

Real-Time Sweetspot: The Multiple Meanings of Game Company Playtests

Simon Niedenthal

Malmö University

School of Arts and Communication

Malmö, Sweden

simon.niedenthal@k3.mah.se

ABSTRACT

Game design, like gameplay, is situated. Though we find ourselves in a period of global growth and consolidation in the games industry, marked by broad changes in how design work is organized, our understanding of game design as it is currently practiced needs to be rooted in local contexts of production. One useful way to explore the situated-ness of game development is by tracing the implementation of playtesting of prototypes in game companies. The implementation of playtesting serves as an acknowledgement of the complexity of designing for the emergent properties of games, and also reveals attitudes towards the player. This case study of playtesting a real-time strategy (RTS) game under development at a Swedish game company is based upon observations of test sessions and interviews with employees from March 2006-February 2007. Specifically, this study will trace the various outcomes of a single game-balancing (“Sweetspot”) playtest conducted in March of 2006. This test serves as a locus of playtest meaning, and demonstrates that playtesting at the company is used to achieving clarity in the game design process, to support an evolutionary design methodology, and as a means of communicating the state of the game to outside actors. In short, playtesting has meaning in several contexts, both within and beyond the immediate design task at hand. Whether the results of a playtest session take the form of a numerical figure, a written report, or a fast scrawl in the lead designer’s notebook, they need to be interpreted carefully in the light of their complex nature.

Author Keywords

Game development, playtesting

In the spring of 2004, members of the local IGDA chapter gathered at Gloria’s, a sports bar, to hear two game company CEOs talk about what was wrong with game design. More specifically, the speakers took aim at the milestone development model as it was then practiced in the games industry. According to this model, a period of specification leading to a project plan is followed by a production period. If the game has been specified correctly, the separate components

come together at the end of the development period and produce a functioning game. The problem with the milestone model, they pointed out, is that it presumes that the final design outcome—an enjoyable game—can be adequately specified in pre-production. In fact, experience suggests otherwise: “quality cannot be defined in advance; it can only be determined once there is something tangible to test” (Walfisz, 2002). In opposition to the milestone method, the speakers championed an evolutionary game development process, which had been implemented in different ways at both of their companies. Here the design effort is focused not on the project plan and meeting milestones, but upon the production of a bi-weekly “build” of the game. Individual design contributions take place within a 10 day cycle, and on the ninth day of the cycle the current build is played by all employees (Walfisz et al, 2006).

This emphasis upon the progressive iteration of playable prototypes in the game development process reflects a larger interest in playtesting that has emerged in recent years from both positive and negative experiences within the games industry. It has become axiomatic that playtesting is the most important means for understanding the current state of a game under development; it is simply not possible to anticipate all of the emergent qualities of gameplay in advance (Salen and Zimmerman, 2003). Fundamental game design texts such as *Rules of play* even suggest rules of thumb that emphasize playtesting and design iteration, such as producing a playable prototype no more than 20% of the way into the production period. Production overruns (both time and budget) and high-profile failures within the games industry have increased the incentives for playtesting, as conducted by publishers and outside consultants.

But most articles on playtesting assume that the process is meaningful only in relation to design activities. The fact is that the current uses of playtesting go well beyond the simple goal of achieving clarity in the development process. Besides the significance of playtesting for design ends, new uses of playtesting have appeared, including, for example, the reporting of playtests in online fora and blogs as a form of pre-release publicity for games. On the Bungie.net site, the playtesting of Halo 3 serves as a teaser for the upcoming game:

Situated Play, Proceedings of DiGRA 2007 Conference

© 2007 Authors & Digital Games Research Association (DiGRA). Personal and educational classroom use of this paper is allowed, commercial use requires specific permission from the author.

Multiplayer, single player, you name it, folks are playing it. The lab formerly known as the multiplayer lab is now also the single player lab, where folks may go in at just about any time and start playing through a selection of single player Campaign missions. At this point, they're testing out the new AI and encounters as much as anything else. The graphics range from completely untextured placeholder surfaces, too (sic) brilliantly lit and detailed objects. But again, this is a gameplay test, not a graphics one. (Bungie, 2006).

Interviews with staff and observations of playtest sessions at the game development company from March 2006-February 2007 suggest a multi-faceted function of playtesting. Not only do individual playtests help answer design questions concerning usability, game balancing and other issues related to the features and playability of the game under development, but playtesting as a general practice has organizational significance as a cornerstone of the company's evolutionary development methodology. Finally, the ability to conduct in-house playtests—instead of relying upon the resources of the publisher—affords the game development company a degree of independence in relation to its corporate parent.

The game development company in question is a medium-sized Swedish company that is part of an international conglomerate. In 2002 the company was acquired by a multinational entertainment group based in Los Angeles. During middle of 2006, the game company had roughly 55 employees, though it is currently in a period of very rapid growth. Like many digital production houses, the company organizes itself by design function: there are departments for cinematics, level design, graphics, programming, consumer testing, quality assurance, sound and concept development. Besides these departments are supporting functions such as HR, bookkeeping, and marketing/consumer testing. Consumer testing is the only department which isn't actually a part of any specific development project. According the company's CEO, "it is deliberately held a bit separate in order to not taint the tests due to influence from the development team. In other words, we are trying to keep it as objective as possible in that regard."

The game currently under development is a real-time strategy game (RTS) for the PC platform. Like many RTS titles, the game involves managing resources while conducting combat missions and occupying strategic control points, primarily (though not exclusively) through a "god's eye" perspective, in single or multi-player modes. The game builds upon the design expertise and game engine of the company's previous RTS titles. Interest in the game (from the game press, as well as internally at the parent company) could be described as high, with successful appearances at E3 2006, as well as generally positive notices in PC gaming magazines and online fora. Currently the release date is set at late summer 2007.

Playtesting of the RTS game at the game company bears the traces of two related practices: usability testing and consumer testing. Usability testing has grown from the discipline of Human-computer interaction (HCI), and, when applied to games, is directed primarily towards minimizing player frustration at the interface (Laitinen, 2005). Usability evaluations can be performed as heuristic evaluations by experts, or through user testing. The usability approach, however, is in itself inadequate for game design; preventing a negative experience does not necessarily translate into a positive play experience (Lazzaro and Keeker, 2004).

Besides providing a range of methods for extracting information about the user experience, usability testing is often conducted under controlled conditions in a usability lab, which, in the case of game companies, allows not only for establishing conditions for player observation, but also serves as a means of securing the conditions of the playtest in a manner that is in keeping with the company's need for confidentiality. One of the first commitments to playtesting that the game company made was the establishment of a test lab. In the spring of 2006, eight networked machines were arranged in two rows of four, back to back, in a central space on the bottom floor of their design offices (fig. 1). This space has since been replaced by a larger facility rented in a nearby building, allowing playtesters to come and go without gaining access to the design offices.



Fig. 1 The playtest lab. c. summer 2006.

Playtesting at the game company is conducted under the aegis of consumer testing, a practice that is oriented towards customer satisfaction, not just usability. Prototype tests of consumer attitudes involving focus groups were conducted by the parent company before the greenlighting of the RTS game. But consumer testing does not require a prototype; thus, consumer testing is not limited to the middle and end of the design process, when pre-alpha, alpha and beta versions of the game exist, but can also be conducted in the conceptual stages at the very beginning of the design process, or during the design process upon other representations of the game, such as packaging imagery. A consumer testing perspective does not focus necessarily upon the game development task itself, but can also serve as a means of putting the designers in touch with the customer, getting to know their likes and dislikes, and thus avoiding the sort of tunnel vision that can occur on any

long design project. As the Consumer testing and Community manager (hired Nov. 2005) put it, consumer testing in a test lab allows game developers to “get closer to customers,” and to “develop a product that they don’t just “believe” that people will like, but rather one that they know is liked.” (Manual, 2005).

The immediate results of playtests are most useful to—and largely directed towards—the lead designer of the game. As with several of the other employees I spoke to, the lead designer was not trained explicitly as a game designer, but came to the practice through modding. The lead designer is in what the company’s CEO calls a “strong” design position, and he represents the game to media, at conventions such as E3, in online fora, and internally to the parent company. Although the lead designer does not design with an “ideal player” in mind, nor does he attempt to previsualize what players will do with modding tools, his advocacy for the player can be seen in his concern for creating the best possible tools for player content creation:

Lead designer: The biggest mods, the biggest ones that really shake the foundations and create new genres, they were the ones that nobody could ever think of. . . I think you are in deep trouble if you design your mod tools or your mapmaking tools in such a way that we want people to create this kind of mod. . . we just want to get (mod tools) out there and let Darwin take care of things (Interview, 2006-03-27)

At its most effective, playtesting of the RTS game has yielded immediate impressions to the lead designer, allowing him to answer specific questions about gameplay that can be integrated into the build of the game. In March and April 2006, several “Sweetspot” playtests were organized to help tune the balance of the game. Specifically, the lead designer wanted to find out whether increasing the hit points of combat units in the game (and thus effectively extending their lives) would lead to the development of more interesting strategies and give teams time to recover from disadvantageous positions:

Lead designer: “What happens if we give the units a lot more health, will that help people plan and execute strategies before all their units are killed? (Interview, 2006-03-27)

A second variable was also tested, game tempo, as influenced by the time taken for reinforcements to be delivered to the players. One build of the game was prepared in which the time for reinforcements was cut in half, from twenty to ten seconds.

The first Sweetspot test occurred in the playtest lab on the evening of Friday March 17th, 2006. Eight men, ages 21-38 participated in the test. Of those eight, 6 described themselves

as dedicated gamers, two as casual gamers. All were interested RTS players, and most indicated that they also played MMO, FPS and RPG games. After signing non-disclosure agreements, and filling out forms on gameplay habits, they played a warmup round for about 20 minutes, during which time they were coached, with one company employee for every two playtesters (three of the playtesters had previously played the game). Four different builds of the game were then tested in multiplayer mode, the base build of the game followed by three variants: a version with increased hit points for combat units, a version with faster tempo (reduced time for reinforcements to be delivered), and a version of the game with both increased hit points and reduced reinforcement time. Between each mission (each of which took about 20 minutes to play to conclusion) the testers filled out questionnaires. The lead designer was present for the entire test, moving behind the players and taking notes in a small notebook. After the final build was played, the players relaxed with soft drinks and gathered around the lead designer.

That evening, before going home, the lead designer implemented changes into the build of the game. He doubled the hit points and accordingly extended the life of the units:

Lead designer: The answer was a pretty resounding yes, it does help strategy and it does increase the feeling of control when they survive a bit longer and you can sort of get out of situations and resolve situations with wit, etc, that you could not do earlier, because the situation would arise and be resolved so quickly due to the fact that they died so quickly, there was no time to react in a tactical way. . .

Interviewer: What would the outcome be if the (unit) life was too long?

Lead designer: . . . It’s a balancing thing . . . (the RTS game) is a fast-paced action game . . . if you increased the health, it would slow down the game and be less action-packed. . . . pacing would be my biggest concern in terms of raising health further.

Interviewer: Have you implemented some of the changes you discovered?

Lead designer: The very night of the test, since it was so conclusive, I didn’t have to wait for a report to see it . . . in the test we did triple health, so I doubled health that very night, rebalanced all the units, and did some adjustments of the support weapons . . . so I have already implemented that . . . and that fared well when the team tested that all day this Thursday . . . I still think it can take more. . . there was one instance when people realized that rocket based artillery didn’t

have the same effect . . . so I did get some feedback on that . . . we're trying to do really broad strokes now . . . I want to do power-of-two balancing . . . (Interview, 2006-03-27).

A follow-up interview with one of the playtesters indicated that the lead designer's perception of the test was in alignment with the playtesters' experience of the different versions of the game. According to the playtester, extending the life of the units did in fact lead to more interesting game challenges:

Playtester: It felt like the (game) was more balanced when they had more life, because the game got more even between the teams. . . . in the beginning, if you were well organized, you could win by just holding three points and just having the right tactics when they came . . . but when your enemies had more life too, you had to really think about where to attack them from which angle, so they couldn't defend themselves. (Interview, 2006-03-24)

The playtester's favorite build was actually the last one, which combined extended unit life and quicker reinforcements (faster tempo). In the final match, his team came from a substantial deficit to eke out a narrow victory.

The new build—with extended life for combat units—was tested by the development team on Thursday March 23rd, and the results, with a few exceptions (see above) confirmed the lead designer's balancing decision. Playtesting by the entire production staff is a feature of the design process at the game company. Building upon agile software development methods, the in-house evolutionary game development process is based upon short-duration design iterations that take place over a two week period. The aim, as with agile development, is to establish a design process that is adaptive to the changes that tend to occur in complex development projects. The process at the game company is not just focused upon the continual development of the game build; it also integrates playtesting by the entire production staff into the process on the 9th day of the cycle. The company's CEO sees this as a natural means of taking advantage of the organization's resources:

CEO: most of the staff here are gamers, they play games . . . we have a full resource here of fantastically knowledgeable people that have very strong and most of the time very good opinions, it would just be wasted if we didn't listen to them. (Interview, 11-24-06)

The March 17th Sweetspot test yielded more than just impressions to the lead game designer that were immediately used to alter the game balance; it also resulted in a report (dated March 20th) containing quantitative and qualitative

results from test written by the Consumer testing manager. The report served to document the test results, as well as to communicate the results internally and externally. The playtest reports that are produced at the company follow a common format, beginning with test objectives and a description of test methodology, followed by quantitative results of the questionnaires, the test manager's observations and analysis of the session, and concluding with detailed information on the game outcomes (which side won) and the compiled comments from the questionnaires.

Besides the written comments, two types of quantitative figures were produced for the Sweetspot report. The first figures represented a summing of the ratings that each player gave to the different builds that were tested. Unlike the normal 10 point scale used in other playtests (in which 10 represents "perfection" and 1 represents "unacceptable"), the Sweetspot tests used ratings of 5 as representing optimal balance of tempo and challenge. Deviations from this figure represented elements of the game that were experienced as being out of balance. The second figure generated for the report was a Test index value (TIV), a figure that represents a benchmark for the test as a whole. The TIV is used to measure the overall result of the test, and is calculated on the average quantitative score of key test areas. The TIV for the Sweetspot test was 78.9.

The consumer testing manager's conclusions from the test drew upon his own observations as well as the written comments by the playtesters, and were very much in keeping with what was noticed by the lead designer (and experienced by the playtesters) regarding unit health and game balance. The report also demonstrates the sort of careful interpretation that is necessary to understand gameplay test results. It is not uncommon for the results of any user test to be inconclusive, or to point in contradictory directions, and this was the case with the build in which tempo and reinforcement time was varied. The quantitative results suggested that the playtesters preferred the build in which tempo was increased by decreasing reinforcement time (rated 202 with a rating of 200 representing optimal balance). However, two of the testers were very critical of the tempo build, and followed up with negative comments, because they felt that rapid access to new combat resources led to "wave after wave (of) boring massacres." The consumer test manager's conclusion prioritized the insights of the individual players at the expense of the summed collective experience of the group.

The consumer test manager's handling of this contradictory feedback indicates some of the uses of quantitative data, and suggests that numerical figures need to be carefully interpreted in the light of player comments and game context. Quantitative figures in playtest reports at the company have the value of focusing attention and helping set priorities. The consumer test manager has indicated that because playtesters tend to give rather high ratings when filling out questionnaires, he is most interested in ratings that vary substantially from the mean, especially very low ratings. These offer the opportunity for reflection and follow-up questions:

Consumer test manager: I would say that I am only interested in HUGE differentials here. Most people tend to give 7 - 9's for example, so it's interesting when you see a 3 or 4. Then you want to ask them in detail why they ranked it so "low". But the data AFTER the text isn't very useful in that sense. That has to be caught during the test, between the written feedback part and the discussion. It's like a warning light, telling us to pay attention to a specific question during the discussion. (Email 2-22-07)

There are, however, other limitations to the quantitative figures are generated for the playtest reports. Besides the obvious statistical shortcomings (the fact that playtest groups represent no more than 8 players' experience), the benchmark TIV figure is calculated on different test areas with every test; thus, the figure does not really allow direct comparison of one test with another.

A broader caveat when it comes to interpreting the Sweetspot playtest results is related to the composition of the test group, and the motivation of the playtesters. All of the testers were men (and it must be noted that, at the time of writing, not one woman has served as a playtester for the game). Moreover, despite the fact that the RTS game was initially greenlighted by the parent company because of its appeal to casual RTS players, three-quarters of the testers represented themselves as dedicated, rather than casual gamers. This is also problematic for the test process, as one finding that emerged from our observations of many tests at the company is that hard-core gamers are more feature-oriented, and less sensitive to usability problems than other players.

The predominately hard-core composition of the test group reflects a larger question about the motivation of the testers. The consumer test manager locates testers in several ways. First, he can draw upon database of 600 players that have indicated through the company website that they would be interested in serving as a tester for the game. He also works through referral. He doesn't have a marketing budget for recruiting testers, and compensation tends to take the form of in-kind merchandise (copies of the game when published, gift certificates to game stores). Without a more substantial "carrot," he has had difficulty motivating casual players to come in to test the game (email 2-1-07). But other motivations might be at play besides acquiring game merchandise. Conventional wisdom has it that playtesting is perceived as one way of moving into employment at a game company (Lindley and Sennersten, 2006), and some of our observations would seem to confirm this. Several of the playtesters we saw were capable of working at the company and have since applied for work there; at least two former playtesters now have positions working at the company. The CEO is aware of this dynamic, and its significance for the playtest results:

CEO: There is no way of getting untainted consumer data from the tests. The persons who are interested (in serving as playtesters) are a particular group of

persons, and they know they are at (the game company). (Interview, 11-24-06)

This motivation became apparent at the end of the Sweetspot playtest session, when the testers gathered reverently around the lead game designer.

Once the Sweetspot playtest report was finished, it was circulated internally, and then sent to the parent company. At this point, the playtest results took on a communicative role, subject to interpretation at the publisher/parent. Everyone connected with consumer testing at the game company—and the consumer test manager at the parent company as well—expressed sensitivity to the fact that playtest results—especially quantitative results—are open to misinterpretation. Not only do the numbers abstract the play experience, removing it from the context of the game, but they often bear a resemblance to the sort of scales used to rate games in game media, suggesting some sort of transferability. Although the CEO of the company has strongly backed the formation of the playtest lab, it is clear that he is skeptical towards any presumed objectivity in the playtesting process. He speaks often of the limitations of "tainted" consumer test data, and continually emphasizes the need to interpret the playtest results within the proper play context. It is apparent that his concerns are related to the way in which playtest results could function as a flawed medium of communication with the parent company:

CEO: Playtesting in the form we are making it . . . is never the truth, even if we try to quantify a lot of the data into numbers; it is never the truth. . . . that also means we need to be a little careful with exactly what data we send to L.A., not to hide anything from them, but just to make sure that they don't use what I call tainted data or skewed data and think about that as the truth, and then challenge us and ask us why we are not conforming to that truth. (Interview, 11-24-06)

It should be noted that, according to the CEO, playtest results are but one means by which the parent company keeps tabs on the progress of the RTS game, but not the most important; site visits by the producers and hands-on opportunities are considered much more telling. And, despite the CEO's concerns, there has never been a problem with misinterpretation of the quantitative results of the playtests by the parent company:

CEO: We have been afraid that they would, we have been at times careful about what we send to them . . . but so far no, we have been scared of them coming down on us like that, but so far they haven't. (Interview, 11-24-06)

These concerns point to a complex playtest relationship between the company and its parent. In fact, the consumer testing department at the parent company conducts its own playtests of the RTS game. Indeed, either the parent/publisher or the game company can request tests of the game, to be conducted at either site. The March 2006 Sweetspot test was commissioned by the game company, as befits the game balancing and design objectives for the test. The consumer testing manager in L.A. is also interested in specific gameplay dynamics of the game, such as complexity, depth and replayability, though usually with an eye towards gauging the appeal of the game, with a marketing rather than design objective. And the parent company also takes responsibility for testing other representations of the game, such as packaging imagery. The CEO of the game company sees the strategic benefit of having consumer testing resources in-house, rather than relying on the resources of the publisher/parent. The company is one of the few studios that doesn't outsource its playtesting to the publisher:

CEO: When it comes to other studios, I don't think there are that many in the world . . . that have their own consumer testing department, I think most of them rely on the publisher doing that kind of consumer testing, which for me is the completely wrong way of doing it.. because what happens is that the publisher does the consumer testing, they get the quantified data, without having the right context, without being able to filter the data. .. what happens is that they come with the test results and say "hey, your multiplayer experience score got lower, what happened?" . . . and we have no idea how the test was conducted. . . (Interview, 11-24-06)

In this sense, playtesting is one of the means by which the company negotiates its relationship to the parent company, and maintains a degree of independence.

The main lesson of this project is that playtesting is not just about the game—it also reflects how the company organizes its work, expresses its core values, as well as how it communicates with external actors. The complex significance of playtest results requires companies to keep a clear focus on what they hope to accomplish by playtesting; there is always a danger that the design and communications functions of playtests could work at cross purposes. Our interviews and observations allow us to draw a few tentative conclusions. Playtesting in game companies appears to work best when curious and perceptive game designers who believe in the process use playtests as a platform for conducting their own gameplay observation and informal ethnographies. Conversely,

increased expenditure on dedicated test facilities and observation rooms may result in more secure playtest conditions, but will not in and of itself lead to the design of better games. Much more importantly, playtesting must be embedded in a design culture that values diverse input to an iterative process. We are, in the end, doubtful that playtests can attain the status of a scientific practice. Increased abstraction and quantification of playtest results can only contribute to an undesirable situation in which playtests become the new milestones, serving as a means of reporting progress on the game to publishers and other outside actors, rather than as a means of better understanding the current state of the game.

ACKNOWLEDGEMENTS

This study was supported by funding provided by the Swedish Sparbankstiftelsen. Thanks to Elisabet Nilsson, Gunilla Svingby, Åsa Harvard, and the rest of the gang at the Malmö University Center for Game Studies for their feedback on an early version. I owe a debt of gratitude, as usual, to my colleague Micke Jakobsson, for his perceptive comments. Thanks to my students Markus Dahlström, Martin Bergöö and Teddy Persson for being good observers of playtest process.. Finally, many thanks to my reflective and clever colleagues at our game industry partner—you know who you are.

REFERENCES

1. Bungie (2006), <http://www.bungie.net/News/TopStory.aspx?cid=8517>
2. Laitinen, S (2005), Better Games through Usability Evaluation and Testing. Gamasutra. URL: http://www.gamasutra.com/features/20050623/laitinen_01_shtml
3. Lazzaro, Nicole and Kevin Keeker (2004), What's my method? A game show on games. *Proceedings of CHI 2004*.
4. Lindley, Craig and Charlotte Sennersten, An Innovation-Oriented Game Design Meta Model Integrating Industry, Research and Artistic Design Practices (unpublished manuscript).
5. *Manual* (2005), Spelare testar tidiga versioner, *Manual* #17, p. 14.
6. Salen, Katie and Eric Zimmerman (2003), *Rules of Play*, MIT press, Cambridge Ma.
7. Walfisz, Martin (2002), Evolutionary Game Development (unpublished manuscript).
8. Walfisz, Martin, Peter Zackariasson and Timothy Wilson (2006), Real Time Strategy, Evolutionary game development, *Business Horizons* 49, pp. 487-498.