Understanding user behavior at three scales: The AGoogleADay story

Daniel M. Russell

January, 2014

Abstract: How people behave is the central question for data analytics, and a single approach to understanding user behavior is often limiting. The way people play, the ways they interact, the kinds of behaviors they bring to the game, these factors all ultimately drive how our systems perform, and what we can understand about why users do what they do. I suggest that looking at user data at three different scales of time and sampling resolution shows us how looking at behavior data at the micro-, meso-, and macro-levels is a superb way to understand what people are doing in our systems, and why. Knowing this lets you not just understand what's going on, but also how to improve the user experience for the next design cycle

Introduction

While there are many motivations for creating games, *serious games* are usually vehicles for teaching and learning. While serious games are important in many educational settings, they sometimes suffer from a lack of attention to the details of game design. While goal is to teach and instruct, the *game* experience sometimes suffers. Face it, some of those serious games aren't so much fun to play. How is it possible that a game can be created and deployed without anyone noticing that it has some playability issues?

Many people have reported on ways to instrument and monitor a game, we have found that a particularly useful approach to understand the overall user experience has been to analyze game-player behavior at three different time scales of behavior, from very short millisecond-by-millisecond behaviors, up to the time scale of millions of players as they use the game to learn over weeks and months.

The AGoogleADay.com game had a simple goal, we simply wanted to show the public some more sophisticated ways to use the Google search engine. While Google is simple to use, there many features within Google that are not widely used. By building a "trivia question" style game where the use of Google was *required* (and not prohibited, as in most such games), we hoped to introduce new features to the players by creating questions that were difficult (and obscure) enough to motivate their use.

Originally planned as a 3-month experiment, the AGoogleADay (AGAD) game continues to run more than 2 years after its launch, serving millions of game players each month, and improving players ability to seek out answers to questions by searching.

Here we describe some of the analyses we did to understand what was happening with the players—what they did, and what effect changes to the game would have.

Although this paper is about AGAD, the game, the approach of understanding complex user behavior at these three different time scales, and using three different kinds of studies, is applicable to software systems with complex user interfaces in general.

Background

There has been a practice of developing games with an eye towards testing and analysis. [Pinelle, 2008; Nacke, 2009; Kofeel, 2010] Typically, the analysis of games has followed traditional usability analysis methods or logging player behavior.

HCI researchers have grown increasingly interested in studying the design and experience of online games, creating methods to evaluate their user experience [Bernhaupt, 2010;], "playability" heuristics for their design [Schaffer, 2008], and models for the components of games [Schaffer, 2008] and game experience [Sweetser, 2005]. Each of these approaches, while useful, is primarily situated within a single scale of analysis—playability heuristics are defined at the micro-scale, while game experience is primarily studies at the meso-scale.

However, in an age when online games are played by a large number of gamers, from potentially many places around the world, a more comprehensive method for understanding game use early on in the design process is often fruitful. This desire to bridge the gap between the lowest levels of use with the behavior of players in the very large led us to apply a broad-spectrum analysis to our game.

Our Game: AGoogleADay.com

The AGoogleADay.com (AGAD) game was originally developed with a dual purpose in mind: to both be a serious game to teach people more advanced search skills, such as how to use the **filetype:** or **site:** or **filter-by-color** search methods, and also a marketing presence for Google search in new markets in which we wanted to experiment. While Google has a good deal of experience in advertising, this was the first attempt at creating an engaging experience with such a dual purpose.

The goal from the outset was to create a very visual game that would be "sticky," but not all-consuming. In other words, it had to be the kind of game one would play for

a relatively short amount of time, be exposed to the message, and learn a bit along the way.

We designed a "trivia question" game that would be significantly more difficult than ordinary trivia answering games. Typical trivia games walk a delicate balance between being too difficult to answer (except by a small fraction of the game-playing population) and being too easy (and therefore unsuitable for competitive play). For AGAD, we selected questions that are purposefully unlikely to be known by even a small fraction of the gamers, yet answers that are discoverable with a few Google searches. The questions also had to be of some intrinsic interest—truly trivial questions, such as what is the 23rd digit in the decimal expansion of pi?—are unsuitable for game play.

AGAD is a fairly simple game to construct. The UI presents a few questions each day, with daily updates to the set. Players work through the questions, doing Google searches in a separate iframe that allows the search engine to be visible alongside the question. Each question has a simple answer that can be recognized by a regular expression. Writing the regexp is slightly tricky, as most questions have multiple variants on an answer (e.g., "three" "3" or "the number 3").

As shown in Figure 1, the questions are presented in a frame at the bottom of the page, with typical game mechanics such as unlocking subsequent questions, buttons to let the player skip a question, ways to compare your score (and time to answer the question) with friends via the G+ social network.

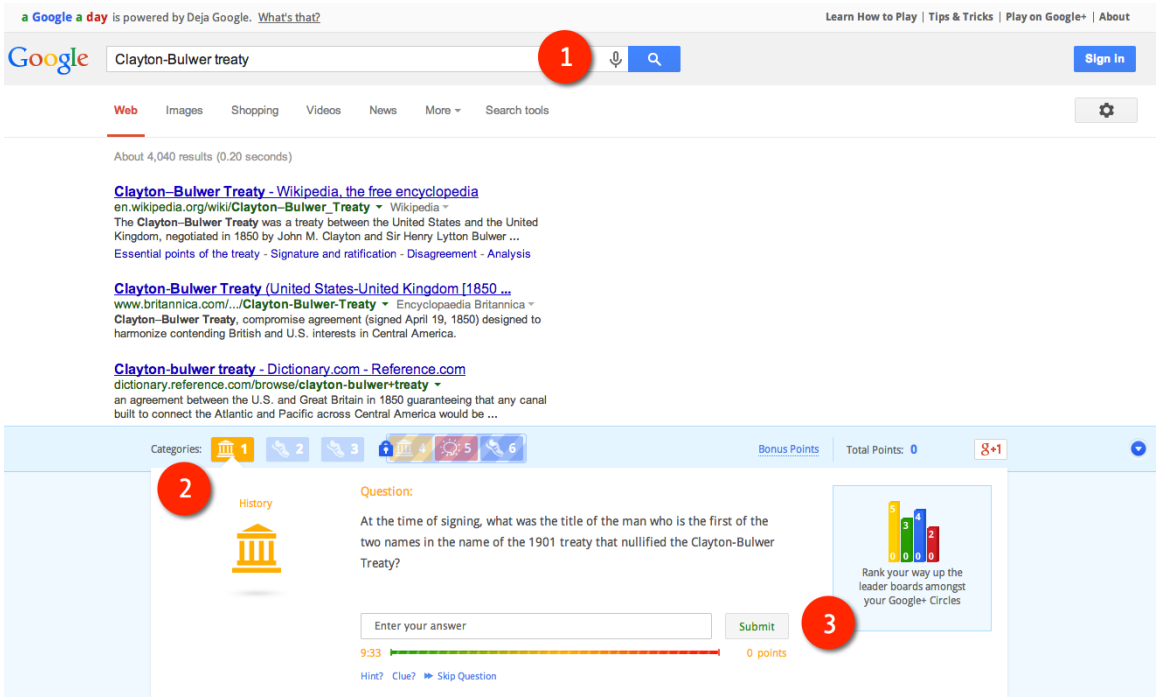

Figure 1 - The AGAD game shows a search question (3), then lets you search for the answer with Google (1). You work through the questions (2) scoring points as you unlock and solve additional questions.

As we developed AGAD, it quickly became clear that user testing would have to be an important part of the design cycle. This is true for all games, but as we proceeded, it became clear that we actually had to do more than standard usability testing in the lab. One large bias was that in the lab setting, people are highly motivated to perform to the best of their ability. [REF] As a consequence, we were seeing a divergence between the lab testing and what we anecdotally observed when watching people in the wild. How could we improve the accuracy of our design-driven testing?

To make our understanding of the players more accurate, we needed to understand what they're doing when the started up the game, how they played the game, and what happened over a broad population. This is what led to the idea that the studies should encompass three different approaches. We needed to understand not just the low-level usability issues, but also verify that the game was fun to play and would leave people with a positive attitude about Google.

Thus, the goal of our design testing strategy became to observe the players in three different ways in order to create a fully rounded picture of what they're doing in (and how they're enjoying) the game.

Three views of the user: Micro, Meso, Macro

After a bit of iteration we settled on a three different analyses of player behavior. First, we wanted to understand the play behavior at a second-to-second time scale. As useful as that is for usability issues (e.g., why do players sometimes never click on a particular game option), this level of analysis is too detailed for understanding user reactions to the game, and doesn't provide any information about their affective responses. To provide that kind of information, we ended up doing analyses of game play over the period of minutes-to-hours. And similarly, that level of analysis (while useful) wouldn't provide the kind of large population data we also wanted to collect and understand—data that would tell us from week-to-month what was happening.

We realized that our three scales of analysis were very similar to Newell's timescales of human behavior [Newell,??]. Newell's analysis framework has 4 bands; neural (1 ms – 10 ms); cognitive (10ms – 10 secs); rational (minutes – hours); social (days to months). His division of cognitive behaviors along different time scales was to emphasize that different kinds of effects can be explained by fundamentally different mechanisms. I take a similar approach here.

Micro scale: measures behaviors from milliseconds up to minutes of behavior, usually with a small number of people, usually in a lab setting. With these studies we want to gain insight into the mechanics of the game—where players look, how they perceive the game elements, and what is left unnoticed and unexplored.

Meso scale: measure behaviors from minutes to hours. This is the realm of field studies, watching people play the game in natural settings (e.g., at home or in the coffee shop). The meso scale provides insights into why people choose to play, and why they end up stopping their play, as well as understanding about what makes the game interesting (or not).

Macro scale: measures behaviors from days to weeks and months. Typically, this involves large numbers of players, and is usually an analysis of the logs of many people playing the game.

Logs of user-behavior has been a standard practice for some time. Traces of behavior have been gathered in psychology studies since the 1930s [Skinner, 1938], and with the advent of web- and computer-based applications it became common to capture a number of interactions and save them to log files for later analysis. More recently, the rise of web-based computing platforms has made it possible to capture human interactions with web services on a large scale. Log data lets us observe how to compare different interfaces for supporting email uptake and sustained use patterns [Dumais et al. 2003; Rodden and Leggett, 2010]

Let's look at each of these time-scale analysis methods in more detail.

Micro level: How players behave over short time scales

While there are many methods to study human behavior over short periods of time, the simplest and most practical method for usability studies is eye tracking (aka "eye gaze") studies. [REF]

Eye tracking studies require bringing a subject into the lab to use a special monitor that has an eye tracking system built into the bezel of the monitor (there are many companies that sell such systems, e.g., Tobii, or SMI, [REF]). These systems calibrate the game player's eye movements on the monitor and output where the eye moves on the screen at time resolutions down to the millisecond. Essentially, they create X, Y, T data streams (X and Y position of the gaze focus on the display, where T is the amount of time the eye dwells on that X, Y location), along with any user actions taken (such as a click, typing, or scroll event).

As shown in Figures 2 and 3, perhaps the most useful way to visualize the millisecond-by-millisecond behavior stream is as either eye tracks on the display,

With this kind of very detailed information about what the player is doing, we can identify distractors and invisible portions of the interface. For instance, in an earlier version of the game, we did not have the "Learn how to play" and "Tips & Tricks" tabs in the upper right corner. As can be seen by the eye movement chart in Figure 2, while they were rarely used, they would be scanned by game players from time to time, ensuring that they knew about their presence, even if only rarely actually used in game play.

More importantly, understanding how a player visually scans the display, and where they spend most of their time (especially when game play is proceeding poorly) is a valuable resource for tuning the game play mechanics. When the player gets stuck, do they exhibit visual search behaviors, or are they able to quickly determine what the next plausible course of action should be?

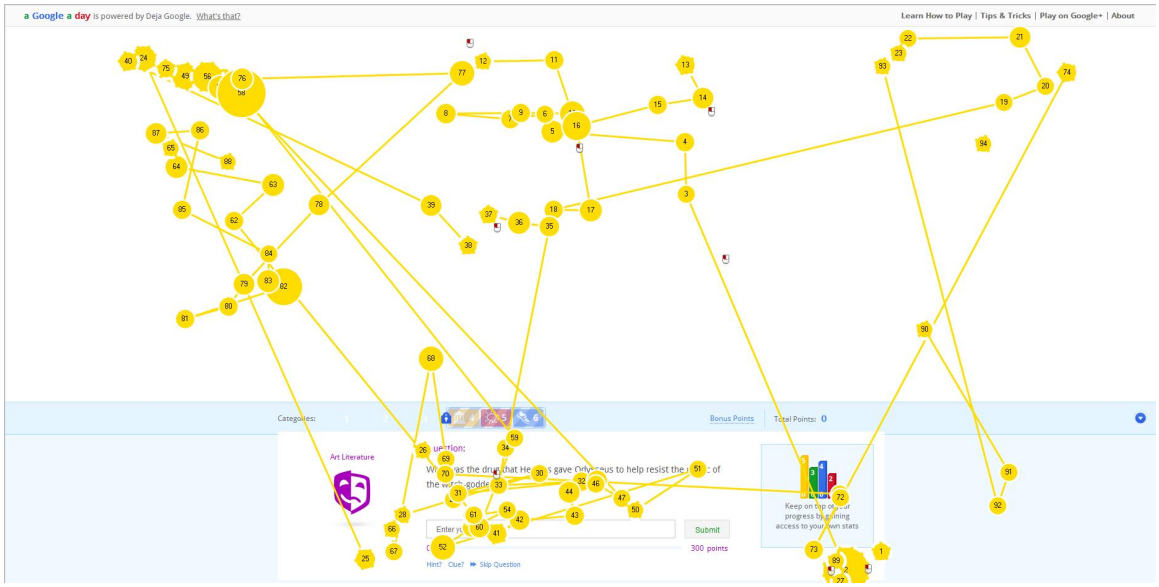

Figure 2 - The movement of the eye on the display is shown by a connected graph. The numbers indicate which fixation, or pause, it represents and the size of the circle represents the duration of the eye fixation at that point. #1 is near the bottom right, where the player clicked the button on the previous page.

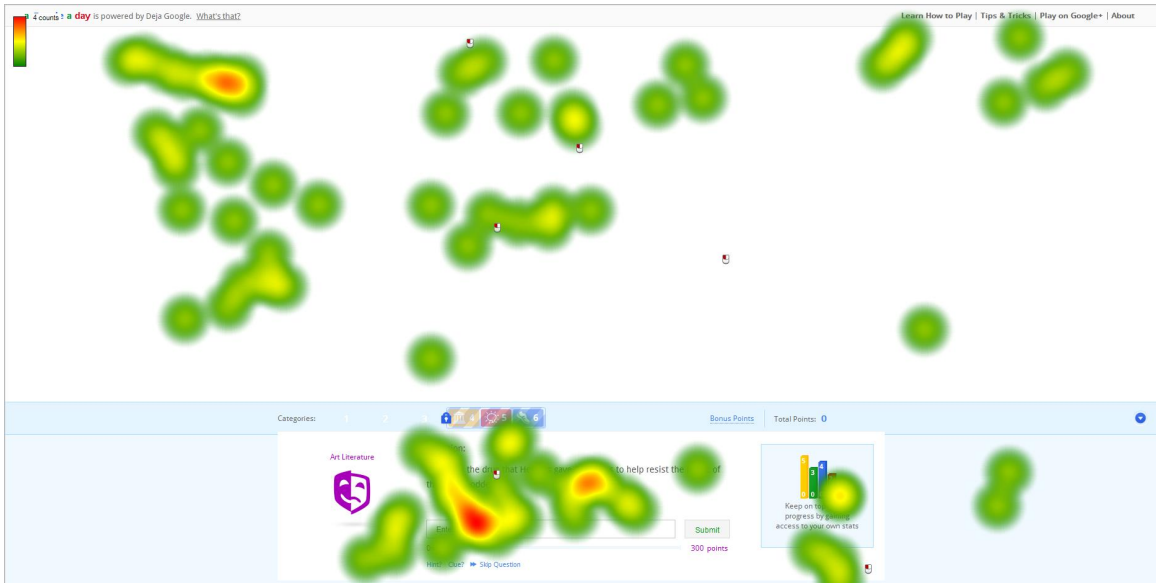

Figure 3 - The heat map display shows how much time the eye spent at any one spot during the sample time. Here we can see the player spent most of the time reading the question (below, next to the theatrical mask glyph) and near the top, at the search query.

The lab setting is useful, allowing demographic and other data to be easily be collected about participants and allowing control over variables that are not interest, while allowing instrumentation of novel systems that could not be easily deployed broadly. However, we have found that in the lab, researchers ask behavioral questions that do not originate with the study participant. While it seems like a small thing, in fact, questions and behaviors that are natural to ask the participant may never arise in a lab setting. [Grimes & Russell, ??] Researchers can learn a good deal about participants and their motivations in this way, but the observed behavior happens in a controlled and artificial setting and may not be representative of behavior that would be observed "in the wild".

In the lab-setting study, a person may invest more time to complete a task in the lab than they might otherwise to please the investigator [Dell et al., 2012]. In addition, laboratory studies are often expensive in terms of the time required to collect the data and the number of different people and systems that can be studied.

The micro-level of study is useful, but doesn't answer all of the questions one might have about game play in a more ordinary setting.

Meso level: How humans behave minute-by-minute

At the meso-level of study, research is done to determine how system users perform tasks of interest (play the game) and to collect information that cannot be gathered in any other way (e.g., direct observation of affective responses to a computer game).

The meso-level research approach is primarily to perform field studies of user behavior. That is, to collect data from participants in their natural environments as they go about their activities. Data collected in this manner tends to be less artificial than in lab studies, but also less controlled, both in positive ways (as when the participant does something completely unexpected), and negative ways (as when the presence of the researchers influences their natural behaviors).

In particular, we were interested in natural use phenomena (the full complex of interactions, ads, distractions, multitasking, etc.) in the AGAD game. Unlike a more deeply engaging game such as quest game, AGAD was always intended to be a short-duration, low-overhead, small-investment game play. So we wanted to understand how people would play the game: that is, what would cause them to start the game, why would they stop playing, what did they find engaging about the game, what was disruptive during game-play, and what were the drivers for returning to the game.

In a series of studies we would interview AGAD players as they used the game for the first time (to identify initial reactions and learnability questions), and later, we interviewed repeat players to identify reasons for returning to the game.

The 15 interviews (fortunately) identified almost no learnability issues—the game was simple enough that the design was straightforward.

However, the interviews also asked various AGAD questions of varying degrees of difficulty. One question we asked was "This member of the Lepidoptera increases its weight nearly 10,000-fold in a fortnight. What is its name?"

As we watched, we learned quickly that this question (as it was for many others) had challenging vocabulary and concepts. People didn't know what a "Lepidoptera" was, or what a fortnight was. This lead to interesting misreadings and varying reactions. This, we learned, is a problem for all such "question asking / trivia question" games.

In this particular case, the term "Lepidoptera" led many people into highly technical articles (e.g., from entomology journals) which in turn use words like "instar" and "mitochrondial transformation." But highly motivated game players would work through these issues and get to the answer, feeling a distinct sense of victory.

Other, more casual players, would simply give up at that point, causing us to add in the "Skip" question feature and "Clue" feature, which would show the number of letters in the answer with one or two letters shown. (The "Hint" button was there from the original design, but given that so many of our observations were about people getting frustrated, this suggested the "Clue" modification to help people who were having difficulty, but not to disclose anything to players who really wanted to solve the full challenge.)

From a game design perspective, this kind of meso-scale user behavior in a natural setting was invaluable. It's the kind of information that can deeply influence a system design by providing feedback *during* the course of play. It's the kind of player behavior that's impossible to abstract from micro-scale behaviors, and difficult to infer from the macro-level.

Macro level: How humans behave over days and in the large

While meso-scale studies are very natural, macro studies collect the most natural observations of people as they use systems in large quantities over longer periods of time. This is generally done by logging behavior from in-situ, natural, uninfluenced by the presence experimenters or observers. As the amount of log data collected increases, log studies increasingly include many different kinds of people, from all over the world, doing many different kinds of tasks. However, because of the way log data is gathered, much less is known about the people being observed or the context in which the observed behaviors occur than in field or lab studies.

Games also have begun using analytics over logs, especially in age of web-based games. [Thawonmas, 2008] used logs analysis to find bot players in MMORPG games, while [Itsuki, 2010] looked for "real-time money trading" in games. [Dow et

al., 2010] used both a meso-scale interview approach in conjunction with a macro-scale logs analysis to demonstrate how interactive game operators changed their use preferences over time.

In the case of AGAD, we closely examined logs for behaviors we expected, but also kept an eye open for behaviors we did not anticipate. (This become particularly important for behaviors that are impossible to see in our micro- or meso-scale studies. Marketing events and system errors being the most common, and most interesting. See Figure 7.)

There were several unexpected behaviors that were found through macro-scale logs analysis—behaviors that were, by definition—only possible to see once the game had been available for some time. One instance of this was the cumulative effect of people returning to the game, and then backing up to play previous days games. (This feature was removed when the game was redesigned to support social play.)

This led to a roughly 2 week rolling sliding window in the game data as players would go backward in time. We noticed this when the log data for a given day's challenge would strangely change *after the fact*. In essence, what this did was to give a very long tail of data for any particular search question. (Figure 4)

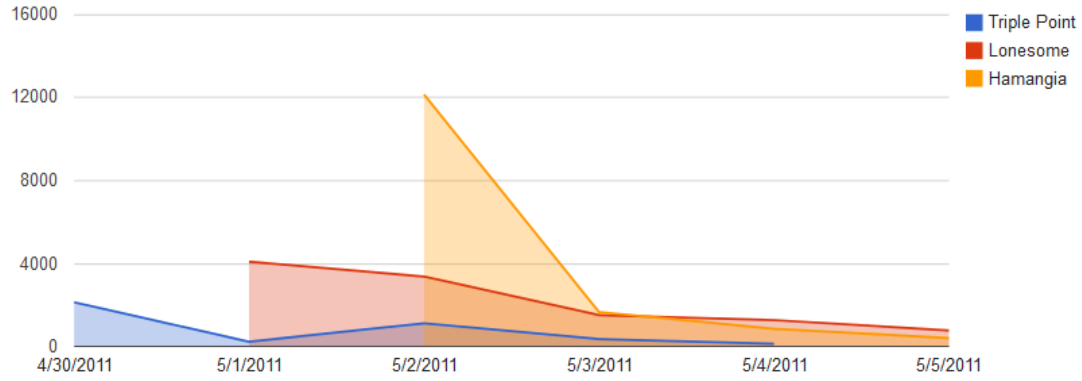

Figure 4 - Tracking data on questions by day. The graph shows player volume for 6 days for three different questions ("Triple point" "Lonesome" and "Hamangia"). "Triple Point" was introduced on 4/30/11, yet players kept backing up to that question 4 days later.

We also found through logs-analysis that players were doing a significantly larger number of queries / visit. Players had a nearly 2X query volume increase overall in distinct queries. (Figure 5) This increase in total search volume would not be explained simply by playing the game, but due to additional searches outside the scope of AGAD. Since one of the goals of the game was to help people become more

effective users, this might seem to be evidence that searchers were becoming *less* effective, not more. However, in a survey of 200 AGAD players in September, 2011, we discovered that after playing for several days, the perception was that they were taking on much more sophisticated search tasks, requiring additional searches. In fact, this was suggestive that the approach of additional time spent practicing actually improved the quality of their search skills.

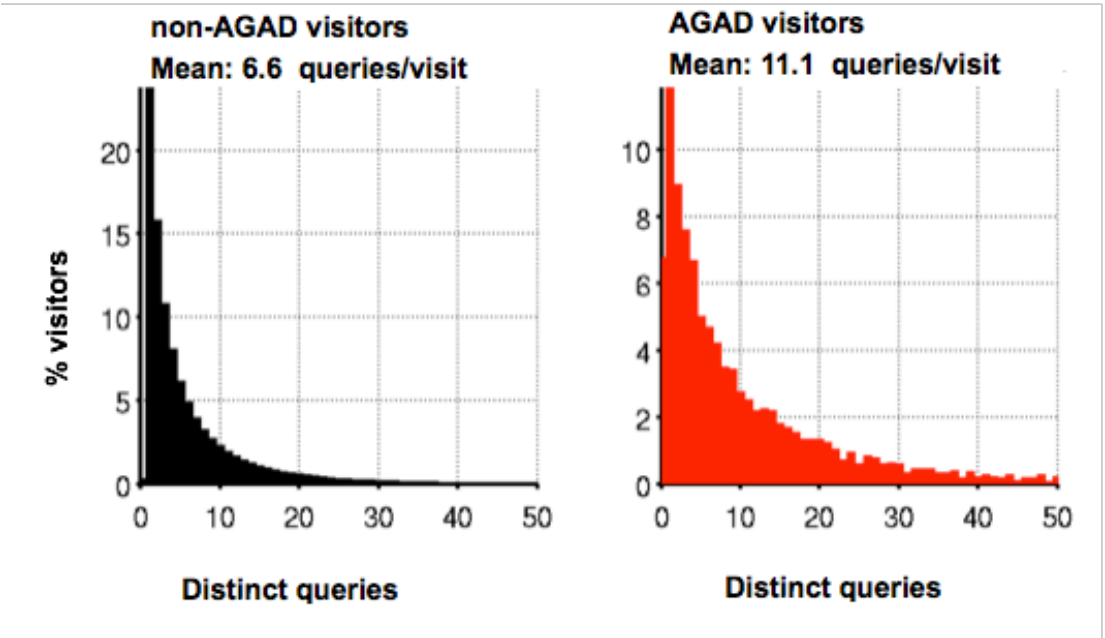

Figure 5: The number of distinct queries / search session, comparing AGAD-players vs. non-AGAD-players (that are matched to AGAD players by language, time-of-day).

Surprisingly, several months after launch, there was a start-of-year error in the way our co-marketing ads were being launched (in particular, in the New Year, our partners were not getting the feeds that were set up). (See Figure 6) Since this was over the holiday, we only checked the logs every two weeks, it was a fortnight before the error was noticed in the logs. It was through this error that we discovered that while AGAD has a large returning visitor rate, the genesis of their return was the presence of a reminder in a media stream. Once a player saw the reminder, they would go to the site and play (and often, as noted above, play more than one day at a time).

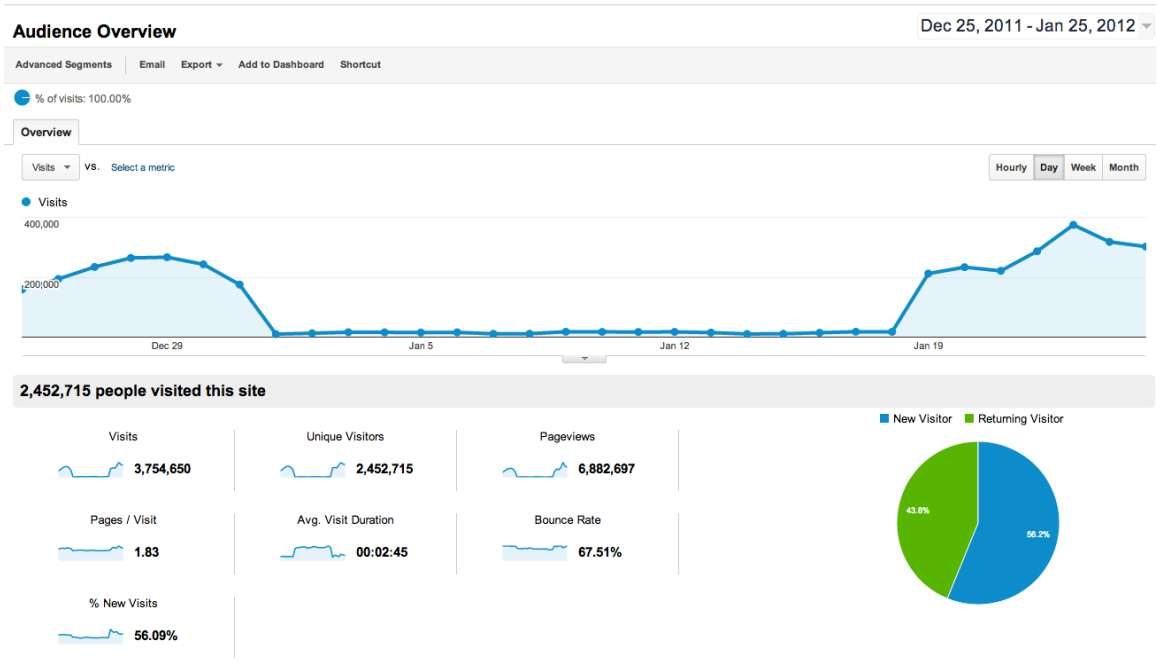

Figure 6. An example of the unexpected. The Analytics view of AGAD over the 2011 New Year's holiday. A mistake in advertising leads to a huge drop in players for several weeks.

Integrating research and design across the three levels: Each of the three levels of analysis gives a particular kind of data to drive development. Micro-level data informs decisions about the user interface and the operation of gameplay. For AGAD, this validated the particulars of the UI and guided item placement on the display.

Meanwhile, meso-level data gives designers critical feedback information about how a game is perceived and used in the real world. Some games (especially mobile games) function very differently depending on the physical and social context in which they're played.

From this kind of information, we learned that it is difficult to write questions that would be engaging without being intimidating. Player reactions are crucial. As a side-effect of the meso-level data analysis, we decided to add a "Feedback" button on the UI so players could connect directly with the design team. (And although it generates a fair number of empty messages (when people click on the button but don't enter any commentary) we found that complaints about mistakes in the question/answers could be fixed rapidly, once we knew about the issue. In effect, the feedback button became a meso-level data feed that

Macro-level data tells the story of how the game is performing in the large and in the wild. With macro-level data we knew that the game began performing well almost immediately. We understood the ways in which players returned to the site, how often they returned, and how long they would engage. As shown in Figure 7, we

could see the effects of various marketing efforts. Plateaus and spikes in the data told us we were on the right track as we tested various ways of getting the word out.

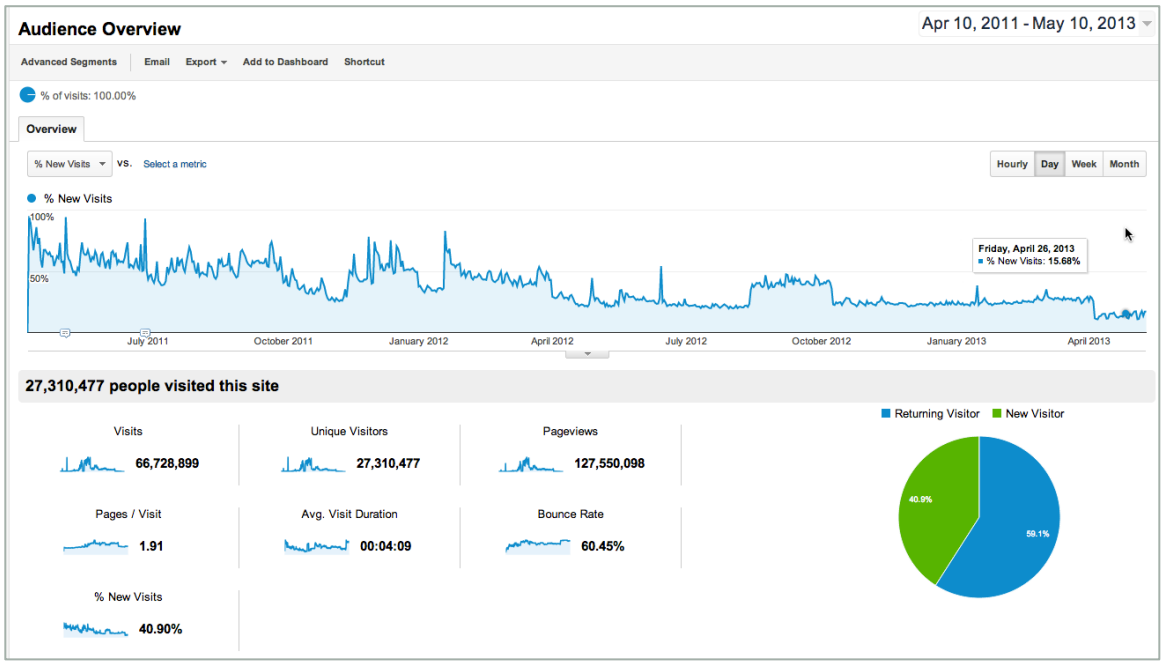

Figure 7 - Total audience participation in AGAD over a two year period. Various spikes and plateaus correspond to marketing events. Note the returning visitor rate in the lower right; this is an extremely high returning player rate.

Summary

A useful way to look at user behavior is at three different cognitive levels—first, the fast/rapid/millisecond level; second, the minute-to-minute behavior; and third, the effect of long/slow cognition over days / weeks. Each time scale reveals substantially different kinds of information about how the user/player uses, thinks-about, and responds to the game.

Examining user behavior across these three levels is as productive for understanding applications as it is for games. The same underlying UX research methods apply and can be immensely useful when developing a game.

The design space for a complex game is huge. User experience research, seen from a multi-scale perspective, gives insights into what opportunities and issues will arise in the game. What's more, looking at insights found at one scale often give insights into behaviors at another.

In general, a good software engineering practice is to tightly weave together not just great software engineering, but also great attention to the user experience at the moment-to-moment, and the play experience over an extended period.

Great games have these characteristics, performing on all of these levels of design simultaneously. They have great visuals; they have great audio; they have great backend engineering, interactions, controllers, and stories. Great games go on to have multiple editions and last for years. You can see how analyzing the game at multiple levels, with attention to different kinds of user interactions, will lead to improved design and user enjoyment overall.

References

Adar, E., Teevan, J. and Dumais, S.T. Large scale analysis of web revisitation patterns. In *Proceedings of CHI 2008*, 1197-1206.

Bernhaupt, R., ed. Evaluating User Experience in Games: Concepts and Methods. Springer, London, 2010

Dow, Steven P., et al. "Eliza meets the wizard-of-oz: blending machine and human control of embodied characters." Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, 2010.

Dumais, S.T., Cutrell, E., Cadiz, J.J., Jancke, G., Sarin, R. and Robbins, D.C. Stuff I've Seen: A system for personal information retrieval and re-use. In *Proceedings of SIGIR 2003*, 72-79.

Fullerton, T. *Game Design Workshop: A Playcentric Approach to Creating Innovative Games.* Morgan Kaufmann, Amsterdam (2008)

Itsuki, Hiroshi, et al. "Exploiting MMORPG log data toward efficient RMT player detection." Proceedings of the 7th International Conference on Advances in Computer Entertainment Technology. ACM, 2010.

Koeffel, Christina, Wolfgang Hochleitner, Jakob Leitner, Michael Haller, Arjan Geven, and Manfred Tscheligi. "Using heuristics to evaluate the overall user experience of video games and advanced interaction games." In Evaluating User Experience in Games, pp. 233-256. Springer London, 2010.

Nacke, Lennart E., Anders Drachen, Kai Kuikkaniemi, Joerg Niesenhaus, Hannu J. Korhonen, van den WM Hoogen, Karolien Poels, W. IJsselsteijn, and Y. Kort. "Playability and player experience research." In Proceedings of DiGRA. 2009.

Newell, Alan "Putting it all together." (Chapter 15) In Klahr, David, and Kenneth Kotovsky, eds*. Complex information processing: The impact of Herbert A. Simon*. Psychology Press, 1989.

Pinelle, David, Nelson Wong, and Tadeusz Stach. "Heuristic evaluation for games: usability principles for video game design." Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, 2008.

Rodden, K. and Leggett, M. Best of both worlds: Improving Gmail labels with the affordance of folders. In *Proceedings of CHI 2010*, 4587-4596.

Schaffer, N. Heuristic Evaluation of Games. In K. Isbister and N. Schaffer, eds., *Game Usability: Advice from the Experts for Advancing the Player Experience*. Morgan Kaufman, Amsterdam et al., 2008, 79-89.

Skinner, B.F. *The Behavior of Organisms: An Experimental Analysis*. Appleton-Century, Oxford, England, 1938.

Starbird, K. and Palen, L. Pass it on? Retweeting in mass emergencies. In *Proceedings of ISCRAM 2010*, 1-10.

Sweetser, P. and Wyeth, P. GameFlow: A Model for Evaluating Player Enjoyment in Games. Computers in Entertainment 3, 3 (2005), Art. 3A.

Thawonmas, Ruck, Yoshitaka Kashifuji, and Kuan-Ta Chen. "Detection of MMORPG bots based on behavior analysis." Proceedings of the 2008 International Conference on Advances in Computer Entertainment Technology. ACM, 2008.