Score One for Quality!

Using Games to Improve Product Quality

Joshua Williams
Senior Software Design Engineer in Test
Windows Security
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052
(425) 703-1059
joshw@microsoft.com

Ross Smith
Director of Test
Windows Security
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052
(425) 706-3982
rosss@microsoft.com

1 Abstract

In this paper we describe how using a game can improve both the quality of a product, but the quality of life of the employees as well. We call this kind of game a "Productivity Game."

Productivity Games[i], as a sub-category of Serious Games[ii], attract players to perform work that humans are good at, but computers currently are not. Although computers offer tremendous opportunities for automation and calculation, some tasks, such as analyzing images, have proven to be difficult and error-prone and therefore lower the quality and usefulness of the output. For tasks such as this, human computation can be much more effective. Additionally, by framing the task in the form of a game, we are able to quickly and effectively communicate the objective, and achieve higher engagement from a community of employees as players of the game.

We will showcase a real Productivity Game taken directly from the Windows development process to highlight this integration and its benefits. The "Windows Language Quality Game" encourages native language speakers to perform the job of traditional software localizers and enhances an otherwise difficult and expensive business process with a "serious game". This has resulted in players who enjoy the opportunity to participate and contribute. It has also resulted in a cost-effective way to improve the quality of native language editions of Microsoft Windows.

The use of Productivity Games has broad implications across how employees are managed, and how employers communicate organizational objectives to their staff. Games in the workplace can be used as substitutes for leadership, which are more applicable and engaging to younger employees.

2 Introduction

The global business challenges of the 21st century require creative approaches and innovative solutions. Traditional methodologies for solving problems are evolving to create hybrid solutions that embrace new collaborative roles for humans and their use of computers. Technology is facilitating these hybrid solutions by enabling a large number of humans to focus on a problem and then easily aggregate their input. This has opened up the opportunity to innovate and creatively solve many business challenges.

In tandem, a generation gap has begun to appear between the established workforce and the Gen-Y[iii] and Millennial generations which are now filling the ranks of young employees and college hires. This younger generation brings its own priorities, communication patterns and perspectives to the workplace, as have previous generations, but in this case the gap is larger, and the challenges even greater[iv]. This younger generation is often referred to as "the gamer generation", as video games have been central to their lives[v]. The influence of games on their expectations of work and life cannot be underestimated. So, leveraging games to engage this generation seems an obvious path to increasing engagement of young employees. This is not to say that games only apply to the younger generations. Employees of all ages find games engaging and fun.

The challenge comes when creating a hybrid business solution which relies on the use of games to encourage increased participation and productivity from employees. Productivity Games are designed to increase productivity through the use of gaming elements and engaging game play. Play is part of being human and can help bring people together to have fun, work as a group and accomplish a task[vi].

Often, this is done within the context of a game. Stuart Brown's research into the concept of play highlights the fundamental elements of human play and showcases the essential roles of trust and community[vii].

A business process can be viewed as a sequence of activities and tasks that are performed to accomplish a specific organizational goal. As we looked at the characteristics of serious games at work it become apparent that these games were actually variants of business processes. In their August 2008 report, Forrester notes, "The strongest ROI and ultimate adoption will be in serious games that help workers do real work. We are already seeing this with the use of games in product development and collective intelligence, but the real dynamic idea is to pull out the incentive structures and tools of games to boost productivity and employee morale[viii]." All of this helped make the case for an increased investment in games.

In a classic statement on the power of working together, Eric Raymond stated in his seminal document *The Cathedral and the Bazaar* that "Given enough eyeballs, all bugs are shallow[ix]." While finding software defects is easy when many are involved, the challenge for many tasks is how to motivate group participation. If a person gets involved in a software beta program or open source project, they have shown an intrinsic interest in participating. However, if they are not involved in efforts like this, other types of motivation to encourage participation are required. We felt that by designing games that incorporate the fundamental elements of play, people could be enticed to participate. Even better, if the game play was interesting enough to the players, they would be willing to perform productive tasks in order to participate whether they had an intrinsic motivation to accomplish the goal or not. In our experience and game deployments, this has proven to be true.

In this paper, we will describe in some detail a Productivity Game deployed by the Windows engineering team to address a complex software localization problem that could not have been solved in a cost-effective way without massive participation. Additionally, we will describe briefly additional Productivity Games deployed to aid with other efforts during the Windows 7 development timeframe. These examples, and the results generated provide a strong case for greater use and research into Productivity Games.

3 Basics of Productivity Games

Productivity games are related to crowd-sourcing or human computation efforts, but with some key differences. Similar to recognized crowd-sourcing efforts like Wikipedia, or human computation initiatives such as the ESP Game[x], Productivity Games enable employees to have fun participating and feel good about accomplishing productive tasks in the process. The key difference between Productivity Games and crowd-sourcing is the use of gaming concepts to motivate participation in work-related tasks. The evolution of the ESP Game into the Google Image Labeler[xi], and the subsequent production of actual business data for Google is an example of a Productivity Game.

Productivity games are not a universal solution for every business process or task. Games introduce an alternative incentive system into the workplace as a byproduct of the game architecture and scoring of play. Since the workplace usually already has an incentive system in place – usually in the form of a paycheck, Productivity Game designers must be careful when, where and how they deploy games that can potentially impact existing incentives and rewards.

3.1 Game Applicability

Work tasks draw upon employee skills that can be grouped into one of three categories: core, unique, or expanding. Employees share "core" skills, such as the ability to type, that may be specific to their industry, but do not differentiate employee A from employee B. Some employees have "unique" skills that require specialized training or experience. "Expanding" skills are what employees aspire to and acquire over time to help them perform their jobs better.

From an organizational perspective, there are two categories of tasks that relate to the goals of the organization: "in-role" tasks and "Organizational Citizenship Behaviors" (OCBs)[xii]. In-role tasks are the tasks that employees are paid to perform. Organizational Citizenship Behaviors are the behaviors that an organization would like employees to voluntarily exhibit to enhance the workplace culture and environment.

From a Productivity Games viewpoint, the employee categorization and the organizational classification overlap in a way that can help identify whether or not a game will be successful in modifying behavior and having people "play".

Table 1 illustrates the areas where Productivity Games can be the most successful. Focusing either on expanding skills in role, or OCB's that require core skills are the best way to ensure the success of the game. Examples of why specific segments work or don't work are described below.

	Core	Unique	Expanding
In-Role Behavior			👍
Organizational Citizenship Behavior	👍		

Table 1. Successful Game Deployment

3.1.1 Thought Examples: Where Games Work

Based on our game experiences, described somewhat below, games which encourage good corporate behavior (or OCBs), but rely on core skills that all users share, are the most valuable space for Productivity Games. Since the games rely on core skills, all employees in an organization are able to participate. Additionally, since the behavior is not closely linked to any individual's job, no one's employment is threatened by the success of another team member.

For example, imagine a game that helped sort a complicated list of items. All employees in a given department are familiar with the items, and with how the organization prioritizes it's work. This provides a great place for everyone to participate on equal footing. But wrapping the sorting and prioritization work in a game-like interface, all players are given a fair chance to contribute and potentially win.

Games for Learning is a well-established genre of software development, and many examples are available in the marketplace for children of all ages. Game in this space work because they focus on the development and growth of the individual. Games are designed to encourage learning, and then test for the learning within the context of play. Players are best rewarded in this space by showing how they have improved themselves, rather than comparing raw completion numbers, which can quickly show disparity between students, but the element of the value of play is not lost.

3.1.2 Thought Examples: Where Games Don't Work

To further illustrate where Productivity Games can be successful versus less so, let us provide some example scenarios which might better illustrate possible games.

First, imagine a game which encompasses the daily tasks and work of a single employee, Joshua. In the "Joshua Game", which maps to the 'core' and 'unique' skills that Joshua performs for his work, players are given points for doing tasks Joshua would normally do. Some players are able to do all the tasks Joshua is capable of, and some are limited because they do not have the same 'unique' skills that Joshua has.

This presents our first problem: games which exclude players are not in the best interest of the organization. Since Productivity Games rely on a broad number of players, the objective of most games must be to add as many players as possible. Games which rely on actions from the bucket of 'unique' skills inherently limit the breadth of players available to play the game.

Back to the example, we find another challenge. If the end of the "Do Joshua's Job Game" comes and Joshua hasn't won, how does that fit in with his performance review? One thing for certain is that Joshua does not feel secure in his job anymore.

These two issues provide examples why games focused on 'unique' skill sets are difficult to deploy. Additionally, we see how competitive games focused on 'in-role' behaviors can introduce some awkward situations into the workplace and existing performance review processes.

3.2 Engagement

One indirect consequence of Productivity Games is the increased engagement of employees in the organization. From literature referenced above we know the "gamer" generation have invested a significant portion of their lives in playing games. And it is interesting to identify some of the attitudes and lessons which this younger generation has taken from playing these games. For example, gamers have learned from games that the cost of failing is very low, and they can always retry, yet from this they expect clear feedback as to what they need to do to change their play in order to succeed later on. From this we can see that the younger generation values a feedback loop and transparency in the consequences. Gamers always expect the game to be fair; otherwise, they will not continue to play. They map these same expectations in a game into their job, expecting the

workplace to have transparency and a clear feedback loop. They also expect fairness in how they are treated and in how they should treat others. Finally, games don't demand lengthy reading or studying of manuals in order to play. Most games provide an introductory training mode where the player is given the opportunity to learn what they must know in order to move forward into the game. Similarly, in the workplace, the lengthy corporate memo outlining detailed reasons for organizational priorities carries less impact than is desired.

Productivity Games provide an opportunity for an organization to communicate an organization objective or priority in a method that easily meets the needs of this younger generation. In a properly designed game, fairness and transparency are in place. A feedback loop demonstrating success or failure clearly teaches and trains employees how to change their behavior. And finally, instead of a lengthy manual or memo, an employee has the opportunity to engage quickly and easily in a "training" mode which provides the basic information required for the employee to play the game. This isn't to imply that employees are more apt to receive criticism (constructive or otherwise). Rather, because the "teaching" or "coaching" is framed in a game, they receive the feedback in a manner they are accustomed to learning from already.

With so many of the needs met for younger employees to understand and learn from organizational priorities, productivity is higher, morale is higher and employees' engagement is stronger. We have witnessed this first hand within our test team by monitoring existing productivity metrics (such as average number of defect reports produced each week) and noticing that through several game cycles the metrics stayed constant or improved. This is significant because many games ranging in size and scope were played. Some of the games had output which was additional defect reports, but not all. Though a seeming conflict can be created for employees between their paycheck (primary reward system) and the game (secondary reward system) when the game is sufficiently motivational, the increase in teamwork, morale and engagement are valuable.

4 The Language Quality Game

The Windows Language Quality Game has been a successful Productivity Game. It addresses organizational citizenship behaviors by calling on employees within Microsoft to apply their core native language skills to help assess the quality of Windows translation efforts.

The traditional business process uses specific language vendors to perform translation work, and then a secondary vendor to assess the quality. The business challenge has been that, for some languages and locales, finding two independent vendors can be difficult and costly. To address this problem, the Language Quality Game was developed to encourage native speaking populations to do a final qualitative review of the Windows user interface and help identify any remaining language issues. The goal was to ensure a high quality language release and using the diverse population of native language speakers within Microsoft has enabled the pre-release software to be validated in a fun and cost-effective way. The list of Windows languages can be found on Microsoft.com[xiii]. Table 2 illustrates the success that the Language Quality Game achieved as run against interim builds of Windows 7. A more detailed description of gameplay can be found below in a later section, but the goal of the game was to achieve reviews of screenshots and dialogs for translation accuracy and clarity. Native language speakers were encouraged to play from across Microsoft's diverse, international population. The results here demonstrate an immense amount of effort applied to the game.

Game Duration	One Month
Total Players	> 4,600
Total Screens Reviewed (Points Earned)	> 530,000
Average Screens per Player	119
Top Player Reviewed	> 9,300
Total Defect Reports	> 6,700

Table 2. Language Quality Game Statistics

Success in the game was defined as the amount of coverage of screens across the 36 languages tested. With the incredible response, most languages had several reviewers provide feedback per screen. Because of the latency in reviewing the feedback, defect reports were not included in players' scores. But, for the Windows International Test Team, defect reports were the most valuable output of the game.

Logistically, the massive amounts of feedback were handled by the international team with tools specially designed to display aggregated feedback. The "Moderator" role was filled on a per-language basis from the ranks of the international team, and allowed

the review of multiple pieces of feedback per screen quickly and easily. Where there was obvious consensus from the game players, a defect report would be created. Reviewed screens lacking consensus were quickly reviewed, but at a lower priority and more quickly, such that the screens with the highest likelihood of fixable defects were handled quickly and efficiently.

4.1 Business Process Challenges

The Windows Language Quality Game provided a solution to challenging business problems that could not be easily solved through traditional processes.

Software development, particularly at the scale of Windows, requires sensitivity towards cultural and political issues. While language issues like this may not impact the reliability of the application, users may react negatively and seek alternatives. In addition, government purchases can also be impacted by mistakes in language translation. As a result of these risks, it is imperative that the Windows Team develops software in a robust way that eliminates cultural and political defects.

The typical process involves finding two vendors; one to do the translation work, and the other to help with quality assessment. As an example, Galician is the language of Galicia, in Northwestern Spain. Portuguese speakers can understand Galician and sometimes refer to it as a dialect of Portuguese. However, there are cultural and dialectal differences that must be accounted for specifically in the Galician version of Windows.

Translation or geopolitical errors can impact the quality, perception, and sales for a region. In Windows XP, for example, a user can set up a profile by entering details such as their age, sex and number of children. A version distributed in Latin America asked users their gender, giving their options as No especificado (unspecified), varon (male) or hembra (female). Unfortunately, in some Latin American countries the term hembra has a negative connotation[xiv]. As a result, additional care must be taken to ensure that localized versions of Windows can be distributed to all countries where that language may be spoken.

4.2 Game Architecture

The Language Quality Game is built using a SQL Server database of images that are rendered in the game using Silverlight. The Windows International Team uses an automated process to copy dialog images from the Windows source code into the SQL server database. The dialogs are then augmented with metadata about the language and usage of the image in question.

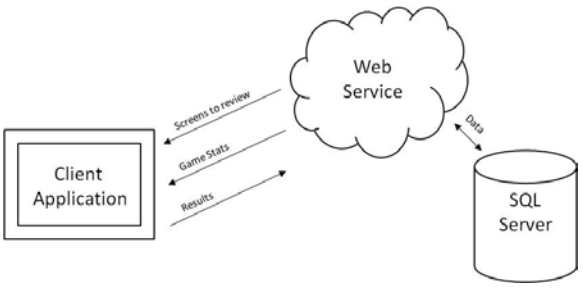

Figure 1 - Language Quality Game Architecture

The dialog images are broken up randomly into groups of 25 to provide multiple "levels" for the player to achieve. As players works their way through the game, each dialog is presented. Players can use their mouse or a digital pen to circle errors using electronic ink. Ink feedback is stored efficiently along with the dialog ID in error reports. This not only saves space in the database, but it also improves performance and helps with results reporting.

4.3 Player Population Selection

Finding players to perform the human computation work of reviewing dialogs in the Language Quality Game can be a challenge. It is critical to find native speakers for all the languages supported by Windows versions. For the Language Quality Game, players were recruited by sending broadcast email announcements to native language speaker social aliases, or email distribution lists available internally at Microsoft. Invitations were sent via email to groups such as "Persian Speakers at Microsoft", asking members to visit the Language Quality Game web site and play the game.

Finding the right aliases of potential players was critical to the response rate. We also found that native language speakers typically have friends and relatives who will be using localized copies of Windows. Therefore, it is in the speaker's best interest to play the game and help ensure the quality of the localized version that is important to them.

4.4 Data Quality and Cheating

While it's not possible to completely prevent cheating in a way that scales and keeps people actively participating, it is possible to inject "known defects" and ensure that players find and record them. This helps assess the reliability and validity of an individual player's answers and allows for filtering. In addition, for

the Language Quality Game, there is an assumption that players' personal desire to improve product quality for their own native language outweighs the desire to cheat. This is furthered by producing a game where no prizes were offered. Leaderboards within the game certainly provided some motivation and competition among players, but between national pride and the limited value of prizes, we believe the incentive to cheat was minimized. Further study is certainly warranted to understand whether successful participation in these kinds of Productivity Games influences annual reviews, etc.

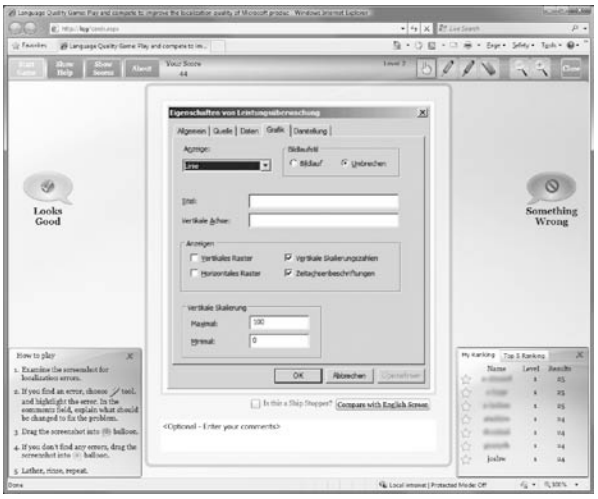

Figure 2 - Language Quality Game Screen Shot

4.5 Feedback Loop

In order to allow users some knowledge about their contribution to product quality, a report was provided that displayed their scores on a per language basis, but also provided a count of defect reports filed based on screens they had reviewed. Additionally, they were also provided a count of bugs filed on screens which they had reviewed as "good". This dual-feedback didn't cover every logical possibility for the combination of outcomes, but did provide to the user a simple method of knowing whether they were being sufficiently critical in their feedback or not. This kind of feedback adds to the experience of the player, as they are able to learn from feedback about their own performance.

4.6 Game Elements

While the language quality screen review work is not tremendously difficult for native language speakers, it is also not the most interesting or engaging, particularly with a large volume of screens. Consequently, game elements and enticing game play were designed and used to attract players and help motivate them to "play". These are the characteristics of Productivity Games that help differentiate them from other crowd-sourcing efforts.

4.6.1 Game Levels

The dialogs are broken up into groups of 25 images and presented as "game levels". Once players review all the images in one level they move to the next higher level and are presented with a new set of 25 images.

4.6.2 Earn Markup Pen Colors

There are multiple markup pen colors. As a player reviews more and more dialogs, they can earn and use a different color pen.

4.6.3 Graphical Image Movement

After a player marks up a dialog, they move it to either the "Looks Good" or "Something Wrong" pile. This movement and displaying the next dialog involve some basic Silverlight animation which adds visual interest and a gaming feel to the experience.

4.6.4 Leaderboard

Each person can view a leader board showing all players, their current game level and how many dialogs they have reviewed. Not only does this allow each person to assess their relative effort, but it also provides the basis for some friendly competition. The leader board is divided up in a variety of categories – by language for instance – to encourage participation.

4.7 Longevity

Like many games, Productivity Games have a limited lifespan for the work that needs to be done. But, in addition, there is a risk of burnout among the players in doing the task involved in the game. It is not safe to assume infinite play from all games. So, as a method to rejuvenate participation in the Language Quality Game, a set of new screens was provided to the game players during week three of play. These new screens were the result of early defect reports and fixes provided during the game. With the new screens, a new series of announcements via email were released to inform players that their feedback had been heard, and now we needed their participation again to help review the repaired screens. This did drive a second surge of participation.

Most Productivity Games can benefit from this same strategy. As the game progresses, there are cases where the priorities of the organization have changed, or the rules of your game have been taken advantage of to the

benefit of one or few players. Sometimes "resetting" the game with amended rules, or providing new content can help reinvigorate players and bring additional life to your game.

4.8 Language Quality Game Results

There has been 100% language participation – all 36 languages have been sent out for linguistic review and reviews have been received for all of them. The participation ranged from Korean, with over 82,000 screen reviews to Finnish, with under 1,000. Across the board, over 7,000 defects were reported across all the 36 languages.

After validation and data quality assessment, an average of 85% dialogs were found to be completely correct – the highest was Slovakian with 92% of screens reviewed marked as correct and the lowest was Bulgarian with 65% of screens marked as correct.

There have been over 4,600 players. The language with the most had 615 players and the least had 10 players.

5 Other Productivity Games

Microsoft has also tried other styles of Productivity Games in a variety of forms and sizes over the past few years. The games with the greatest participation were the games used in the Windows Vista Beta program. This experience is covered extensively in chapter 5 of "The Practical Guide to Defect Prevention[xv]".

More recently, a Productivity Game was created and used to classify freeform text comments as "actionable" or "not actionable". This feedback was generated by beta testers of Windows 7, and returned to Microsoft using a built-in tool which gathered this kind of text-based comments and feedback. Traditionally, this feedback categorization has been performed manually by the software team and is time-consuming and labor intensive. In some cases, automated machine translation and "text-crunching" tools have been tried with limited success, and still required a human step for final validation.

The strong interest in college basketball tournaments was used to attract potential players. The Feedback Productivity Game was structured as three phases, one before the tournament started and the other two phases related to subsequent rounds. The goal was to keep the duration of each "sub-game" short, vary the format slightly, and keep interest levels in the game high.

To participate in the Feedback Productivity Game, the player had to gain credits by classifying text comments into "actionable" or "not actionable". A screen was displayed with the ability for users to categorize each comment. For each comment classified, one game play credit was received. The credits could later be used in each round to play different "games".

The pre-tournament phase provided each player with a random pair of basketball teams (from the 64) and they could then select the one they thought would win between this hypothetical pairing. Players made selections by clicking on the name and logo of the school they preferred. Each selection required one comment classification credit, and immediately another choice was placed before them. Once credits were consumed, the player was again encouraged to classify additional text comments.

The next phase of the Productivity Game mapped to the teams who remained in the playoffs, and had real matchups displayed. The player could then select who they thought would win in that matchup. Each selection again required one comment classification credit.

The final phase of the game focused on the final teams in the tournament. To play, each player exchanged four credits for a "team ticket" indicating that team would win it all. Multiple tickets could be obtained for each team and tickets could be obtained for multiple teams. The objective of the game was to obtain tickets for the team that actually won. All players with tickets for the winning team would earn points in proportional to the number of tickets they had.

A total of 150 players participated in classifying 4723 feedback comments and 53% were assessed to be "actionable". These results saved the Windows team a tremendous amount of effort by distributing the work across basketball fans with these core skills.

This Productivity Game differed somewhat from the Language Quality Game where the relationship between play and work was more unified. But, the motivational factors were similar in that play of a game (or in this case, the ability to make my picks for games) was enabled by accomplishing a task useful and valuable to the organization.

6 Conclusion

In this day and age, many business challenges can benefit from groups of people working together to provide solutions. Recently, crowd-sourcing has been used to distribute tasks that can benefit from human

computation. This same concept can be utilized in corporations to tackle tasks that they are not resourced to support or that require unique skills such as native language proficiency.

A challenge in any of these efforts is how to entice and motivate people to participate. The Productivity Game concept utilizes gaming elements and engaging game play to help generate that motivation. Through Productivity Games like the Language Quality and Feedback games, we have shown that people can become engaged in a game and willing to exchange "real work" in order to participate. These results have demonstrated to us the tremendous potential of Productivity Games to help solve problems that are difficult or impossible to solve within traditional organizations and business processes. We look forward to the continued pursuit of that potential.

7 Acknowledgements

Our thanks to the following who continue to be instrumental in the success of Productivity Games:

- Darren Muir
- Harry Emil
- Robert Musson
- Dan Bean
- Robin Moeur
- James Rodrigues
- Karen Djoury
- Jian Chen
- Delton Porter

References

[i] Ross Smith (2008), "Productivity Games – Using Games to Improve Quality", Google Testing Blog, http://googletesting.blogspot.com/2008/06/productivity-games-using-games-to.html

[ii] The Serious Games Initiative, http://www.seriousgames.org

[iii] Choonghoon Kim, Douglas Michele Turco (1999). "The Next Generation in Sport: Y", Cyber-Journal of Sport Marketing, http://fulltext.ausport.gov.au/fulltext/1999/cjsm/v3n4/lim34.htm

[iv] Wikipedia, http://en.wikipedia.org/wiki/Generation_gap

[v] John C. Beck, Mitchell Wade (2004). Got Game: How the Gamer Generation is Reshaping Business Forever. Harvard Business School Press. ISBN 1-57851-949-7

[vi] Michael Elliott (2008), "The Games that Bring Us Together", Time Magazine, http://www.time.com/time/specials/2007/article/0,28804,1815747_1815707_1815705,00.html

[vii] National Institute for Play: www.nifplay.org

[viii] T.J. Keitt, Paul Jackson (August, 2008) "It's Time to take Games Seriously", Forrester

[ix] Eric S. Raymond (1999, 2001). The Cathedral & the Bazaar: Musings on Linux and Open Source by an Accidental Revolutionary. O'Reilly. ISBN 1-56592-724-9.

[x] Wikipedia, http://en.wikipedia.org/wiki/ESP_Game

[xi] Google's Image Labeler, http://images.google.com/imagelabeler/

[xii] Wikipedia, http://en.wikipedia.org/wiki/Organizational_citizenship_behavior

[xiii] Microsoft Help and Support: Knowledge Base, "List of languages supported in Windows 2000, Windows XP and Windows Server 2003", http://support.microsoft.com/kb/292246

[xiv] Nic Fleming (2004), "It's a tricky world in computers, says Microsoft Chief", Telegraph, http://www.telegraph.co.uk/news/worldnews/1469733/Its-a-tricky-world-in-computers-says-Microsoft-chief.html

[xv] Marc McDonald, Robert Musson and Ross Smith (2008). The Practical Guide to Defect Prevention. Microsoft Press. ISBN-13: 978-0-7356-2253-1. www.defectprevention.org or http://productivitygames.blogspot.com