Exploring Severity of Gameplay Issues from Players' Perspective

Nour Halabi
Ontario Tech University
nour.halabi@uoit.ca

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

Lena Uszkoreit*
Ontario Tech University
lenau@google.com

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

ABSTRACT

Understanding how different players experience gameplay is of vital importance in game development to ensure that the games are enjoyable and rewarding for a diverse audience. Since each user research method has its own strengths and weaknesses, one of the key questions for user researchers is to determine which methods to use in their study design. In this paper we studied the severity of first-person shooter game issues identified using Electrodermal activity based biometrics in comparison to observation and in relation to the type of issues. Our results provide an understanding and a supporting argument about the importance of severity from a player's perspective. This incremental work provides an important contribution to the Games User Research field by aiding data-driven decision making during the game evaluation process.

CCS CONCEPTS

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

KEYWORDS

games user research, severity, methods

ACM Reference Format:
Nour Halabi, Pejman Mirza-Babaei, Lena Uszkoreit, and Günter Wallner. 2020. Exploring Severity of Gameplay Issues from Players' Perspective. In *Extended Abstracts of the 2020 Annual Symposium on Computer-Human Interaction in Play (CHI PLAY '20 EA), November 2–4, 2020, Virtual Event, Canada.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3383668.3419932

1 INTRODUCTION

Games User Research (GUR) has become an essential part of the game development process. It describes the evaluation of the player experience (PX) in terms such as enjoyment, accessibility, and usability [7, 28] to help developers identify and fix design issues before release and to bring the game closer to its design intent. According to Medlock [11], GUR can encompass a variety of user research methods that range from expert studies (e.g., heuristic evaluations) to surveying players and analyzing in-game data. Central to GUR is the playtest in which players are invited to play (an often pre-released) game prototype in a lab or remotely to generate data for analysis (e.g., gameplay videos or answers to survey questions). Several user research methods are often used to gather data during playtests. Understanding the advantages and limitations of the different user research methods is critical for games user researchers as they need to design studies and select appropriate mixed-methods designs for a playtest based on their evaluation focus (e.g., evaluating a tutorial level or balancing game difficulty) and other considerations they may have to take into account (e.g., budget, development time).

Our previous work [13] showed that Electrodermal activity (EDA) based biometrics approaches (triangulating EDA data to structure post-session interviews) may help researchers to identify more latent issues related to players' immersion and gameplay experience compared to using observation alone. However, it is important to stress that identifying more issues would not necessarily mean that those issues are severe enough to warrant additional time and resources to fix them. Moreover, reporting large numbers of issues can also be overwhelming and distracting for the game development team. Therefore, one of the key tasks of user researchers is to decide which issues are severe enough to be prioritised and reported to the developers for fixing (cf. [12, 24]). There are a number of factors involved in making this decision, some based on available resources (e.g., availability of staff members or time left in production) and some are based on the effect the issue may have on PX (e.g., severity and impact of the issue on gameplay). Hence, an important step in the GUR process is to decide on issue severity in order to prioritize them and assist developers in allocating their resources in a best way to optimise the game [12, 24]. In this paper, we are looking to take a step back and investigate issue severity from another perspective: The perspective of players who did not personally participate in a playtest as testers but are seasoned players of similar games. Obviously, these players cannot identify issues in a game they have never played. They can, however, be used as a group of experienced users to provide us with an idea of which issues players generally rate as severe.

To study this question, we use the same dataset (gameplay issues identified through a series of playtests) that was created by Mirza-Babaei et al. [13]. In the paper at hand, we focus to understand the severity of those issues from the players' perspective. This is an important contribution as it opens the discussion around the

*Also with User Experience Research, Google Zürich, Switzerland.

CHI PLAY '20 EA, November 2–4, 2020, Virtual Event, Canada

ACM ISBN 978-1-4503-7587-0/20/11.
https://doi.org/10.1145/3383668.3419932

severity of issues identified using different user research methods and in relation to the type of issues (e.g., gameplay, usability).

2 BACKGROUND AND RELATED WORK

To decide on issue severity, user researchers often need to rely on data they collect during playtesting which directly depends on the user research methods they employ. For example, when using observation-based approaches, researchers may observe and record how many times players would experience a certain issue [23]. Similarly, in using self-reporting methods (e.g., interview, questionnaires) researchers may look at how players rate certain issues in a survey or how much weight they give to specific issues in their interview comments [19].

Nielsen [20] provided a scale, ranking issues from 0 (a *non-issue*) to 4 (a *usability catastrophe*) mainly using criteria such as the frequency in which the issue occurs, its persistence, and the impact of the problem, i.e. how easy it will be for the user to maneuver around this problem. Moreover, Molich and Dumas [16] conducted a series of studies investigating how reliable usability reviews and ratings are for a website. They instructed 17 teams of usability professionals to analyze the user experience of a hotel website. Across all teams, they identified 340 issues with only nine of these issues being reported by more than half of the teams. No two teams discovered the same issues. In addition, due to the lack of agreement between evaluators, they were found to neither agree with other evaluators regarding the severity of the issues nor with the evaluators' expert reviews [16].

Stråat et al. [27] studied if violations found by a set of heuristics had an effect on the assessed quality of video games. They examined design issues of 10 low and high scoring games, determining frequency of issues, heuristics violations the games had, and assigned a severity rating for the found issues through a heuristic evaluation conducted by four evaluators. Their findings revealed that the severity rating was indicative of the quality of the games as determined by their Metacritic[1] scoring. However, these severity ratings were again presented from the evaluators' perspectives, with the authors acknowledging that their own evaluation may differ from what a player may experience or find annoying.

In a similar attempt, Stråat and Warpefelt [26] studied which design heuristics developers should prioritize during the design process based on the idea that certain heuristics may be more important from a user's perspective to adhere to than others [25, 27]. They approached their investigation by presenting a survey to experienced players and instructed them to classify the heuristics into a *Motivator* or a *Hygiene Factor* based on Herzberg's two-factor theory [6].

Our work, in contrast, aims to explore the severity of issues by using predetermined gameplay issues (taken from [13]) that have been categorized based on the user research methods by which they were detected (biometrics-based and observation-based as explained in [13]), and the type of issues based on PLAY categories [3]. Better understanding of issue severity from a player's perspective is not only a goal for academics but also for the games industry [8, 14, 17] as it can translate towards optimizing GUR processes, assist user researchers in selecting appropriate methodologies for designing their evaluation sessions, and impact the final product by prioritizing issues to fix.

[1]Metacritic is a review aggregator that gathers reviews from different sources to create an average score. http://www.metacritic.com/ (Accessed: April, 2020)

3 STUDY DESCRIPTION

Our study in this paper was guided by two overall goals:

(1) First, we were interested in comparing between the severity of issues identified using observation-based and EDA-based biometrics evaluation approaches.
(2) Second, we wanted to investigate which broader types of issues (based on the PLAY categories such as gameplay and usability as described in [3]) do more severe issues fall under.

We approached these goals by preparing an online survey consisting of a set of predetermined gameplay issues taken from [13]. In the following, we outline the details of our survey design, our participants, and measured variables.

3.1 Dataset

The dataset consisted of 89 unique issues identified as a result of a series of playtests on two commercial first-person-shooter (FPS) games in [13] using different methods. Specifically, the following information was taken from [13]: (1) list of issue descriptions, (2) how the issues were found (either observation-based, EDA-based biometrics[2], or both methods), and (3) their assignment to one of the three overall PLAY categories adapted from [3], that is Cat1: Gameplay, Cat2: Coolness, Entertainment, Humour, & Emotional Immersion, Cat3: Usability & Game Mechanics.

We then reviewed all 89 issues (and, if required, rephrased them slightly) to prepare them for our survey. This step took place to ensure that the issues would be understandable to a broader group of players participating in our survey and to have a uniform language and presentation. We refrained from making major changes so that the unique issues arising from the playtesting approaches used in [13] were preserved. We also did not modify the PLAY categories they were assigned to. Overall, we adapted 86 issues (three were omitted due to their similarity to other included issues) for our survey.

To avoid fatigue and dropouts, we decided to only present each participant with a subset of the issues. We separated the total set of issues (86) into 3 subsets each containing about 35 issues to be rated. Of those 35, 10 randomly chosen issues were included in each subset as a baseline, while the other (also randomly assigned) 25 (± 2) were unique to each subset. With the baseline measure we aimed to minimize potential bias arising from ordering effects or responding to issues in comparison to others (e.g., being presented with a subset of highly severe issues would possibly prompt participants to rate still critical issues less severely in comparison to others).

3.2 Survey

The survey was administered using Qualtrics [21], participants were recruited through our lab's social media channels (such as *Twitter*, *Facebook*, *Reddit*, and *Discord*) as well as *Mechanical Turk*. The structure of the survey was as follows: (1) online consent form,

[2]EDA data was used to visually identify game events that caused a physiological reaction by the participants. These game events were used to structure post-session interviews to inquire about the participants' experiences during those events.

(2) basic demographic information and experience with games, including the questions to screen for the inclusion criteria of fluency in English and sufficient experience with games in general, (3) briefing scenario (asked to imagine they are playing an FPS game), and (4) rating 35 issues of the randomly selected subset on a scale from 0 (*not a problem at all*) to 5 (*makes the game unplayable*), implemented as a slider for convenience.

3.3 Participants

We recruited players of different experience levels ranging from self-reported levels of novice to expert. Participants needed to have at least played video or computer games for over a year and with a frequency of at least several times per month. This was important to make sure participants had familiarity with the FPS genre to understand the issues presented in our survey.

We received 164 survey responses of which 154 met the inclusion criteria. This means that every non-baseline issue was rated by at least 45 participants (more than 50 for most issues) and each baseline issue was rated 154 times. There were 103 males and 48 females with the remaining participants defining themselves as non-binary/non confirming. Mean participant age was 29 (STD = 7.8, range 18 to 58 years). 84 (54%) played video games every day and 64 respondents (41%) said they were advanced gamers. 93 participants (60%) said they played FPS regularly.

Predictor	B	OR	95% CI	p
PLAY Category				
Cat1		*— reference —*		
Cat2	.129	1.138	[-.084, .342]	.235
Cat3	**1.359**	**3.894**	**[1.084, 1.634]**	**< .001***
Method of Discovery				
Biometrics		*— reference —*		
Both	**.418**	**1.519**	**[.140, .696]**	**.003***
Observation	**.931**	**2.538**	**[.585, 1.277]**	**< .001***
PLAY Category × Method				
Cat3 × Biometrics		*— reference —*		
Cat3 × Both	**-.750**	**0.472**	**[-1.048, -.453]**	**< .001***
Cat3 × Observation	**-1.205**	**0.300**	**[-1.545, -.865]**	**< .001***
Cat2 × Biometrics		*— reference —*		
Cat2 × Observation	.022	1.022	[-.357, .401]	.910
Cat1 × Biometrics		*— reference —*		
Cat1 × Both	**.750**	**2.118**	**[0.453, 1.048]**	**< .001***
Cat1 × Observation	**1.205**	**3.337**	**[0.846, 1.545]**	**< .001***

Note: B = coefficient, OR = odds ratio (calculated as e^B)
CI = confidence interval, * significant at $\alpha = .01$
reference indicates the reference category chosen for the GEE model

Table 1: Results of the GEE model predicting the effect of method of discovery and PLAY category on participants' severity rating of an issue (significant results are bold).

4 ANALYSIS & RESULTS

To check if there are differences in the responses to the three subsets we conducted individual Kruskal-Wallis tests on the 10 baselines issues, using a Bonferroni corrected α-level of .005 to account for multiple comparisons. The tests showed that there were no statistically significant difference in response values between the different groups of participants, with $p > .005$ for each issue. As such we do not distinguish between the three participant subsets in the remaining analysis.

To assess the effect of method of discovery (Observation, Biometrics, Both) and PLAY categories (Cat1, Cat2, Cat3) on the participants' severity responses we calculated a generalized estimation equations (GEE) model using an ordinal logistic response scale, issue ID as within-subject variable, method and PLAY category as factors (categorical predictors), and considering main effects as well as two-way interactions. We have opted for a GEE model due to our data not being normally distributed, being measured on an ordinal scale, and to account for the fact that participants rated multiple issues. GEE model estimates are summarized in Table 1. GEE is a regression technique to predict the dependent variable based on the independent variables. For that purpose, GEE uses a reference category – marked as 'reference' in Table 1. In the following, we discuss the results obtained with the GEE model in more detail.

Method of discovery, PLAY category, and the interaction of these two variables all had a statistically significant effect on the prediction of the severity of the issues (Wald $X^2(2) = 46.180$, Wald $X^2(2) = 133.787$, and Wald $X^2(3) = 75.297$, respectively, all with $p < .001$).

The results reveal an interaction effect between the PLAY category and the method of discovery. Specific to Cat3, Biometrics was more successful than Observation and Both in identifying severe issues, and observation received significantly lower odds of being rated more severe to issues discovered by Both. In case of Cat2 the interaction effect between Biometrics and Observation was not significant. Lastly since Cat1 was chosen as a reference category in Table 1, the GEE model was recalculated using another reference category to explore CAT1 interactions, showing that for issues belonging to Cat1, Observation was more successful than Biometrics and Both in identifying severe issues, and both having significantly higher odds of being rated more severe than issues identified by Biometrics.

Regarding the significant main effect of method of discovery, averaged across all categories issues identified by Both methods ($p = .003$) and Observation ($p < .001$) have significantly higher odds of being rated more severe than issues identified through Biometrics alone. The main effect of PLAY category shows that independent of the method of discovery, that issues categorized as Cat3 have significantly higher odds of being rated more severe than issues categorized as Cat1 ($p < .001$). Issues belonging to Cat2, on the other hand, were considered to be of similar severity than those from Cat1 ($p = .235$).

Descriptive statistics of the severity ratings for the different methods and categories can be found in the supplementary material.

4.1 Severity of issues per PLAY Category

To better understand participants' responses towards their perceived severity of issues, we group similar issues and discuss low and high severity issues below. We consider issues with a median rating of 0 to 2 as low and 3 to 5 as high. Issues were regarded as moderate when the median is 2.5 and having mixed ratings when the IQR (Q3 - Q1) is greater than two. In the following, we organize our results under the three PLAY categories to examine the

differences between the two methods of discovery in more depth. For a detailed break-down of the scores (and description) for each issue please refer to the notched box plots and issue list in the supplementary material.

Cat1 – Gameplay. In terms of Cat1, we observed that a subset of the issues that were considered to have low severity (e.g., Q25_1, Q25_2, Q26_7) were related to players' experience over time, revealing that players do not necessarily mind the idea of having to perform repetitive or dull tasks.

Additionally, we observed that issues related to pacing and balancing of challenges received mixed ratings with some being regarded as moderate (e.g., Q33_2) and mixed severity (Q32_8) and others as low severity (e.g., Q27_11, Q32_5, QR_4) – these issues were all discovered using biometrics with the exception of QR_4 which was discovered by both methods.

Cat2 – Coolness, Entertainment, Humour, & Emotional Immersion. In terms of Cat2, we observed that issues pertaining to the use of audio and visual content for promoting immersion were common. However, they received mixed severity ratings. For example, two severely rated issues, one – found by observation – was concerned with a lack of audio and visual content (Q28_3) and another – identified through biometrics – occurred when there was a mismatch between the setting of the world and the narrative description of it (Q33_1). However, issues were leaning towards low severity in cases when players' expectations did not match the game world interactions (e.g., QR_9, Q32_10, Q32_3) or real life physics (Q32_2) – all these issues were revealed through biometrics. Biometrics discovered more issues in this category, but only Q33_1 was of higher severity. Observation, on the other hand, discovered less issues, with only Q30_1 being of low severity.

Cat3 – Usability & Game Mechanics. Regarding Cat3, the first set of issues that were deemed highly severe were related to the user interface (UI) with 1) UI descriptions being too extensive, confusing, or unnecessary (e.g., Q25_8, Q27_5, discovered by both), and 2) UI art being unrecognizable or not clear (e.g., Q26_4, Q25_3, discovered by observation). There were also two issues (Q30_10, QR_8) which were related to UI elements fading out during game events. These issues were of low severity with the first being discovered by observation and the second by both.

The second set of issues considered highly severe were associated with feedback, with five issues dealing with audio feedback such as a lack of distinctive audio feedback (Q27_4) and voices (Q29_6), a mismatch between audio cues and information provided (Q25_7), as well as when the audio was not clear and overshadowed by other audio sources while lacking supportive text (Q27_7, Q30_3). All audio feedback issues were discovered through both methods and were of higher severity. Additionally, feedback issues with high severity were related to missing or delayed messages (Q28_4, found by observation; Q26_5, found by both) and when completing a task or action successfully (Q29_3, discovered by observation). The third set of issues related to when feedback occurred at a time when a player is occupied or performing a fast-paced action (Q28_6 found by both) – this was of high severity – and when there were unclear instructions (Q26_9, found by both; QR_6 and Q32_6, found by biometrics) – they were of high, high, and low severity respectively.

When it came to understanding controls, results surprisingly show that issues considered to have low severity are issues that came up with players not knowing or being confused with the controls (e.g., Q27_9, found by biometrics; Q26_2 and Q30_5, found by observation). However, not understanding how to use resources or perform a specific necessary task such as healing was regarded as severe (e.g., Q26_1, found by observation).

In terms of biometrics, the remaining issues not mentioned yet and pertaining to this category were all of high severity and were related to enemies being difficult to spot in a scene due to their visual placement (Q33_3, Q27_10). Another issue was concerned with the controls feeling unnatural and uncomfortable to use (Q33_5). On a similar note, both methods were able to discover a slightly different issue in terms of controls which had to do with controls being easy to misclick (Q30_8), which was rated higher in severity.

5 DISCUSSION

In the following, we will discuss our results in relation to how severity relates to the PLAY categories, the method of discovery, and the interaction of the two.

5.1 Severity based on Issue Category

Usability and user experience (UX) are two highly relevant and interlinked areas in GUR that are studied and assessed through various methods to improve the overall game design [1]. Our results offer a comparison between these two important areas through the distinction in the severity ratings of the three categories. Usability is reflected by issues of the Usability & Game Mechanics category (Cat3), while UX is reflected by the other two categories: Gameplay (Cat1) and Coolness, Entertainment, Humour, & Emotional Immersion (Cat2).

Overall, Usability & Game Mechanics (Cat3) issues were regarded as more severe compared to Gameplay (Cat1) and Coolness, Entertainment, Humour, & Emotional Immersion (Cat2) issues. This could be due to the subjective nature of UX [2], reflected by the varying levels of severity as presented in Section 4.1 (see also the supplementary material). Additionally, Cat1 had lower and mixed severity results when it came to challenge and pacing issues which can be explained by players desiring a certain degree of difficulty for strategy and balancing of challenge portraying a more positive outlook compared to viewing these as undesirable issues [3].

Furthermore, McAllister and Long [10] presented a layered model used in the industry for when to apply different user research methods. The layers depict the dimensions of play experience and how each layer has an impact on the other. Two of the layers correspond to usability and PX. They showcase how usability is a predecessor to a positive PX, meaning that by addressing usability they are able to impact and optimize the PX. Based on our overall results, issues concerned with usability were perceived more severe compared to issues concerned with UX, reinsuring that usability is an important aspect that should not be overlooked as professionals move from usability towards prioritizing UX [22].

5.2 Severity based on Method of Discovery

A number of works (e.g., [4, 9, 18]) has investigated biometrics-based techniques with the interest of utilizing them and understanding their benefits within the GUR field. The results indicate that severity ratings differ depending on the method used for discovering the issues. Overall, issues only found by observation or issues found by both (observation and biometrics) were regarded as more severe compared to issues found through biometrics alone in general when the type of issue is not taken into consideration. This is an interesting finding revealing that a mixed-methods approach is a viable option for providing researchers with a comprehensive set of results that are also perceived as more severe. However, further research using different biometric sensors and examining different genres of games will allow for a deeper understanding as to where the use of biometrics is the most successful.

5.3 Severity based on Method in each Category

Previous work by Mirza-Babaei et al. [13] quantified the difference between observation and biometrics-based approaches for exposing unique issues across the three categories. We rely on their results to better assess the different methods in combination with severity. Our results indicate that there is a difference in the severity rating of issues based on the type of method they were revealed by and the PLAY category the issues belong to.

Firstly, in case of Cat1 issues, Mirza-Babaei et al. [13] showed that a biometrics-based approach identified more issues in this category. However our results show that the issues found by observation were regarded as having higher severity. Secondly, in case of Cat2 issues, while they noted that biometrics found more issues, our results showed that severity ratings were regarded as similar between biometrics and observation. Lastly, in case of Cat3 issues, while they highlighted that an observation-based approach detected the majority of issues, our results show that the issues found by biometrics were regarded as having higher severity.

Referring back to McAllister's and Long's work [10], their layered model of game evaluation demonstrates its significance by providing developers with a framework for when to use certain GUR methods during development. For example, in order to improve game usability, they suggest applying the following evaluation methods: UX Competitor Analysis, Collaborative Design, Usability Expert Analysis, and Usability Playtest. This approach aligns with our investigation on how changing the methods used for discovering issues can reveal variations in severity and issues found, and how severity varies based on the type of issue. As part of their model, they suggest that observation can be applied as part of the usability playtest method. However, biometrics was not considered for assessing usability whereas our findings offer new insight into how biometrics is able to discover severe usability issues.

These findings offer an interesting contribution to the field, offering a new perspective as they allow researchers to choose an efficient method tailored for their specific design questions which, in turn, optimizes their GUR process.

5.4 Why focusing on Severity?

Game development is a laborious and resource-intensive process. One of the key responsibilities of a game producer is to manage the resource allocation and optimize the process to achieve the best possible result while maintaining the budget and production timeline. GUR, as an embedded part in game production, has a direct impact on the project's resources, hence it strives to support game production by providing the needed feedback (often in form of changes to be made to the game) in a timely and cost-effective manner [5]. Therefore, supplying developers with information on the severity of an issue will be beneficial in understanding if and how to tackle a problem, and when to tackle it.

In the work of Mirza-Babaei et al. [15], developers who were participants of the *Biometric Storyboard* tool study described that they want a tool that provides them with information on the order of issues to be fixed based on their severity or priority. The results of the study also showed that trust is an important part in determining whether developers will address a certain problem or not. If developers are not aware of the severity of an issue and how many users it affects, then the issue is likely left unaddressed [14]. This is also evident in the work of Stråat et al. [27] where their assessment of published games revealed that low scoring games contained a high degree of severe issues that were not addressed during the development process. Our work offers detailed insights into why studying issue severity is an important aspect to be considered. By approaching the rating of severity from the perspective of the players, issues can be prioritized more adequately allowing developers to enhance the game experience more effectively and goal-directed.

5.5 Limitations and Future Work

In order to perform our study, we relied on data from [13]. We used the same issues and their assigned categories in our paper. Therefore, our results are based on the gameplay issues identified and categorized by other researchers. Preferably, a follow-up study which implements an updated data gathering phase involving more participants is ideal. Including more varied game genres that could expand the list of issues and build upon this work is advisable as well. Moreover, the fact that the issues were driven by two FPS games may impact generalizability to some degree, but we expect that our findings could be extended to other genres as well. Additionally, as the focus was on examining observation and biometrics (in the form of EDA), other user research methods, especially questionnaires, and other biometric measures can be investigated in future work.

Since the issues were associated with individual gaming experiences, a future area of research can examine the differences in issue severity between individual and social gaming experiences. This can also include the differences in severity between novice and expert PX. Lastly, this work can impact and be of benefit for visualizations aimed at delivering actionable insight to developers. In that sense, techniques for classifying and prioritizing problematic game areas can be adapted from our findings to support this direction of work.

6 CONCLUSIONS

The study presented in this paper supports employing mixed-methods evaluation approaches in GUR. However, to maximize the benefits of understanding and finding high severity issues, tailoring the method of discovery towards the type of issue is suggested. Our results show that a simple observation-based approach can

reveal severe issues in *gameplay experience* whereas biometrics-based is distinguished in finding severe *usability issues*. Arguably, a successful game would achieve both usability and user experience goals to keep the player engaged and entertained. Our findings can help user researchers on deciding when to invest in utilising a complex biometrics-based approach and when they can answer their evaluation goal with a observation study. This is an important contribution as game development resources are limited and optimizing the development process is a vital responsibility for game companies.

Assessing user research methods based on the severity of issues they can identify brings a new perspective to GUR that was not previously available to developers and researchers. We believe our key contribution is introducing this new lens for selecting research methods in evaluation settings accordingly.

REFERENCES

[1] Omar Álvarez-Xochihua, Pedro J Muñoz Merino, Mario Muñoz Organero, Carlos Delgado Kloos, and José Ángel González-Fraga. 2017. Comparing Usability, User Experience and Learning Motivation Characteristics of Two Educational Computer Games. In *International Conference on Enterprise Information Systems*. SciTePress, Setubal, Portugal, 143–150.

[2] Eduardo H. Calvillo-Gámez, Paul Cairns, and Anna L. Cox. 2015. Assessing the Core Elements of the Gaming Experience. In *Game User Experience Evaluation*, Regina Bernhaupt (Ed.). Springer, Cham, 37–62. https://doi.org/10.1007/978-3-319-15985-0_3

[3] Heather Desurvire and Charlotte Wiberg. 2009. Game Usability Heuristics (PLAY) for Evaluating and Designing Better Games: The Next Iteration. In *Online Communities and Social Computing*, A. Ant Ozok and Panayiotis Zaphiris (Eds.). Springer, Berlin, Heidelberg, 557–566. https://doi.org/10.1007/978-3-642-02774-1_60

[4] Anders Drachen, Lennart E. Nacke, Georgios Yannakakis, and Anja Lee Pedersen. 2010. Correlation between Heart Rate, Electrodermal Activity and Player Experience in First-Person Shooter Games. In *Proceedings of the 5th ACM SIGGRAPH Symposium on Video Games - Sandbox '10*. ACM, New York, NY, USA, 49–54. https://doi.org/10.1145/1836135.1836143

[5] B. Fulton, M. Ambinder, and J. Hopson. 2012. Beyond Thunderdome: Debating the Effectiveness of Different User-Research Techniques. Presented at the IGDA GUR SIG Summit 2012. Retrieved from http://vimeo.com/groups/gursig/videos/26733185, Accessed: April, 2019.

[6] Frederick I Herzberg. 1966. *Work and the Nature of Man.* World, Cleveland, OH, USA.

[7] W.A. IJsselsteijn, Y.A.W. Kort, de, K. Poels, A. Jurgelionis, and F. Bellotti. 2007. Characterising and Measuring User Experiences in Digital Games. In *Proceedings of the International Conference on Advances in Computer Entertainment Technology (ACE 2007), June 13-15, 2007*, R. Bernhaupt and M. Tscheligi (Eds.). ACM, New York, NY, USA, 1–4.

[8] 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* (Vancouver, British Columbia, Canada). ACM, New York, NY, USA, 48–55. https://doi.org/10.1145/1920778.1920786

[9] Regan L. Mandryk and Lennart E. Nacke. 2016. *Biometrics in Gaming and Entertainment Technologies.* CRC Press, Boca Raton, FL, USA, Chapter 6, 191–224. https://doi.org/10.1201/9781315317083-7

[10] Graham McAllister and Sebastian Long. 2018. A Framework for Player Research. In *Games User Research*, Anders Drachen, Pejman Mirza-Babaei, and Lennart E. Nacke (Eds.). Oxford University Press, Oxford, UK, 281–299. https://doi.org/10.1093/oso/9780198794844.003.0017

[11] Michael C. Medlock. 2018. An Overview of GUR Methods. In *Games User Research*, Anders Drachen, Pejman Mirza-Babaei, and Lennart E. Nacke (Eds.). Oxford University Press, Oxford, UK, Chapter 7, 99–116.

[12] Pejman Mirza-Babaei. 2018. Reporting User Research Findings to the Development Team. In *Games User Research*, Anders Drachen, Pejman Mirza-Babaei, and Lennart E. Nacke (Eds.). Oxford University Press, Oxford, UK, 323–332. https://doi.org/10.1093/oso/9780198794844.003.0017

[13] Pejman Mirza-Babaei, Sebastian Long, Emma Foley, and Graham McAllister. 2011. Understanding the Contribution of Biometrics to Games User Research. In *Proceedings of DiGRA*. DiGRA/Utrecht School of the Arts, Tampere, Finland, 329–347.

[14] Pejman Mirza-Babaei, Lennart Nacke, Geraldine Fitzpatrick, Gareth White, Graham McAllister, and Nick Collins. 2012. Biometric Storyboards: Visualising Game User Research Data. In *CHI '12 Extended Abstracts on Human Factors in Computing Systems* (Austin, Texas, USA). ACM, New York, NY, USA, 2315–2320. https://doi.org/10.1145/2212776.2223795

[15] Pejman Mirza-Babaei, Lennart E. Nacke, John Gregory, Nick Collins, and Geraldine Fitzpatrick. 2013. How Does It Play Better? Exploring User Testing and Biometric Storyboards in Games User Research. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1499–1508. https://doi.org/10.1145/2470654.2466200

[16] Rolf Molich and Joseph S. Dumas. 2008. Comparative Usability Evaluation (CUE-4). *Behav. Inf. Technol.* 27, 3 (May 2008), 263–281. https://doi.org/10.1080/01449290600959062

[17] Pablo Moreno-Ger, Javier Torrente, Yichuan Grace Hsieh, and William T. Lester. 2012. Usability Testing for Serious Games: Making Informed Design Decisions with User Data. *Advances in Human-Computer Interaction* 2012 (2012), 13. https://doi.org/10.1155/2012/369637

[18] Lennart E. Nacke. 2013. An Introduction to Physiological Player Metrics for Evaluating Games. In *Game Analytics: Maximizing the Value of Player Data*. Springer, London, UK. https://doi.org/10.1007/978-1-4471-4769-5{_}26

[19] Lennart E Nacke. 2015. Games User Research and Physiological Game Evaluation. In *Game user experience evaluation*. Springer, Cham, Switzerland, 63–86. https://doi.org/10.1007/978-3-319-15985-0_4

[20] Jakob Nielsen. 1994. Severity Ratings for Usability Problems. https://www.nngroup.com/articles/how-to-rate-the-severity-of-usability-problems/

[21] Qualtrics. 2020. Qualtrics. www.qualtrics.com Accessed: July 2020.

[22] Cristian Rusu, Virginica Rusu, Silvana Roncagliolo, and Carina González. 2015. Usability and User Experience: What Should We Care About? *International Journal of Information Technologies and Systems Approach* 8, 2 (2015), 1–12. https://doi.org/10.4018/IJITSA.2015070101

[23] Mirweis Sangin. 2018. Observing the Player Experience: The Art and Craft of Observing and Documenting Games User Research. In *Games User Research*, Anders Drachen, Pejman Mirza-Babaei, and Lennart Nacke (Eds.). Oxford University Press, Oxford, 175 – 188. https://doi.org/10.1093/oso/9780198794844.003.0011

[24] Natalie Selin. 2019. An Analyst's Guide to Communication. In *Data Analytics Applications in Gaming and Entertainment*, Günter Wallner (Ed.). CRC Press, Boca Raton, FL, USA, 205–221.

[25] Björn Strååt and Harko Verhagen. 2014. Vox Populi – A Case Study of User Comments on Contemporary Video Games in Relation to Video Game Heuristics. In *GameOn 2014: Simulation and AI in Computer Games*. EUROSIS, Ostend, Belgium, 5–9.

[26] Björn Strååt and Henrik Warpefelt. 2015. Applying the Two-Factor-Theory to the PLAY Heuristics. In *Proceedings of DiGRA 2015: Diversity of play: Games – Cultures – Identities*. DiGRA, Tampere, Finland, 12.

[27] Björn Strååt, Fredrik Rutz, and Magnus Johansson. 2014. Does Game Quality Reflect Heuristic Evaluation? Heuristic Evaluation of Games in Different Quality Strata. *Int. J. Gaming Comput. Mediat. Simul.* 6, 4 (2014), 45–58. https://doi.org/10.4018/ijgcms.2014100104

[28] Veronica Zammitto. 2018. Games User Research as Part of the Development Process in the Game Industry: Challenges and Best Practices. In *Games User Research*, Anders Drachen, Pejman Mirza-Babaei, and Lennart E. Nacke (Eds.). Oxford University Press, Oxford, UK, Chapter 2, 15–30.