A Postmortem on Playtesting: Exploring the Impact of Playtesting on the Critical Reception of Video Games

Pejman Mirza-Babaei
Ontario Tech University
pejman@uoit.ca

Samantha Stahlke
Ontario Tech University
samantha.stahlke@uoit.ca

Günter Wallner
Eindhoven University of Technology
g.wallner@tue.nl

Atiya Nova
Ontario Tech University
atiya.nova@uoit.net

ABSTRACT

Game studios aim to develop titles that deliver a fun and engaging experience for players. Playtesting promises to help identify opportunities to improve player experience and assist developers in achieving their design intent. However, a lack of research on the added value of playtesting means that many studios are still uncertain about its commercial viability and impact on product success. This gap in understanding is further complicated by the vague definition of "success" afforded by sales figures and review scores. In this paper, we assess reported feature quality of three commercial titles by analyzing playtesting reports and game reviews. By comparing themes and design issues expressed in game reviews to the results of pre-release playtesting for each game, we aim to highlight the value of playtesting and propose a set of guidelines for selecting playtest methods based on the needs of a given game evaluation. Through the real-world case studies presented, this paper contributes to the growing domain of games user research and highlights the value of playtesting in game development.

Author Keywords

Games user research; playtesting; game reviews; guidelines.

CCS Concepts

•Human-centered computing → User studies; •Applied computing → Computer games;

INTRODUCTION

Playtesting is increasingly accepted as a standard part of the game development cycle. However, playtesting, as a term, is not always used consistently and can be interpreted differently depending on the development context. In this paper, we will use the term to refer to formative evaluation conducted by external researchers during post-production, before final release, in order to identify adjustments that bring the game closer to the developer's intent. This approach is commonly employed in Games User Research (GUR) to improve game usability and user experience. As a field, GUR builds upon psychology and Human-Computer Interaction (HCI), aiming to help game developers achieve their design intent by performing a series of user evaluations to better understand players and their experience [6]. Many major game studios and publishers have in-house GUR facilities and staff. There are also third-party GUR specialists, which are contracted by a variety of developers of all sizes. However, while GUR is gaining traction, many developers are still hesitant to adapt GUR approaches during development.

While there exist many articles (e.g., [2, 22, 31]) on the adoption of playtesting and other GUR evaluation techniques in game development – which have provided an undeniable advancement of the field [35] – there is a lack of research and postmortem discussion regarding the impact of GUR on the quality of published games. We can identify a number of possible reasons for this deficit. First, it is partly caused by the fact that academic researchers rarely have access to research data of commercial games. Even if studios or publishers conduct the research themselves, this information is often kept internal and not publicly shared. Secondly, it is hard to measure the impact of GUR on a given game's quality and its relation to the game's success in the market. This circumstance is aggravated by the fact that success in the game industry is influenced by a variety of factors which cannot be viewed and assessed in isolation.

The lack of research in this area impacts the utilization of playtesting in game development, where both the effectiveness of playtesting (e.g., assessed through critical reception or return on investment [ROI]) and availability of resources (e.g., access to a playtesting lab) have a direct influence on whether playtesting (or other GUR evaluation techniques) is included in the development process.

This paper contributes to closing this gap by studying the relationship between design features highlighted through playtesting during development (pre-release) and design features pointed out by game reviews post-release. To study this, we partnered with three commercial indie game studios, conducted pre-release playtesting on their games, and provided them with playtesting reports. All three games were com-

CHI '20, April 25–30, 2020, Honolulu, HI, USA.
© 2020 Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-6708-0/20/04 ...$15.00.
http://dx.doi.org/10.1145/3313831.3376831

mercially released between 2016 and 2017. We then waited one year before collecting and analyzing both user and critic reviews on each game. We employed a hybrid deductive and inductive content analysis approach to identify reoccurring design features mentioned in reviews (positive and negative) and grouped them into high-level categories. Features mentioned by reviewers were then compared to the design features highlighted during playtesting. Our goal is to examine the ability of different playtesting approaches to investigate design features important to game critics and reviewers.

While the adaptation and triangulation of underlying user research methods (e.g., observation, interviews) in playtesting has been covered from a methodological perspective (e.g., [21, 25]), our goal is to study the contribution of playtesting (and its underlying methods) in identifying issues that had an impact on critical reviews and players' perception of games. This is an important question to address because it helps to demonstrate the "actual" impact of playtesting on a game. Moreover, based on the insights from these case studies, we propose a set of guidelines for selecting appropriate user research methods in playtesting procedures. To summarize, our contribution lies in studying the impact of HCI research methods in the context of game development. In particular, we make three key contributions: a) We investigate the relationship between playtesting and post-release reviews from users and critics. b) To this end, we propose an analysis approach for the above process. c) Based on our findings, we present guidelines for selecting appropriate methods in the structuring of playtests.

BACKGROUND AND RELATED WORK

Playtesting is arguably one of the most recognized evaluation techniques in GUR. Schell [29] considers playtesting a necessary part of game development to assess whether a game evokes the experience for which it was designed. In playtesting, researchers observe players while they are playing a game, often before its release. Playtesting is a term that is used in different contexts in game development and can take on different forms, including free-form play sessions and participatory design workshops, among others (cf. [7]). These approaches use the same underlying user research methods (e.g., observation, interview), but with a different core focus (e.g., co-designing with players).

In the context of this paper, we refer to playtesting as a test conducted before release where external user researchers evaluate the game, identifying adjustments to bring the game closer to the developer's intent. Researchers usually take detailed notes on player behaviour to compare against the designers' intended behaviour. In addition to observation, other user research evaluation methods (e.g., interview, diary study) are often combined in a single playtest procedure. For example, researchers can include interviews or questionnaires to gather more subjective data about players and their experience with a game [1]. Researchers often produce a report for game developers to inform them about possible usability and user experience issues. The main goals of this process are to assist the game developer in optimizing the design quality of their game and enhancing the player experience.

A common challenge for studios incorporating playtesting in their development cycle is maintaining a strict development budget and avoiding unnecessary expenditures. Large game publishers (e.g., Ubisoft, Sony, Microsoft, Electronic Arts) have in-house playtesting facilities and staff. Playtesting is therefore often included by default in the development cycle of larger studios. Since these companies produce many titles overlapping each other's production, it makes sense to have internal playtesting teams [23]. On the other side of the spectrum are small and mid-sized studios, where playtesting may not always seem feasible, as it requires specialized setup and expertise. Moreover, as highlighted by Choi et al. [5], testing methods are not as well-understood in the industry as one might hope. Playtesting is a challenging process that requires careful and purposeful decision-making and can thus prove problematic for novice game designers [5].

In addition to resource-related issues, there is a lack of knowledge dissemination regarding the impact of playtesting on a game's quality. This lack of knowledge means that studios may hesitate to invest in playtesting, as the potential return on investment is currently not well-studied. Furthermore, the immediate effects of playtesting on key performance metrics (such as sales) are hard to measure and pinpoint. Although there are several articles and conference presentations on *the importance of conducting a playtest* [3], or on *how to conduct a playtest* [24], there are very few publicly available resources (e.g., [32]) discussing the potential impacts of playtesting on a game's ROI, quality, and overall post-release success.

This is also because there are many different ways to define success in game development, and the notion of success itself can also vary depending on the developer or game in question. There are various objective ways to measure the success of a game. For example, games can be considered successful if they have a high volume of sales, high number of daily active users, or high retention and conversion rates [18]. Success can also be determined via more subjective metrics, such as whether a game received positive reviews or a high Metacritic[1] score. Many developers feel a sense of triumph if their games receive a high review score and people enjoy playing their games [27]. Securing a publisher for future projects can depend on the critical success of past efforts, and some studios even link benefits or bonuses to review scores [30].

Other definitions of how developers define success – as highlighted by Ruggiero's GDC talk [27] – range from having a long-lasting legacy, to pushing boundaries, and to providing unique experiences. Some developers feel successful if they manage to create a game that they personally love and feel proud of. Sapieha [28] notes that Ubisoft Montreal measured the success of *Far Cry 5* not only in sales figures, but also in how much the studio learned and grew from the experience of developing the game. Nintendo also considers the legacy built through their many iconic intellectual properties to be a major accomplishment of the company [8].

[1] Metacritic is a review aggregator for games and other media that collects reviews from various sources to create an average score. http://www.metacritic.com/ (Accessed: January, 2020)

Reviews have been demonstrated as capable of impacting game sales. Several studies have investigated the effects of review comments on a player's perception of a game. According to Livingston et al. [16], players are more prone to negatively evaluate their experience if they read negative reviews before playing a game. Following Livingston et al. [17] further, negative reviews are more likely to bias players' impressions than positive reviews. Moreover, other studies, such as the one conducted by Meer [20], showed that when players read positive reviews before playing a game, they are more willing to buy it and are also more likely to spread positive information via word-of-mouth, leading to potentially higher sales. Internet phenomena such as review bombing, a practice where large groups of players purposefully post poor ratings for a game, also demonstrate how reviews are used in an attempt to impact sales (e.g., [9, 19]).

Review content can provide strong insights into how a game is received by players, as they contain details about the strengths and weaknesses of a given game. These comments frequently highlight issues with the game that could have been noticed (and possibly fixed) during development. The value of reviews for researching player experience has thus been recognized in industry and academia alike. Livingston [14] introduced the review analysis methodology, which has been used in large studios such as Ubisoft. This process consists of thoroughly analyzing critic reviews through qualitative coding, similar to the methodology we employed in this paper. Zagal et al. [33] surveyed game reviews from online websites and found that they are rich and varied with respect to the themes they cover. These include, among other things, the personal experiences of players as well as suggestions for improvement. Bond and Beale [4] analyzed game reviews using a grounded theory approach, extracted common features that characterize "good" games and, in turn, deduced heuristics to inform game design. These findings show that reviews can be a valuable source of insights pertaining to issues which may be identified or missed during playtesting. Zagal and Tomuro [34] made use of reviews to study cultural differences between Japanese and American players in their appreciation of games. Besides gaining insights on cultural preferences, they found that qualitative descriptions may not be matched exactly with scores, and that reviewers may award different scores despite pointing out similar issues. In our work, we thus rely on review text only, as we are interested in the particular issues noted by players and critics. In summary, game reviews have been widely studied in a general sense, but to the best of our knowledge, have not yet been applied in the context of playtesting practices.

METHOD

Our goal is to investigate the relationship between design features that are identified through different playtesting methods and compare them to key points highlighted in game reviews. By investigating this relationship, we are aiming to highlight the effectiveness of playtesting evaluation during development (pre-release) on the general reception of games. To study this, we partnered with three commercial independent game studios. The first author was involved in facilitating all playtesting sessions which were conducted external to the development team (i.e., employees of the game studies were neither involved in conducting the tests nor in preparing the reports). In all cases, the playtesting conducted was the only formal external playtesting completed on the games (although informal feedback may have been obtained while presenting the games at exhibitions). All games were in the post-production stage at the time of testing and commercially released between 2016 and 2017. Three different games from different genres were selected, as our intention is to explore the impact of playtesting across a variety of individual cases with different test designs. We anonymised (due to non-disclosure agreements with the developers) and summarized these efforts in the section below. We then waited for one year to allow for the publication of reviews before collecting reviews for each game. We used Metacritic[1] to collect reviews from both users and professional game critics. We then analyzed and compared design issues found through the playtesting reports with design issues pointed out in game reviews. This step was completed in five phases which are explained in more detail after introducing the three cases.

Case Studies

Game A

Game A is a card collection game (CCG) where the player strategically builds a deck of cards to engage in one-on-one battles. The game board is a 3×2 grid where each player can place up to three cards in a round. The goal is to place a better card next to an opponent's card to win the battle. The focus of playtesting for *Game A* was evaluating retention, first time user experience (FTUE), and the user interface (UI). Expert evaluation (based on design heuristics) conducted by the first author was the key method used and it was augmented by a small diary study [12] with 4 mid-core CCG players as participants. Each diary study participant played the game for around 10 hours individually. The first hour was conducted at the studio's playtest lab and was followed by face-to-face interviews conducted by the researcher. This was done to evaluate the FTUE. The players then continued to play the game (for around 9 more hours) on their own terms and submitted their feedback via weekly emails to the researcher. Once they completed 10 hours of gameplay, each player had a phone interview with the researcher. The longer play duration and concluding interview were used to evaluate the game's retention and gather players' feedback on UI design.

Game B

Our second case is a multiplayer top-down western style shooter. The game provides multiple character classes known as outlaws, each with different damage, firing rate, stamina, range, speed, health, and character abilities. The developers were interested in learning about their game's UI flow to see if players understood their menus. The developers also wanted to evaluate their tutorial and FTUE. Similar to *Game A*, expert evaluation was the key method used and it was supported by a small gameplay session followed by a focus group with four mid-core players. Participants played through the tutorial level in a co-located format, each on their own individual computer. Once all players completed the tutorial, teams of two were formed, where each team played against the other for four matches (each match lasted about five minutes). Teams were mixed after each match was completed. Four researchers observed the players during gameplay (one researcher at each computer station). After playing the game for 30 minutes, all four participants were first individually interviewed by their assigned researcher before participating in a focus group discussion led by the main researcher.

Game C

Our final case is a hardcore platformer where players control two characters simultaneously to navigate through platforming puzzles. The game utilizes the PS4 controller touchpad to directly interact with items in the game. The developers were interested in evaluating the entire player experience and balancing difficulty. Two rounds of playtesting were conducted on the game. In round one, nine players individually played through the first two worlds of the game (playtest focused on FTUE) totalling about one hour of gameplay; the researcher observed the gameplay and interviewed each player after the session. In round two, eight players (four returning from the first round) were recruited to play the first four worlds totaling around three hours of gameplay (playtest focused on FTUE and balancing difficulty). The returning players completed Worlds 3 and 4, where the new players started the game from World 1 and played until the end of World 4. Similar to the first round, the researcher observed all gameplay and interviewed each player individually after their session. In order to balance difficulty, the second round used analytics, specifically tracking the players' avatar position, jumps, death locations, and time to complete each level.

Review and Playtest Report Analysis

To contrast the types of design features identified through playtesting with those areas focused on by critics and players, we conducted a comparative analysis of review data and playtesting reports for the case studies listed above.

For each of the three games studied, we collected reviews from Metacritic[1]. We chose Metacritic for this process due to its popularity and its compilation of both professional critics and user reviews from all over the world. In the following analysis, we will not distinguish between reviews from users and critics (in order to capture all review comments) and will collectively refer to them as *game reviews*. The reviews downloaded represented all of those available on Metacritic for the games chosen as of the collection date (June 2018). The gathered reviews were then checked for suitability by excluding those that we deemed unusable for analysis due to a poor English translation (sometimes Metacritic reviews originate from non-English language websites), not relating to the game at all (sometimes reviews are added for the wrong game), or for which the full text was not available on the Internet (sometimes Metacritic adds the score from a review which was only published in a printed magazine). Lastly, we combined reviews into three distinct corpora, one for each game, as summarized in Table 1.

Game	R_{total}	$R_{cleaned}$	$\overline{wc}$
Game B	54	44	1110
Game A	17	15	1050
Game C	30	21	1055

R_{total} = total number of reviews, $R_{cleaned}$ = number of reviews after cleaning
$\overline{wc}$ = approximate average word count

Table 1. Details on each of the three sets of collected reviews.

Extracting Game Features from Review Data

To identify and assess the themes and issues discussed in each collection of reviews, we employed a hybrid deductive and inductive content analysis approach similar to that described by Fereday and Muir-Cochrane [10]. Our process consisted of five main phases:

Phase 1 – Collaborative deductive category/feature generation. Four researchers were involved in the initial brainstorming of codes. All four researchers are experts in GUR and/or game design. We defined features as specific components or qualities of a game or its subsystems (e.g., combat mechanics, engagement, art style) falling within broader descriptive categories (e.g., GAMEPLAY, USABILITY). During our brainstorming session, we referenced the reviews loosely without reading them individually in detail. The codes generated were based primarily on researchers' design expertise and the formal elements of game design (see, e.g., [11]). These initial features and categories served as the basis for inductive refinement in Phase 2.

Phase 2 – Individual inductive feature generation. Three of the researchers from the initial session individually read through every review corpus to familiarize themselves further with each game and its reception. During this process, each researcher independently edited their own copy of the initial feature and category list to better reflect the content of each game. Researchers were permitted to modify the list differently for each game to account for differences emerging from a game's specific genre and mechanics (e.g., "platforming" is a feature in *Game C* but not in *Game A* or *Game B*).

Phase 3 – Collaborative finalization of categories and features. Following individual review, the researchers met again to compare their insights and merge their changes into a master list of categories and features which was agreed upon for each game. The finalized collection of categories and features for each game are summarized in Table 2, 3, and 4. Little disagreement occurred between the researchers during the process of finalizing the feature lists; any conflicts were resolved through discussion until a consensus was reached.

Phase 4 – Independent coding. After finalizing our feature lists, we performed a single round of independent review coding to gain an approximate assessment of the relative importance of each issue. Two of the researchers involved in the code generation process independently read through the reviews after a short meeting to ensure sufficient common ground and understanding was present to interpret features similarly. Individual reviews were assigned codes for each feature, deeming its treatment of the feature as positive, negative, neutral ("neutral" was assigned for outright statements of a feature being unremarkable as well as cases where both positive and negative points were made), or not present (if the review did not mention a feature at all).

Phase 5 – Collaborative review of coding results. After coding, the two researchers involved met again to discuss discrepancies in their analysis and correct any errors in the process of combining their results. We elected not to perform a statistical analysis of inter-coder reliability on a per-item basis, due to the large number of features coded, limited sample size (particularly for *Game A* and *Game C*), our work's focus on the identification rather than quantization of features, and above all, the fact that our codes are not mutually exclusive (i.e., a comment in a review can relate to multiple codes). However, we did consider the total number of instances identified for each code (e.g., "number of reviews mentioning level design positively", "number of reviews not mentioning game narrative") for each review corpus, to give a coarse indication of the perceived importance and quality of features for each game. These totals were calculated as the average of the totals found by each coder. Anecdotally, we found that more specific features (e.g., "monetization" in *Game A*) tended to have higher agreement between coders than features which had inherently more ambiguous interpretations (e.g., "engagement"), despite meeting to discuss feature definitions before coding. While this part of our process serves only as a supplement to our main focus, the challenges involved with generating a coding scheme suitable for quantifying issues in game reviews may form an interesting basis for future work.

Comparative Feature Analysis

After completing our analysis of the collected reviews, we reviewed the playtesting reports generated from each of the case studies described earlier. For each of the features we inferred from the review data, we determined whether or not the feature was addressed by the playtest through analyzing the issues covered in the report. We then assessed the presence or absence of these features (and more broadly, their respective categories) to understand how the different user research methods and playtesting formats employed in generating each report contributed to its ability to address the areas highlighted in game reviews.

We also performed an exploratory assessment of feature quality and importance to the reviewers for each game, based on the results of the coding described in the previous section. To provide an approximate prioritization of features, we consulted similar work done by Livingston and Mandryk [15] on quantifying the impact of game issues based on heuristic evaluation scores and genre weightings, as well as the issue prioritization metrics of severity and frequency suggested by Nielsen [26]. Since we are operating on a much different dataset, we devised our own method using the number of reviews mentioning each feature in a certain sentiment (positive, negative, or neutral) to calculate indices representing a feature's quality, importance, and overall priority. These indices are calculated for each feature f based on simple proportionate comparisons of different mentions as follows:

$$I^f_{quality} = \frac{R^f_+ - R^f_-}{R^f} \qquad I^f_{importance} = \frac{R^f}{R}$$

$$I^f_{priority} = I^f_{quality} \cdot I^f_{importance}$$

Where R is the total number of reviews, R^f the number of reviews mentioning a feature, and R^f_+ and R^f_- the number of reviews being positive or negative about a feature. Thus, $I_{quality}$ is a measure of the polarity and extent with which a feature was received, while $I_{importance}$ quantifies how much attention a feature received across all reviews. A positive priority index indicates that a feature was well-received, while a negative priority index means that a feature was poorly received. A larger absolute value indicates that a feature was mentioned in more reviews, and/or the feature had a greater critical consensus. We also normalized the priority indices independently for each game based on their maximum value in order to provide a more uniform indication of relative feature importance. Thus, a value close to 1.0 signifies that a feature was mentioned by many reviews and was generally liked, while a value close to -1.0 indicates that a feature was mentioned frequently but was generally disliked. Values close to 0.0 can arise from either neutral or polarized opinions of a feature, or relatively little presence in reviews at all. In these cases, consulting the quality and importance indices can provide further insight. As a supplement to our analysis of features included or omitted from playtesting reports, we considered the priority indices as indications of which features should receive additional attention in playtesting reports.

RESULTS

For each game, we have compiled the complete feature list, playtesting report coverage, and quality, importance, and priority indices into a summary table (Table 2 to 4). Additionally, the tables list the total positive, negative, and neutral mention tallies for each feature. Note that half values can occur since instance counts were averaged between the two coders. We are especially concerned with identifying those areas highlighted in reviews which were not present in the playtesting reports, as well as indicating gaps between the methods chosen and the needs of the games studied.

Game A

The playtesting report for *Game A* highlighted features from every category with the exception of ART & SOUND. Issues identified related to *Engagement* (the game's ability to hold players' attention and promote immersion), *Depth* (referring to complexity in tactics and game mechanics), *Singleplayer* (the game's solo campaign outside competitive play), *AI* (the quality of play from computer-controlled opponents), *Monetization* (the game's implementation of a free-to-play model), *Polish/Bugs* (general technical competence), *UI/Menus*, and *Player Learning*. Of the eight features highlighted by the report, three were deemed problematic by the overall opinion of reviewers (*AI*, *Polish/Bugs*, and *UI/Menus*).

The most negatively received feature overall, *Sound* (the game's auditory cues and sound design), was not mentioned in the playtesting report. All other negatively critiqued issues were highlighted by the report, with only a couple of other excluded features approaching possible contention on the part of the reviewers (e.g., *Balance*, or the game's strategic fairness).

Game B

The features included in the playtesting report for *Game B* fell within the USABILITY and LEARNING & DIFFICULTY categories. We identified three features highlighted by the report: *Controls* (game input), *Tutorialization* (how in-game mechanics are taught to the player), and *Practice Mode* (a game mode specific to *Game B* for players to hone their skills outside of competitive play). One of these features was noted by the reviewers to require improvement (*Tutorialization*). *Practice Mode* was well-received but not noted in a large proportion of reviews, while *Controls* were somewhat contentious and mentioned in approximately ⅕ of all reviews.

Five features were absent from the playtesting report but poorly received by the reviewers, falling within the CONTENT and USABILITY categories. These features were *Level Variety* (the diversity of levels available to play), *Game Modes* (the different match types available), *Story*, *Polish/Bugs* (general technical competence and glitches), and *Networking* (practical function of the game's online features and matchmaking). Reviewers generally did not take issue with features in the GAMEPLAY and ART & SOUND categories, which were also not commented on in the playtesting report.

Game C

The playtesting report of *Game C* highlighted features from each of the five categories. Issues mentioned were identified as relating to *Boss Fights* (periodic combat-focused sections of the game), *Art Style* (the game's overall visual impression), *Puzzle Design*, *Controls* (game input), *Touchpad* (use of the touchpad on compatible game controllers specific to *Game C*), *Tutorialization* (how in-game mechanics are taught to the player), and *Difficulty Curve* (sequencing and progression of game challenge). Of these seven features, three were poorly received (*Boss Fights*, *Controls*, and *Tutorialization*). Two others, *Touchpad* and *Difficulty Curve*, were both highly contentious, being mentioned in a significant number of reviews with both complaints and praise.

Polish/Bugs (the game's general technical competence) was the only feature that was both negatively received and not highlighted by the playtesting report. Generally speaking, all other features were either well-received, addressed by the playtesting report, or both.

DISCUSSION

In reflecting upon the results of our review analysis, we can observe some general trends beginning to emerge. For each of the three games studied, features related to the games' core mechanics were discussed frequently in game reviews. Many of the features among those with the highest importance indices for each game were directly related to its core; such as *Platforming* and *Dual Characters* for *Game C*; *Engagement* and *Multiplayer* for *Game B*; and *Balance* and *Crafting* for *Game A*. This reinforces the idea that, as the foundation of a player's experience and thus a key factor in critical reception, a game's playtesting is well-advised to focus at least partially on the quality of its core gameplay mechanics. We might expect that this trend extends to those features more strongly tied to a game's genre (e.g., on average, features related to combat would see more weight in reviews of a first-person shooter as opposed to a platformer, while features related to controls would be seen as more important in a platformer when compared to a strategy game). However, further work is needed to investigate or attempt to quantify such patterns in a more general context.

Another observation pertinent to each of the games studied is that aspects outside of the game's core play mechanics can also form the focus of critical analysis – at times, arguably more so than the mechanics themselves. In *Game C*, for example, *Story* was discussed to some extent in nearly every review, with references to *Art Style* and *Sound* also popular among reviewers. All three of these features were observed to appear in more reviews than *Puzzle Mechanics*, despite the fact that *Game C* is a puzzle-platforming game. Similarly, reviews for *Game A* contained frequent discussions of *Story* and *Graphics*, just as *Art Style* was often mentioned in reviews of *Game B*.

This tendency of game reviewers to focus not only on a game's core may serve to bolster or damage its critical reception. While a game might be lauded for features outside of core gameplay, the most severe issues identified by reviewers may also stem from factors beyond play mechanics. For instance, technical polish was a consistent concern for each of the games studied. Features related to longevity – such as the variety of a game's content – were also pertinent in reviews, though these features were received with marked differences in opinion between the games studied. The manner in which these patterns were expressed in each case can help us to learn how specific playtesting protocols can affect the ability of a playtest to address specific game features. In the following we discuss our results with respect to the three games in more detail.

Game B

Of the three games studied, *Game B* had the greatest mismatch between game review analysis and playtest reporting, with five critically problematic features "missing" from the playtest report. When reflecting on the testing focus and protocol used, the origin of this disparity becomes apparent. Among the three games, *Game B* had the shortest playtest by far, with the fewest participants, and a stated focus on tutorial and first-time user experience. Examining the nature of the test's structure, it follows naturally that these five problem features would not be addressed by the subsequent report.

Two of the poorly received features absent from the playtest report are related to the game's technical proficiency – *Polish/Bugs* and *Networking*. One could argue that playtesting and quality assurance (QA) are two separate disciplines – that it is not the responsibility of user researchers to find and report bugs. While this is a valid argument, this quandary begs the question of whether user researchers should bother reporting bugs when they are observed, citing them as "a problem for the QA team". However, this mindset bears the risk of rendering the test as somewhat of a sieve for technical issues. Sometimes playability issues identified during playtesting may also point to technical issues of relevance for QA, which should be shared with developers accordingly. Such an approach could offer an additional benefit, for instance, for independent developers with a limited budget available for QA. While this

Category	Feature	in report	$I_{quality}$	$I_{importance}$	$I_{priority}$	$\tilde{I}_{priority}$	$\#_{positive}$	$\#_{negative}$	$\#_{neutral}$
Gameplay	Engagement	✓	0.83	0.40	0.33	0.77	5	0	1
	Depth	✓	0.76	0.57	0.43	1.00	7.5	1	0
	Balance	–	0.36	0.47	0.17	0.38	4	1.5	1.5
	Multiplayer	–	0.80	0.33	0.27	0.62	4	0	1
	Singleplayer	✓	0.73	0.37	0.27	0.62	4.5	0.5	0.5
	Ranking	–	0.75	0.13	0.10	0.23	1.5	0	0.5
	Crafting	–	1.00	0.40	0.40	0.92	6	0	0
	AI	✓	**-1.00**	**0.07**	**-0.07**	**-0.15**	0	1	0
Art & Sound	Graphics	–	1.00	0.40	0.40	0.92	6	0	0
	Art Style	–	0.50	0.13	0.07	0.15	1.5	0.5	0
	Animation	–	1.00	0.13	0.13	0.31	2	0	0
	Sound	–	**-0.60**	**0.17**	**-0.10**	**-0.23**	0	1.5	1
Content	Monetization	✓	0.55	0.67	0.37	0.85	7	1.5	1.5
	Story	–	0.79	0.47	0.37	0.85	6	0.5	0.5
	Content Variety	–	1.00	0.27	0.27	0.62	4	0	0
Usability	**Polish/Bugs**	✓	**-0.25**	**0.13**	**-0.03**	**-0.08**	0.5	1	0.5
	UI/Menues	✓	**-0.40**	**0.17**	**-0.07**	**-0.15**	0.5	1.5	0.5
Learning & Difficulty	Player Learning	✓	0.17	0.20	0.03	0.08	1	0.5	1.5

Features with a negative priority index are written in bold face. $\tilde{I}_{priority}$ = normalized priority index.

Table 2. Results of comparative review and report analysis for *Game A*.

Category	Feature	in report	$I_{quality}$	$I_{importance}$	$I_{priority}$	$\tilde{I}_{priority}$	$\#_{positive}$	$\#_{negative}$	$\#_{neutral}$
Gameplay	Engagement	–	0.59	0.72	0.42	1.00	23	4.5	4
	Depth	–	0.88	0.36	0.32	0.76	15	1	0
	Balance	–	0.33	0.24	0.08	0.19	7	3.5	0
	Multiplayer	–	0.24	0.38	0.09	0.22	9.5	5.5	1.5
	Ranking	–	0.19	0.18	0.03	0.08	4	2.5	1.5
	Combat	–	0.85	0.30	0.25	0.59	11.5	0.5	1
Art & Sound	Graphics	–	0.73	0.30	0.22	0.51	11	1.5	0.5
	Art Style	–	1.00	0.31	0.31	0.73	13.5	0	0
	Animation	–	0.33	0.03	0.01	0.03	1	0.5	0
	Sound	–	1.00	0.09	0.09	0.22	4	0	0
	Atmosphere	–	1.00	0.08	0.08	0.19	3.5	0	0
Content	**Level Variety**	–	**-0.54**	**0.15**	**-0.08**	**-0.19**	1	4.5	1
	Characters	–	0.65	0.45	0.30	0.70	14.5	1.5	4
	Game Modes	–	**-0.19**	**0.31**	**-0.06**	**-0.14**	4.5	7	2
	Story	–	**-1.00**	**0.07**	**-0.07**	**-0.16**	0	3	0
Usability	**Polish/Bugs**	–	**-0.65**	**0.19**	**-0.13**	**-0.30**	1.5	7	0
	Controls	✓	0.00	0.18	0.00	0.00	3	3	2
	Networking	–	**-0.65**	**0.26**	**-0.17**	**-0.41**	1.5	9	1
Learning & Difficulty	**Tutorialization**	✓	**-0.35**	**0.30**	**-0.10**	**-0.24**	2	6.5	4.5
	Practice Mode	✓	0.56	0.10	0.06	0.14	2.5	0	2

Features with a negative priority index are written in bold face. $\tilde{I}_{priority}$ = normalized priority index.

Table 3. Results of comparative review and report analysis for *Game B*.

Category	Feature	in report	$I_{quality}$	$I_{importance}$	$I_{priority}$	$\tilde{I}_{priority}$	$\#_{positive}$	$\#_{negative}$	$\#_{neutral}$
Gameplay	Engagement	–	0.57	0.50	0.29	0.44	7.5	1.5	1.5
	Platforming	–	0.58	0.79	0.45	0.70	10	0.5	6
	Dual Characters	–	0.75	0.48	0.36	0.56	7.5	0	2.5
	Boss Fights	✓	**-0.86**	**0.17**	**-0.14**	**-0.22**	0	3	0.5
	Puzzle Mechanics	–	0.72	0.43	0.31	0.48	7.5	1	0.5
Art & Sound	Graphics	–	0.87	0.36	0.31	0.48	6.5	0	1
	Art Style	✓	0.00	0.45	0.45	0.70	9.5	0	0
	Animation	–	1.00	0.10	0.10	0.15	2	0	0
	Sound	–	0.64	0.52	0.33	0.52	8.5	1.5	1
	Atmosphere	–	0.82	0.26	0.21	0.33	4.5	0	1
Content	Story	–	0.60	0.95	0.57	0.89	13.5	1.5	5
	Level Design	–	0.82	0.79	0.64	1.00	13.5	0	3
	Puzzle Design	✓	0.63	0.45	0.29	0.44	6	0	3.5
Usability	**Polish/Bugs**	–	**-0.73**	**0.26**	**-0.19**	**-0.30**	0.5	4.5	0.5
	Controls	✓	**-0.13**	**0.55**	**-0.07**	**-0.11**	2.5	4	5
	Touchpad	✓	0.00	0.67	0.00	0.00	5	5	4
Learning & Difficulty	**Tutorialization**	✓	**-0.38**	**0.19**	**-0.07**	**-0.11**	1	2.5	0.5
	Difficulty Curve	✓	0.11	0.43	0.05	0.07	4	3	2

Features with a negative priority index are written in bold face. $\tilde{I}_{priority}$ = normalized priority index.

Table 4. Results of comparative review and report analysis for *Game C*.

debate is far beyond the scope of our current work, measures may be taken to allow playtesting to function as an additional stage of identifying technical issues – for example, ensuring that the QA team has access to recorded gameplay footage.

The usability of the game's *Networking* and understandability of its *Ranking* system – another feature receiving low review reception – is a different story. Since the playtest of *Game B* used co-located participants in the game's local multiplayer mode, it would have been impossible for the test to evaluate any aspect of networked multiplayer. Validation of these features could have been pursued through an appropriate test later in development, such as a remote playtest, or closed beta with player questionnaires.

The three remaining negative features not present in the playtesting report are *Level Variety*, *Game Modes*, and *Story*, all falling into the broader category of CONTENT. Generally, comments in reviews related to these features cited their failure to promote the game's longevity and replayability, primarily lamenting a lack of diversity in the combat modes and maps available. Given that these concerns only reveal themselves after players have engaged with the game for a time, it makes sense that they simply would not arise in an hour-long session focused on FTUE. For issues related to a game's longevity, longer-term test structures, such as diary studies, may be advisable. Shorter-term studies with a focus on content variety – such as focus groups evaluating a game's catalogue of levels or characters – may also help to evaluate these concerns.

Those features that were present in the report arise naturally from the focus of the test – *Controls*, *Tutorialization*, and *Practice Mode*. Since participants were engaging with the game for the first time, one would expect that the issues observed would stem primarily from initial impressions of the game's interaction scheme, learning its mechanics, and practising those mechanics as they became proficient. Thus, we might conclude that a short-term FTUE evaluation is best suited for the evaluation of features related to core interactions (e.g., controls and user interface) and player learning (e.g., tutorialization).

Game C

Game C's playtest addressed most of the features which were seen as problematic in game reviews. At a high level, this likely stems in part from the larger sample size (14 players in total) and longer total session length (approximately 3 hours). The resultant reports addressed concerns relating to each of the five categories identified (GAMEPLAY, ART & SOUND, CONTENT, USABILITY, LEARNING & DIFFICULTY). The only feature absent from the testing reports that received negative reception was technical *Polish* – as discussed above, this is less of a question regarding the testing methods employed and more a debate about the goal of playtesting and user research in general. Looking at the features that were addressed, we can form connections between the nature of the test design used to evaluate *Game C* and the issues identified.

The *Game C* playtest identified issues related to the features associated with FTUE discussed above – *Tutorialization*, *Controls*, and *Touchpad*. Arguably, the report comments associated with *Art Style* may be better categorized as relating to user interface (another key factor in users' initial impressions of usability), though UI was not identified as a notable feature discussed explicitly in the game's reviews. It follows that a longer testing session will naturally cover many of the concerns related to FTUE. Thus, when a game is at a suitable stage in development, a longer test is advisable to increase the breadth of features that can be tested, assuming sufficient resources are available.

The more protracted nature of *Game C*'s test design allowed it to identify issues with features related to long-term play, such as *Boss Fights*, *Puzzle Design*, and the game's *Difficulty Curve*. Concerns with the level progression and difficulty over several hours of play would be impossible to identify with a shorter test format. The identification of these issues was also supported by the data collection methods employed, namely the inclusion of metrics. This data made it possible to assess player performance in much greater depth, with certain measurements especially pertinent to each of the features mentioned above (e.g., using player death counts to identify difficult boss fights, consulting a map of attempted jumps and deaths to flag frustrating puzzle segments, and examining how these metrics change as the game progresses).

Game A

Game A's playtest report covered most of the design issues echoed by the game reviews. *Game A* had the longest total playtime per player overall, though with just four participants, quite a small sample size. Extended playtime proved effective in ensuring the test's overall breadth, addressing features from every category with the exception of ART & SOUND. Unlike the other two tests, technical *Polish* was addressed in the report for *Game A*, as were the expected FTUE concerns of *User Interface* and *Player Learning*. This provides further support for the idea that long-term user studies will naturally allow for the evaluation of many FTUE elements. While the timing of the *Game A* test in the development cycle prohibited the evaluation of multiplayer, it addressed issues related to the game's single-player campaign and consequently the strength of AI opponent design.

The extended length of the playtest also facilitated the identification of issues relating to long-term *Engagement* and strategic *Depth*, in line with the test's stated goal of assessing retention and engagement. The test also addressed problems with the game's *Monetization* strategy, suggesting a rebalance to improve the perceived value of in-game rewards. This contrasts the inability of shorter tests to address features contributing to longevity, as was the case with *Game B*. However, it should be noted that the report for *Game A* still does not comment on many aspects of the game's content (e.g., variety of game elements) – though these features were well-received in the game reviews. Direct probing of such questions may be advisable for longer-term tests to gain a more complete understanding of a game's longevity.

Only one feature with negative overall reception was absent from the playtest report – *Sound*. This omission is interesting; sound was not discussed in a very sizeable number of reviews, and likewise its absence from the playtesting report may be due to the small sample of players selected and the nature of the game in question. Since players were left to play the game during their everyday routine for the majority of total playtime, they may have simply played with the sound off, a common habit when playing on a mobile device. This oversight occurs perhaps all too frequently – sound design is notorious for its often late inclusion in the development cycle. However, the importance of sound, or perhaps more accurately poor sound, should be considered in the evaluation of the complete user experience.

Summary and Selection of Playtesting Protocols

Overall, we can make some additional observations about the ability of typical playtest reports to address features important to a game's critical reception. For instance, we have consistently observed a lack of comments in the playtest reports studied on aspects of a game's creative direction – perhaps since qualities such as graphics and art style are seen as too subjective to be evaluated with a few participants. However, features such as animation and sound design can form critical components of a game's feedback system, and the importance of game aesthetics in game reviews is hard to ignore. It is arguable that user research in general may benefit from broadening its focus beyond usability to include more questions of how a game's presentation may improve user experience. In doing so, it may be possible to extract deeper insights from players when probing their initial impressions (e.g., if a player comments negatively on a character's movement, asking only about how the controls could be improved may overlook potential issues with the game's animation design). It is important to point out that larger sample sizes (i.e. more participants) are needed when attempting to answer more subjective questions.

Another consideration to be made is the tendency of playtesting reports to focus on problematic aspects of a game, rather than also highlighting the features players enjoyed. A game's reception is ultimately both additive and subtractive. While problems will impact its reputation negatively, its strengths serve equally to improve its overall image. Identifying these strengths as a part of playtest reporting may help developers to focus on and polish those features which players enjoy the most, and potentially inform aspects of post-production, such as marketing strategy.

In general, it is not easy to comment on the connection between certain user research techniques (such as semi-structured interviews) and the types of features that can be addressed through them due to their open-ended nature. Here, we can only advise that user researchers directly address suspected areas of critical importance as directly as possible – for example, by following up with players on their perception of core mechanics even if they do not comment on those mechanics during a testing session. Certain techniques lend themselves naturally to the identification of specific issues as discussed when noting the ability of *Game C*'s test to pinpoint difficulty problems by using game metrics. The format of a test – including session length, environment, and participant sample size – is also critical in determining its ability to validate certain features, a principle which consistently arose in our analysis of the three games studied. Placement and timing of a test in a game's development cycle is also crucial – testing certain features may simply be impossible depending on when the test is planned. If a game is not yet feature-complete, it may be advisable to split a test into multiple phases to ensure proper coverage.

The lessons learned from dissecting how playtesting results in several cases correlate with the features focused on by game reviewers can be used to inform playtest design according to developer needs and the intended focus of testing. Based on the insights gained from the cases studied here, we propose a set of guidelines for selecting test structures and methods based on the goals of the evaluation, summarized in Table 5.

FUTURE WORK

In this work, we combined reviews written by professional critics and users into a single corpus for each game to increase sample size and promote an overview of issues identified in all reviews. On the other hand, this mixing of user and critic reviews, may have resulted in some 'noise' as users may have different concerns than professional reviewers. A future direction would thus be to differentiate between these two categories, as patterns may emerge in the types of features discussed and the relative importance given to different features. It should also be noted that, given our focus on playtesting, we chose to examine smaller projects from independent studios, as playtesting reports from larger studios are often less accessible due to stricter confidentiality policies. Despite this, we expect our results and our methodology for assessing the effectiveness of playtesting techniques to also generalize to other games, at least to a certain degree. Players may, for example, have different expectations and may be less lenient when evaluating AAA titles. This in itself may also be an interesting avenue for further research to establish an understanding of how players' expectations differ depending on factors such as genre, production budget, publisher size, and so on. Such research could help provide pointers on what aspects to focus on in playtesting for different projects and studios. Our focus on indie games resulted in a lower number of reviews available for each title when compared with large AAA releases. This had the benefit of resulting in a manageable number of reviews for manual inspecting and coding. At the same time, this means that our approach may not scale directly to games with a significantly larger number of game reviews. However, we can envision a semi-automated process which takes advantage of natural language processing (NLP) techniques to assist in this process. When using this method, one should also take precautions regarding "troll" reviews and abusive practices such as review bombing (cf. [19]). In our case, this was not an issue, as manual inspection could identify such problems with relative ease. However, when employing an automated approach and/or combing through a larger corpus of reviews, these potential pitfalls should be kept in mind.

Lastly, we would like to stress that our goal was to identify whether design issues pointed out by playtesting reports match those pointed out in game reviews. In doing so, we hoped to gain insights regarding which aspects of a game playtesting should focus on and the suitability of different evaluation methods under different circumstances. We did not investigate how and if issues noted in a playtesting report were addressed by the developers and if the report itself was of sufficient qual-

Game Features	Suggested Test Structures	Supporting Methods	Comments
Core gameplay (e.g., platforming, shooting, combat)	Short-term sessions with a demo to showcase required mechanics, long-term sessions	Semi-structured interviews, thinking-aloud, observation	It is important that players are able to experience all mechanics during the play session. All mechanics should be followed up on, even if players do not comment on them during the playtest.
Long-term engagement (e.g., depth, strategy, achievement system)	Long-term sessions (preferably in a player's home environment)	Player diaries, questionnaires, analytics, semi-structured interviews	Leaving players to their own setup can help to ensure that their impressions are organic, giving a better idea of what real players will feel like when playing for hours after purchasing the game.
Multiplayer (e.g., ranking system, matchmaking)	Co-located or remote short-term sessions, soft launches & beta testing	Group interviews, questionnaires (for remote tests/beta tests), analytics	Depending on the type of interactions permitted and the game's scale, a local (co-located) test may not provide sufficient coverage.
Graphics, art style, animation, sound, etc.	A/B testing (e.g., comparison of different sound designs), focus groups, any session type supported by interviews	post-gameplay review of specific images, animations, sounds, semi-structured interviews, questionnaires,	Questions related to visual/auditory feedback may be used in probing issues stemming from usability or the feeling of mechanics. Practical concerns, such as large sample size and speaker/display setup, should not be neglected.
Narrative, level variety, game modes, etc.	Long-term sessions, soft launches & beta testing, content-based focus groups	Semi-structured interviews, questionnaires (e.g., ranking/preference), analytics	Since content variety may only become an issue for players after a longer play time, a soft launch is especially relevant for games with a larger scope or longer intended playtime. Large sample sizes are recommended.
General usability (e.g., controls, user interface)	Any session, particularly FTUE evaluations to capture users' first impressions	Observation, interviews	Issues identified in this category may relate back to core gameplay features or feedback (e.g., animation/sound design).
Short-term challenges (e.g., tutorials, onboarding)	Short-term FTUE evaluations, short/long-term gameplay sessions	Skill-check interviews, observation, analytics	Many longer sessions will be able to naturally identify FTUE issues, as long as appropriate questions are asked.
Long-term challenges (e.g., difficulty curve, mastery)	Long-term gameplay sessions	Analytics, questionnaires	Long-format sessions are essential. Employing metrics or simple measurements of player performance greatly assists evaluation of difficulty. Large sample sizes are recommended.

Table 5. Proposed guidelines for structuring playtests and selecting methods according to their ability to address specific game features.

ity to allow them to act upon it. For example, an issue which was identified but was still negatively received in the released game may have received less attention than warranted because of limited resources or other development constraints. This is a worthwhile area to consider for future work, examining how developers utilize playtesting reports and which issues are fixed or neglected and why, can shed further light on the viability, plausibility, and persuasiveness [13] of playtesting.

CONCLUSION

Attempting to correlate development practices with game quality and post-release success in the game industry is inherently complicated, as many different metrics (e.g., revenue, number of downloads, review scores) can define and contribute to success. One key metric is game reviews, as reviews provide an indication of product quality and can have a direct impact on sales and other metrics. Game studios aim to develop high-quality games that deliver a fun experience for players and gain positive reception from the community. Therefore, understanding how players interact and behave during gameplay is of vital importance. User research (often via playtesting) aims to assist developers in achieving their design intent and help to identify and resolve design issues during development. However, a lack of research on the value of playtesting and its impact on the game quality means that many studios question the return on investment and usefulness of playtesting. In this paper, we reported on three case studies where we analyzed playtesting reports for three commercial games (pre-release) and reviews published for each game after its release. Through these unique and real-world case studies, our paper contributes to the growing domain of games user research and highlights the value of playtesting in game development. In particular, we detailed an analysis approach for extracting key game features from game review data. Through comparison of the identified features in reviews (based on our case studies) with the features highlighted in different playtesting reports with different underlying evaluation methods, we explored the relationship between different playtesting procedures and key game features they can investigate. Moreover, we proposed guidelines for structuring playtests and selecting methods according to their ability to address specific game features. However, we believe the key contribution of this paper resides in assessing the impact of HCI and user research on the game industry. Considering that the field of GUR continues to adapt HCI techniques within the entertainment sphere, this work advances the discussion by addressing how different user research methods inform developers and consequently impact the overall success of their products.

ACKNOWLEDGMENTS

We would like to thank the game studios who participated in this study and the Natural Sciences and Engineering Research Council of Canada (NSERC Discovery Grants Program - GPIN-2014-05763) for supporting this research.

REFERENCES

[1] Andrea Abney, Brooke White, Jeremy Glick, Andre Bermudez, Paul Breckow, Jason Yow, Rayna Tillinghast-Trickett, and Paul Heath. 2014. Evaluation of Recording Methods for User Test Sessions on Mobile Devices. In *Proceedings of the First ACM SIGCHI Annual Symposium on Computer-human Interaction in Play*. ACM, New York, NY, USA, 1–8. DOI: http://dx.doi.org/10.1145/2658537.2658704

[2] Regina Bernhaupt, Manfred Eckschlager, and Manfred Tscheligi. 2007. Methods for Evaluating Games: How to Measure Usability and User Experience in Games?. In *Proceedings of the International Conference on Advances in Computer Entertainment Technology*. ACM, New York, NY, USA, 309–310. DOI: http://dx.doi.org/10.1145/1255047.1255142

[3] Regina Bernhaupt, Wijand Ijsselsteijn, Florian 'Floyd' Mueller, Manfred Tscheligi, and Dennis Wixon. 2008. Evaluating User Experiences in Games. In *CHI '08 Extended Abstracts on Human Factors in Computing Systems*. ACM, New York, NY, USA, 3905–3908. DOI: http://dx.doi.org/10.1145/1358628.1358953

[4] Matthew Bond and Russell Beale. 2009. What Makes a Good Game?: Using Reviews to Inform Design. In *Proceedings of the 23rd British HCI Group Annual Conference on People and Computers: Celebrating People and Technology*. British Computer Society, Swinton, UK, UK, 418–422.

[5] Judeth Oden Choi, Jodi Forlizzi, Michael Christel, Rachel Moeller, MacKenzie Bates, and Jessica Hammer. 2016. Playtesting with a Purpose. In *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play*. ACM, New York, NY, USA, 254–265. DOI: http://dx.doi.org/10.1145/2967934.2968103

[6] Anders Drachen, Pejman Mirza-Babaei, and Lennart E. Nacke. 2018. *Games User Research*. Oxford University Press, Oxford, UK.

[7] Mirjam P. Eladhari and Elina M. I. Ollila. 2012. Design for Research Results: Experimental Prototyping and Play Testing. *Simulation & Gaming* 43, 3 (2012), 391–412. DOI: http://dx.doi.org/10.1177/1046878111434255

[8] Rob Fahey. 2013. Nintendo in the Firing Line. (2013). https://www.gamesindustry.biz/articles/2013-12-05-nintendo-in-the-firing-line Accessed: September, 2019.

[9] Wes Fenlon. 2017. Steam Review Bombing is Working, and Chinese Players are a Powerful New Voice. (2017). https://www.pcgamer.com/steam-review-bombing-is-working-and-chinese-players-are-a-powerful-new-voice/ Accessed: September, 2019.

[10] Jennifer Fereday and Eimear Muir-Cochrane. 2006. Demonstrating Rigor Using Thematic Analysis: A Hybrid Approach of Inductive and Deductive Coding and Theme Development. *International Journal of Qualitative Methods* 5, 1 (2006), 80–92. DOI: http://dx.doi.org/10.1177/160940690600500107

[11] Tracy Fullerton. 2008. *Game Design Workshop: A Playcentric Approach to Creating Innovative Games*. Morgan Kaufmann, Burlington, MA, USA.

[12] Serena Hillman, Tad Stach, Jason Procyk, and Veronica Zammitto. 2016. Diary Methods in AAA Games User Research. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1879–1885. DOI: http://dx.doi.org/10.1145/2851581.2892316

[13] Effie Lai-Chong Law. 2011. The Measurability and Predictability of User Experience. In *Proceedings of the 3rd ACM SIGCHI Symposium on Engineering Interactive Computing Systems*. ACM, New York, NY, USA, 1–10. DOI: http://dx.doi.org/10.1145/1996461.1996485

[14] Ian Livingston. 2018. Post-launch in Games User Research. In *Games User Research*, Anders Drachen, Pejman Mirza-Babaei, and Lennart Nacke (Eds.). Oxford University Press, Oxford, UK.

[15] Ian J. Livingston, Regan L. Mandryk, and Kevin G. Stanley. 2010. Critic-proofing: How Using Critic Reviews and Game Genres Can Refine Heuristic Evaluations. In *Proceedings of the International Academic Conference on the Future of Game Design and Technology*. ACM, New York, NY, USA, 48–55. DOI:http://dx.doi.org/10.1145/1920778.1920786

[16] Ian J. Livingston, Lennart E. Nacke, and Regan L. Mandryk. 2011a. The Impact of Negative Game Reviews and User Comments on Player Experience. In *ACM SIGGRAPH 2011 Game Papers*. ACM, New York, NY, USA, 4:1–4:5. DOI: http://dx.doi.org/10.1145/2037692.2037697

[17] Ian J. Livingston, Lennart E. Nacke, and Regan L. Mandryk. 2011b. Influencing Experience: The Effects of Reading Game Reviews on Player Experience. In *Entertainment Computing – ICEC 2011*, Junia Coutinho Anacleto, Sidney Fels, Nicholas Graham, Bill Kapralos, Magy Saif El-Nasr, and Kevin Stanley (Eds.). Springer, Berlin, Heidelberg, 89–100.

[18] Trevor McCalmont. 2015. 15 Metrics All Game Developers Should Know by Heart. (2015). https://gameanalytics.com/blog/metrics-all-game-developers-should-know.html Accessed: September, 2019.

[19] Kirk McKeand. 2017. A Brief History of how Steam Review Bombing Damages Developers. (2017). https://www.pcgamesn.com/history-of-steam-review-bombing Accessed: September, 2019.

[20] Alec Meer. 2010. EEDAR Study: Review Scores do Affect Sales. (2010). https://www.gamesindustry.biz/articles/eedar-review-scores-do-affect-sales Accessed: September, 2019.

[21] Pejman Mirza-Babaei, Naeem Moosajee, and Brandon Drenikow. 2016. Playtesting for Indie Studios. In *Proceedings of the 20th International Academic Mindtrek Conference*. ACM, New York, NY, USA, 366–374. DOI: http://dx.doi.org/10.1145/2994310.2994364

[22] Pejman Mirza-Babaei, Veronica Zammitto, Jörg Niesenhaus, Mirweis Sangin, and Lennart Nacke. 2013. Games User Research: Practice, Methods, and Applications. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems*. ACM, New York, NY, USA, 3219–3222. DOI: http://dx.doi.org/10.1145/2468356.2479651

[23] Naeem Moosajee and Pejman Mirza-Babaei. 2016. Games User Research (GUR) for Indie Studios. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, New York, NY, USA, 3159–3165. DOI: http://dx.doi.org/10.1145/2851581.2892408

[24] Lennart E. Nacke, Pejman Mirza-Babaei, and Anders Drachen. 2019. User Experience (UX) Research in Games. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, C25:1–C25:4. DOI: http://dx.doi.org/10.1145/3290607.3298826

[25] Lennart E. Nacke, Christiane Moser, Anders Drachen, Pejman Mirza-Babaei, Andrea Abney, and Zhu (Cole) Zhenyu. 2016. Lightweight Games User Research for Indies and Non-Profit Organizations. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, New York, NY, USA, 3597–3603. DOI: http://dx.doi.org/10.1145/2851581.2856504

[26] Jakob Nielsen. 1995. Severity Ratings for Usability Problems. *Papers and Essays* 54 (1995), 1–2.

[27] Dana Ruggiero. 2017. How Game Developers Define Success. Presentation at GDC 2017. (2017). https://www.gdcvault.com/play/1023960/How-Game-Developers-Define Accessed: September, 2019.

[28] Chad Sapieha. 2019. Far Cry New Dawn Review: Eminently Playable, Predictably Repetitive. (2019). https://business.financialpost.com/technology/gaming/far-cry-new-dawn-review-eminently-playable-predictably-repetitive Accessed: September, 2019.

[29] Jesse Schell. 2014. *The Art of Game Design: A Deck of Lenses* (2nd ed.). Schell Games.

[30] Jason Schreier. 2015. Metacritic Matters: How Review Scores Hurt Video Games. (2015). https://kotaku.com/metacritic-matters-how-review-scores-hurt-video-games-472462218 Accessed: September, 2019.

[31] Magy Seif El-Nasr, Heather Desurvire, Lennart Nacke, Anders Drachen, Licia Calvi, Katherine Isbister, and Regina Bernhaupt. 2012. Game User Research. In *CHI'12 Extended Abstracts on Human Factors in Computing Systems*. ACM, New York, NY, USA, 2679–2682. DOI: http://dx.doi.org/10.1145/2212776.2212694

[32] Matt Streit. 2018. ROI of UX: Making the Business Case. White paper. (2018). https://docs.google.com/document/d/1J3jRbqKUjlwt4el82KL3k-Z5y6mGHOOJY8jbrusYW8k/mobilebasic Accessed: January, 2020.

[33] José P. Zagal, Amanda Ladd, and Terris Johnson. 2009. Characterizing and Understanding Game Reviews. In *Proceedings of the 4th International Conference on Foundations of Digital Games*. ACM, New York, NY, USA, 215–222. DOI: http://dx.doi.org/10.1145/1536513.1536553

[34] José Pablo Zagal and Noriko Tomuro. 2013. Cultural Differences in Game Appreciation: A Study of Player Game Reviews. In *Proceedings of the 8th International Conference on Foundations of Digital Games*. 86–93.

[35] Veronica Zammitto, Pejman Mirza-Babaei, Ian Livingston, Marina Kobayashi, and Lennart E. Nacke. 2014. Player Experience: Mixed Methods and Reporting Results. In *CHI '14 Extended Abstracts on Human Factors in Computing Systems*. ACM, New York, NY, USA, 147–150. DOI: http://dx.doi.org/10.1145/2559206.2559239