Games User Research (GUR): Our Experience with and Evolution of Four Methods

George Amaya has been doing usability work since 1989. He has a Ph.D. in cognitive psychology with a minor in social psychology from Wayne State University in Detroit, Michigan. He started in usability during graduate school, working as an intern at General Motors. Upon graduation he accepted a teaching position at Seattle University and moved to the Pacific Northwest. Since then he's had the opportunity to work on many different products at a few different companies, leading to his current role as a user researcher for Microsoft Game Studios. His focus in recent years has been casual and social/party games.

John P. Davis is a user experience researcher in the entertainment experience group, where he works on products related to the Xbox and Windows gaming. He was also a user researcher in the games user research group in Microsoft Game Studios for several years. He has worked on several titles large and small, including Jade Empire, Inside Drive, and Poker Smash. John has co-instructed a successful tutorial on conducting usability research for games at the Game Developers Conference, and has been a regular contributor to the Conference.
on Computer-Human Interaction (CHI). Before coming to the games user research group, John was a usability engineer in the social computing group at Microsoft Research and an assistant professor of psychology at Seattle University. John earned his B.S. in psychology from Texas A & M University and his Ph.D. in experimental social and personality psychology from the University of Washington.

Daniel V. Gunn is a user research engineer within the user research group at Microsoft Game Studios and has been an avid gamer for over twenty-seven years. Prior to joining the group, he taught college courses for over seven years including research methods in psychology, psychology of learning, physiological psychology, and statistics. His work at MGS has been on titles such as Forza Motorsport for the Xbox and Forza Motorsport 2 for the Xbox 360 as well as Rise of Nations: Rise of Legends and Viva Piñata PC for the PC. Daniel received his B.S. in psychology from Wright State University and his M.A. and Ph.D. in experimental psychology, with an emphasis in human factors, from the University of Cincinnati.

Chuck Harrison has been active in the user research community since the early 1990s and has industry experience ranging from consumer products to back office and business applications. Before joining Microsoft Game Studios, Chuck worked at several other software companies such as Siebel, BMC, Intel, and Netscape. Chuck's primary gaming research focus has been on Microsoft's digitally distributed gaming experiences, such as Xbox LIVE Arcade and Web-based games. Over the years, Chuck has presented research at several HCI and game-related conferences and has published researched focused articles and book chapters. Chuck has a B.S. in human factors from the University of Idaho, an M.A. in engineering psychology from New Mexico State University and an M.B.A. from the University of Houston.

Randy J. Pagulayan manages the user research group at Microsoft Game Studios. As one of the early members, he helped pioneer user testing methods in the entertainment space while leading user research efforts on numerous games, including Top Spin, RalliSport Challenge, Age of Empires III, and Halo 2 and 3. Randy also has co-authored several book chapters on user-centered design in games, and has been published
in several scientific journals. Most recently, he was featured in a cover story in *Wired* (September 2007) representing the work of his team. Before joining Microsoft, Randy was part of the human factors and ergonomics group at Motorola. Randy has a B.A. in psychology from the University of Maryland, and a Ph.D. in experimental psychology from the University of Cincinnati.

**Bruce Phillips** is a user research engineer at Microsoft Game Studios and has worked there since 2001. He received his B.A. in psychology from Carleton University and his Ph.D. from the University of Victoria. Bruce has a special interest in multiplayer games and matchmaking systems.

**Dennis Wixon** joined Microsoft Surface Computing in February 2008. As the user research manager, Wixon leads a team responsible for directing, planning, executing, reporting, and evangelizing the research that measures the usability, usefulness, and desirability of Microsoft Surface. He believes his work on user interaction complements traditional market research because his team focuses specifically on the user's experience with working hardware and software prototypes. According to Wixon, user research is an important aspect of understanding and measuring the user's experience with Surface.

Wixon has worked at Microsoft Corp. for more than ten years. Before working for Microsoft Surface Computing, he was the user research manager for Microsoft Game Studios, where he and his team played a pivotal role in the success of Microsoft's popular games including the *Halo* franchise.

Before Microsoft, Wixon was a usability manager at Digital Equipment Corp., where he worked on and helped develop a number of important usability methods. A member of the human factors community for many years, Wixon also was one of the five founding members of the Association for Computing Machinery (ACM) Special Interest Group on Computer-Human Interaction (SIGCHI) Boston Chapter. Wixon holds a doctorate in social psychology from Clark University, and he has written more than twenty-five articles, columns and talks on Human-Computer Interaction (HCI) methods.
4.1 Objectives in the Chapter

For twenty-five years, user research methods have been applied to software in order to make it easier and more effective for people to use, and in some cases adopt and buy it. These methods have also been advocated as part of a movement to “humanize” technology and involve workers in decisions about their workplace (Ehn, 1988). The methods used have been derived from methods used in the social sciences, psychology, anthropology, and sociology. However, they have been adapted (sometimes to a point that obscures their roots) to the unique goals, constraints, and opportunities presented by development of commercial products and in-house tools. As these methods have evolved, they have become both more widely adopted and more effective in making products usable and useful.

The application of research methods to entertainment products and games in particular is relatively more recent (last ten years) and is evolving rapidly. Games represent an important domain with its own unique challenges and opportunities. In this chapter, we will discuss the methods that we have developed at Games User Research as part of Microsoft Game Studios.

In this chapter:

- We'll discuss some of the unique challenges and opportunities when performing research on games.
- We'll review a set of case studies of the application of research on games. For each method we will discuss
  - the context
  - the research problem/question
  - the research method for approaching the problem
  - learnings for both the method, and the game design implications
  - the broader context for this method and future developments
- We'll conclude with our take on the state of user research on games and its future.

4.2 The Opportunity and Challenge of Games Research

Applying research to game design presents unique challenges and opportunities. These stem from the following factors:

1. The purchase and use of games is discretionary. Unlike productivity applications no one has to learn how to play a game in order to get work done (Pagulayan, Keeker, Fuller, Wixon, & Romero, 2007). As such, research in
gaming must focus both on what users do and how they feel about what they are doing. In contrast, productivity applications can focus primarily on usefulness (does the application support required tasks) and usability (what are the costs, ease of learning, efficiency of use, minimizing errors). For discretionary products like games, these factors are secondary and only relevant to the extent that they affect the likelihood of a rich and engaging experience. That is, the user's aesthetic/emotional experience is paramount. While such considerations apply to all consumer products (for example, cameras, cell phones, etc.), consideration of the user's experience is primary for games since their sole purpose is to provide a positive emotional experience to the player.

2. **Games are a highly competitive space.** Several truisms reinforce this conclusion. Most games (almost 80 percent) lose money. A few games (20 percent) make most of the money (80 percent). Successful games are highly profitable. For example, *Goldeneye*, the movie, cost $10 million to make and generated $200 million in gross revenue. *Goldeneye* (1997), the game, cost somewhere around $25 million to make and generated $200 million in revenue. Games can also drive additional sales ranging from specialized controllers to action figures.

3. **There are many game studios and publishers.** Unlike most other markets for software (business applications, databases, etc.) in which a few producers have established a dominant position in each market, there are many suppliers and publishers of games. While the games industry is consolidating rapidly, it lags behind most other software markets in which two or three suppliers dominate.

4. **Games are extremely complex.** Many modern games push the technology envelope in terms of graphics, artificial intelligence, and system integration. For games played over the internet, the games can demand a very high level of performance.

5. **Games are very popular.** According to Entertainment Software Association (ESA, 2007) 67 percent of American heads of households play videogames. Twenty-four percent are over fifty. Thirty-eight percent are female. Internationally, the popularity of game can also drive national network infrastructure. Some writers have attributed the development of high-speed bandwidth in Korea to the popularity of computer games such as *Starcraft*.

6. **Games are big business.** In 2006, the revenue for video and computer game sales combined was $7.4 billion (ESA, 2007). These numbers do not include hardware sales for consoles. The growth in games has been strong for the past ten years increasing from 2.6 billion (1996) to today's figure. Game sales from last year have outpaced the growth of the U.S. national economy and have eclipsed film box office sales for a number of years. In addition to new consoles, games are increasingly appearing on new platforms such as cell phones.

7. **Games have an increasingly wide range of application.** Games are being considered as a platform not only to teach specific skills but also to teach transferable skills, and impart values and attitudes.
8. **The essence of a game rests in creative design.** Games do not attempt to automate an existing task the way many productivity products and applications do. Instead, they attempt to create unique, novel, and appealing experiences for their customers and users.

9. **Gaming has a development cycle that is well-adapted to user research.** A typical development cycle can begin by creating tools to build the game. Then the game is built in bits and pieces until a “playable” build is created. Once a playable build exists, the team focuses on refining and tuning the game play. This tuning period is an ideal time to conduct user research and a window during which findings can be readily incorporated into the game play. In contrast, it is all too often the case for productivity applications that as soon as the functionality is working and the product relatively crash free, management will press to ship the product. So, in this respect, games development represents a more fertile environment for user research.

10. **Most game developers and designers are dedicated to providing a great user experience.** One of our design partners put it well when asked what convinced him to work with user research—“we needed to know you cared as much about the game as we did.” By “the game” he meant that the game is great for the players. Every studio we have ever worked with has been dedicated to their vision of a great user experience. Unfortunately, all too often the designers and developers of productivity products are satisfied when the functionality is present.

11. **Gamers demand novelty.** While there are exceptions, people who use productivity applications rarely welcome innovation in UI design. They know how to do their work using a given tool and a new interface often represents both a learning cost and a productivity risk. In contrast, gamers welcome new challenges and new capabilities.

12. **There is an excellent framework for considering how user research can contribute to games.** Commercial games research does not have to live in a vacuum. The Mechanics, Dynamics, Aesthetics (MDA) framework developed by Mark LeBlanc (Hunike, LeBlanc, & Zubek, 2007) postulates the experience of games is based on the game mechanics—the elements of the game, the rules by which they interact and the goals of the game. As he put it, this is “what you can put in the box” after that it is out of your hands. When the user interacts with the mechanics he/she produces dynamics. These are patterns of behavior. These emerge from the interaction of user with the game and the combination of game mechanics. They are influenced by the player’s history and expectations. As the user plays the game he or she will draw conclusions about the game. These conclusions can take many forms. They could be judgments about the game, such as “this game is too easy.” They also represent “emotional” conclusions, like “I am having fun,” “I am frightened,” “I’ve accomplished something.” This framework is a good starting point for thinking about research in games and any product that has discretionary use.
In this chapter, we will give a brief overview of some well-established test methods and some new applications of the testing process. These methods are by no means definitive or complete. Instead, it is intended to provide inspiration to researchers and practitioners to build on and improve what the games user research group at Microsoft Game Studios has done. This is also by no means a complete inventory of our work. The methods we will discuss are:

1. Researching Play in the first hour: Playtest. An early and evolved method of assessing users' reactions to initial game experience.
3. Researching Play in the Real World: Collecting user data from beta
4. Researching Trials and Demos

4.3 Researching Play in the First Hour: Playtest

4.3.1 Behavioral vs. Attitudinal Data

In general, usability testing and Playtest were the first two methods we originally employed at Microsoft in the games user research (GUR) group. Although there are many ways to classify these methods (Olson & Moran, 1995), it’s useful to think of them at a general level in terms of the type of data they generate: behavioral and attitudinal. In usability testing, we focus on players’ behavior to reveal areas of the game in which the player experience does not match with the design intent. By measuring players’ behaviors, such as time to completion or frequency and nature of errors for a task, we are able to make explicit where these mismatches occur. In Playtesting, we focus on players’ opinions to illuminate areas of the game in which player experience does not map onto design intent. By explicitly tapping players’ opinions and attitudes, we hope to reveal more about what the subjective experience of playing the game is like for people and ensure it matches with what the designers hoped it to be. The demarcation between these two methods based on the type of data we collect is important because, as we will discuss in this section, some research questions that speak to the core of any game (for example, Is it fun?) cannot be answered by behavioral data alone.

4.3.2 The Problem

Conducting usability research on games required us to adapt the research methods used on productivity applications to account for the differences between these types of software. The most general difference between productivity applications and games is the overall design goal behind them: productivity applications are tools that help consumers be productive whereas games are “tools” that help consumers
have fun. While applying traditional usability methods to games allows us to answer many similar types of research questions we would have with a productivity application (for example, can users equip a weapon? can users turn on highlighting text?), we don’t often focus on whether a spreadsheet is fun to use. Indeed, the discount usability methods we commonly employ in our group are not particularly well-suited for quantifying user opinions in a reliable fashion (Nielsen, 1993; Pagulayan, Kecker, Fuller, Wixon, & Romero, 2007; Davis, Steury, & Pagulayan, 2005). Because the overarching goal users have when approaching a game is to have fun, we needed a way of quantifying it.

4.3.3 The Solution

Playtest is a large-sample, survey-based methodology we developed to measure and quantify users’ perceptions, attitudes, and opinions about a game. Like many survey-based methodologies, Playtest is one way for us to get reliable attitudinal data from people. That is, if we ran the same Playtest (with different groups of participants) on the exact same game multiple times, our findings would be relatively consistent across tests. This characteristic of Playtest is very important for three reasons: (1) it allows us to have some degree of confidence in our results from any particular Playtest; (2) over multiple tests it allows us to see the effects that specific changes to the game have on users’ enjoyment of that game; and (3) it allows us to reliably compare Playtest scores between games.

Playtest History at MGS

The application of survey research, with the sole intent on improving games that were in production, began over a decade ago at MGS in 1997. At that time, Playtest consisted of several PCs arranged in an open conference room. Participants would play a game for a short period of time and then answer a paper-and-pencil survey regarding their subjective impressions of the game. Although this first instantiation of Playtest was quite different from how it is today at GUR, and the data analysis process was quite cumbersome (that is, we had to meticulously transfer the paper-and-pencil survey data to a spreadsheet before we could even start analyzing it), it was an important first step towards measuring gamers’ attitudes and opinions.

Over the next several months, we made radical changes to the process to ensure the validity and reliability of our Playtest data. Our survey question format, participant recruitment process, participant instructions, and lab configuration all were refined and standardized. For example, some of the first few PC game tests were conducted on computers that had different hardware configurations. Because we wanted to ensure that the gaming experience from one PC to the next was the same from a performance and hardware perspective, we decided to equip the lab with identically configured PCs (in other words, same sound cards, video cards, and monitors). In addition, the overall goal of Playtest changed in these first few months: The new
goal of Playtest focused on how to make a game better as opposed to evaluating how good or bad a game was. Hence, early Playtesting and early iteration in a game's production cycle were stressed, and Playtests were focused on more experimental and exploratory questions (for example, do users like weapon A better than weapon B; which vehicle do users like the most, etc.) as opposed to strict evaluation.

The early pioneers of Playtest at MGS (Bill Fulton and Ramon Romero) envisioned it to be less of a method and more of a research situation in which several participants could be run quickly and cheaply. Unlike usability testing, Playtest didn't require a trained engineer to directly observe each individual participant. Without that restriction, the User Research Engineer could easily, quickly, and efficiently run numerous participants at once—enough to make quantification of survey data possible. This was, and still is, one of the key utilities of Playtest: Fast and efficient quantification of user opinions about a game.

**Playtest Today at MGS**

Our Playtest of today starts in a similar fashion to any other user research (UR) activity—defining the user. Thus, if we are testing a racing game, we want to bring in racing gamers. If we are testing a real-time strategy game, we want to bring in real-time strategy gamers. After consulting with team members for a particular title, we agree on a user profile and then recruit participants from our database based on that profile. Participants then proceed through the following sequence of events:

1. After arriving at our Playtest labs, participants are greeted by a Playtest Moderator. Playtest Moderators are highly trained and skilled individuals who specialize in running Playtests. Because each Playtest is essentially a miniature research study, all of the normal precautions that would go into typical research involving human participants need to be heeded (for example, ensuring ethical treatment of participants and minimizing potential sources of bias that can creep into any study).

2. Participants are asked to sign a nondisclosure agreement form. This is a legal document in which the participant agrees that they will not talk about the nature or content of the Playtest outside of Microsoft.

3. Moderators give participants information and instructions on how the session will proceed and then answer any questions participants may have.

4. Participants play the game for a specified amount of time (typically 1 hour) and then answer a series of structured questions about their experience on an electronic survey.

5. After completing their survey, each participant is thanked for their time and given their gratuity for participation.

The electronic collection of survey data allows the User Research Engineer to quickly collate the data after a session and begin analyzing almost immediately. We typically send out a preliminary report to the team members within a day
after the Playtest and a full report within a week. The full report is much more extensive than the preliminary report and includes a point by point reference to each of the issues uncovered in the Playtest as well a set of recommendations from us on how to fix those issues in the game. Usually, after consulting with the team and debrief sessions with the full report, we agree upon fixes and wait for implementation into the game. Often, after the fixes have been implemented, we will test again.

4.3.4 Things We’ve Learned Over the Past Ten Years

Over the past ten years we have identified numerous factors within Playtest that require close attention to get the most out of the process. However, there are factors we consider fundamental. Assuming that your sampling process (which recruiting is subsumed under) is robust (a topic that is beyond the scope of this chapter), the following three things are critically important to a successful Playtest program.

1. **Standardization.** This is a fundamental aspect of research methods and because most of our backgrounds are in experimental psychology, it became apparent to us very early on how much impact this principle can have on the data. Without standardization, the reliability (and validity) of the data is severely compromised. Imagine you wanted to compare users’ enjoyment between two games, so you run two separate Playtests: Playtest for Game A and Playtest for Game B. After analyzing the data, you come to the conclusion that Game B is rated more enjoyable than Game A. However, let’s also say you failed to standardize how you asked players the question; in fact, you worded the enjoyment question differently in each test. Now, the dilemma becomes: Are the differences between ratings of the two games because one is more enjoyable or because of the confounding factor of differences in question wording?

   Keep in mind that standardization is not confined solely to question wording. Other components, including participant instructions, communication between participants and moderators, length of gameplay, hardware used, etc. all have to be carefully standardized in order to help ensure the integrity of Playtest data.

2. **Reference Data.** This aspect ties into standardization and reliability in that in order to make informed decisions about data gathered from a Playtest, you need some sort of “yardstick” with which to compare it to. In other words, it is difficult to run a Playtest in a vacuum and generate actionable data. For example, if 70 percent of users indicated they really liked the graphics in your game, is that good, bad, or indifferent? It becomes much easier to make a decision about a data point when you have something with which to compare it. Thus, if you had data from a released game that was highly acclaimed for its graphics indicating it scored a 72 percent on the same graphics question, you would know that you were within the ballpark in your game. Over the past decade, we have built up a database full of Playtest scores from released games in several genres. This
allows us to compare scores on any Playtest to any game in our database, on a variety of different gameplay topics.

3. **Focused Research Questions.** Good research questions are a requisite for most research situations, and this principle applies to Playtest as well. “Good” questions are both answerable with the tools at the user research engineer’s disposal and are focused. With focused research questions, the user research engineer can more efficiently design the Playtest, write the survey questions, analyze the data, and generate recommendations for fixes. Unfocused research questions can lead you down the wrong road in which you become lost in possibilities while wasting precious resources (in other words, your time and money).

   Having focused research questions before a Playtest has many benefits:

   a. It helps you to ask the “right” Playtest questions and avoid asking unnecessary questions that bog the participant down.

      Asking the right Playtest questions (which are derived from your research questions) ensures that you are not adding a bunch of unnecessary questions to your survey; with each question added, the survey length increases as does the potential for respondent fatigue.

   b. It helps you determine when you are finished writing your survey.

      You know you are finished generating your Playtest survey when you can look at it and feel secure that the responses participants will give will answer all of your focused research questions.

   c. It helps you determine when you are finished analyzing your data.

      You know you are finished analyzing your data when you have answered each of your research questions with your analysis. Although you are free to do additional analyses beyond the scope of your original research questions, first and foremost those have to be answered.

   d. It helps you more efficiently write your recommendations.

      With focused research questions guiding your survey design, the data typically generated are much more actionable.

4.3.5 **Future of Playtest**

In our traditional description, Playtest is a tool used to quantify users’ attitudes, opinions, and perceptions about a game; yet, we have been rapidly expanding its breadth and deriving additional user research tools from it. For example, we found that collecting ancillary behavioral data during a Playtest provides a more holistic view of the user experience. Around 2002, we began building out our system for tracking user behavior in real-time as they played games in our Playtest labs. We call this TRUE instrumentation (Tracking Real-time User Experience) and, over the past several years, have successfully used it on numerous MGS titles across multiple genres (Kim et al., 2008; Romero, 2008; Schuh et al., 2008). In parallel, we have
been developing methods for Playtesting multiplayer game modes. These methods have also been buoyed by our TRUE instrumentation (see Chapter 15 of this book for an in-depth discussion of TRUE instrumentation).

Over the past decade, we have been continuously refining and improving our Playtest methodology. The processes have changed, survey questions have changed, scales have changed, and types of analyses have changed. But at the core, Playtest remains the same as in its early instantiation; Playtest gives us the ability to get data from users quickly, cheaply, and easily, in large enough groups to make quantification of user opinions possible. It gives us a reliable way to get evaluative feedback from users on various aspects of a game and allows us to compare user ratings of one game to another. In the end, Playtest gives us a very powerful tool with which to help improve games in development, before they are released into the real world. For additional examples of how Playtest research was used to improve games, see Pagulayan, Steury, Fulton, & Romero (2003) and Davis, Steury, & Pagulayan (2005).

4.4 Researching Social/Party Games

4.4.1 The Problem

Social/Party games are games where multiplayer mode is a significant part of the user experience. These games are designed to be entertaining for social gatherings of various sizes. Karaoke games (such as Singstar, 2004, or Karaoke Revolution, 2004), music games (such as Rock Band, 2007), and trivia or board games (such as Scene It? Lights, Camera, Action, 2007) are often placed in this genre. The terms “couch experience” or “shoulder-to-shoulder” are sometimes used to highlight the fact that these games may be played by a group of people who are in the same physical space, using the same gaming console. Social interaction is an important feature, considered as important, and sometimes more important, than the single player experience.

For example, our research into karaoke games like Singstar, Karaoke Revolution, and Rock Band (although Rock Band is not a karaoke game, it does have a karaoke element) has shown that most people who play karaoke games do not like to do so alone. A typical Singstar purchaser, for example, might play the game alone once or twice, but no more than that. There seems to be a stigma against playing a singing game alone, which is reflected in player comments that solo play “feels weird.” People who enjoy karaoke games prefer to play with groups of six people on average.

Compared to games played solo, social games present a new and unique set of research challenges. For example, if we know that people are playing a game in groups most of the time, or if the game is designed to be played primarily in a group/party setting, is it appropriate to test that game using only individuals? Traditional usability testing focuses on the experience of a single user. How can we adapt those methods to effectively gather data from groups of people?
While providing support for an Xbox 360 trivia game called *Scene It? Lights, Camera, Action*, we had an opportunity to address these questions. We learned that it is, indeed, very important to test using groups of people rather than individuals, and that traditional usability testing methods must be adjusted to be more effective in a group testing setting.

### 4.4.2 Why Test with Groups?

During usability testing of *Scene It? Lights, Camera, Action* and other social/party games for the Xbox 360, we observed several phenomena that led us to believe that group testing is indeed worth the cost and effort when working on a game in this genre.

1. Social/party games are designed to be played with groups of people in the same physical location. Testing such games in a group or social setting has higher ecological validity.

2. Playing with other people is more fun than playing alone. During our benchmark test of *Scene It? Lights, Camera, Action* we found a statistically significant increase in Overall Fun Rating among group players as opposed to solo players. We also found that solo players rated the game as being slower-paced than the group players did. The graphs below show the rating data from 135 benchmark participants, 35 who played the game solo and 100 who played in groups. The Overall Fun Rating asks players to rate “How fun was this game?” using a 5-point scale, where 1 means “Not Fun,” 3 means “Somewhat Fun,” and 5 means “Very Fun.” (See Figure 4.1). The Overall Pace Rating asks players to rate “How was the pace of this game?”, using a 5-point scale, where 1 means “Much too slow,” 3 means “About right,” and 5 means “Much too fast” (See Figure 4.2)

3. Many important behaviors occurred in group testing sessions (cheering, moving around, dancing, etc.) that we rarely saw, if ever, in individual testing sessions or group sessions with strangers.

4. The group presence led to emergent game strategies that we did not see in individual test sessions. For example, during *Scene It* testing, we saw a subset of players shouting out incorrect answers as a strategy, and also “pump faking” buzzing in.

5. Design elements of the game may facilitate or discourage social interaction. The designers need to be aware of how their work affects the social interaction because it could help or hurt the game.

### 4.4.3 Group Size

In a traditional usability test, the recruiting process is focused on gathering individuals who meet the selection criteria of the project. Recruiting groups of people for testing
**Figure 4.1**

Overall Fun Rating

**Figure 4.2**

Overall Pace Rating
social/party games required us to answer two important questions. What group size did we need for our research? Should these be groups of strangers or people who know each other?

We found that the best group size for a research project depended heavily on the type of game we were studying. A game like Scene It? Lights, Camera, Action could, potentially, be played by a very large number of people at a party, depending on how many controllers they have and whether or not they form teams. We decided to test groups of three to four as well as groups of six to eight players. The three-to-four-sized groups represent people who are playing the game without teams, while the six-to-eight-sized groups allowed us to observe the dynamics of team play (Scene It only supports four player entities but many more controllers, so once you have more than four controllers, they are limited to being on a maximum of four teams). Studying team play was very helpful to the game designers because it allowed us to learn about how players shared the controllers, how they communicated with their teammates, and how the different teams interacted with each other.

Karaoke games, like Singstar, have different group dynamics than Scene It. Our observations of people playing karaoke games revealed that groups with less than four people are not as animated or interactive as groups with four or more. The smaller groups are also quieter and less likely to move around. The nature of the genre seems to be such that people are much more comfortable singing with larger groups of people than they are in smaller groups. The average group size when playing these games in real life is six people, according to our participants. As a result, we decided to recruit groups of four to six people for any singing game studies, and also for any games in which social interactions are critical for the gaming experience.

Karaoke games typically allow two to play at any given time. This means that the remaining two to four people we recruited for the test session were sitting and watching someone else play. A fun social/party game is often one which entertains both players and observers, or facilitates entertaining social interaction so that these people can entertain themselves. We decided that these observers are an important part of the social/party game experience both from an ecological validity and a game design perspective, and their behaviors may add to or detract from the fun of the game.

### 4.4.4 Group Composition

After some trial and error, we decided that it was best to use groups of people who know each other, rather than random groups of strangers. We made this decision for several reasons:

1. There is more ecological validity testing with groups of acquaintances than with groups of strangers if studying social/party games.

2. Players seem to be more relaxed when in a group of people that they know. This makes it more likely that they will behave as they do naturally, talking out loud,
cheering, jumping up and down, etc. Players grouped with strangers are generally more reserved.

3. Emergent behaviors are more likely to be observed when the players are relaxed and with their friends.

4. Players are more likely to speak out loud with people they know than with strangers.

Once we decided to focus on recruiting groups of people who knew each other, we found that the best way to do so was to differentiate between core players and cohort groups. The core player is the person who the recruiters call. This person must meet all recruiting criteria. We then ask the core player to come up with a list of friends/family/acquaintances to act as the cohort group. The list of criteria for the cohort group is usually much more lenient than those used for the core player. The core player is like a party host, while the cohort group is the set of people the core player has invited to the party.

4.4.5 Adapting Usability to Group Settings

One of the first changes we had to make when testing social/party games in a group setting was to remove any think-out-loud instructions (Editors' note: for more information about think-aloud techniques, see Chapter 5 in this book). We quickly realized that traditional think-out-loud instructions just don't work when you are observing a group of four people playing a game. Four people thinking-out-loud at once would be chaos, and it is difficult (or even impossible) to stop one person's comments from influencing another person's thoughts in this group setting. Removing the think-out-loud instructions, however, did not mean that we sacrificed the option of gathering verbal data from our test participants.

We found that groups of people who know each other will ALWAYS talk about the game and what they think of it while they are playing. After testing 200 to 300 groups of people on a variety of social/party games we have found that all the groups that were made up of people who knew each other would talk without any instruction to do so. Plus many of the comments were fairly high quality and related to the design of the game.

Another change that we had to make to our testing methods while working with groups was that we had to refine our task lists. Structured task lists that included individual tasks focused on specific design elements or features did not work very well with groups of people. Task lists that are more general, and focused on the game play elements we wanted the group to experience were much more effective. For example, when testing Scene it? Lights, Camera, Action, we found that we gathered the most useful data when we had the group sit down and play through the game with little or no interaction with the researcher. This allowed the group to focus on the game, forget that they were being observed, and behave more naturally. The observers noted any behaviors of interest so they could ask the group about them after the game play session.
Free play at the beginning of a test session has turned out to be a good way to allow the test group to settle down and adjust to the test environment. For example, when testing music games like Singstar, we found that allowing the participants to choose their own play mode and song in the beginning helped them relax and adjust to the lab environment. Once everyone had a chance to sing a song of his or her choice, the moderator could then ask them to play using specific modes or songs that we were interested in observing.

Interrupting a group during a play session with questions about their experience or comments very often seems to distract the group from the game play and decreases their comfort with the lab setting. Such interruptions also seem to decrease the frequency of the banter that occurs when groups of acquaintances play a game. We decided that the comfort of the group and the banter were important data sources, so all of our moderators were instructed to note any questions that came up during the play session and ask them after group play was over. When we were done with our testing the researcher would go into the participant room and facilitate a discussion of questions and issues, very much like a focus group. We have found that spending time on discussion at the end of the test session allows the group to play comfortably without interruption, while still providing us with the data that we needed from the test session.

4.4.6 Future Developments

Today we are continuing to gather data on social/party games and refining our group methods. There are still many open questions to explore, such as how we can adapt group-testing methods for other audiences, such as children. Do our results generalize to other cultures, which may differ in the value they place on social interaction? These are just some of the issues we are exploring. The data from these group studies have helped our development teams think about new ways of enhancing social interaction during gaming. Thinking about the experience of observers as well as players is not something that has often taken place in the past and requires a new approach to game design. In general, with relatively simple adjustments to traditional usability testing methods, we’ve been able to increase the ecological validity of our work, while providing our designers and developers with data that helps them to envision new and exciting social gaming experiences.

4.5 Researching Play in the “Real-World”: Beta

4.5.1 The Problem

Playtest and usability are valuable tools for collecting consumer feedback on games. However, some game research questions may require hundreds or even thousands of participants to answer. For example, when trying to balance character progression...
in a massively multiplayer online game you may need to study many permutations or combinations of variables, often over an extended period of play. Sometimes you need many players interacting simultaneously to test matchmaking processes or other aspects of the game that require a critical mass of players. Beta testing can be very useful in these circumstances.

A beta version of a product (sometimes in the productivity world called community technology previews) is usually the first public, though often restricted, release of the product to consumers. Developers typically release beta versions of products towards the end of the development lifecycle, when the software is nearly feature complete but still needs fine-tuning. Typically, the test (or QA) departments on development teams have been responsible for running and collecting feedback during beta programs. Test teams employ betas to investigate hardware/software compatibility, to identify difficult to find bugs or other technical issues with the software. Putting prerelease versions of a product into the public's hands can help assess compatibility with a much wider range of computer hardware and software configurations than testers could reasonably replicate in a lab. Even a large test team cannot cover the same ground that hundreds or thousands of beta testers can.

While the public beta testing of PC games and productivity software is relatively common, the beta testing of console games is relatively new. Newer still is the use of beta not just for finding bugs, but for improving the core experience of the game itself, through either the tweaking of game mechanics or for game play balancing, to ensure that the game is neither too hard nor too easy. At Microsoft Game Studios, the user research team manages the beta program and we are developing methods to make beta a valuable consumer feedback tool, like usability and Playtest, for improving games. There are, however, unique benefits and challenges when collecting consumer feedback in a beta test. In this section, we describe some of these and present a short example of using beta on a game in development.

Besides the number of players you collect feedback from, beta testing differs in several other ways from the methods of collecting consumer data that we have discussed. First, the game has to be more stable. Because the game is provided to players outside of our testing facilities, usually downloaded from a secure website or through Xbox Live, we have much less control over the testing environment. This means that the game should not crash very often and there should be few bugs that frustrate or confuse players. Further, because it is the first exposure of the game to the public, it has to be sufficiently polished and fun to play. For these reasons, beta testing usually occurs late in the development cycle.

Second, because the testing occurs in the players' home, not in the Playtest or usability labs which are onsite, it is more difficult to collect qualitative feedback. Therefore, when analyzing the instrumentation data we get from beta participants we often do not have a lot of the context for interpreting the data that we can collect during in-house testing.

Third, because the testing occurs outside the lab, we have less control over the test itself. We have no control over how (or even whether) the beta participants play the game. In-house Playtesting enables us to provide instructions and monitor
participant behavior while controlling for external sources of variability, such as instructing players to use only the sniper rifle or to race only with the Austin Mini. Therefore, we must lower our expectations about managing game play behavior in beta. Further, while we may be able to update build versions of the game we are testing weekly (or even daily) when testing in-house, it is difficult to update builds of the game while testing 20,000 or more beta participants.

Fourth, and one of the main benefits, is that beta testing affords extended testing. Typically, when participants come to our Playtest labs they play only once, for between two and six hours. During beta tests, players play many times over an extended period. Most games involve a significant amount of learning and experimentation. By collecting data from players in their homes over the course of weeks or months, we are able to look at longer-term trends and patterns in player behavior.

In the remainder of this section we discuss how we used beta testing for *Shadowrun* (2007), a game published by Microsoft Game Studios. *Shadowrun* is a multiplayer first-person shooter game for the Xbox 360 and PC. At the beginning of the game, players are able to choose from one of four different character classes, each with different abilities, such as trolls who are strong but slow, elves that are quick and adept at magic, dwarves or humans. At the beginning of each new round of the game, players are able to purchase weapons and additional skills with money collected during prior rounds. This design permits players a great deal of customization for their character. Not only are there many initial options for the creation of their character, but over the course of the game the player can further tweak and enhance their attributes in many different ways.

### 4.5.2 Learnings from the Method, and Learnings Applied to Game Design Itself

One of the questions we were interested in answering in the beta test was the relative effectiveness of the weapons available to the players. There were several different weapons, each intended to serve a specific purpose. There was a shotgun for short-range combat, a sniper rifle for long-range encounters, and few weapons in-between. Ideally, the various weapons would enable players to engage in a variety of different combat strategies, providing for several different play-styles. For example, some people like to play stealthy and shoot from long range; other players like to run-and-gun and jump straight into the action. However, there was some concern on the development team that some of the weapons were less effective and fun than the others.

To answer these questions we used both questionnaire feedback and automated data collection. We used the questionnaires as we typically use them in Playtest, to collect attitudinal data about participants’ experiences. Questionnaires, however, have not proven reliable for collecting behavioral data. For a variety of reasons, we have found that participants are not good at accurately reporting data such as how often they play, what items they used when playing, or their success at the game.
Further, participants would not be able to report much of the data we needed, such as how far they were from opponents they “killed,” how often they shot a particular weapon, etc. To collect this data, we used a system that automatically recorded participant behaviors (Schuh et al., 2008). To collect data relevant to our question about weapon use and effectiveness, the game automatically logged every time a player purchased a weapon, every time a player scored a “kill” with a weapon, as well as logging the coordinates of the players in the game world when these events occurred.

When examining the data at various intervals during the beta test, we discovered that the majority of the time the players were purchasing only one gun, the submachine gun, and there were some weapons that players rarely purchased. Could this be a matter of players not having learned to use the other weapons? Were the other weapons too expensive? To begin answering this question, we broke down the analysis of weapon selection by the number of games a participant had played. As shown in Figure 4.3 (Weapons purchases by games played), the pattern in weapon preference was consistent across different experience levels. In fact, it appeared that the preference patterns for the weapons were even more exaggerated as players had more experience with the game. This allowed us to rule out inexperience with the game as a plausible explanation for purchase patterns and effectiveness with particular weapons.

Was the submachine gun overpowered, or perhaps it was more popular because they were cheap to purchase? To investigate further, we calculated a kills-to-purchase ratio, the number of “kills” logged for that weapon divided by the number of times players purchased that weapon. As shown in Figure 4.4 (Kills-to-purchase
4.5 RESEARCHING PLAY IN THE "REAL-WORLD": BETA

![Bar chart showing weapon usage](image)

**Kills-to-purchase ratio**

(not only was the submachine gun the most popular purchase, it also appeared to be the most effective weapon in the game. For every 100 submachine guns purchased, there were 39 "kills," compared to 6 "kills" for every 100 rifles purchased. These types of analyses would be difficult or impossible using Playtest or usability as players need extended interaction with the game.

In the end, the beta test was useful for answering the weapon balance questions we had, as well as many other questions we had going into the beta test. In the end, we were able to provide this data to the development team, which then tweaked the weapon parameters to achieve their desired result—a variety of different weapons effective in different combat situations. However, the beta test provided many challenges we had to overcome that are not typically present in usability or Playtesting.

### 4.5.3 Future Developments for Beta

As discussed earlier, beta differs from Playtest and usability in several ways, many of which create unique opportunities and hurdles for beta testing. In order to improve beta as a consumer feedback and game design tool, there are several areas we need to improve. First, we need to move beta testing to occur earlier in the development cycle. Development and marketing teams are rightly concerned about putting early builds of the game into beta. If a potential consumer of the game has a bad initial impression, if it looks or plays poorly, it can affect consumer expectations of the final product and, potentially, affect sales of the game. Currently, most developers err on the side of polish over getting the game into beta for early consumer feedback.
Invariably, however, our experience has been that once teams see the benefit of consumer feedback, they wish they had started beta testing earlier in the development cycle. One of the main benefits of early beta testing is the opportunity for iteration. As we collect feedback from participants and designs are changed, we can distribute new builds to verify fixes or to iterate on new issues that arise.

Second, as we move towards focusing on game design and balance in beta, and relatively less so on technical issues such as stability and bug finding, how we recruit and select participants for beta becomes more important. Traditionally, beta testing has been conducted by technically savvy consumers who understand, and can cope with, instability and less user-friendly beta builds. This makes a lot of sense if you are primarily interested in technical feedback. However, when you start using beta to collect gameplay feedback, such as difficulty levels, you need to be more cognizant of who you recruit for your betas. If you want to look at balancing the game for a wide variety of different types of players, such as those with little experience to those who play a great deal, you will need to come up with a recruiting plan to ensure that these different gaming profiles are represented in your participant pool.

Another consideration for participants in a beta is attrition. In Playtest and usability, participants typically participate in one test: they come to the lab, play the game for a few hours, and the test is complete. Depending on the type of data you are collecting in a beta, like the weapon preference question for Shadowrun, you may need participants to keep playing anywhere from a day to several months. Further, some tests require a critical mass of simultaneous testers, while others only require a particular number over the course of the beta. How you keep players engaged in the beta and participating in the feedback process will vary, but managing the beta community to keep participants playing the game is critical for data collection.

One of the methods we have developed to minimize attrition is to invite existing communities of gamers and friends-of-friends. We have found that having other familiar players to game with helps maintain interest and participation in the beta. Other tools include contests and tournaments, opportunities to play with and interact with the development team, and so on. Lastly, it is very important to be responsive to participant feedback. Participants who feel like the developers care about their participation and feedback are much more likely to continue participating that those who do not.

4.6 The Importance of First Impressions: Trials and Demos

4.6.1 The Problem

Computer and videogame publishers have a limited number of ways to get their games into customers’ hands. Historically, game publishers have sold their games
through traditional retail channels (for example, selling a game off the shelf or through an online storefront) and nontraditional ones, like the Internet (in other words, downloading a game to a PC or playing through a Web browser).

Until recently, console-based games were sold exclusively through retail channels. However, the latest generation of gaming consoles has opened up a new purchase path for the console audience, one in which consumers can digitally download games and game-related content (for example, game expansion packs) directly to their console hard drives through broadband Internet connections.

Microsoft’s Xbox Live Arcade (XBLA) was the first console-based service to allow consumers to download and try a free, limited trial version of a game, and then immediately purchase and download the full version if they wished. For all practical purposes, this was the same basic model that publishers and consumers had used for many years in the PC space.

Using this model, Xbox Live Arcade found a sweet spot with the Xbox 360 customer base and exceeded almost every business expectation in terms of sales. A high percentage of consumers tended to buy the full version of the Xbox Live Arcade game after they had played the game’s trial. These high “conversion rates” far exceeded those typically seen in the downloadable PC game market, even for XBLA games with the lowest conversion rates on the platform.

As the XBLA service matured and more and more games were made available to consumers, conversion rates began to flatten out with instances where games that were considered very good, in terms of consumer feedback during production and or positive industry reviews post release, occasionally experienced low conversion rates and sales that did not meet expectations, even though the game’s trial was downloaded in very high volumes. Put another way, the game design was solid, marketing and merchandising were effective in ensuring downloads but the trials were not “closing the deal.”

Potential purchasers of retail console videogames can gather a lot of information before they decide whether or not to purchase (for example, read online game reviews, play games on an in-store kiosk or using a game demo disk, or rent a game through an online service like GameFly.com or at the local Blockbuster Video). Xbox Live Arcade customers, on the other hand, have to rely almost exclusively on the game’s trial experience to collect information about the game to inform their purchased decisions. Consequently, even a great game is at risk of selling poorly if the trial experience is not good. The clear, critical connection between the trial game experience and consumers’ purchase decisions on XBLA led to an important question for the business: “How do we create good trial experiences for Xbox Live Arcade games that lead to better sales?”

4.6.2 Methodology

The XBLA team understood that many factors, including the quality of the trial experience, figure into a consumer’s purchase decision when considering whether
to buy an XBLA game. Is the genre interesting? Is the game known or is it brand new? How much does it cost? Are there other new games on the market? Was the game marketed well? The team also understood that of all these variables, the one that they have the most control over is the quality of the trial game.

When User Research was asked to look into the question of trial quality and how to improve it, there were already several dozen XBLA games in the Live ecosystem. This already-available information gave the team the opportunity to look at both the games’ trial experiences and sales data to understand how each game performed in the market. The existing data allowed us to focus on particular games, such as those with sales and conversion rates that either exceeded or fell short of expectations.

However, before attempting to understand how and why specific games performed well or poorly, we sought to better understand the elements that make a good trial game by first hashing out an understanding of the general goals of a trial, and then thoroughly reviewing the trial experiences for existing XBLA games. The goals for trial games include showcasing some of the “cool” and interesting features of the game, and giving the player enough content and features to give him or her a sense of what the full game would be like, while not giving away so much that the player will not be enticed to buy the full version. We have seen that in games which tested very well, won awards and then failed to meet projected sales numbers. While we can’t be sure of the cause, one highly plausible hypothesis is that the demo gave away too much. Armed with a better understanding of what trials are supposed to do and a comprehensive knowledge of the content of existing XBLA trials, we developed a set of “trial game design heuristics.” The heuristics, which were intended to be used as a “first pass” assessment of the quality of trial games, were based on principles culled from publisher best practices, observations from past usability tests and Playtests that had included game trials, longstanding usability guidelines, and designer intuition.

Two sets of trial game heuristics were created. The first set contained questions about the trial that were designed to be objective and unambiguously answered for each trial game (for example, “Does the trial contain a controller map?”). The second set of heuristics was less objective, but each heuristic contained clear goals and examples (for example, “Is the trial’s upsell message written in a positive tone?”). The heuristics were then informally tested for inter-rater reliability by having several user research engineers apply the heuristics to an identical set of games. Each of the individual heuristic questions was then assessed to identify those where evaluators differed in their assessments and modify them, remove redundant questions and refine the descriptions and examples of how each of the heuristics should be applied.

While the heuristics offered a way for user researchers to perform a “first cut” professional evaluation of trial games, a concurrent effort was underway to collect data from consumers to gauge their perceptions of trial experiences. To do so, user research engineers modified existing MGS Playtesting methods to make them more appropriate for a trial experience, which differs in important ways from a full
game. Both full games and trial games should be “fun,” appropriately paced, and appropriately challenging. Trial games, however, are unique in that they are also sales tools that, like advertisements, must contain enough game content (for example, features and gameplay) to allow the player to have a good idea about what the full game will be like. In addition, the trial game must provide clear “upsell” messages that describe the important (and valuable) features and gameplay elements that the full game will have that the trial does not, and an easy path to purchase the full game. Modifications to our traditional Playtest questions included paring down the number to be more appropriate to shorter trial games. We also included questions designed to tap the trial-specific aspects of the experience, such as whether players felt they had adequate information to allow them to make a purchase decision and how they felt about the frequency and content of the upsell messages.

In our trial Playtests, players were given a choice of trials that they could play as much or as little as they wanted. After playing a trial, they then answered questions about some basic elements of the experience (for example, fun, challenge, pace). In addition, they answered several of the trial-specific questions described earlier (for example, “Did the trial provide you with enough information to make an informed purchase decision?” and “Did the upsell messages provide clear information about what additional features were available in the full version of the game?”). The goal was to create a tool to measure the overall quality of a trial from the consumer’s perspective, and its effectiveness at meeting its goals as a “sales tool” for the full version of the game.

The modified Playtest method was then piloted using ten trial games that had already been released to market via XBLA. There were several goals for the pilot test. First, we wanted to assess the testing method to ensure consumers were experiencing and evaluating the trial games in a manner that mimicked, to the extent possible in a lab environment, how they might at home. Additional goals of the pilot test were to assess the specific trial games in the Playtest, extract general issues that consistently arose in the trials and recommendations to address them, and to assess the validity of the heuristics described earlier.

An additional goal for collecting data from consumers was to determine whether the multiple metrics could be combined and presented to the development team as a single score to make it easier to consume and understand. A factor analysis of the questionnaire responses from consumers established that there were four components composed of related items (these were Fun/Engagement; Challenge; Enough Information; Up-sell quality/frequency). Weightings for each of the factors was assigned based on the amount of total variance each accounted for, the weighted factors were combined and converted to a score out of 100 (for example, Game “Foo’s” combined trial score is 67); this score was referred to as the game’s “C-Score” or Consumer Score. Each individual metric was also reported, but providing the data in this fashion helped the team quickly understand the results because it was in a format familiar to them from Gamerankings and Metacritic-measures that are generally considered industry standards of a game’s quality.
Reporting the C-Score and metrics together allowed the game development team to quickly understand the current state of a game's trial user experience or if the trial had been tested multiple times, if its user experience was improving.

Figure 4.5 shows how the second iteration of a trial showed significant improvement on both the core Playtesting metrics as well as the trial's overall score.

By the end of the method development phase, the user research team had created and refined two sets of heuristics that could be applied to any XBLA trial, a new Playtesting method designed specifically to determine the trial's quality, and a composite score for trials that was easily distributed to and understood by the development team. In addition, the team generated a set of general trial design best practices based on the themes that emerged from the consumer data and had a running start on creating a rich comparison database of consumer assessments of a trial games. The next step was to continue to apply these tools to released games in order to collected additional data on trial experience to further grow the comparison database, refine the consumer trial game quality metric and expand on the trial game best practices document that can aid developers as they create trial game experiences. Importantly, the methods were then applied to in-development games while there was still time in the schedule to iterate on them to help ensure that their trials were the best they could be.

4.6.3 What We Learned

Over the course of this research program, we learned some valuable lessons about both trial games and general development practices. Although we were
able to create a set of best practices and iteratively evaluate a game’s trial experience prior to release, it became evident that a well-developed trial could not guarantee that a game would sell well; after all, if the game itself is not fun or well-designed, a good trial experience is lipstick on a pig. We also discovered that certain games seemed robust enough to overcome a non-optimal trial. Games with very well-known intellectual property or an existing following can overcome a poor trial. Although those types of games may sell well, they likely could have sold even better if their trials were better; those games probably left money on the table.

On the flip side of the games that sell well regardless of the quality of the trial, is the reality that certain games are at a higher risk of poor sales if they have a non-optimal trial. Specifically, original games or games that fit into more of a niche market have a more critical need to showcase their wares in a good, if not spectacular, trial. As the trial experience is likely to be the consumers’ only mechanism to determine the game’s “fun” or “quality,” there is very little room for error for these unfamiliar games.

4.6.4 Broader Context

This work not only empirically demonstrated the importance of game trials but also that their effectiveness can be measured with consumer data. What can be measured can be improved. This work has led to a subtle but important paradigm shift in the publishing model within Xbox Live Arcade. Trials were once little more than an afterthought in the development cycle, often not being considered until shortly before a game was released to the public. Now, there are multiple check points that focus on the game trial. In addition, it is now considered customary that trials are regularly investigated with usability tests and Playtests in order to ensure that they deliver the optimal user experience. This new focus on the trial experience has also required the game developer to better plan their trial experience and build it out much earlier in the development cycle so that it’s playable enough to gather player feedback early, while there is still time to make substantive changes. This new focus is well summarized by this quote from an XBLA executive “it is marketing’s and merchandizing’s job to get people to the trial, to download it; it is the trial’s job to close the sale.”

This work also has broader implications beyond Xbox Live Arcade games. The new console generation has given a much larger user population access to demos for full retail games that players can freely download to their hard drive and play; what was once a niche focused on the most hardcore audience (for example, those who purchased gaming magazines) is now available to the masses.

Although the goals and utility of a full game demo may be somewhat different than those of an XBLA trial, many of the design goals and best practice are similar. To this end, user research is expanding the scope of the trial research to cover full game demos as well.
4.7 Conclusion

User research at Microsoft Game Studios has evolved over the last ten years, and we will continue to explore methodological innovations that produce reliable, valid, and timely data that informs design and makes for exciting products. This chapter briefly describes four of our methods at a "high" level. Our accompanying chapter on instrumentation covers that topic in depth. Throughout our evolution we have been guided by some simple yet fundamental tenets which we call "the golden six."

1. Empirical data about users is the core of our contribution to product success.
2. Partnership with designers is essential when the product goal is a compelling experience.
3. Judicious investment in tools and techniques pays off by allowing us to generate large amounts of empirical data in a timely way.
4. Creating, evaluating and then standardizing methods is the key to reliability, validity, and efficiency.
5. Addressing management goals, for example, how much fun are our games and how can we make them better is the path to long term success.
6. Methods and their output must map onto both the culture and the development processes of your partners and your organization.

The past ten years have been highly rewarding for all of us involved in the development of tools and methods for Games User Research. We have been especially pleased not only by the success of MGS games but also by the recognition in the games and popular press:

"Still the best in terms of developer support understanding how to make great games and a strong vision. Their usability is by far the best in the industry."

Quoted from Game Developer magazine discussing Microsoft Game Studios in their annual publisher poll issue (Wilson, 2007, p. 12). Halo 3 How Microsoft Labs Invented a New Science of Play” (Wired magazine, cover story September 2007)

We look forward to the next ten years.

4.8 Acknowledgments

We want to thank all the members of the games user research group (past and present). They all contributed to the evolution of our thinking about game evaluation and research. We also want to thank Microsoft Game Studios Management, especially Ed Fries and Shane Kim, who supported this work over the years. Thanks
to David Holmes, our director, who has led us effectively over the last five years. We also thank the numerous studios and designers we have worked with over the years. We have learned much from them and enjoyed close collaboration. Finally, we thank the thousands of people who have participated in Playtests, our moderators who have actually run the Playtests, and the gamers who have loved our products and given us constructive feedback.

The opinions expressed here are those of the authors and do not necessarily reflect the views of Microsoft Game Studios or Microsoft Corporation.

4.9 References


