Using the RITE method to improve products; a definition and a case study

Michael C. Medlock, User Testing Lead, Microsoft Games Studios (mmedlock@microsoft.com)
Dennis Wixon, Usability Manager, Microsoft (denniswi@microsoft.com)
Mark Terrano, Game Designer, Ensemble Studios **(**mterrano@EnsembleStudios.com)
Ramon L. Romero, User Testing Lead, Microsoft Games Studios (ramonr@microsoft.com)
Bill Fulton, User Testing Lead, Microsoft Games Studios (billfu@microsoft.com)

ABSTRACT

This paper defines and evaluates a method that some practitioners are using but has not been formally discussed or defined. The method leads to a high ratio of problems found to fixes made and then empirically verifies the efficacy of the fixes. We call it the Rapid Iterative Testing and Evaluation method – or RITE method. Application to the tutorial of a popular game, Age of Empires II, shows this method to be highly effective in terms of finding and fixing problems and generating positive industry reviews for the tutorial.

INTRODUCTION

Traditionally the literature on sample sizes in usability studies has focused on the likelihood that a problem will be found [11, 12, 13, 16, 17, 18, 20]. This literature suggests:

- Running zero participants identifies zero problems.
- The more participants used, the fewer new problems are discovered.
- That calculating the number of participants needed to uncover "enough" problems can be done via a formula based on the binomial probability distribution –but that this number will vary depending on what the experimenter sets as the likelihood of problem detection. It is important to note that this calculation is based on the assumption that the experimenter might see the problem at least once. For example,
    - Observing 4-5 participants will uncover approximately 80% of the problems in a user interface that have a high likelihood of detection (0.31 and higher) [10, 11, 18].
    - Problems in a user interface that do not have a high likelihood of detection (for whatever reason) will require more participants to detect [16, 20].

When the researcher is interested in problems that have a high likelihood of detection, the suggestion has been made that it is more efficient to test with 4-5 users and test more often compared to running fewer, large-sample studies [11, 13]. Depending on the goals and context of the test, there are situations in which running even fewer than 4-5 participants is appropriate and more efficient. Lewis [11] noted that as long as the likelihood of problem detection was very high (0.50 and higher) that 87.5 % of these problems will be uncovered by at least 1 of 3 participants.

However, the usability literature on sample size has often not focused on what we as practitioners view as the primary goal of usability testing in an applied commercial setting: shipping an improved user interface as rapidly and cheaply as possible. We stipulate that when the determination has been made that a discount usability method is appropriate it is more important to get the team to fix problems and to determine the likelihood that a "fix" has solved a problem than to agonize over if every problem has been uncovered. The same likelihood of detection calculation based on the binomial probability distribution can be used for this purpose (and all the same caveats apply). It is noteworthy that relatively few studies have focused on the likelihood that change recommendations will be implemented [7, 15]. A small number of studies have focused on the magnitude of improvement in the user interface of a shipped product or tool, or the relative effectiveness of these improvements in affecting commercial sales or user efficiency [1, 5, 9, 19].

The following are 4 reasons that we encounter that can explain why usability issues that are uncovered do not get fixed:

1. Usability issues are not "believed". The decision-makers on the product do not think that the issues uncovered are "real" or worthy of a fix.
2. Fixing problems takes time and resources. Development and design resources are scarce, and when faced with the decision between fixing a "working" feature or putting another feature in the product, the choice is often made to add the new feature.
3. Usability feedback arrives late. Feedback that is available when product feature decisions are being made is far more likely to be taken into account than that which arrives after the decisions have already been made. The delay between when a feature is implemented and when usability feedback is delivered to the team is a barrier to those recommendations being used.

4. Teams are uncertain whether a proposed solution will successfully fix the problem. The team doesn't want to undertake a potentially difficult or time-consuming fix if they are not certain to fix the problem. The lack of verification that the fix is working forces the team to implement the usability recommendations on faith, rather than a demonstrated history of accuracy.

This paper defines and evaluates a discount usability method to minimize the 4 problems above, and thus maximize the likelihood that usability work results in getting problems fixed. To promote future discussion and analysis, we have named this method the Rapid Iterative Testing and Evaluation method – or RITE method. The number of participants used is based on how many participants are needed to reasonably determine that an applied fix has solved a problem that was previously uncovered. This method is not "new", practitioners are already using it [2] but it has not been formally discussed or defined. This method focuses on making very rapid changes to the user interface. More significantly, it evaluates the efficacy of user interface changes immediately after (sometimes within hours) of implementation. What follows is the basics of how to perform the RITE method, and a case study of its use while testing an interactive tutorial for the game Age of Empires II.

WHAT is the RITE Method?

A RITE test is very similar to a "traditional" usability test [4]. The usability engineer and team must define a target population for testing, schedule participants to come in to the lab, decide on how the users behaviors will be measured, construct a test script and have participants engage in a verbal protocol (e.g. think aloud). RITE differs from a "traditional" usability test by emphasizing extremely rapid changes and verification of the effectiveness of these changes. Some of the notable differences are:

- Changes to the user interface are made as soon as a problem is identified and a solution is clear. Sometimes this can occur after observing 1 participant. Once the data for a participant has been collected the usability engineer and team decide if they will be making any changes to the prototype prior to the next participant. In the end the process by which this occurs is up to the usability engineer and team. What follows are some general rules that we found useful for the RITE method:

  We classified the issues seen for the participant into four categories:

  1. Issues that appear to have an obvious cause and an obvious solution that can be implemented quickly (e.g., text changes, re-labeling buttons, rewording dialog boxes, etc.).
  2. Issues that appear to have an obvious cause and an obvious solution that cannot be implemented quickly or within the timeframe of the current test (difficult new features, current features that require substantial design & code changes, etc.).
  3. Issues that appear to have no obvious cause and therefore no obvious solution.
  4. Issues that may be due to other factors (e.g. test script, interaction with participant, etc).

  For each category 1 issue, we implement the fix, and test it with the next participant. For each category 2 issue, we start implementing the fix. As soon as the fix has been implemented, use the revised prototype. For each category 3 and 4 issue, we collect more data to see if they can be upgraded to category 1 or 2 issues, or reclassified as "non" issues.
- There must be agreement prior to the test on tasks every user of the system must be able to perform without exception when using the product. This is critical as it sets the importance of issues seen in the test.
- There must be the ability to make changes very rapidly, (e.g., in less than 2 hours to test with the next participant, or before the next day of testing). This means that development time/resources must be set aside to address the issues raised in the test and the development environment is such that rapid changes can be made.
- Time must be set aside at the end of each participant or day of testing to review results with decision makers and decide if issues raised warrant changes for the next revision of the prototype.
- There must be the ability to run a sufficient number of new participants to verify that the changes made to the user interface have alleviated the issues seen with previous participants and have not caused other unforeseen issues. There is no "set" number for the number of participants needed to verify a fix –the probabilities can be calculated via Lewis's tables of the binomial probability distribution [10].

Simply meeting the requirements for the method as outlined above will not guarantee its success. To effectively use this method the following conditions must be met:

- The usability engineer must have experience both in the domain and in the problems that users typically experience in this domain. Without this experience it is difficult to tell if an issue uncovered is "reasonably likely to be a problem for other people". For someone who is inexperienced with a domain a more traditional usability test is more appropriate.
- The decision makers of the development team must participate to address the issues raised quickly (note that this is more than a commitment to attend and review).
- The design team and usability engineer must be able to interpret the results rapidly to make quick and effective decisions regarding changes to the prototype.

CASE STUDY: AGE OF EMPIRES II RITE TEST

PARTICIPANTS

The primary audience for the Age of Empires II tutorial was identified as individuals who had little experience with real-time strategy (RTS) games but who were interested in trying them. The participants were 5 females and 11 males aged 25 to 44 years old. None of the participants had ever played a RTS game but indicated an interest in trying them. All of them had played at least one retail computer game on the PC in the last year.

MEASUREMENT

The primary source of data was the tester's observation of the participants. Specifically:

- Failures. Errors made that resulted in user failure to continue the tutorial–in particular those made on any item identified prior to the test by the team and usability engineer.
- Errors. Errors made that resulted in user confusion–in particular those made on any tasks identified prior to the test.

PROCEDURES

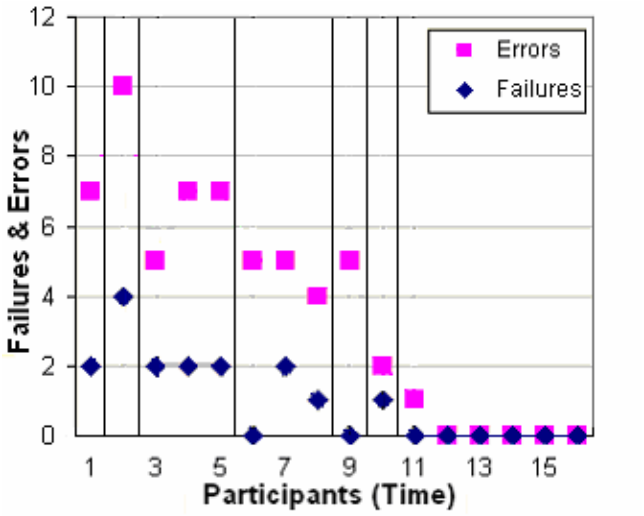
Prior to the testing the usability engineer and the team developed a list of tasks and concepts that the participants should be able to do and/or understand after using the Age of Empires II tutorial. Some of the issues that were identified for which there would be zero error tolerance included: unit movement, performing an "action" on something (gathering resources, attacking, etc.), multi-selection of units, understanding the different kind of resources, showing where resources are accumulating (in the menu areas up top), understanding the "fog of war", scrolling the main screen with the mouse, using the mini-map, and how to build and repair buildings.

At least one member of the development team (e.g. the Program Manager, Game Designer, Development Lead or User Assistance Lead) was present at every session. After each participant the team would quickly meet with the usability engineer and go over issues seen and place them into one of the 4 categories. For each category 1 issue a fix was implemented, and then the new build was used with the remaining participants. For each category 2 issue a fix was started, and the new build was used with the remaining participants as soon as it was available. This process allowed the team to determine if the "fix" worked or caused new/different problems. For all other issues (category 3 and 4) more data were collected (e.g. more participants run) before any fix was attempted.

AGE OF EMPIRES RITE TEST RESULTS

The results are summarized in Figure 1, which is a record of all the failures and errors over time (as represented by the participants) on the Age of Empires II tutorial. In addition, the graph shows the points at which the tutorial was revised. Every vertical line on the graph indicates a point at which a different version of the tutorial was used. Changes were implemented between participants 1, 2, 5, 8, 9, and 10 (6 iterations).

Figure 1. A record of errors over time as changes to the Age of Empires II tutorial were made using the RITE method.



From Figure 1 it can be seen that the build was changed after the first participant. It is instructive to examine an issue that caused the team to make a fix. In the second part of the tutorial participants are supposed to gather resources with their villagers. One of the resources that they are instructed to gather is wood by chopping trees. However, initially there were no trees on screen and as a result the first participant spent a long time confused as to what to do. The problem was clear and so

was the solution –place some trees within view and teach users how to explore to find trees off-screen. Both of these were done and the issue never showed up again in the next 15 participants.

It's also clear in Figure 1 that the number of failures and other errors generally decreases over time as iterations occur and eventually go to 1 error after the final iteration. Sometimes there were "spikes" after a revision. This was due to the fact that participants could interact with more of the game after previous blocking issues were removed. For example, once the wood chopping issue was removed after the first participant subsequent participants were able to move farther into the tutorial, thus encountering other issues along the way. This illustrates an important strength of the RITE method. By fixing errors early, additional problems can be uncovered.

In addition, we also can calculate the probability that the fixes we made actually fixed the issues we uncovered. After the last participant there were 6 additional participants run with no detection of any previously fixed issue reoccurring (the one issue that turned up never received a fix). Based on Lewis's tables of the binomial probability distribution [10] we are at least 88% likely to have fixed all the issues addressed if their likelihood of occurrence was 0.30 and 47% likely if their likelihood was 0.10. In point of fact the issues that were fixed earlier had an ever greater degree of verification –for example the wood chopping issue fix was verified by 15 participants (almost 100% at 0.30 and 79% at 0.10).

As stated in the introduction, finding problems in a user interface is only half the battle for a practitioner –the other half is getting them fixed. Sawyer, et al [15] proposed a metric they called the "impact ratio" which they used to determine the effectiveness of a test. The impact ratio was defined as:

$$\text{Impact ratio} = \frac{\text{Problems receiving a fix}}{\text{Total problems found}} * 100$$

Table 1 below shows that the impact ratio was very high using the RITE method.

Table 1. Impact ratio using the RITE method for the Age of Empires II tutorial.

Total problems found	Problems receiving a fix	Impact ratio
31	30	97%

One of the weaknesses of traditional usability tests is that there is often not enough time or resources to discover which "fixes" were not effective. In contrast, the RITE method allows designers to uncover "fixes" that need re-fixing. As such we have come up with a new measure of interest which we will call the "re-fix ratio" and we will define as:

$$\text{Re-fix ratio} = \frac{\text{Re-fixes}}{\text{Total fixes (including re-fixes)}} * 100$$

Table 2 below shows the re-fix ratio in the RITE test of the Age of Empires II tutorial.

Table 2. Re-fix ratio using the RITE method for the Age of Empires II tutorial.

Total Fixes (includes "re-fixes")	Changes that needed "re-fixing"	Re-fix ratio
36	6	20%

An example of a re-fix; participants had problems with terminology in Age of Empires II that required repeated fixes. Specifically, participants thought that they needed to convert their villagers units into swordsmen units –when in fact the two unit types were completely unrelated. When the tutorial directed the user to "Train 5 Swordsmen", participants would move their villagers over to the Barracks (a building that creates swordsmen) assuming that the villagers could be converted into swordsmen. In fact, swordsmen had to be created at the barracks using other resources (e.g. food, gold). When this issue manifested itself, there were inconsistencies in the terminology used in the product. In some places the user interface said "train units" and sometimes it said "create units". After watching six of eight participants fail to produce new units, the team decided that the text inconsistency was the problem and chose to use the term "train units" throughout the user interface. When the next participant was run the problem reoccurred in the exact same way. As a result the team changed the terminology to "create unit" throughout the rest of the user interface. Once it was "re-fixed" this problem was not seen again after seven participants.

INDEPENDANT reviews and financial success of Age of Empires II

While the RITE method can clearly eliminate user interface problems during a user test, assessing its overall effectiveness for practitioners requires looking at additional information outside of the lab setting, such as team rapport, awards, reviews and sales.

- Mark Terrano –(a lead designer for Age of Empires II) was highly satisfied with the method and wrote about it here, http://firingsquad.gamers.com/games/aoe2diary/.

- The Age of Empires II tutorial received an "Excellence" award from Society of Technical Communication in 1999.
- The game play and the tutorial of Age of Empires II received critical acclaim from the gaming press. In many cases the tutorial was singled out as excellent. To see aggregated reviews of Age of Empires II online go to http://www.gamerankings.com/htmlpages2/804.asp
- It is practically impossible to relate sales of a product to any single change, feature, or any particular method or technique. However, it is worthwhile noting that Age of Empires II was very successful and that it clearly reached a broader audience than the previous product. Since its release in last August of 1999 until the end of October 2000 Age of Empires II never left the top 10 in games sales according to PC Data. In addition, it continued to break back into the top 10 in 2001 from time to time almost two years after its release. The original Age of Empires was also very successful selling 456,779 units between Oct 1997 – Dec 1998. Age of Empires II sold almost double the number of units of the original Age of Empires in a similar time frame (916,812 copies between Oct 1999 – Dec 2000). The sustained sales of Age of Empires II over the original demonstrates that it is reaching broader markets –many members of which may not have had experience with RTS games before (the segment at which the tutorial was aimed).

CONCLUSIONS

The goals of the RITE method are to identify and fix as many issues as possible and to verify the effectiveness of these fixes in the shortest possible time. These goals support the business reason for usability testing, i.e. improving the final product as quickly and efficiently as possible. The results presented here suggest that at least in the context of its use for the Age of Empires II tutorial, the RITE method was successful in achieving its goals. RITE's goal is very similar to other iterative methods (cognitive walkthroughs and heuristic reviews [14], GOMS analysis [3], usability tests in general, etc.). It differs from other methods in the rapidity with which fixes occur and are verified. In the introduction we postulated 4 reasons that could explain why usability recommendations do not make it into products. The RITE method did a good job addressing all of these issues in the Age of Empires II case study.

1. The usability issues were "believed". The decision-makers had often pre-defined what tasks participants should be able to accomplish. In addition, through their constant involvement the decision-makers "believed" issues for which there were no previous tasks (issues they or the usability engineer had not anticipated).
2. Fixing the discovered issues was planned for and agreed upon prior to testing.
3. The usability feedback was delivered as soon as it possibly could be –right after the issues occurred.
4. The team had measurable assurance that the solutions were successfully fixing the problems because the fixes were tested by the subsequent participants. In addition the team caught "poor" fixes for problems and corrected them.

It is important to reiterate the reasons that the RITE method succeeded for the Age of Empires II tutorial. They were:

- The attendance of product decision makers at the tests (developers, usability engineer, program manager, etc.).
- The rapid identification of issues and potential solutions.
- The identification of tasks that users must be able to do.
- The usability engineer had a great deal of experience watching participants in similar situations in the past, and was very versed in the product itself.
- The developers had a strong knowledge of design and a deep understanding of the system architecture.
- The opportunity to brainstorm different fixes as testing occurred.
- There was agreement between decision makers on changes to be made.
- Age of Empires II had a powerful and flexible architecture of the scenario creation editor, which allowed for rapid changes to be made to the product.

As stated in the introduction, the RITE method is not "new"—practitioners do this (and other activities like it) all the time. But there is surprisingly little said about methods like it in the literature. The practice of making changes after having run 1-3 participants appears to be common. A quick internal survey of usability engineers and practitioners at Microsoft and on a public list service (Utest) found that 33 of 39 respondents had used a similar method of very rapid iterations and fixes at least once. Lewis [11] recommends essentially running 3 participant usability studies with fixes in between. In addition, Nielsen [13] has pointed out that given the option it is better to run more tests with fewer participants because that is the most efficient method for both uncovering problems and verifying fixes. The RITE method differs from Lewis and Nielsen's recommendations in that changes do not occur after a set number of participants, and verification of those changes are planned during the course of testing.

If usability engineers are considering using the RITE method for themselves they should first consider whether they have the right mix of "ingredients" similar to those listed above. In addition there are dangers to using the RITE method. Some of them are:

- Making changes when the issue and/or the solution to the issue are unclear. Poorly solved issues can "break" other parts of the user interface, or user experience. This happened a couple of times in the Age of Empires II test –although to the best of our knowledge we were able to catch these issues and fix them because of the structure of the RITE method.
- Making too many changes at once. If one of these changes degrades the user experience it may be difficult to assess which of the changes is causing the problem.
- Not following up at the end of the test with enough participants to assess the effectiveness of the changes made. Without this follow up there is no guarantee that the changes made were any more successful than the previously tested user interface.
- Missing less frequently occurring, yet important, usability issues. By using very small samples between iterations it is possible that less frequently seen issues will slip through unnoticed. This is particularly true if the task is broad and the domain is known to have many such issues [16], or the topic is one in which the researcher can not afford to miss an error.

When asked about the RITE method some usability practitioners we surveyed either wholeheartedly endorsed it, or reluctantly admitted to using it. The reluctant practitioners expressed the following reservations with the method:

- Reliability/Validity. The issues found by 1-3 participants might not be seen ever again. They might be abnormal or not the true phenomenon of interest.
- Power. With a small number of participants (1-3) they would not uncover all the issues that the user interface had.

With respect to reliability/validity, making a change based on minimal participants is a risk but one we are willing to take in certain situations. When the problem and solution are "obvious", (as we would stipulate our wood gathering example was) – seeing an issue once is sufficient. In addition it is important to note that the vast majority of issues were not obvious and therefore did not get fixed until we had more clarity (e.g. seen more participants). A quick thought experiment should make our position clear. Consider the following two cases where the one participant in a usability test doesn't seem to notice the difference between two color codes.

- Case 1: The color codes are Red and Green.
- Case 2: The color codes are Black and White.

In Case 1, if you know that Red-Green color blindness affects about 10% of males, and we discover through the verbal protocol that our participant is red-green color blind then it is clear that the red-green color states are a problem and it is clear how to fix the problem (don't use both Red and Green, or add an additional cue like shape or size, etc.). In Case 2, it isn't clear what is going on—there isn't any such thing as Black-White color blindness (to our knowledge). If you confirm that the participant can see normally, then you're bound to conclude this is random, inconclusive or simply bizarre. The key difference between the two cases is that the observation is evaluated against the knowledge and critical thinking ability of the observers—additional data aren't needed in the Red-Green case, but they are in the Black-White case. In addition, in our view in a business context it is more important to establish the sufficiency of a solution rather than the reliability of observed problems. Thus, problems should be fixed as soon as they can be identified, agreed upon and a plausible solution proposed. These solutions can then be tested with subsequent participants to establish some confidence that they would work in practice.

With respect to power, we are far more concerned with the power needed to assure that a fix has actually solved a problem. Within the definition of the RITE method the only guideline is to run as many participants as is needed to be confident of the fixes your team has made. Having said this, it is absolutely correct that a small number of participants will uncover a small number of problems that are only likely to occur in a high frequency of the population. But this rests on the assumption that the researcher will only run 1-3 participants and that the interface for each batch of participants is entirely different. Our experience in using this method was that the likelihood of detecting problems actually increased by rapid iteration (note: figure 1 shows an increase in problems and failures in the early stages of the testing). This is due to the fact that much of the interface was still the same and as problems are fixed the participants completed more of the tasks without intervention and in effect "tested" more of the user interface.

Finally, Gray and Salzman [6] have argued that usability methods should be assessed using multiple techniques. We have applied that approach and although the results of each assessment (user response, and reviews) can be explained in other ways (and in fact are probably the product of multiple factors), our interpretation is that overall they provide substantial evidence in favor of the RITE method.

In conclusion, we believe that an understanding of the overall effectiveness of applied usability methods in different situations is still evolving. Previous work has established the relative effectiveness of different usability methods to efficiently uncover problems. However, some comparative studies have been criticized for their methodology [6]. We agree with Bonnie John [8] who suggests that an understanding of applied methodologies will develop through the careful reporting and discussion of a series of case studies of a variety of methods combined with clear and precise definitions of the method

used. We hope that this study helps foster such a tradition. We encourage others to replicate this work in other areas, find additional strengths and weaknesses and report them.

REFERENCES

1. Bias, R. G. and Mayhew, D.J. (1994). *Cost Justifying Usability*. Academic Press, New York.
2. Butler, M. B. and Erlich, K. (1994). Usability Engineering for Lotus 123 version 4. In M. Wiklund. (ed.) Usability in Practice: How Companies Develop User Friendly Products. Academic Press, New York, 1994. 293-327.
3. Card, S. K., Moran T.P. & Newell, A. (1983). The Psychology of Human-Computer Interaction. Lawrence Erlbaum Associates, Hillsdale NJ.
4. Dumas J., and Redish J.C. (1993). A Practical Guide to Usability Testing. Ablex, Norwood, N.J.
5. Gray W.D. John, B.E. & Atwood, M.E. (1993) Project Ernestine: Validating a GOMS Analysis for Predicting and Explaining Real-World Task Performance. *Human Computer Interaction*, 237-309.
6. Gray, W. D. and Salzman, M.C. (1998) Damaged Merchandise? A review of Experiments that compare usability evaluation methods. *Human Computer Interaction, 13* (3). 203-261
7. Gunn, C. (1995) An example of formal usability inspections at Hewlett-Packard Company. In *Proceedings of CHI '95 Conference Companion* (May 7-11, Denver, CO) ACM Press., 103-104.
8. John, B.E. (1998). A Case for Cases, *Human Computer Interaction, 13* (3), 279-280.
9. Klemmer, E. T. (1989) *Impact in Ergonomics: Harness the Power of Human Factors in Your Business*. Ablex, Norwood, N.J..
10. Lewis, J. R. 1990. Sample sizes of observational usability studies: Tables based on the binomial probability formula. *Tech. Report 54.571*. Boca Raton, FL: International Business Machines, Inc.
11. Lewis, J. R. 1991. Legitimate use of small samples in usability studies: three examples. *Tech. Report 54.594*. Boca Raton, FL: International Business Machines, Inc.
12. Lewis, J. R. 1993. Sample Sizes for Usability Studies: Additional Considerations. *Tech. Report 54.711*. Boca Raton, FL: International Business Machines, Inc.
13. Nielsen, J., and Landauer, T. K. (1994). A mathematical model of the finding of usability problems," *Proceedings of ACM INTERCHI'93 Conference*. Amsterdam, The Netherlands, 24-29 April, 1993. 206-213.
14. Nielsen, J., and Mack. R. (1994). *Usability Inspection Methods*. New York: John Wiley and Sons.
15. Sawyer, P., Flanders, A., and Wixon, D. Making a Difference – The Impact of Inspections. In *Proceedings of CHI' 96*, (Vancouver, B.C., April 1996) ACM Press, 376-382.
16. Spool, J. and Schroeder, W. "Testing Websites : Five Users is Nowhere Near Enough. In Proc. CHI 2001, Extended Abstracts, ACM 285-286
17. Virizi, R. A. (1990). Streamlining the design process: running fewer subjects. In Proceedings of the Human Factors Society 34th Annual Meeting (Orlando, FL.) Human Factors Society, 291-294.
18. Virizi, R. A. (1992). Refining the test phase of usability evaluation: How many subjects is enough? Human Factors, 34, 457-468.
19. Wixon, D. R. and Jones, S. (1996). Usability for Fun and Profit: A Case Study of the Design of DEC Rally Version 2. In. M. Rudisill, C. Lewis, P. Polson, T. McKay (eds.) Human Computer Interface Design: Success Stories, Emerging Methods, and Real-World Context. Morgan Kaufman, New York, 3-35.
20. Woolrych, A. and Cockton, G., "Why and When Five Test Users aren't Enough," in Proceedings of IHM-HCI 2001 Conference: Volume 2, eds. J. Vanderdonckt, A. Blandford, and A. Derycke, Cépadèus Éditions: Toulouse, 105-108, 2001