Weak Inter-Rater Reliability In Heuristic Evaluation of Video Games

Gareth R. White
University of Sussex, Brighton,
East Sussex, BN19RH, UK
g.white@sussex.ac.uk

Pejman Mirza-babaei
University of Sussex, Brighton,
East Sussex, BN19RH, UK
pm75@sussex.ac.uk

Graham McAllister
University of Sussex, Brighton,
East Sussex, BN19RH, UK
g.mcallister@sussex.ac.uk

Judith Good
University of Sussex, Brighton,
East Sussex, BN19RH, UK
j.good@sussex.ac.uk

CHI 2011, May 7–12, 2011, Vancouver, BC, Canada.
ACM 978-1-4503-0268-5/11/05.

Abstract
Heuristic evaluation promises to be a low-cost usability evaluation method, but is fraught with problems of subjective interpretation, and a proliferation of competing and contradictory heuristic lists. This is particularly true in the field of games research where no rigorous comparative validation has yet been published. In order to validate the available heuristics, a user test of a commercial game is conducted with 6 participants in which 88 issues are identified, against which 146 heuristics are rated for relevance by 3 evaluators. Weak inter-rater reliability is calculated with Krippendorff's Alpha of 0.343, refuting validation of any of the available heuristics. This weak reliability is due to the high complexity of video games, resulting in evaluators interpreting different reasonable causes and solutions for the issues, and hence the wide variance in their ratings of the heuristics.

Keywords
Heuristic evaluation, usability, user experience, video game.

ACM Classification Keywords
H.5.2 Information Interfaces and Presentation: User interfaces – Evaluation/ methodology

General Terms

Design, Experimentation, Human Factors, Measurement, Reliability, Verification.

Introduction

Heuristic evaluation (HE) has an established tradition of being used to assess productivity software [21]. As a discount usability evaluation method, these simple and easy to use lists not only appear to speak with authority, but also hold a great deal of potential for teams under time and financial pressures. They offer the advantage of not requiring participants for user testing, nor even necessarily a fully functional prototype as they can be used to guide the pre-production design process.

However, extensive work has been conducted exploring the evaluator effect in usability testing [13] and HE [10]. For example there is evidence to suggest that different teams identify different issues [15,16], as well as number of issues detected [14] and their severity rating [1]. Institutional tendencies may also be problematic [17].

Some studies attempted to validate heuristics by comparing the number of issues they revealed to the number of issues identified through user testing, however arguments have been raised as to the number of participants needed [6, 23, 20], as well as evaluator expertise and the type of task involved [12]. A related issue is matching of usability problems, where similar problems identified during HE are grouped together to produce a reduced and generalised set of problems [11]. Considerably different results are produced depending on the matching technique employed, and in most studies this matching process is implicit and rarely commented on, and thus is a hidden source of variability when trying to compare different approaches.

Thus we see that even in traditional domains of relatively simple productivity applications, there are still considerable disagreements about how to perform HE and to what degree the results can be trusted.

Heuristics in games

The proliferation of heuristic sets raises the question of which to use, and how to compare one to another. In order to explore this question, the heuristics in the literature were considered for suitability, and those intended for different platforms, domains, or genres were excluded from further consideration, as were a number of subjective or otherwise non-validated design guidelines. In addition, Nielsen's list was included in order to compare our validation of traditional and game specific heuristics. The following six remained, totalling 146 heuristics when duplicates were removed.

- Desurvire and Wiberg (PLAY) [4]
- Desurvire and Wiberg (GAP) [5]
- Federoff [7]
- Korhonen et al. [16] (not mobile components)
- Nielsen [19]
- Pinelle et al. [22]

User Test

A single-player first-person shooter (FPS) console game (*Aliens Vs Predator*, 2010) was subjected to user testing prior to the game's release in order to help guide development. At the request of the development team, six male participants (19, 22, 20, 20, 22, 30 yrs) were recruited from two informally identified demographic groups: "mainstream" (more casual gamers who play occasionally) and "core" (experienced gamers who play FPS games frequently.) The game was a high fidelity interactive prototype at "vertical slice" quality, meaning that only a portion of the game was complete to a level of quality indicative of the final product, and only these sections were tested. Participants played for approximately one hour each, and testing was conducted over 2 days under laboratory conditions on an Xbox 360 connected to widescreen HD television. Video cameras recorded the player, and real time footage from the game console was simultaneously streamed to the observation room next door. All feeds were composited together on a widescreen HD display, and saved to disk for later analysis. The game's producer and a senior user experience consultant monitored the participants' play from the observation room. The UX consultant had spent some time familiarising himself with the game before the test sessions, and the producer was able to identify when players were not playing the game as intended.

Following the user testing, a report was produced by the senior UX consultant who ran the session. The report listed the usability and playability issues encountered by each participant, as well as some additional design recommendations.

The main focus of this paper is on the following stage, which was to evaluate which heuristics were violated by each issue, and hence to validate or refute their applicability beyond their original studies.

Validation

Each of the issues identified through user testing were considered against the 146 heuristics identified earlier. Following Nielsen's approach [19], each issue was rated against each heuristic on a scale of 0 to 5:

0. Does not explain the problem at all.
1. May superficially address some aspect of the problem.
2. Explains a small part of the problem, but there are major aspects of the problem that are not explained.
3. Explains a major part of the problem, but there are some aspects of the problem that are not explained.
4. Fairly complete explanation of why this is a usability problem, but there is still more to the problem than is explained by the heuristic.
5. Complete explanation of why this is a problem.

Three evaluators were used: a video game user experience doctoral student with professional experience of conducting user tests; a further video game user experience doctoral student, considered as a ``double expert'' with professional experience of conducting user experience tests and professional game

development; and a junior HCI researcher. Each evaluator participated in a training session, where each heuristic was reviewed, and trial ratings were assigned to sample issues. Uncertainty about the meaning or intention of particular heuristics was discussed and consensus agreed upon.

88 issues from the user test session were randomly ordered and presented to each evaluator. Each issue was presented with a uniquely randomised ordering of the 146 heuristics. Each evaluator rated each issue against each heuristic using Nielsen's explanatory scale.

Once each of the evaluators had completed their evaluation of each issue the data were collected together for statistical analysis, presented in the following section.

Results

All ratings were analysed for variance between the three evaluators. Variances of up to 8.33 were identified in cases such as when one evaluator rated a heuristic as 5 (``Complete explanation of why this is a problem'') and the other two evaluators rated as 0 (``Does not explain the problem at all'.') Krippendorff's Alpha [9] was computed for our ordinal data set using an online calculator [8] at a value of 0.343 (nCoders = 3; nCases = 12848; nDecisions = 38544.) This represents very poor reliability, and similar levels were calculated across all six heuristic sets used.

Discussion

Evaluators were interviewed and explained their decision making process for their ratings. They tended to see the cause of issues from a different perspective to one another, but it was clear that there were multiple reasonable interpretations for which heuristic best explained each issue. This appears to be due to the complexity of the tasks involved, featuring a high volume of simultaneous multi-modal data for the player to process, dynamic emergent interaction, and temporally distant events who's influences overlap one another.

For example, during a gunfight the player was reminded how to use a skill which had been taught earlier, but in the heat of the action he respond incorrectly. There are arguably violations of heuristics about controls, the reminder, and the tutorial, as well as more subjective heuristics about enjoyment and challenge. Our evaluators showed clear disagreement in many similar cases. In future work, further qualitative analysis will be conducted to fully understand the reasons for these points of disagreement, so that the decision making process can be formalised into objective criteria for the violation of each heuristic.

It is pertinent to consider this example in relation to the taxonomy of HE problem discoverability proposed by Cockton and Woolrych [2]. Their "PAC" scheme identified three increasing levels of complexity: "perceivable" problems which can usually be discovered without interaction, by simply looking at the display; "actionable", which typically only require a few clicks on the interface to expose; and "constructable" which involve several interaction steps with multiple application objects before they become apparent. They show that in non-trivial tasks most HEs are incorrect, and that these errors increase with task complexity. The issues in our study required a substantial degree of interaction, and would likely exceed the intended scope of PAC. Future work will extend and formally define

these degrees of complexity, and explore their affect on inter-rater agreement.

In a further study by Cockton et al. [3] evaluators identified which heuristics best explained issues encountered in a user test, but in only 31% of cases was there agreement that these assignments were considered appropriate. Our alpha is indicative of a similarly low level. It's noteworthy that in their later studies, which employed structured reporting formats, this level increased to 60%. Similar results were reported for the percentage of problems predicted that were discovered during user testing. In order to mitigate the subjectivity and low inter-rater agreement in our study, we plan further work to explore how structured reporting formats can help evaluators reach consensus.

The original evaluation teams of the six heuristic sets considered here achieved agreement in their own studies through private discussion during evaluation. Without repeatability of those discussions instantiated as formal, objective evaluation processes in the methodology, repeatability and validation of their results is not possible. We do accept that individual teams may find some utility in HE for video games, particularly for formative evaluation during pre-production with non-interactive prototypes. However, in order to objectively assess which heuristics would be useful for complex interactive games, a more nuanced analysis of the issues affecting inter-rater reliability still needs to be conducted and validated. Currently evaluators have little objective data upon which to base a reliable decision for which heuristics to employ, or indeed precisely how they should be used. Furthermore, comparison of the results between HEs from different teams remains difficult, and clear, common protocols need to be established to make this possible.

Conclusion

From a user test session of a commercial first-person shooter console video game, 88 issues were identified and rated for degree of explanation against a comprehensive selection of 146 heuristics. Despite evaluator training and a tightly focussed study, we observed systematic low inter-rater reliability across all of the heuristic sets. This appears to be due to subjectivity in evaluators' interpretation of complex, multi-modal events, and calls into question the use of heuristic evaluation as an HCI practice for video games. Given these findings, further work needs to be conducted in order to demonstrate a significant level of reliability in order for it to continue to be used as a widespread method.

Acknowledgements

The authors would like to thank Emma Foley for her assistance, without which this paper would not have been possible.

References

[1] B. Bailey. Judging the severity of usability issues on web sites: This doesn't work. http://www.usability.gov/articles/newsletter/pubs/102005news.html, Last accessed July 9 2010.

[2] G. Cockton and A. Woolrych. Understanding inspection methods: Lessons from an assessment of heuristic evaluation. *Joint Proc. HCI 2001 and IHM 2001: People and Computers XV*, ACM (2001), 171–191.

[3] G. Cockton, A. Woolrych, and M. Hindmarch. Reconditioned merchandise: extended structured report formats in usability inspection. Proc. *CHI '04 extended abstracts on Human factors in computing systems*, ACM (2004), 1433–1436.

[4] H. Desurvire and C. Wiberg. Game usability heuristics (PLAY) for evaluating and designing better games: The next iteration. *Proc. OCSC*, Springer-Verlag (2009), 557–566.

[5] H. Desurvire and C. Wiberg. User experience design for inexperienced gamers: GAP - game approachability guidelines. In R. Bernhaupt (ed.), *Evaluating User Experience in Games - Concepts and Methods*, Springer-Verlag (2010), chapter 8.

[6] L. Faulkner. Beyond the five-user assumption: benefits of increased sample sizes in usability testing. *Behavior Research Methods, Instruments, & Computers 35*, 3 (2003), 379–383.

[7] M. Federoff. *Heuristics and usability guidelines for the creation and evaluation of fun in video games*. Department of Telecommunications, Indiana University, USA, 2002. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.89.8294&rep=rep1&type=pdf, Last accessed 18 February 2011.

[8] D. Freelon. Recal for ordinal, interval, and ratio data (OIR). http://dfreelon.org/utils/recalfront/recal-oir/, Last accessed 22 September 2010.

[9] A. F. Hayes and K. Krippendorff. Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures 1*, 1 (2007), 77–89.

[10] M. Hertzum and N. E. Jacobsen. The evaluator effect: A chilling fact about usability evaluation methods. *International Journal of Human-Computer Interaction 15*, 1 (2003), 183–204.

[11] K. Hornbæk and E. Frøkjaer. Comparison of techniques for matching of usability problem descriptions. *Interact. Comput. 20*, 6 (2008), 505–514.

[12] W. Hwang and G. Salvendy. What makes evaluators to find more usability problems?: a meta-analysis for individual detection rates. In *Proc. HCI'07*, Springer-Verlag (2007), 499–507.

[13] N. E. Jacobsen, M. Hertzum, and B. E. John. The evaluator effect in usability tests. In *Proc. CHI '98*, ACM (1998), 255–256.

[14] B. E. John. The evaluator effect in usability studies: Problem detection and severity judgments. *Proc. Human Factors and Ergonomics Society Annual Meeting 42*, 5, (1998), 1336–1340.

[15] M. Kessner, J. Wood, R. F. Dillon, and R. L. West. On the reliability of usability testing. In *Proc. CHI extended abstracts*, ACM (2001), 97–98.

[16] H. Korhonen, J. Paavilainen, and H. Saarenpää. Expert review method in game evaluations - comparison of two playability heuristic sets. In *Proc. Mindtrek*, ACM (2009).

[17] R. Molich, M. R. Ede, K. Kaasgaard, and B. Karyukin. Comparative usability evaluation. *Behavior and Information Technology 23*, 1, Taylor & Francis, Inc. (2004), 65–74.

[18] J. Nielsen. Finding usability problems through heuristic evaluation. In *Proc. CHI '92*, ACM (1992), 373–380.

[19] J. Nielsen. Enhancing the explanatory power of usability heuristics. In *Proc. CHI '94*, ACM (1994) 152–158.

[20] Nielsen and T. K. Landauer. A mathematical model of the finding of usability problems. In *Proc. CHI '93*, ACM (1993) 206–213.

[21] J. Nielsen and R. Molich. Heuristic evaluation of user in-terfaces. In *Proc. SIGCHI*, ACM(1990), 249–256.

[22] D. Pinelle, N. Wong, and T. Stach. Heuristic evaluation for games: usability principles for video game design. In *Proc. SIGCHI*, ACM (2008), 1453–1462.

J. Spool and W. Schroeder. Testing web sites: five users is nowhere near enough. In *Proc. CHI '01*, ACM (2001), 285–286.