What makes a good game? Using reviews to inform design

Matthew Bond
Advanced Interaction Group
School of Computer Science
University of Birmingham, Edgbaston
Birmingham, B15 2TT, UK
+44 (0) 121 414 3729
ug56msb@cs.bham.ac.uk

Russell Beale
Advanced Interaction Group
School of Computer Science
University of Birmingham, Edgbaston
Birmingham, B15 2TT, UK
+44 (0) 121 414 3729
R.Beale@cs.bham.ac.uk

ABSTRACT

The characteristics that identify a good game are hard to define and reproduce, as demonstrated by the catalogues of both successes and failures from most games companies. We have started to address this by undertaking a grounded theoretical analysis of reviews garnered from games, both good and bad, to distil from these common features that characterize good and bad games. We have identified that a good game is cohesive, varied, has good user interaction and offers some form of social interaction. The most important factor to avoid is a bad pricing. Successfully achieving some of these good factors will also outweigh problems in other areas.

Categories and Subject Descriptors

H.1.2 User/Machine Systems

General Terms

Design, Human Factors.

Keywords

Games, grounded theory, analysis

1. INTRODUCTION

At present, video game development accounts for billions of dollars of US expenditure, and US video game companies directly employ 24,000 people. The market for games has also been growing disproportionately fast [10]. Clearly, not all of these games have the same degree of success; a game needs certain properties to make it successful and to make the most of this growing market. Gamers often rely on review sites to make purchasing decisions, and so a game that receives a high average review score will have considerably better sales than a bad game. Therefore it is desirable to know what makes a game get a good review rather than a bad or mediocre review.

The aim of this paper is to determine what earns a game a good review, and, by extension, what features should be prioritised during the production of a game to make it successful.

This paper tries to determine practical categories which should be considered during game development and, to an extent, marketing and distribution. It is different from most of the other approaches in that it uses Grounded Theory [5] to develop a hypothesis from the data source, instead of using an existing hypothesis. We determine the elements that characterize good and bad games from an analysis of the reviews of a wide range of games, providing a bottom-up, practical approach to determining factors that make for a decent gaming experience.

2. RELATED WORK

Csikszentmihaly's paper on optimal experience (Csikszentmihaly, 1991) first used the word flow. Flow describes a phenomenon experienced in many fields, often described as "being in the zone". While experiencing flow, a person will feel absorbed, lose sense of time, and focus entirely on the task in hand. This state can be brought on by an intrinsically rewarding task with clear goals, immediate feedback, balanced difficulty and a sense of control over the situation.

GameFlow, Sweetser & Wyeth's [11] adaptation of Csikszentmihaly's model of flow [1] for game experience, groups game-design criteria into each of these categories, and shows that gaming is an example of flow experience. Their approach makes for an interesting comparison, since it takes a top-down approach, but emerges with similar conclusions to ourselves.

Our own work developed in the opposite direction to GameFlow, with no prior knowledge of flow theory. Both GameFlow and our categories share similar 'leaf node-criteria'; it is the higher level categories which are significantly different. Also, some of the specificity of our leaf nodes is not captured by GameFlow's criteria.

As such, we see our categories as a more practical abstract level approach to game design than GameFlow, due to the more domain specific terminology used. However the same low level details were revealed by both studies.

Both GameFlow and our own work identified social interaction to be a key part of gaming experience that was unaccounted for by traditional flow theory. Whether this is something that should be revised into flow or whether it is entirely domain specific is still an outstanding research question. The GameFlow model never brings into question the price of a game, something we found to be of critical importance, although it is debatable whether price contributes to the design process of game development.

Kristian Kiili's work on Game-based learning [3] applied the idea of flow to educational game design. It determined gameplay to be the most important part of game design, with storytelling, graphics, sound and balance to be auxiliary factors.

Federoff [3] also outlines gameplay as the most important factor, followed by 'game mechanics'. Interestingly, our findings contradict both of these studies in this respect.

Back in the early 1980's, Malone [7] determined three criteria for the creation of an enjoyable game and discussed applying his findings to the creation of educational games. He found challenge, fantasy and curiosity to be the most important features in good game design. His findings are less relevant today, overlooking the importance of variety and a social aspect, and the importance and meaning of 'fantasy' is questionable.

In Simulation and Gaming, Myers' identifies challenge, social and meditation as the core elements of a game [9]. These findings, especially the importance of a social aspect, tie in well with our own findings.

3. METHODOLOGY

Grounded theory is an empirical research strategy which works from some raw data to form hypotheses about some situation [2]. In this case, the exploration was to discover what makes a game 'good' or 'bad'.

Some introductory work was done to find an acceptable data source. The criteria for the source were that it had to be up to date and have numerical scores, as well as having some form of text for coding. It is assumed here that a successful game is one that will sell a lot of copies: however, sales data for games is not published for many games, and has no body of text from which to find a hypothesis, so we need to find a more detailed source than just sales figures.

By comparing sales of Playstation 2 games as published by VG Chartz [12] with their respective review scores from Metacritic [8] there was a Pearson correlation of 0.3289 (i.e. reasonable strength, but not strong), suggesting that a high review score is reasonably indicative of strong sales. Metacritic is a review aggregator, and only provides a numerical score. Gamespot provide both a numerical score, and a body of text related to each score. A strong correlation between Gamespot scores and Metacritic led to the decision to use online game reviews from Gamespot UK [4] as the data source, since it was the text and phrases in the reviews that we needed to analyse.

Samples of particularly good and bad reviews from Gamespot UK were analysed and coded to the near point of saturation. Note that we are certain these categories are incomplete, because full saturation was not reached: saturation in Grounded Theory occurs when the data no longer adds to the current hypothesis.

The coded criteria for success and failure were then organised into a hierarchical structure, based on emergent relationships between those categories. This gave a set of core categories summarising success and failure. These were then strengthened by coding a second smaller sample of reviews into the hierarchy. In total 25 reviews were used for the original data coding.

Having established these core categories, the importance of the categories was determined. This was done by comparing the latest game releases at the time to the core categories, and recording whether the review mentioned a category.

4. RESULTS

4.1 Categories

For game success, 13 core categories were found, and for game failure, 12 categories were found. These give a good intuitive overview of what to aim for in game design.

Good Factors	**Bad Factors**
Gameplay	Poor gameplay
Environment	Poor environment
Storytelling	Poor storytelling
User interaction	Poor user interaction
Customisation	
Social	Lack of social
Variety	Lack of variety
Technical	Technical issues
Cohesion	Lack of cohesion
Maintenance	
Price(value for money)	Price (worthless)
Franchise	Failure of Franchise
Quantity (lots)	Quantity (little)
	Annoyance

Table 1: Good and bad factors in game design

Table 1 shows the overview of what to aim for, and what to avoid, in game design. However it offers no idea of the relative importance of each category.

Each category has a sub tree of criteria, of which the category heading is a summary. A sample of these criteria for 'good' is given in Table 2. The full versions of these trees are very large; however the table below gives a good feel for the criteria found within each category.

4.1.1 Interpretation of good and bad factors

With the exception of quantity and franchise, a "good or bad x" means that some or all of the criteria of x are covered in a good or bad way, where x is a category.

The good and the bad factors, although apparently 'opposites', have different importance data associated with them. Fulfilling the criteria for a good category may outweigh (or be outweighed) by fulfilling the criteria for the bad 'opposite' category. See the section on importance for more details.

Quantity and franchise are both unusual categories, in that the 'good-bad' divide is not as clear. Quantity refers to the amount of a category that a game has, the more of a good thing, the better it is, likewise the more of a bad thing, the worse it is. Franchise being good is defined by using the franchise well, staying true to it and improving. The bad side is betraying the franchise, adding bad elements and changing things which underpinned the original's success.

The 'opposites' of customisation and maintenance do exist, and were encountered during the importance ranking stage, but were not coded during the original coding stage. They were the inability to customise, and difficulty to maintain.

Category	Sample Criteria
Good Gameplay	Engaging, fair, balanced, progressive, fun, innovative, easy to play, hard to master, objective based, freedom, compelling, dynamic, various possible solutions
Good Environment	Impressiveness, eye catching, good lighting, lifelike effects, good soundtrack, good sound effects, good music
Good Storytelling	Mature, progressive, tense, engrossing, embedded in gameplay
Good User Interaction	Fast feedback, customisable, invisible controls, realistic, functional
Customisable	Powerful easy personalisation, character modification
Good Social Interaction	Multiplayer co-op, multiplayer competition, communication, sharing
Variety	Non linearity, choice, differences, dynamic combat, varied AI, emergent tactics, varied delivery media
Good Technical Implementation	Well designed camera, unobtrusive adverts, smooth framerate, uniformity, freedom to behave as expected
Cohesive	Seamless integration, story related to gameplay, cohesive story, consistent style
Easily Maintained	Low hardware requirements, easy to maintain, independent of external software
Price(value for money)	Value for money, cost, add on cost, hardware cost
Prior Franchise	Franchise, established genre
Quantity	Lots of good, not much bad

Table 2: Sample tree content for the 'Good game' tree

Annoyance never developed an 'opposite' despite being important to avoid. It may be that 'annoyance' is a cumulative effect of other bad factors, especially high quantities of bad criteria.

4.1.2 Relative Importance

This set of categories assumes all categories are equal, but during coding certain categories were clearly more important than others, occurring more frequently, for example. To investigate this, for 33 different, equally varied reviews, it was recorded whether a category was mentioned during a review, and a first table of importance was developed shown in Table 3.

Quantity is not included here because it is quantity in a way that is being counted to rank importance. Franchise is also not included because despite being mentioned in nearly every review, it was usually only in passing, and with no proper knowledge of the franchise we could not establish whether it was a positive or negative effect.

Gameplay and environment came out on top, being mentioned most in the reviews, for both good and bad games. These results did not quite reflect our expectations, as they put more importance on these two items in relation to other categories than we had expected. We therefore looked for a modified measure.

4.1.3 Improved measure

We suspected the complexity of the categories was causing this effect, with some categories encompassing far more criteria than others, making them far more likely to be mentioned than those with relatively few criteria. In a rough attempt to overcome this, the count was divided by the number of criteria for each category.

This produced Table 4, which was much closer to our expectations from the coding stage. We do question whether this is a valid approach. It addresses the fact that, for some categories, there are many synonyms and alternative ways to express the same higher level concept, and so some reviews will use more words to explain the same thing. However, it may also be that a review addresses subtleties and hence multiple mentions of words in one category are commenting on slightly different things, and so deserve being counted twice. The point of grounded theory is not to be quantitative, however, so we need to treat these measures as indicative, rather than precise.

	Variety		Cohesion		Social		Customisation		Technical		Storytelling		Environment		User interaction		Maintenance		Gameplay		Price	Annoyance
Positive	Negative	Positive	Negative	Positive	Negative	Positive	Negative	Positive	Negative	Positive	Negative	Positive	Negative	Positive	Negative	Positive	Negative	Positive	Negative	Positive	Negative	Negative
16	16	12	4	14	7	8	2	10	12	11	7	24	18	15	14	1	1	25	18	5	5	4

Table 3: Categories discovered through grounded theory analysis

	Variety		Cohesion		Social		Customisation		Technical		Storytelling		Environment		User interaction		Maintenance		Gameplay		Price	Annoyance
Positive	Negative	Positive	Negative	Positive	Negative	Positive	Negative	Positive	Negative	Positive	Negative	Positive	Negative	Positive	Negative	Positive	Negative	Positive	Negative	Positive	Negative	Negative
2	1	2	1	2	1	1	0	1	1	1	0	1	0	2	1	0	0	1	1	1	2	1

Table 4: Normalised categories

4.1.4 Interpretation

Table 5 summarises the our findings into 3 categories, where categories that received a '2' are of most importance, a '1' is of high significance and a '0' is relatively unimportant.

	Most importance	Moderate importance	Relatively Unimportant
Feature	Variety, Cohesion, Social, User Interaction	Customisability, technical soundness, good storytelling, good environment, good gameplay, value for money	Easy maintenance
Avoid	Bad pricing	Lack of variety, lack of cohesion, lack of social, lack of customisation, technical issues, bad user interaction, bad gameplay, annoyance	Lack of customisation, poor storytelling, poor environment, difficult maintenance

Table 5: Simplified summary of key elements in game design

If a game features an element of most importance, it will overwhelm the effect of moderate and unimportant categories. An example of this in practice is 'F.E.A.R. 2: Project Origin', which has storytelling and environment problems, but these are overwhelmed by the excellent variety and cohesion demonstrated, and so receives a decent overall score.

The four most important factors are **variety, cohesion,** a **good social aspect** and **good user interaction**. The most important thing to avoid is an unreasonable asking price.

One of the more interesting findings is that a bad environment and bad storytelling do not have a very significant impact on game success.

5. ANALYSIS

The intention of these results is to provide thought-provoking heuristics for game development. They are in no way facts, and should not be read in that way – they are summaries of human opinion, with the inherent variability that that implies.

These results could be useful to game designers, to ensure they are including the most important criteria in their game design and by implications producing a product that scores highly in reviews; to reviewers, to check they are covering the key criteria well enough in their reviews; and finally from a buyers perspective, both in assessing the expressiveness of a review, and then in assessing whether a game itself is worth buying.

These heuristics are very general. A wide range of games, with different genres, on different platforms, at different prices and from different eras was coded. The core categories seemed to apply across all of these ranges, though we feel that the importance values for each category probably vary.

One of the most unexpected findings was that gameplay was not featured as one of the most important categories to fulfil. Other work has found gameplay to be the most important factor [3, 6]. We will focus on the discussion of gameplay, however these points also apply to the other categories.

There are three interpretations of the discrepancy over gameplay importance. Gameplay may not be as important as previously thought; we have developed a different set of criteria; or there may be a mistake in our findings. It is likely that all three interpretations make a contribution.

Our criteria for gameplay are certainly different to those found in common literature. Early in the Grounded Theory coding process, cohesion and variety were criteria for each of the other categories. They were mentioned so frequently, in relation to so many things, it was quickly apparent they formed categories of their own. We removed them from all the other categories, including gameplay. This lowers the chance gameplay will be mentioned in a review because that element may well be captured in a variety or cohesion criterion. A natural effect of Grounded Theory is that the results 'make sense'. This is good, in that it does make sense, and is therefore easy to accept as right. However, this may also make it harder to dispute where it is wrong. It can also be seen as 'stating the obvious' to those with prior knowledge in the field.

The biggest fault with our application of Grounded Theory is that the categories did not reach a high enough level of saturation. Normally, when a core category becomes apparent, coding for other categories stops, and the core category is then saturated, coming back to saturate other categories later.

The aim of our work was to identify the broad categories, and not focus on one topic. Having identified 13 core categories, there was simply not enough time to saturate each of them. This means that the categories may be incomplete, and more importantly, that there may be categories missing.

Because the GameSpot resource that we used has many authors, this lessens the problem of a individual source being used for this study. However this study could easily be extended to include other review sites, and this may potentially add some

new criteria, strengthen the categories found, and better balance the importance findings.

Although the current importance ranking works as a rough estimate of importance, it was massively oversimplified, and does not use a large enough data set. The current scheme only marks whether a review does or does not include a factor. This does not seem to account for games in which a certain category has many or all criteria fulfilled, instead providing a rough estimate..

Importance also varies from criteria to criteria. If the calculation was done on a criteria level, and propagated up to the higher level categories, the findings would be more far more useful in game design, as for the most important criteria, it could be asked "Have we included/prevented x criteria", whilst also providing a high level overview.

At present the calculation does not take into account review scores. With the correct algorithm to calculate roughly how much each category contributed to the final review score, the general importance of each category could be better generated. One possible implication of an improved calculation of the relative importance criteria is that new reviews could be coded based on whether the categories are fulfilled, and a review score generated automatically for the review. It would be very interesting to see whether this is possible or reliable, and even more interesting to compare to the subjective scoring system of the GameFlow scoring system.

At present the data set for importance was 33 different reviews, however this should be far higher to improve the reliability of the importance algorithm.

It is worth noting that improving the importance rating would be futile without coding all the categories to saturation first.

6. CONCLUSIONS

The application of grounded theory to game reviews seems a worthy one, and this research provided some similar findings to prior work. It also identified some new criteria, and provided an intuitive set of high level categories for analysing games. It suggests that the most important elements of good game design are cohesion, variety, good user interaction and some form of good social interaction. The most important factor to avoid is bad pricing.

These findings provide a cursory set of heuristics for use while developing, reviewing, or buying a game. These findings should be of particular interest to new development teams that do not have the backing of a franchise, especially in the early stages of game design.

This paper is primarily intended to inspire further work in the field. Suggested work includes fully saturating the criteria for each category, and using some form of learning algorithm to determine the significance of each criteria in relation to the review score given.

If importance data could be reversed and applied to new reviews, it may provide a fairly objective method for generating review scores.

If this were successful, another possible challenge would be for existing games developers to try and generate a successful game that disobeys the importance rankings found, in an attempt to find innovative new techniques for making enjoyable games.

7. REFERENCES

[1] Csikszentmihaly, M. Flow: The Psychology of Optimal Experience. Harper Collins, New York, 1991.

[2] Dick, B. Grounded Theory: A Thumbnail Sketch. http://www.scu.edu.au/schools/gcm/ar/arp/grounded.html, March 1, 2009

[3] Federoff, M.A. Heuristics and usability guidelines for the creation and evaluation of fun in video games. . Bloomington, 2005.

[4] Gamespot. Video Game Reviews. http://uk.gamespot.com/reviews.html, February 27, 2009

[5] Glaser, B.G. and Strauss, A.L. The discovery of grounded theory. Aldine, Chicago, 1967.

[6] Kiili, K. Digital game-based learning: towards an experiential gaming model. 8 (1).

[7] Malone, T.W., What makes things fun to learn? Heuristics for designing instructional computer games. . in, (Palo Alto, California, 1980), Association for Computing Machinery.

[8] Metacritic. Game Reviews from Metacritic. http://www.metacritic.com/games/, March 1, 2009

[9] Myers, D. A Q-Study of game player aesthetics. Simulation and Gaming. 375-396.

[10] Siwek, S.E. Video Games in the 21st Century: Economic Contributions of the US Entertainment Software Industry. . Entertainment Software Association, Washington DC, 2007.

[11] Sweetser, P. and Wyeth, P. GameFlow: A Model for Evaluating Player Enjoyment in Games. 3 (3).

[12] VGChartz. VG Chartz.com. (VG Chartz). http://uk.gamespot.com/reviews.html, January 15, 2009