

Critic-Proofing: Robust Validation Through Data-Mining

Ian J. Livingston

Department of Computer Science
University of Saskatchewan
Saskatoon, SK, Canada
+1 306 966 2327
ian.livingston@usask.ca

Lennart E. Nacke

Department of Computer Science
University of Saskatchewan
Saskatoon, SK, Canada
+1 306 966 6593
lennart.nacke@acm.org

Regan L. Mandryk

Department of Computer Science
University of Saskatchewan
Saskatoon, SK, Canada
+1 306 966 4888
regan@cs.usask.ca

Abstract

Critic-proofing is a modified heuristic evaluation technique, specifically designed to provide a fine-grained, prioritized list of heuristic violations. The critic-proofing technique weights the severity of a problem based on the frequency that similar problems are found in similar games. The severity ratings are calculated using data collected from game reviews, and the severity assigned to a problem during the heuristic evaluation process. However, heuristic techniques have had limited adoption within the video game industry. One reason for this is the perceived lack of validity and robustness of game specific heuristic principles. In this paper, we introduce and outline a new data-mining project designed to validate game-specific heuristic techniques, especially the critic-proofing technique by using the popular game-review aggregation website Metacritic.

Keywords

Critic-proofing, heuristics, usability, playability, data mining.

1. Introduction

In the gaming industry - especially in casual game development - the need for a quick turnaround from playtests often does not allow for the implementation of sophisticated and advanced biometric measurement techniques (e.g., Mandryk, Atkins, & Inkpen, 2006; Nacke, Grimshaw, & Lindley, 2010), because they require personnel training and expensive tools. A possible alternative for getting reliable and robust data in playtesting is the utilization of easy-deploy inspection methods, some of which could be considered discount evaluation methods because of their low cost and short time requirements, making them ideal for application in the casual games industry. One example of a discount evaluation method is heuristic evaluation, where expert evaluators examine system actions and features that violate heuristic principles. After the inspection has been completed, a severity rating is assigned to each violation. The evaluators' ratings are then combined. Jakob Nielsen developed the most well known heuristic principles (Nielsen, 2005). His heuristics were designed for software usability, and since have been shown to be both valid and effective (Nielsen, 1994). In recent years, many heuristics collections have been developed for game specific purposes. Generally, these game heuristics are designed to evaluate usability, playability, or both (Federoff, 2002; Korhonen & Koivisto, 2007; Korhonen & Koivisto, 2006; Pinelle, Wong, & Stach, 2008a; Pinelle, Wong, Stach, & Gutwin, 2009). In some cases, attempts have been made to combine different collections of heuristic principles into more comprehensive collections of heuristics (Desurvire, Caplan, & Toth, 2004; Koeffel et al., 2010). However, as with many advanced evaluation techniques the industry uptake of these heuristics has been lackluster.

Recently, Livingston, Mandryk, & Stanley (2010) published work that used video game review data (Pinelle, Wong, & Stach, 2008b) to develop a new heuristic evaluation technique called Critic-Proofing (CP); an evaluation technique that assigns a rating to heuristic violations (called a genre rating) based on the game's genre and the heuristic being violated. The genre ratings are combined with severity ratings assigned during the heuristic evaluation to form a prioritized list of heuristic violations.

Figure 1. The metacritic website

The technique was tested in a case study, where it was used to evaluate the game usability of Capsized (AlienTrap Software, 2010), which is currently under development. Feedback was collected from the development team, who found the CP results to be much more valuable than those from traditional heuristic evaluations. In addition, the easy deployment and quick turnaround time of these heuristics would make them ideal for evaluating casual games, since shorter development times are common there. What CP currently lacks is a larger database of reviews upon which CP weightings can be based, since it currently uses only reviews from GameSpot. The natural next step to improve CP would be to use a larger aggregation of reviews. One website dedicated to aggregating and accumulating reviews for games is the website called *metacritic* (see Figure 1).

Metacritic is a review aggregation website that collates articles and scores of games, movies, TV shows, DVDs, and music. The scoring mechanism implemented by the website converts each review into a percentage before taking a weighted average and then listing different numbers of reviews. For reviews with no explicit scores, metacritic staff manually evaluates the tone of the review before assigning a subjectively fitting grade. Weighting is applied to reviews depending on the public impact of the review source, so that reviews from major periodicals may have a greater effect on the average than niche ones. The weight numbers are however not publicly available (“About Metascores,” n.d.).

Game publishers and developers are currently using metacritic in a variety of different ways. Some are using the system in contract negotiations as a metric for bonus distribution; others are using it as a management metric - in some cases using metascores to influence developer royalties, or measure of desired quality for the final game product (“Is Metacritic Damaging the Games Industry? | Edge Magazine,” n.d.).

In this paper, we will discuss our ideas for using the review base of metacritic as the foundation for a tool that validates CP and other game heuristic methods based on this large review database.

2. Requirements Analysis

Before we start the development of our game heuristic evaluation tool, we need to consider that it needs to meet certain requirements of a low-cost, easy-to-deploy, and fast evaluation tool for the game industry. For any type of evaluation technique to be valuable in the game industry the technique must meet five requirements based on our past experience in contractual work for game companies:

1. It must have demonstrable value.
2. It must be cost-effective.
3. It must be valid.
4. It must be easy to integrate into the current pipeline.
5. It must have clear and interpretable results.

The heuristics developed by Nielsen (Nielsen, 2005) meet all of these requirements. However, these heuristics were developed for the evaluation of traditional desktop user interfaces. Video games are more complex (e.g., they value process over outcome), which means that Nielsen’s heuristics can only be directly applied to game features that are similar to desktop interface features.

2.1 Game Heuristics

Some heuristics have, however, been developed specifically for games. These heuristics fall into one of three categories: (1) usability, (2) playability, or (3) a combination of both. In contrast to Nielsen's heuristics, which were developed for user interface analysis, game usability heuristics apply to the usability of the gameplay, such as camera control or artificial intelligence issues, which are fundamentally different problems than those found in traditional software development. Playability heuristics apply to the dramatic and formal elements of games. Playability heuristics might apply to problems with the game narrative or to player fatigue.

Figure 2. The technique was used to evaluate Capsized (AlienTrap Software, 2010), a 2D platform shooter game

Federoff (Federoff, 2002) did pioneering research on game heuristics. Through observations and conversations at a San Francisco game company, and a comprehensive review of literature, Federoff was able to compile a list of heuristics, which focus on three game areas: (1) interface, (2) mechanics, and (3) play. Federoff's heuristics are quite broad, while we are specifically interested in problems pertaining to game usability.

In 2004, Desurvire et al. (2004) developed a set of heuristics called Heuristic Evaluation for Playability (HEP). Although similar to Federoff's heuristics, Desurvire provided a simpler description and organization of the heuristics. Recently, Desurvire has published a newer version of the HEP called PLAY (Heuristics of Playability) (Desurvire & Wiberg, 2009). The PLAY heuristics attempt to make the underlying principles more generalizable by evaluating the principles across three different genres.

Other forms of heuristics have been developed to fill different needs of game evaluation: Korhonen et al. created heuristics for mobile multiplayer games (Korhonen & Koivisto, 2007); Pinelle et al. generalized this idea with a set of heuristics for all multiplayer games (Pinelle et al., 2009). Many of these heuristics overlap and share similarities, and some attempts have been made to

compare techniques. Korhonen, Paavilainen, & Saarenpää (2009) compared two different sets of playability heuristics – HEP and their own mobile games heuristics (Korhonen & Koivisto, 2006) – to examine the relative strengths and weaknesses of each approach. Koeffel et al. (2010) attempted to create a comprehensive list of heuristics from those found in literature, and to assess the collection's effectiveness using a comparison to video game reviews.

Pinelle et al.'s (2008a) heuristics list is specific and short—there are ten principles—making the evaluations simple and quick to perform. In addition to this, prior work has been extended in the exploration of usability specific to game genres (Pinelle et al., 2008b). These heuristics are different from others because they were developed through the analysis of game reviews found on the game review website GameSpot. According to Larsen (2008), it is reasonable to consider game reviewers as unscientific user experience evaluators with expert domain knowledge. Thus, the approach of Pinelle et al. already provides an implicit measure of validation because their heuristics are based on user reports.

However, all of these heuristic principles - with the exception of Nielsen's - are not in use within the video game industry. This is primarily because they do not fulfill the five basic requirements outlined at the start of this section. Most importantly, game specific heuristics have not been shown to be valid, clear and interpretable, or demonstrably valuable to game developers. Conversely, Nielsen's heuristics have been shown to be valid in other software fields, the results are clear and interpretable, and the value of improving user interface elements is obvious. By providing means to validate review-based heuristic techniques, such as CP, we are able to make a first step to scientifically support these game evaluation techniques, making them more robust. Once the support for CP is established, interface problems that have historically caught critics' attention can be identified early in the development process. This is especially critical for casual games, since a game's success depends largely on how easy its interface is understood and whether it allows for easy-in/easy-out gameplay cycles, which are most common in this field.

2.2 *Game Critics*

The voice of a professional critic is powerful. A single influential critic can have a significant effect on any product, influencing consumers' purchasing decisions. Larsen (2008) explored how a critic's review is in essence an unscientific user experience evaluation. Larsen identifies that game reviewers generally provide both a qualitative (textual review) and quantitative (score) component. Some reviewers evaluate different aspects of a game separately. In many cases, these aspects are comparable to hedonic and pragmatic aspects of user experience research.

Koeffel et al. (2010) also offers support for the use of the reviews written by game critics. In their recent study, they compare results of heuristic evaluations to the average critic scores of games, and suggest that a similar trend between review scores and the number and severity of usability problems exists.

The term critic-proofing has recently been used to describe the technique used at BioWare in *MassEffect 2*. In an interview with Gamasutra (Nutt, 2010), producer Adrien Cho described how the development team mapped critic and player feedback from the first game onto the design objectives for the sequel.

Game reviews provide us with an excellent way of dealing with the limitations of current game heuristics. Unfortunately the current review-based heuristics (Pinelle et al., 2008a) were limited and biased in their scope, focusing on only 6 game genres and 108 reviews from GameSpot (Pinelle et al., 2008b). Also, the entire evaluation was conducted manually, with researchers visually searching the reviews for criticisms of usability problems. Since metacritic provides an aggregation of a large sample of game reviews, it is perfectly suited to provide a basis for improving and extending review-based heuristics. metacritic has also gained traction in the industry as a metric used by publishers to determine the success of a title and the potential return from a given game. We plan to use metacritic similarly to how Pinelle used GameSpot, to a) validate existing heuristics, b) use a tool already used by the games industry to provide understandable evaluation results, and c) demonstrate the potential value of using game specific evaluations.

1. Implementation Strategy

Examining game reviews manually is impractical on a larger scale. Instead, the process should be automated as much as possible so a much larger review sample set can be utilized. To develop a more automated process we have developed a three-step implementation strategy (Figure 3).

Figure 3. System design

3.1 Web Crawler Functionality

Before any analysis can begin, review data must first be collected. This entails two types of data in addition to the more specific textual qualitative reviews and game scores (as well as metascores). The web crawler will collect complete game reviews from the same websites used by Metacritic in the calculation of their 'metascores'. The web crawler is an automated system that will run continuously, collecting new data as it's posted and periodically rescanning to ensure that the data collected is up to date.

3.2 Data Collection Strategy

Metacritic is a rich source of data. In addition to the actual text reviews, we will also collect: platform, genre, release date, publisher, developer, number of players, rating, number of reviews, reviewer's name, metascore, and any relevant or potentially important data from the specific game review sites we mine. We will store this data in a database for further analysis.

There exist two primary challenges that must be addressed in this data collection strategy. First, data must be collected from each review site that metacritic provides access too. However, each site provides different information formatted in different ways. The web crawler will need to be

customized to ensure that it is capable of sorting data correctly when presented in different formats. This challenge is exacerbated because metacritic aggregates across more than one hundred game reviews. Second, metacritic uses some publications that are only available in print mediums, restricted to subscribers, or in a language other than English. The web crawler cannot mine these publications; therefore we will have to omit them from our evaluation database. Fortunately, the number of restricted or print publications is small in comparison to those that remain, so a substantial number of reviews are still accessible.

3.3 Analysis Strategy

Perhaps the most challenging aspect, and one of the primary goals, for this project is the development of an automated analysis tool that will mine information from the text of game reviews. The analysis strategy will validate game specific heuristics by fulfilling the five requirements that we have outlined above. Since game reviews are essentially untrained user experience reports, we can use the information from them to validate or reject game heuristics.

Analysis of the review data will be performed in a number of different ways. Initially the textual data will be analyzed using text analysis tools such as LIWC (Pennebaker, Chung, Ireland, Gonzales, & Booth, n.d.) and ANEW (Bradley & Lang, 1999). The results from LIWC can be examined to determine how words are used within a specific text. The use of heuristic-specific words could be used to determine the frequency of a specific heuristic problem, for example artificial intelligence issues. ANEW is a similar tool to LIWC except focusing specifically on the emotional content of words.

4. Discussion

Casual games are known to have smaller development budgets and quicker turnaround times. Thus, a low-cost and simple game usability evaluation approach is needed. Many of the expert reviews and heuristic techniques in use for evaluating the playability of games provide such a quick turnaround, however not always at small budgets. In addition, the major challenge of current playability evaluation approaches is that they are usually only based on evaluator perception and lab-constrained user studies. Many of these approaches have no quantitative backing by structured scientific studies using large numbers of evaluators. However, to make these approaches widely accepted in the games industry, a wide range of globally situated evaluators can provide the credibility needed for the acceptance and deployment of these techniques in the game industry. Since Metacritic is already an accepted source of criticism in the game industry, using the reviews that build the foundation of Metacritic to validate the CP game evaluation technique will provide a unique opportunity to establish this technique for quick and easy game interface and usability evaluation. Having such a technique available that is half-automated or assisting in evaluating games, the game quality of casual games will be easy to improve.

5. Conclusion

In this paper, we have presented preliminary work into a novel game evaluation technique we have dubbed critic-proofing and a possible evaluation technique to improve its validity and foster application in the game industry. Previous work (Livingston et al., 2010; Pinelle et al., 2008a; Pinelle et al., 2008b) in this area has relied on a small sample of game reviews to validate this method. If CP and game heuristics in general are to be applied in a game industry context, they need to have established validity, ideally based on a source that is trusted in the industry. Thus, they are likely to receive large-scale adoption by the game industry. In summary, to validate previous game-specific heuristics and make them relevant to the game industry, we are developing an

automated game review data collection and analysis tool that can be used to validate existing game-specific heuristics and will be used to further strengthen existing CP techniques.

6. References

About Metascores. (n.d.). Retrieved July 29, 2010, from <http://www.metacritic.com/about/scoring.shtml>

Bradley, M., & Lang, P. (1999). *Affective norms for English words (ANEW): Instruction manual and affective ratings* (Technical Report C-1). University of Florida: The Center for Research in Psychophysiology.

Desurvire, H., Caplan, M., & Toth, J. A. (2004). Using heuristics to evaluate the playability of games. In *CHI'04 extended abstracts on Human factors in computing systems* (p. 1512).

Desurvire, H., & Wiberg, C. (2009). Game usability heuristics (play) for evaluating and designing better games: The next iteration. *Online Communities and Social Computing*, 557–566.

Federoff, M. A. (2002). *Heuristics and usability guidelines for the creation and evaluation of fun in video games*. Citeseer.

Is Metacritic Damaging the Games Industry? | Edge Magazine. (n.d.). . Retrieved July 29, 2010, from <http://www.next-gen.biz/features/is-metacritic-damaging-games-industry?page=0%252C0>

Koeffel, C., Hochleiter, W., Leitner, J., Haller, M., Geyen, A., & Tscheligi, M. (2010). Using Heuristics to Evaluate the Overall User Experience of

- Video Games and Advanced Interaction Games. In *Evaluating User Experience in Games* (pp. 236–260). Springer.
- Korhonen, H., & Koivisto, E. M. (2006). Playability Heuristics for Mobile Games (pp. 9-16). Presented at the MobileHCI.
- Korhonen, H., & Koivisto, E. M. (2007). Playability heuristics for mobile multi-player games. In *Proceedings of the 2nd international conference on Digital interactive media in entertainment and arts* (p. 35).
- Korhonen, H., Paavilainen, J., & Saarenpää, H. (2009). Expert review method in game evaluations: comparison of two playability heuristic sets. *Proc. of MindTrek 2009*, 74–81.
- Larsen, J. M. (2008). Evaluating User Experience—how game reviewers do it. In *Evaluating User Experiences in Games, Workshop at CHI*.
- Livingston, I. J., Mandryk, R. L., & Stanley, K. G. (2010). Critic-Proofing: How Using Critic Reviews and Game Genres can Refine Heuristic Evaluations (pp. 72-79). Presented at the Future Play 2010, Vancouver, Canada.
- Mandryk, R. L., Atkins, M. S., & Inkpen, K. M. (2006). A continuous and objective evaluation of emotional experience with interactive play environments. In *Proceedings of the SIGCHI conference on Human Factors in computing systems* (pp. 1027-1036). Montréal, Québec, Canada: ACM. doi:10.1145/1124772.1124926
- Nacke, L. E., Grimshaw, M. N., & Lindley, C. A. (2010). More Than a Feeling: Measurement of Sonic User Experience and Psychophysiology

in a First-Person Shooter Game. *Interacting with Computers*.

Nielsen, J. (1994). Guerrilla HCI. Retrieved July 29, 2010, from http://www.useit.com/papers/guerrilla_hci.html

Nielsen, J. (2005). 10 Heuristics for User Interface Design. Retrieved July 29, 2010, from http://www.useit.com/papers/heuristic/heuristic_list.html

Nutt, C. (2010). Back In Space: BioWare On Mass Effect 2. *Gamasutra*. Retrieved July 29, 2010, from http://www.gamasutra.com/view/feature/4251/back_in_space_bio%20ware_on_mass_.php,%202010.

Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., & Booth, R. J. (n.d.). LIWC. Retrieved July 29, 2010, from <http://www.liwc.net/>

Pinelle, D., Wong, N., & Stach, T. (2008a). Heuristic evaluation for games: usability principles for video game design. In *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems* (pp. 1453–1462).

Pinelle, D., Wong, N., & Stach, T. (2008b). Using genres to customize usability evaluations of video games. In *Proceedings of the 2008 Conference on Future Play: Research, Play, Share* (pp. 129–136).

Pinelle, D., Wong, N., Stach, T., & Gutwin, C. (2009). Usability heuristics for networked multiplayer games. In *Proceedings of the ACM 2009 international conference on Supporting group work* (pp. 169–178).

Figure 1 The metacritic website

Games

Best of 2009 | 08 | 07 | 06 | 05
Advanced Search

Sony
P3 | PSP | PS2 | PS One

Microsoft
Xbox 360 | Xbox

Nintendo
Wii | DS | DSi | DSi XL | DS Lite

Other Platforms
PC | N-Gage | Dreamcast

What is a Metascore?

Upcoming & Recent Releases

Sort by name | Sort by score

- 2010 FIFA World Cup South Africa
- 5 Arcade Gems
- 5 in 1 Solitaire
- Alice in Wonderland
- All Star Wars
- Art of Balance
- Art Style: Light Trax
- Atlas of the Movies 3.0
- Big Top Heroes
- Blaster Master Overdrive
- Blood Beach
- Bloons
- Brain Train
- Cabela's Monster Buck Hunter
- Cadeg
- Cave Story
- Chess Challenger
- Dance Sensation
- Def Rigo
- Data East Arcade Classics
- Derp Dash
- Diagon: Unleashed Awakening
- Endless Ocean: Blue World
- Family Go-Kart Racing
- Family Party: 20 Quests
- Games Master Fun
- Fishy Fishy
- Flip's Twisted World
- Fragile Dreams: Faraway
- Fun of the Moon
- Free Running
- Ghostlayer
- Green Day: Rock Band
- Grid: Off with Ultra Hand
- GTT Club Supermil: Final
- Happy Homebrew!
- How to Train Your Dragon
- Hubert the Teddy Bear: Winter Games
- Iron Man 2
- Kung Fu Furi: Everybody is Kung Fu Fighting!
- Last Airbender: The Legend of the Avatar
- League of Legends
- Major League Baseball 2k10
- Max & the Magic Marker
- Mega Man 10
- Military Madness: Nectaris
- Monster Hunter Tri
- Moon House
- My Body Coach
- Never! Fitness First Mod
- Body Yoga & Fitness Workout
- Phoenix Wright: Ace Attorney - Justice For All
- Phoenix Wright: Ace Attorney - Justice For All

Super Mario Galaxy 2

Wii

98 Universal acclaim
Based on 38 critic reviews
How did we calculate this?

User Score: **9.5**
Based on 452 votes
Read user comments
Rate this game >

Game Info

Publisher: Nintendo
Developer: Nintendo
Genre(s): Action, Platform
Players: 2
ESRB Rating: E (Everyone)
Release Date: May 23, 2010

Summary

Super Mario Galaxy 2, the sequel to the galaxy-hopping original game, includes the gravity-defying, physics-based exploration from the first game, but is loaded with entirely new galaxies and features to challenge players. On some stages, Mario can pair up with his dinosaur buddy Yoshi and use his tongue to grab items and spit them back at enemies. Players can also have fun with new items such as a drill that lets our hero tunnel through solid rock. [Nintendo]

Also On Metacritic
GAMES: Super Mario Galaxy
Also On The Web: Metacritic Feature: Best & Worst Mario Games | Official Website

What The Critics Said

All critic scores are converted to a 100-point scale. If a critic does not indicate a score, we assign a score based on the general impression given by the text of the review. Learn more...

- 100** **Edge Magazine**
This isn't a game that redefines the genre: this is one that rolls it up and locks it away.
Read Full Review >
- 100** **1UP**
Takes everything its predecessor did and maintains that urgent sense of wonder and discovery, creating a world that copies its past self, yet keeps the freshness that made it so exciting when the first game debuted two years ago.
Read Full Review >
- 100** **IGN**
Consider me "wowed." Again. Yes, Super Mario Galaxy 2 is more of what made the original so amazingly good: brilliant gravity platforming. Incredibly tight and responsive controls. Fantastic musical score. But it's not a "me too" experience -- a significant majority of this sequel is brand new.
Read Full Review >
- 100** **Games Radar (in-house)**
Expanding on the original's brilliance, this ingenious platformer only gets better by embracing the Wii's limited capabilities and adds so many ideas, big and small, that we can't help but love it. Even if you (mistakenly) believe it's too similar to the first game, what's wrong with more of an incredibly great thing?
Read Full Review >
- 100** **The Escapist**
Super Mario Galaxy 2 doesn't tinkler with the established formula very much, but we didn't really want it to. It's huge, brilliant, and gorgeous. It's why you started playing videogames in the first place.
Read Full Review >
- 100** **Digital Chumps**
Even the harshest critics will struggle to find fault in its symphony of imagination, execution, and polish, while fans will cherish its brilliance for years to come as one of the very best games in history. Quite simply put, Super Mario Galaxy 2 is the

Figure 2 The technique was used to evaluate *Capsized* (AlienTrap Software, 2010), a 2D platform shooter game



Figure 3 *System design*

