Decision-Based Testing for Casual Games

Llewyn Paine
Media Research Labs
11904 North IH-35
Austin, Tx 78753 USA
llewyn@utexas.edu

CHI'12, May 5–10, 2012, Austin, Texas, USA.
ACM 978-1-4503-1016-1/12/05.

Abstract

Several tools are available in the effort to better understand player experience and to uncover game design issues. In evaluating casual games for industry clients, the key question is often "Will people play my game?" Incorporating decision-making into playtesting provides a way to answer this question behaviorally and to quantify improvements in designs.

Author Keywords

Decision-making; methodology;

ACM Classification Keywords

H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous. See: http://www.acm.org/about/class/1998/ Mandatory section to be included in your final version.

General Terms

Measurement

Introduction

I am a cognitive psychologist in Austin, Texas. For over five years I studied aspects of perception and performance in an academic research lab. My major area of research dealt with understanding the way we organize perceptual information that is distributed in time -- in other words, temporal grouping. Other projects I worked on involved reaction time, success/failure, visual search, object perception, and coordination between sounds and trackball movement. Since moving to media research I've been grateful for this background, as it's given me insight into various perceptual issues that crop up over and over -- such as pop-out effects when a user is searching for one type of object among others.

The other value of my academic background has been understanding how to conduct a strong experiment. This is the main approach I take in my current position. Unlike many industry media labs which rely largely on focus groups or qualitative methods, I answer questions for clients by isolating variables and measuring differences in outcomes. While this is de rigueur in academia, it represents a new movement in this sector of industry that seems to be gaining momentum.

Most of my game user research is performed for clients who are not primarily game-makers. They are television networks, advertisers, or educators who want to take advantage of "gamification" trends to increase brand loyalty, sales, or education. The result of this is that the games I test are largely simple, casual games on web or mobile platforms. In about half of the cases, testing is post-release, and changes based on the results of testing are issued as updates. Because of challenges in coordinating teams across departments, however, it is often only those gameplay issues which affect the bottom line (e.g., time on site, ad exposure, test scores) which stand a good chance of being remedied. This means that a significant part of my job is measuring those key performance indicators and demonstrating their tangible connection with user experience issues.

I have taken several approaches to evaluating user experience in games, but they can be generalized as qualitative observation, quantified observation (that is, coding and counting behaviors), restricted and open-ended feedback, error log analysis, heuristic evaluation, and technologically-assisted measures (e.g., eye tracking). I have also used biometric, reaction time, and implicit association measures for other studies, but I have not used these techniques in game evaluation. In the course of all this, I have found that for the type of testing I do, no method is more effective than coding and quantifying behaviors. Tracking successes and failures at sub-tasks within a game provides incontrovertible evidence of an issue's prevalence, and perhaps more importantly, it permits clients to see whether the changes they make are having any effect. Despite my cognitive psychology background (where I was taught to be highly distrustful of self-report), I also find that user feedback is particularly valuable in the game user research sphere. Again, my approach tends to be quantitative (comparing numbers of people who mention a particular issue across iterations, for instance), but user feedback provides helpful qualitative insights as well.

Although I've worked with various technologically-assisted measures -- several types of eye trackers, electrodermal activity, perception analyzers, and so on -- I have never found them to be capable of taking the place of simple observational behavioral metrics and user feedback. Eye tracking, for instance, has not shown a great deal about enjoyment or users' engagement in games. However, it is a great way to understand other reactions people have to a game. In one case, participants complained about a lack of instructions at various stages of a game, when in fact instructional elements were clearly displayed onscreen. Eye gaze recordings are valuable in this situation to clarify whether participants were simply not seeing instructional elements, or whether they were seeing them but misinterpreting their meaning.

Despite the number of available tools, I have found that for several games I've tested the single most useful metric has been to give participants an actual choice. Sometimes we ask users to rate their enjoyment of a game, or to predict whether they would choose to play it again. But I have found that for my clients' goals, it is much more telling to actually give users the option at the end of their sessions either to leave early or to play another round. This is somewhat gimmicky and participants may see through it, but the effectiveness lies in the fact that time is valuable [1]. Despite possible concerns about demand characteristics [2], my experience has been that participants will happily move on when a task becomes too burdensome. It has been demonstrated that participants in visual search tasks will risk responding incorrectly in order to save a few-hundred milliseconds [3], so it is not surprising that participants are willing to disappoint an experimenter in a playtest gain an extra fifteen minutes of free time. However, if demand characteristics remain a concern, it is also possible to offer a choice of games to mask the true intent of the test.

The decision technique is appropriate for casual games where a client's best bet is often to impress a user quickly and keep him playing for a few minutes longer. If a user stumbles onto a website and finds a game he was not originally intending to play, he may not be willing to invest more than a few minutes to check it out. If the game draws the user in, he may spend a lot more time there. He may bookmark it and return later, or he may send it to his friends. However, if the game fails to engage him, he will decide to move on. By bringing a similar decision into the lab, we attempt to capture this behavior in a controlled way. Plus, like other metrics, decision data can be tracked across iterations in order to determine whether changes are making games more or less addictive.

This simple technique has made a major difference in the results I report to clients. I have found that when users self-report their likelihood of returning to a game, it correlates highly with other subjective self-reported experience measures such as enjoyment. But even well-meaning users will often misunderstand what makes them return to a game. Enjoyment is an important factor, certainly. But there are other factors as well which often fail to be incorporated into the self-reported rating. One of these is simply how much there is left to do in the game. If a user has already played all of the available levels in a game, then it sometimes takes more to make him choose to replay. In testing scenarios, users will often pass on spending more time with these games, opting instead for a different game or leaving early. This is true even when enjoyment and reported replay likelihood are high.

This method does have several limitations. It is an aggregate measure that is influenced by several aspects of game user experience. This means that it is poorly suited for research evaluating the effect of specific variables on individual dimensions of the game experience. Nevertheless, it is a true measure of user experience, inasmuch as players will be significantly less likely to replay if their experience has been inferior. The technique is also poorly suited for evaluating serious games where users have pre-existing expectations and have already invested financially by purchasing the game. In this situation, players may be more willing to invest their time for a longer-term reward, and the decision to keep playing or to take a break may be more complicated.

While it is something of a blunt instrument, allowing participants to make the decision to replay or not inside the lab is a simple and effective way to provide an answer to the core question that many casual game developers are asking with user research: will people play my game?

References

[1] Alexander, R.M. *Optima for Animals*. Princeton University Press, Princeton, NJ, USA, 1996.

[2] Orne, M. On the social psychology of the psychological experiment: With particular reference to

demand characteristics and their implications. *American Psychologist 17* (1962).

[3] Gilden, D.L., Thornton, T.L., & Marusich, L.R. The serial process in visual search. *Journal of Experimental Psychology: Human Perception and Performance 36*, 3 (2010).