.

CHI PLAY '14, October 19 - 22 2014, Toronto, ON, Canada
Copyright 2014 ACM 978-1-4503-3014-5/14/10…$15.00.
http://dx.doi.org/10.1145/2658537.2658694

[1] http://www.metacritic.com

[2] VentureBeat: http://venturebeat.com/2009/08/27/the-influence-of-metacritic-on-game-sales/

Validated psychometric scales are not necessarily designed to measure the same components of player experience as review scores. However, players use reviews and review scores to choose which games they are going to play and hence any overlap between review scores and the player experience is of interest. This study explores how two, commonly used, player experience scales relate to aggregate game review scores (from both professional reviewers and consumers). Our study explicitly aims to identify which components of player experience relate to game review scores.

Quantifying Player Experience

As a result of the increasing popularity of studying videogames in experimental settings, many questionnaires have emerged, designed to measure the general engagement and enjoyment associated with playing videogames [5, 7, 9, 13, 14, 32, 39] and the extent to which specific states or experiences occur – for example, flow [38]. The Player Experience of Need Satisfaction [32] and the Game Experience Questionnaire [13] were chosen for use in the current study. We selected them on the basis that they offer multiple subscales designed to assess different components of player experience and they have been widely used in previous research (as detailed below).

The Player Experience of Need Satisfaction Questionnaire

Ryan et al. [32] applied an established psychological theory of motivation – Self-Determination Theory (SDT) – to videogame player motivations. SDT is primarily concerned with the potential of social contexts to provide experiences that satisfy universal needs of people. SDT has been successfully applied in research on sports, education and leisure domains. Przybylski and colleagues explored how videogames fulfill or thwart psychological needs and thus promote or discourage sustained engagement and either positive or negative outcomes for players. Based on SDT and other relevant theories (e.g., Presence), Przybylski and colleagues developed the Player Experience of Need Satisfaction (PENS) measure. This questionnaire assesses player experience in the dimensions: Competence (perceived efficacy playing the game), Autonomy (sense of freedom and independence), Relatedness (connectedness within the game), Intuitive Controls (perception of the in-game controls) and Presence/Immersion (sense of being in the game world, as opposed to experiencing oneself as a person outside the game, manipulating controls or characters).

The initial validation of the PENS was performed in experimental contexts with participants, who may or may not have been experienced video game players and in a non-experimental context with a sample of MMO players [32]. The PENS has been used successfully in different settings and with many videogames [15, 22, 27-29, 32]. The five factors measured by the PENS have been shown to be associated with aspects of videogame experiences, including: videogame sales, game series loyalty, duration of videogame interest, positive affect and enjoyment [3, 30, 35]. As part of their initial validation Ryan and colleagues showed that the PENS accurately distinguished between a game that had received very positive reviews and a game that had received largely negative reviews.

The Game Experience Questionnaire

The Game Experience Questionnaire [GEQ: 13] is designed to provide a comprehensive evaluation of the gameplay experience. Unlike the PENS, the GEQ structure is not based around a specific theory. Rather, the GEQ is based on conceptual accounts of player experiences and focus-group explorations with a range of gamers. The GEQ is comprised by seven factors, which are: Positive Affect (experiencing positive emotions during gameplay), Negative Affect (experiencing negative emotions during gameplay), Frustration (irritation from negative experiences of gameplay), Flow (holistic sensation of acting within the confines of the game), Challenge (feelings of being tested within the gameplay experience), Immersion (perception of being absorbed in the game environment), and Competence (perceived efficacy playing the game).

Concerns have been raised regarding the GEQ's psychometric properties because the preliminary validation work (leading to the creation of the scale) has never been published [see 26]. Nevertheless, the GEQ has been applied in many game research studies. The GEQ has been used in psychophysiological studies of player experience [16, 24, 25], studies of social experiences of videogaming [11], studies that experimentally explored how player experiences vary across genres [17] and the designing of videogames for unique user populations [1].

Metacritic Scores

The Metacritic website typically produces a Metacritic *Professional Score* and a Metacritic *User Score* for a videogame. Metacritic professional scores are calculated by collating a number of professional reviews (from various online videogame review sources), converting the associated individual review scores into a score out of 100 (e.g., 4 out of 5 stars becomes a score of 80) and then weighting the scores to produce a final aggregated Metacritic professional review score. It is stated that weighting is given to critics and publications as function of their quality and overall stature (with greater weighting given to those outlets judged to be higher quality [23]). Unfortunately, the weighting that is applied to the collated reviews is not available to the public. In contrast to professional Metacritic scores, Metacritic user scores are simply the unweighted average of all ratings provided by visitors to the Metacritic website. User scores are provided anonymously (or pseudonymously) with reviewers identifiably only by a self-chosen username.

The use of the Metacritic Professional Scores varies, but it is substantial in some cases. For instance, it has been reported that contracts between game developers and publishers include clauses that tie key deliverables and outcomes to Metacritic scores [8, 33]. Metacritic scores that fall short of pre-determined benchmarks can lead to bonuses or royalties not being paid to a game's developer by the game's publisher [36]. Additionally, a game developer's track record of Metacritic scores can be used by game publishers to negotiate reduced royalties and expenditure for the developer [33]. Moreover poor Metacritic scores can lead to a drop in a game development company's stock price [8]. In these ways, Metacritic scores can be a substantial financial influence on game development.

Metacritic scores have become so important to the gaming industry that, in some cases, it has been suggested that game developers are seeking to influence who publicly reviews their games and thus which professional game reviewers can contribute to Metacritic scores [33]. This means that game publishers and developers can hire professional game reviewers to review their unreleased games to provide feedback. Games user research studios sometimes – but rarely – use professional reviewers for early user testing. This feedback is then used to improve the game prior to its public release. However, once the professional game reviewer has been paid to provide feedback on a game, they cannot ethically provide a consumer review of that game. This conveniently excludes them from contributing to the overall Metacritic professional score for that particular game [33].

The utility of professional Metacritic scores is equivocal for benchmarking the videogame experience. Research has shown that non-player characters from videogames with higher professional Metacritic scores are more believable and enjoyable than those from videogames with lower Metacritic scores [19]. Additionally, Metacritic professional scores have been shown to be strongly correlated with videogame sales [12]. While player experience should arguably be the primary driver of both review scores and sales, Metacritic scores should not drive sales independent of game quality. However, it has been suggested that a Metacritic Professional Score, particularly scores over 90 out of 100, can become the primary driver of videogames sales, [4, 8, 40] rather than the influence of the actual quality of player experience.

The reliability of the Metacritic professional scores is also ambiguous. For instance, a reviewer's scores of a videogame sequel can be influenced by both the commercial success and player experience of the prequel videogame [34]. In such cases, experiencing a videogame brand (e.g., a game franchise with many sequels) seems to strongly influence the attitudes brought to experiencing an individual game. Professional videogame reviewers are likely to have a systematic approach for reviewing videogames. Nevertheless, the subjective nature and the potential for bias for or against a videogame's genre, platform, producer, and series [31] are likely to affect the reliability of the Metacritic aggregation. Moreover, the undisclosed procedure by which reviewers' scores are weighted to create the professional Metacritic score could also impact the reliability and validity of the aggregated professional review scores.

With respect to Metacritic user scores, other issues arise that potentially influence validity and reliability. It is possible that those who feel strongly about a game (either positively or negatively) are more likely to take the time to rate it than those who feel less strongly. Thus, it may be that Metacritic user scores disproportionately sample extremely negative and extremely positive opinions about a particular game. Relatedly, there is evidence – in terms of the written reviews submitted by users on the Metacritic website – that negative reviews for some games are written in an attempt to make other games appear relatively more popular in comparison. For instance, there exists rivalry between fans of the Call of Duty and Battlefield franchises, and fans of either series can be seen to post negative reviews for the other series. Regardless of these potential bias sources, it is not clear how Metacritic user scores relate to player experience (as measured by psychometrics) or which facets of player experience are most strongly associated with user scores.

The Current Study

Professional videogame reviewers are an established and influential source of information for consumers. However, to the best of our knowledge, no studies have yet addressed the relationship between videogame reviewer scores and gameplay experience measures. Therefore, we wanted to study how commonly used measures of player experience (i.e., PENS and GEQ) relate to the aggregated review scores of Metacritic. Metacritic scores were chosen over any individual source of reviews as a means of avoiding any idiosyncrasies associated with specific review websites. By examining the relationships between the PENS and GEQ and Metacritic scores, we can identify player experience facets that are most important to professional videogame critics and to Metacritic website users, who review and rate games. Therefore, we pose the following research questions:

RQ1: Which subscales of the PENS and GEQ are associated with the Metacritic Professional scores?

RQ2: Which subscales of the PENS and GEQ are associated with the Metacritic User scores?

METHOD

Participants

For this study, we selected only participants that had an interest in videogames and that played videogames at the time of the study. No other selection criteria were applied for participation in the study. The research protocol received ethical as well as health and safety approval from the host University.

In total, 573 participants took part in the study. The average age of the participants was 20.7 years (*SD* = 5.1; range = 13-54), with the majority of participants being male (81.7%). On average, the participants played videogames 16.6 hours per week (*SD* = 12.5; range = 1-100 hours) and played their favourite game for an average of 9.5 hours per week (*SD* = 9.6; range = 1-100 hours).

An extensive list was generated from the participants' favourite game titles; over 200 different titles were listed. For the sake of brevity, the game titles will not be listed but rather the proportions of game genres that were represented. The most common genre of participants' current favourite game was first-person shooter games (24.9%), followed by action role-playing games (13.6%), action adventure games (11.9%), role-playing games (9.6%), massively multiplayer online role-playing games (8.7%), multiplayer online battle arenas (5.8%), real-time strategy (5.6%), and other various games genres (19.9%).

The measures of player experience used in the current study were not provided by the same sample that provided the Metacritic scores (i.e., the professional or user scores). However, we can expect a relationship between the two measures of player experience on the basis that the same games were being analysed with each measure. The qualities of each game exist independent of who is playing them. This is our core assumption.

Measures

Player Experience of Needs Satisfaction Questionnaire

The PENS questionnaire [PENS: 32] is a self-report measure of an individual's experiences while playing a videogame. Participants indicate their agreement with 21 items on a seven-point Likert scale (1 – "do not agree" to 7 – "strongly agree"). The PENS measures different aspects of player experience with five subscales: Competence, Autonomy, Relatedness, Presence and Intuitive Controls (as described above). Higher scores on each scale indicate greater agreement. The PENS subscales have been found to have good reliability, as determined by Cronbach's alpha. The initial examination study of the utility of the PENS, reported a Cronbach's alpha statistic between 0.7–0.8 [32]. Subsequent work has reported Cronbach's alpha statistics between 0.7–0.9 for multiplayer videogaming sessions [16]. Based on these previous studies, we consider the PENS a reliable instrument.

Game Experience Questionnaire

The GEQ [GEQ: 13] is another self-report measure of an individual's experience during gameplay. Participants indicate their agreement with 33 items on a five-point Likert scale (1 – "not at all" to 7 – "extremely"). The GEQ is designed to measure seven facets of player experience: Positive Affect, Negative Affect, Frustration, Flow, Challenge, Immersion, and Competence (as described above). Higher scores on a particular subscale indicate greater experience of that facet during gameplay.

As discussed, the scale's authors have not published any formal studies assessing the structure and performance of the GEQ. On that basis, exploratory factor analysis was undertaken to examine the performance of the GEQ in the current sample. The full 33 items were subject to exploratory factor analysis via principal axis factoring, using oblique rotation. Initial analyses suggested the existence of 5 or 6 factors, but there were split loadings in both solutions, and the hypothesized scale structure did not clearly emerge. Items with no loading higher than .4, and items with loadings of higher than .3 on two or more factors, were dropped from the analysis. In total, seven items were dropped, and a final 6-factor solution (which explained 50.4% of the variance) was chosen as best reflecting the underlying structure. In contrast to the original factor structure, negative affect and tension/annoyance items were found to load on a single factor, which was renamed *Frustration*. Another item was dropped because it lowered the associated scale reliability. The GEQ competence subscale was not used in the current study given its conceptual overlap with the PENS measure of the same construct.

Procedure

The majority of participants (81.7%) were drawn from a first-year university videogame study course. A snowball procedure was used to gather additional participants via gaming forums, social media web pages and through personal contacts of the researchers. Participants completed the study questionnaires online. There was no way to control for the time that had elapsed since participants had last played videogames, because participants completed the survey at a time of their own choosing. To address this issue, a guided recall process was used to prime respondents before they answered questions about their gameplay experiences. Respondents were asked to recall and describe in detail what was happening when they were most recently playing their current favourite game. Participants then responded to the PENS and GEQ with regard to their experience of playing that game. Participants were offered the chance to win a $100 voucher in return for their participation. For each game nominated by participants, Metacritic Professional and User scores were obtained from the Metacritic website (CBS Interactive Inc, 2014). These variables were used as the dependent variables for the study.

Statistical Analyses

The internal consistency of the scale scores was evaluated with Cronbach's alpha coefficient. The variables of Metacritic Professional Score, Metacritic User Score, and "Hours of Playing Favourite Game" had non-normal distributions. To account for this non-normality, Spearman's rho correlations were used to examine the bivariate associations between study variables.

RESULTS

Descriptive Statistics

The means, standard deviations, and – where applicable – Cronbach's alphas for the study's main variables are displayed in Table 1 (below). The participants nominated favourite games tended to be rated highly by others, as demonstrated by high mean values for both the Metacritic Professional and User Scores.

Study variable	Mean	SD	α	Actual Range
Metacritic Pro Score	86.84	6.95	-	55.00-99.00
Metacritic User Score	7.38	1.43	-	2.40-9.40
Hours Play Fav Game	9.50	9.62	-	1-100.00
PENS Competence	5.80	0.94	.71	2.00-7.00
PENS Autonomy	5.52	1.10	.68	1.00-7.00
PENS Relatedness	3.95	1.49	.71	1.00-7.00
PENS Presence	4.30	1.31	.87	1.11-7.00
PENS Int Controls	5.86	0.90	.57	2.00-7.00
GEQ Positive Affect	4.25	0.56	.84	2.40-5.00
GEQ Frustration	1.48	0.52	.85	1.50-5.00
GEQ Flow	3.50	0.90	.81	1.00-5.00
GEQ Challenge	3.37	0.82	.66	1.00-5.00
GEQ Immersion	3.57	0.92	.72	1.00-5.00

Table 1.Means, standard deviations (SD), Cronbach's alphas (α), and actual range of the study variables.

Correlations

Table 2 (below) displays the Spearman correlation coefficients for the Professional and User Metacritic scores, Hours Playing Favourite Game and the PENS and GEQ subscales. Many significant correlations were found between the study variables. Clear differences between what professionals and users are responding to showed in their ratings of games. Regarding the correlations with the Metacritic Professional Scores, all of the PENS subscales had significant and positive correlations with the Metacritic Professional Scores. The two largest correlations were between the Metacritic Professional scores and Intuitive Controls and Autonomy. Somewhat smaller correlations were found between the Metacritic professional scores and the PENS measures of competence, autonomy, relatedness and presence. The GEQ subscales were less aligned with the Metacritic Professional scores. Specifically, only three of the GEQ subscales (Positive Affect, Immersion and Competence) were correlated with the Metacritic Professional Scores.

Study variable	Pro Score	User Score
Metacritic Pro Score	-	
Metacritic User Score	$.26^{**}$	-
Hours Play Fav Game	-.03	$-.13^{**}$
PENS Competence	$.18^{**}$	$.08^{*}$
PENS Autonomy	$.22^{**}$	$.11^{*}$
PENS Relatedness	$.16^{**}$	.05
PENS Presence	$.20^{**}$	$.13^{**}$
PENS Int Controls	$.25^{**}$	.07
GEQ Positive Affect	$.15^{**}$	$.18^{*}$
GEQ Frustration	-.05	.01
GEQ Flow	.02	.05
GEQ Challenge	-.06	$.09^{*}$
GEQ Immersion	$.20^{**}$	$.12^{**}$

$^{**}p < .01$, $^{*}p < .05$

Table 2. Spearman rho correlation coefficients for the Metacritic scores, Hours Playing Favourite Game, and subscales of the PENS and GEQ

Slightly fewer significant correlations were found between the Metacritic User Scores and the measures of player experience. The PENS subscales of Competence, Autonomy and Presence were significantly associated with the Metacritic Users scores. Regarding the correlations between Metacritic Users scores and the subscales from the GEQ, only three significant correlations were found. These were for Positive Affect, Challenge and Immersion. Interestingly, the variable "Hours of Playing Favourite Game" was negatively correlated with the Metacritic Users score (that is, lower Metacritic User scores were associated with higher hours of play and high Metacritic User scores were associated with lower hours of play).

DISCUSSION

Overall Patterns

Overall, it can be seen that both professional and user Metacritic scores relate to player experience facets that are measured by the PENS and the GEQ. This generally indicates there is overlap between player experience constructs being assessed in videogame research and the videogame aspects to which professional critics respond. Professional Metacritic scores (and to a lesser extent, player Metacritic scores) are more strongly associated with the components of the PENS than the GEQ. All five components of the PENS were positively associated with professional Metacritic scores. This means that higher Metacritic scores are indicative of higher PENS scores. However, only two of the five components of the modified, more reliable, version of the GEQ used were associated with Metacritic Professional scores: Positive Affect and Immersion.

These results could reflect the relative importance of the satisfaction of self-determination theory related needs in creating a rewarding player experience. At the very least, the constructs measured by the PENS measure seem to be of high importance for people (reviewers and users) that rank videogames on Metacritic. It may also be that the PENS is a relatively stronger measure in terms of psychometric properties, having been formally validated and refined in various settings.

With the exception of Positive Affect, all the measured components of player experience show a stronger association with Metacritic Professional scores than with Metacritic User scores. This pattern may reflect a greater ability on the part of professional critics to identify how well a game will satisfy player needs and provide a pleasing experience. Professional reviewers arguably have developed a game literacy through playing many games that allows them to accurately judge aspects of player experience that are important for most people when playing a game. User generated scores, in contrast, might reflect more idiosyncratic opinions, which relate less well to typical player experiences. There might be a difference between the formalized idea of player experience developed in academia and the subjectively voiced player experience of Metacritic users. It might also point to a bias in reporting experience. The number that Metacritic users give in their rating might reflect an opinion or an attitude toward a game rather than a marker of the quality of player experience. What players experience and what they report to have experienced might not be the same thing. However, a quantitative analysis of user game ratings is limited here because it distils their possibly complex feelings about gameplay to a single numeric quality indicator. However, this pattern provides an important starting point for a necessary discussion of the relationship between game ratings and player experience.

Furthermore, it is interesting to note that only a moderate correlation exists between Metacritic professional and user scores. This supports the idea that professional critics and everyday players view the same games in different ways and perhaps respond to different things when rating a game. It may be that professionals are more objective than players, or perhaps, more critical. Alternatively, it may be that professional critics are required to provide a more superficial rating of some games, because of constraints in available playing time. This would not apply in the case of games that primarily consist of single-player campaigns, because most reviewers play such games in full before publishing a final review. Generally, both the required professionalism and available playing time might drive professional reviewers to adopt a more formalized reviewing approach. It may also be that the low correlation is partially a function of the aforementioned issues with user scores related to bias (e.g., scoring less favoured games with extremely negative ratings). Since we looked at aggregate Metacritic scores, outliers could be having a strong influence on these results. Without knowing the score distribution for each game, we cannot know for sure how much influence these factors might be having.

Strength of Associations

The strongest relationships for Metacritic Professional scores were found with Intuitive Controls, Autonomy, Presence and Immersion. The strongest relationships for Metacritic User scores were found with Positive Affect, Presence, Immersion and Autonomy. Thus, some clear differences and similarities exist between what professionals and users are responding to as components of player experience measured in the current study. It seems that the experience of positive affect is similarly influential to both professionals and players. This is not surprising – enjoyment and associated positive emotions are a universal component of videogame play. Similarly, Immersion and Presence are related to the scores given by both professional reviewers and users. It seems likely that these are universally positively regarded components of player experience. For both Presence and Immersion, the relationship is stronger for professional reviewers. It may be that professionals place greater importance on these components of player experience.

Competence and Autonomy are more strongly related with Metacritic Professional scores than Metacritic User scores. These stronger associations with professional scores (Presence, Immersion, Competence and Autonomy) may reflect that professional critics are more clinical or objective than regular players – focussing on these formal aspects of the game, while regular players are relatively more influenced by their emotional response to the game (as shown in the relationship between Metacritic User scores and Positive Affect).

Differences between Professional and User Scores

Some key differences emerged in terms of elements of the player experience associated with one Metacritic score but not the other. Intuitive Controls and Relatedness are related to Metacritic Professional scores but not to Metacritic user scores. With respect to Intuitive Controls, a relatively large relationship with Metacritic Professional scores is evident. This most likely reflects the fact that the understandability and ease of controls in a game is a key issue for a reviewer aiming to provide a rating that is useful and informative for consumers with varying skills levels and prior experience. Given their personal game literacy (built through playing many different games) with genre-based controls, professional reviewers might also be less forgiving when a game does not handle player controls well. Alternatively, if we assume that players spend longer playing a game than professional reviewers, then it may be that players have more time to become familiar with controls even if they are not intuitive. However, it should be acknowledged that the reliability for the Intuitive Controls subscale was relatively low in the current study and this result should thus be interpreted with some caution.

With respect to Relatedness, professional reviewers may place greater importance on games that are enjoyable when played with others, because their focus (when reviewing a game) must in some ways take place with reference to all other games. As a result, a game that offers greater connection with others can be seen as deserving of a higher score. In contrast, a user might not be concerned with this aspect of player experience, as some games (and arguably genres of games) do not attempt to provide relatable characters or to create connections between players. Players, who favour such games, may disregard relatedness when judging a game they like for other reasons.

Challenge was only related to Metacritic User scores. This may reflect the fact that players are more influenced by whether the game offers them a suitable level of challenge. Professional reviewers, on the other hand, are aiming to give a review that is relevant to users of varying skill levels, so the degree of challenge offered is less likely to be directly influential. Alternatively, it may be that greater challenge in a game sometimes interferes with the process of reviewing it in a timely fashion, resulting in reviewers adjusting game difficultly (and thus challenge) to a level that allows them to complete the game quickly. In contrast, for user, the game is a purely recreational task wherein challenge is directly related to the quality of player experience. A challenging game could also be likely to influence player attitudes more strongly, because players have to invest more time into more challenging games. The more time invested could lead to stronger game attachment and a greater likelihood of rating the game on a website like Metacritic. Separately, it may be that professional game reviewers are so experienced with and proficient at playing games that few games provide high levels of challenge, and hence the notion of challenge does not directly inform their rating of a game.

The lack of a relationship between Flow and Challenge with Metacritic Professional scores may reflect a "ceiling effect" in that these components of player experience are "givens" in popular games. Hence, they do not inform the score unless the game is particularly poor (unlikely in a sample made of players' current favourite games). However, this interpretation is not well supported by the relatively low Flow scores found in the sample (see Table 1). Similarly, the lack of relationship between Metacritic scores and frustration may be a "floor effect" in that a low level of frustration is commonly required for a game to become a players' favourite game.

It is interesting to note how "hours of play" relates to the Metacritic scores. The lack of relationship between Metacritic Professional scores and hours of play suggests that professional review scores do not provide an indication of how long people play a particular videogame. It is likely that variations associated with different genres (e.g., MMORPGs are structurally longer games than platformers or action adventure games) hide any connection between review scores and play time. The negative relationship between Metacritic User scores and hours of play is at first glance, counter-intuitive. However, it may be that shorter games are easier to make in a way that appeals to players as (other things being equal) greater development time and energy can be placed into the production of all aspects a shorter game than a longer game. Alternatively, it may be that players of longer games are more critical of them because of the relatively greater amount of time they have invested in playing them.

Limitations and Future Research

Overall, the effect sizes associated with the relationships identified in the present study are relatively small. This suggests that the factors identified by researchers as being key components of player experience are having only a small impact on review scores. It may also be that the influence of these factors is weakened as part of the weighting process of professional scores undertaken on the Metacritic website. For example, if specific critics treated as less influential (and therefore apportioned relatively less weight as part of the overall professional Metacritic score) are more influenced by these factors than critics considered more influential then the strength of the relationship would be partially hidden. This might be the case if more influential critics are, for example, more jaded, more cynical or more subjectively critical than less influential or less experienced critics. However, this explanation is not relevant to the relatively small effect sizes between player experience factors and Metacritic User scores. In this case, the effect sizes are relatively small most likely because of the previously discussed noise or bias associated with user scores (resulting from extreme opinions being more likely to motivate providing a review score and some users posting low scores for games that are perceived to compete with their own preferred games).

However, for both Metacritic Professional and User scores there are three likely explanations for the relatively small effect sizes: Firstly, the sample of people, who completed our player experience measures (PENS, GEQ) is not necessarily the same sample of people, who provided ratings towards the Metacritic scores. The strength of the relationship is inevitably reduced by this difference. Secondly, a large variety of factors can potentially influence the rating of a game. It is probable that different people respond to games in different ways – factors of personal taste, preferences for or against different games types based on prior experiences and personality differences among players and reviewers are all likely to contribute meaningfully to how each individual person responds to any particular game. As a result, the player experience factors measured in the current study – while directly relevant to the experience of play – and associated ratings of a game cannot account for the full range of variability in an individual's player experience. Future research should seek to incorporate a larger range of relevant measures of the player experience to better explore this possibility. An option could be the textual analysis and exploration of the content of user reviews on Metacritic. Finally, it should be noted that the full range of values on the measurement scales was not included in the current study. For example, games with a professional Metacritic score below 55 were not assessed (this is a function of participants evaluating their current favourite game). This reduced range of the scales may have caused the correlations we reported above to underestimate the true relationships. Future research should extend to games that are relatively less popular.

It should also be acknowledged that the player experience metrics used in the current study are far from perfect and the small effect sizes may also relate to noise and variability introduced through the use of these metrics. In particular, the psychometric properties of the GEQ are difficult to judge because of the lack of factor structure validation studies published to date. Future research could attempt to explore these relationships using other measures of player experience or incorporate larger samples, which would allow for more extensive assessment of the validity and structure of the component measures.

Conclusion

Overall, the findings in the current study suggest that there is a degree of overlap and commonality between measures of player experience used by researchers and Metacritic scores. However, Metacritic Professional scores, though potentially useful in some situations (e.g., getting a sense of the majority critical opinion of a game or distinguishing particularly well received games from particularly poorly received games) do not reflect the full complexity of the experience of playing a game. Arguably, Metacritic scores may be being given too much weight in some situations such as negotiations between developers and publishers or purchasing decisions being made by players. Our findings support the notion that Metacritic scores may be influencing videogame sales irrespective of game quality, because people basing their purchasing decisions on Metacritic scores may inadvertently miss out on games they would greatly enjoy. This is particularly likely if players respond strongly to specific aspects of player experience (e.g., feelings of relatedness or the experience of challenge) that are not necessarily fully reflected in Metacritic scores.

REFERENCES

1. Al Mahmud, A., et al. Designing and evaluating the tabletop game experience for senior citizens. In *Proceedings of the 5th Nordic conference on Human-computer interaction: building bridges* 2008, ACM press (2008), 403-406.

2. Anderson, C.A., et al., Violent video game effects on aggression, empathy, and prosocial behavior in eastern and western countries: a meta-analytic review. *Psychological bulletin 136*, 2 (2010), 151-173.

3. Birk, M. and R.L. Mandryk. Control your game-self: effects of controller type on enjoyment, motivation, and personality in game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* 2013, ACM press (2013), 685-694.

4. Bower, B., *Valuing a Video Game: Does Score Determine Value?* 2009.

5. Brockmyer, J.H., et al., The development of the Game Engagement Questionnaire: A measure of engagement in video game-playing. *Journal of Experimental Social Psychology 45*, 4 (2009), 624-634.

6. Cox, J., What Makes a Blockbuster Video Game? An Empirical Analysis of US Sales Data. *Managerial and Decision Economics 35*, 3 (2013), 189-198.

7. Dauphin, B. and G. Heller, Going to other worlds: The relationships between videogaming, psychological absorption, and daydreaming styles. *Cyberpsychology, Behavior, and Social Networking 13*, 2 (2010), 169-172.

8. Everiss, B. Metacritic has changed the games industry. 2008, Available from: http://www.bruceongames.com/2008/06/04/metacritic-has-changed-the-games-industry/.

9. Fang, X., et al., Development of an instrument to measure enjoyment of computer game play. *Intl. Journal of Human–Computer Interaction 26*, 9 (2010), 868-886.

10. Ferguson, C.J., The good, the bad and the ugly: A meta-analytic review of positive and negative effects of violent video games. *Psychiatric Quarterly 78*, 4 (2007), 309-316.

11. Gajadhar, B., Y. De Kort, and W. Ijsselsteijn. Shared fun is doubled fun: Player enjoyment as a function

of social setting. In *Fun and Games* 2008 (2008), 106-117.

12. Greenwood-Ericksen, A., S.R. Poorman, and R. Papp, On the Validity of Metacritic in Assessing Game Value. *Eludamos. Journal for Computer Game Culture 7*, 1 (2013), 101-127.

13. IJsselsteijn, W.A., K. Poels, and Y.A.W. de Kort, *The Game Experience Questionnaire: Development of a self-report measure to assess player experiences of digital games*. Eindhoven University of Technology: Eindhoven. 2008.

14. Jennett, C., et al., Measuring and defining the experience of immersion in games. *International Journal of Human-Computer Studies 66*, 9 (2008), 641-661.

15. Johnson, D. and J. Gardner. Personality, motivation and video games. In *Proceedings of the 22nd Conference of the Computer-Human Interaction Special Interest Group of Australia on Computer-Human Interaction* 2010, ACM press (2010), 276-279.

16. Johnson, D., et al., Cooperative Play with Avatars and Agents: Differences in Brain Activity and the Experience of Play. *Computers in Human Behavior under review*, (2014).

17. Johnson, D., et al. Personality, genre and videogame play experience. In *Proceedings of the 4th International Conference on Fun and Games* 2012, ACM press (2012), 117-120.

18. Jones, C.M., et al., Gaming well: links between videogames and flourishing mental health. *Frontiers in psychology 5*, (2014), 1-8.

19. Lee, M.S. and C. Heeter, What do you mean by believable characters?: The effect of character rating and hostility on the perception of character believability. *Journal of Gaming & Virtual Worlds 4*, 1 (2012), 81-97.

20. Livingston, I., L. Nacke, and R. Mandryk, *Influencing Experience: The Effects of Reading Game Reviews on Player Experience*, in *Entertainment Computing – ICEC 2011*, J. Anacleto, et al., Editors. Springer Berlin Heidelberg. p. 89-100, 2011.

21. Livingston, I.J., L.E. Nacke, and R.L. Mandryk. The impact of negative game reviews and user comments on player experience. In *ACM SIGGRAPH 2011 Game Papers* 2011, ACM press (2011).

22. McEwan, M., et al. Videogame control device impact on the play experience. In *Proceedings of The 8th Australasian Conference on Interactive Entertainment: Playing the System* 2012, ACM press (2012), 1-3.

23. Metacritic. How we create the metascore magic. 2014, Available from: http://www.metacritic.com/about-metascores.

24. Nacke, L. and C.A. Lindley. Flow and immersion in first-person shooters: measuring the player's gameplay experience. In *Proceedings of the 2008 Conference on Future Play: Research, Play, Share* 2008, ACM press (2008), 81-88.

25. Nacke, L.E., M.N. Grimshaw, and C.A. Lindley, More than a feeling: Measurement of sonic user experience and psychophysiology in a first-person shooter game. *Interacting with Computers 22*, 5 (2010), 336-343.

26. Norman, K.L., GEQ (Game Engagement/Experience Questionnaire): A Review of Two Papers. *Interacting with Computers 25*, 4 (2013), 278-283.

27. Przybylski, A.K., C.S. Rigby, and R.M. Ryan, A motivational model of video game engagement. *Review of General Psychology 14*, 2 (2010), 154-166.

28. Przybylski, A.K., R.M. Ryan, and C.S. Rigby, The motivating role of violence in video games. *Personality and Social Psychology Bulletin 35*, 2 (2009), 243-259.

29. Przybylski, A.K., et al., The ideal self at play: the appeal of video games that let you be all you can be. *Psychological Science 23*, 1 (2012), 69-76.

30. Rigby, C.S. and R.M. Ryan, *The Player Experience of Need Satisfaction (PENS) Model*. Immersyve Inc. 2007.

31. Rollings, A. and E. Adams, *Andrew Rollings and Ernest Adams on game design*. New Riders. 2003.

32. Ryan, R.M., C.S. Rigby, and A.K. Przybylski, The Motivational Pull of Video Games: A Self-Determination Theory Approach. *Motivation and Emotion 30*, 4 (2006), 344-360.

33. Schreier, J. Metacritic Matters: How Review Scores Hurt Video Games. 2013, Available from: http://kotaku.com/metacritic-matters-how-review-scores-hurt-video-games-472462218.

34. Situmeang, F.B.I., M.A.A.M. Leenders, and N.M. Wijnberg, History matters: The impact of reviews and sales of earlier versions of a product on consumer and expert reviews of new editions. *European Management Journal 32*, 1 (2014), 73-83.

35. Tamborini, R., et al., Defining media enjoyment as the satisfaction of intrinsic needs. *Journal of Communication 60*, 4 (2010), 758-777.

36. Totilo, S. Low Metacritic scores cause game publishers to withhold developer royalties. 2008, Available from: http://multiplayerblog.mtv.com/2008/05/29/low-metacritic-costs-developers/.

37. Vella, K., D. Johnson, and L. Hides. Positively playful: when videogames lead to player wellbeing. In *First International Conference on Gameful Design, Research and Applications* 2013, ACM press (2013).

38. Wang, L.C. and M.P. Chen, The effects of game strategy and preference-matching on flow experience and programming performance in game-based learning. *Innovations in Education and Teaching International 47*, 1 (2010), 39-52.

39. Wiebe, E.N., et al., Measuring engagement in video game-based environments: Investigation of the User Engagement Scale. *Computers in Human Behavior 32*, (2014), 123-132.

40. Wingfield, N. *High scores matter to game makers, too.* The Wall Street Journal, 2007.