

I 543 Usability Test Methods for Interaction Design

Project 3
Comparison between the results of
Two Usability Test Methods:
Usability Testing and Heuristic Evaluation
For IU Account Management Services

Team 5

Aaron Houssian
HyunSeung Koh
Jamie McAtee
Mingxian Chang

October 3, 2006

Introduction

Usability Testing (UT) and Heuristic Evaluation (HE) are two popular methods used to test systems. This project focuses mainly on comparing results of these two methods vis à vis their goals, rules, processes, costs, and findings. Both UT and HE were conducted on the account management system (AMS) at Indiana University (IU). For this project we used Nielsen's 10 heuristics (Nielsen, 1994) that he has been developing for over a decade. In this project, four experienced evaluators independently did the HE after studying the ten heuristics together. The results of both UT and HE can be found in Table 1 and 2 in Appendix. UT was conducted previously by the same evaluators; please see project2 documents as well as Appendix.

The subject system tested

The Account Management Service was the system used in this project.

The Account Management Service is designed to help students, faculty, staff, affiliates and guests at Indiana University create and manage their computing accounts and passwords (<https://itaccounts.iu.edu/>).

The results of User Testing

UT on the AMS showed there were many problems with the system. Even when given very clear detailed instructions, some users are still unable to accomplish basic tasks. The AMS system was clearly designed in such a way that every function was clearly worded in full sentences. The level of detail in the description of the functions is very deep. Despite all these efforts, UT showed that users don't read long instructions, and are even distracted by them. Almost all of the problems found fell directly into this category. Even logging into the AMS, a task we thought everyone would be able to do easily, was not accomplished by one user.

None of the three users tested were able to complete three of the tasks assigned. We concluded that the factors that went into these failures were due mainly the type and amount of information offered on the web pages. One additional factor in the last task (not being able to reset passwords) was that the link to the actual reset page is very small and not easy to find, but this wouldn't be a problem if the volume of other information wasn't so large. During UT, users also made some suggestions that didn't conform to this such as not being able to delete or modify accounts, and not being able to logout.

The results of HE

The four experienced evaluators discussed and explained the ten heuristics one by one in order to ensure that everyone had the same, clear understanding of the rules. HE tests were then performed independently. In total, 44 problems and 6 positive comments were found. Among the 44 problems, only 4 usability problems were found by all four evaluators, and 32 problems were founded by single evaluator. (See Table 1 in the

Appendix)

The comparison between the results of UT and HE

1. HE reveals more usability problems with more diverse problems than UT.

We obtained 44 usability problems with HE while many of them were the same as the problems revealed in UT as well as a significant number of other problems, whereas 10 usability problems were founded with UT,

2. Many of HE problems would never have been found by most users of UT

It was the amount of time and the intense attention to details, guided by 10 heuristics, which revealed many of these problems. For example clicking on the link to change self-service password reset questions yields a page that looks substantially like the other pages on the AMS site, but with several things that are different, such as the navigation links that are no longer in a bold font. A quick look at the address bar confirms that you are on a different server. The evaluators looked at all the different parts of the AMS and believed that they should be consistent, and so this became a problem.

Another problem that was found by HE that would never be found by UT is that when users login to AMS it never matters what users choose for the first option. One can click student, faculty or whichever items as long as choose create new account it takes the user to the same screen. If a user chose to mange his/her account it takes the user to the same screen. If a user chose 'create guest account' it takes the user to the same screen. One could effectively eliminate choosing who are for this reason alone, although there are others.

3. Personal Design Philosophies Differ

Even with the same set of heuristics, evaluators will impose their own interpretation of the heuristics, will notice different types of things because of different past experiences and expertise, and have their own set of "bêtes noires" that drive them crazy. This is all part of the evaluators' design philosophy which can radically affect the kinds of things that are called problems and the degree of severity assigned to that problem. This is true even for the users of UT in that they each may have things that they would like to see in the systems they use and will mention these as problems during testing.

4. Different opinions about the problems' severity

The equality of the severity of the problems revealed by the testing methods must be questioned. When we compiled the problems found from HE many of them were labeled high or medium severity, but this was based solely on the problem founders' opinion and highly subjective, because 73% questions were founded by only one evaluator. When conducting UT, testers discuss and decided a problem's severity together according the test strategies. For example, only problems that would completely stop the user from completing the task were labeled as highly severe. We must therefore not give equal weight to rating of the severity of the problems found between methods. It may be suggested that some of the problems found by HE are almost trivial, but still are problems as defined by the heuristics.

We believe that a standardized way of classifying severity to be used in UT could solve these kinds of problems.

Conclusion

The kinds of problems revealed by UT and HE overlap to a large extent in that all but one of the problems found by UT were found by HE in one form or another. We thought we must however keep in mind that we used a limited set of data that was produced in a short time (the last three weeks). In addition to all the problems found by UT, HE revealed many other diverse problems. UT in this case clearly showed some of the major problems that an average user may have. The zero success rate in some of the tasks shows how some basic functions of the website are fundamentally obfuscated.

As the paper (Law and Hvannberg, 2002) said, the goal of performing HE is to uncover as many potential usability problems as possible, and the main goal of UT is to find out the effectiveness, efficiency, and satisfaction degree when specified users to achieve a specified goals. We found out 50 problems by HE compared only 10 problems found by UT. Some problems only found by HE are about that the website we evaluated should not provide much irrelevant or rarely needed information which makes the interface boring, unattractive. This kind of problem would never be found by UT, because users can complete the tasks no matter the redundant information is there or not.

The problem about how to allow users to tailor frequent actions found by HE is a very special problem, because let the website personalize is a high level function. By using this function, maybe users would feel the convenience and satisfied, but users don't feel very exhausted with a fixed order of the service list, and can also successfully complete the tasks. The problem about the website doesn't provide a consistent form between different pages is another typical problem which found in HE. Because no matter the buttons are squares or circles or in different colors, users can login by click them in only the words 'login' are on the buttons.

The big difference I learnt from HE and UT is that the UT testers must create good scenarios if they want to get comprehensive feedbacks. That means the testers must know the tested system very well and systemically plan the test before they do the test. Novice users and experts testers work and on different part of UT if the users in UT are novice. Novices and experts work on the same thing in HE, and nobody knows who can contribute more until the tests done.

Law and Hvannberg's findings vs. our findings

The following 8 issues are taken from the article of Law and Hvannberg (Law and Hvannberg, 2002).

1. *Relative higher cost-effectiveness of HE*

Factoring in the time each evaluator used plus the compilation of the data the HE took eight hours. The UT took thirty hours when all of the setup, questionnaires, recruitment of users, analysis of data as well as the actual test is taken into account for each of the evaluators. More equipment was set up for UT than HE. In order to record both quantitative data and qualitative data during the test process of UT, a Webcam and a special software for it were set up, and several more forms and questionnaires were prepared, such as consent forms, data log form and time record sheet. Given all these data we must therefore conclude just as the authors of the article that HE is more cost-effective.

2. *Convergence of results*

We found that HE and UT method seem to complement each other and that there was some convergence of results since many problems found by HE have not been found by UT, but all problems but one has been founded by HE. We believe that the diversity of problems found in HE was largely due to the fact that HE is more intense, involves all aspects of the web site, and the in-depth knowledge of the evaluators. UT was limited to only the portions of the website that were part of the tasks.

3. *Accuracy and objectivity of UT results and misidentification of problems in HE*

We believe that all humans have their own biases therefore total accuracy and objectivity does not exist when human subjects (users and evaluators) are used. This does not mean that UT and HE are not useful instead it indicates objectivity and accuracy can only be used in a limited scope. In addition, people's personal preferences and biases may lead to the identification of problems that do not exist in the system or are of very low severity.

4. *Linking intrinsic feature to payoff performance*

We found that it depends on evaluators' experiences, cognitive styles and expertise since two different evaluators have opposite opinions on this issue. In other words, one evaluator said that UT tends to make evaluators think about intrinsic features, the causes of the problems, more than HE since UT makes evaluators focus on users' whole processes of a task instead of one aspect of a task or heuristics, leading them to think about the reasons users make mistakes. In the meantime, the other evaluator said that HE tends to make evaluators think about intrinsic features more than UT since heuristics guide evaluators to think about the causes of the problems.

5. *Pool of evaluators vs. population testers*

We noted that HE found more diverse problems by different testers. From this result, it seems that the results of HE depend on evaluators' background and expertise so that HE is more constrained than UT. However, since we also found that the results of UT depend on tasks, it seems to be also true that UT is more constrained than HE. In other words, the

selection of tasks will constrain the result of UT.

6. Positive findings

Just as Law and Hvannberg stated, finding positives during testing will depend on how the test is set-up and whether the evaluators are willing to look for them. In our case two of the evaluators were actively looking for at least a few positives, and we found some that fit the heuristics very well. We feel that in most cases there are positive points in any design; it's just a matter of finding them.

7. Predictive power of UEMs

We noted that HE also found 9 out of 10 problems found by UT (90%), which is higher than 48.7% in the result of this paper. However, there is a possibility that this percentage will be decreased by adding more tasks in UT (from 5 tasks to 10 tasks) since as already mentioned earlier, the results of UT depend on tasks and the results of HE on heuristics. We must also note that all the evaluators are essentially biased because they conducted the UT as well as the HE.

8. Accumulative insights into problems

We rather found that HE tends to describe the same problem at a more detailed level from different features and with different severity. For example, from UT, we found the problem of 'Faculty & Staff only options are distracting for student users' and judged that this is not severe since it delays users' tasks, not prohibit. In the meantime, HE found four similar problems from different features with different severity.

Reflections

Overall conducting usability testing is surprising, engaging, and time intensive. Heuristic evaluation is not time intensive, but is also can yield surprising results, and can be very interesting. On the whole it was a learning experience; one that I think was invaluable

1. Usability testing

We were not prepared for the amount of work in conducting usability testing as well as the difficulties encountered. Writing clear scenarios and tasks is more challenging than we would have thought. Keeping consistency between users, resisting the urge to help users complete tasks, and encouraging users to keep trying without forcing them or making them uncomfortable are all things that we struggled with.

One of the most difficult tasks was to recruit the right users. Since this task is very critical and will affect the results of testing, it seems to be a good idea that testers spend more time in selecting and recruiting the right users. Also, it might be a good idea that we recruit more users than we needed to prepare for unexpected circumstances.

It was very difficult to train and promote users to conduct a think-aloud protocol. A few minutes of demonstrating and training seem to be too short. It might be better for a facilitator to demonstrate it with sufficient time until users fully understand it. Also, it might be a good idea to let users have sufficient time to repeat exercising different tasks until users are accustomed to this new protocol in order for testers to get richer data.

We also used a video camera along with a webcam and Camtasia since it is one of popular methods (Nayak, Mrazek, & Smith, 1995) and we wanted to record users' body languages and their movement which are not easy to be captured using either a webcam or Camtasia. However, we could not obtain useful data from it since users' bodies were hidden by a facilitator closely sitting next to users. It might be a good idea to place a video camera in the most appropriate position.

It was difficult to conduct the formal testing in the natural setting. We challenged to combine the formal usability and the ethnography-based testing so that we set up the testing equipment in the library instead of the usability lab. It might make users feel more relaxed, but observers had difficulty to listen to users' verbal protocol due to noisy from people around observers. It might be a good idea for us to conduct testing in the natural setting, but keeping a certain amount of distance away from noise.

2. Heuristics Evaluation

We thought heuristic evaluation would be quick and easy, and wouldn't really show any further problems with the system. It was very wrong, we found a number of problems that can really only be found by working with a system intensively for an extended period of time. I didn't think compilation of data and prioritization would take nearly as long as it did.

When we did the usability test last week, there were two international students in our users group, and one of them is a Korean boy. Sometimes international students are not very good at speaking in English, even though they can do reading and writing very well. We didn't get a lot of verbal protocol data when he was thinking-aloud, because he needed time to think how to express his thought in English, therefore he wasn't able to make all his thought out loudly when he was doing the tasks. I thought it is a disadvantage of UT. If let him do test by using HE method, at least there are no language limited, because there are no time limited, and he can think by whichever language. Maybe we can think about to use HE method do usability tests in which the tested system would be used by international people.

References

- Law, Lai-Chong, and Hvannberg, Ebba Thora. Complementarity and convergence of heuristic evaluation and usability test: a case study of UNIVERSAL brokerage platform. *Proc. NordiCHI*, ACM Press (2002), 71-80.
- Nayak, Nandini P., Mrazek, Debbie, and Smith, David R. Analysing and communicating usability data: now that you have the data what do you do? A CHI'94 workshop. *SIGCHI Bulletin* 27, 1 (1995), 22-30.
- Nielsen, Jakob, http://www.useit.com/papers/heuristic/heuristic_list.html further references from page are:
- Molich, R., and Nielsen, J. (1990). Improving a human-computer dialogue, *Communications of the ACM* **33**, 3 (March), 338-348.
- Nielsen, J., and Molich, R. (1990). Heuristic evaluation of user interfaces, *Proc. ACM CHI'90 Conf.* (Seattle, WA, 1-5 April), 249-256.
- Nielsen, J. (1994a). Enhancing the explanatory power of usability heuristics. *Proc. ACM CHI'94 Conf.* (Boston, MA, April 24-28), 152-158.
- Nielsen, J. (1994b). Heuristic evaluation. In Nielsen, J., and Mack, R.L. (Eds.), *Usability Inspection Methods*, John Wiley & Sons, New York, NY.
- Tohidi, Maryam, Buxton, William, Baecker, Ronald, and Sellen, Abigail. Getting the right design and the design right. *Proc. CHI2006*, ACM Press (2006), 1243-1252.
(<http://doi.acm.org/10.1145/1124772.1124960>)

Appendix

Table 1.

Comparison between the results of UT and HE

UT		HE		
Usability Problems	Severity	Severity	Usability Problems	# of Tester
Unable to login to the AMS	High	High	On the login page get rid of the options and only have login and create a new account.	1
Password reset extremely difficult to find	High			
Unable to find where to activate personal web page	Low	High/Medium/Low	The names of the accounts that can be created are not easy to understand.	4
Extreme difficulty finding URL for personal webpage "mypage.indiana.edu" NO user was able to do this.	Low	Medium	Completion of step 2 on the create new account page is somewhat unclear since it happens without showing the user anything.	1
Faculty & Staff only options are distracting for student users	Low	Medium	When you are logged in as anyone you get all of the options even if you can't use them.	1
		Medium	The system should only show the options you are authorized to use.	3
		Low	When you login you should only see information for the campus that you are affiliated with.	1
Option menus with long explanations in the center screen	Low	Low	The system attempts to be descriptive but gives the user too much information.	4
Set your primary email address' option is distracing for users visiting the site for forwarding emails	Low	Low	Setting the primary e-mail and forwarding of your e-mail should be on separate pages.	2
There is no option to delete or modify pre-existing accounts	Low	Medium	Users can't delete or modify accounts once they have been created.	3

Terminology, passphrase	Low	Medium/ Low	Wording of passphrase vs password.	4
No way to log out	Low	High	The user has no way to logout.	4
		Medium	All options on the first page lead to the same place.	1
		Medium	Change password screen pops up in new window	3
		Medium	The change password and set up of registered questions should be in the same navigation link.	1
		High	On the password change page when the user puts in a wrong username the system tells you that the new password doesn't match the username.	1
		Low	The passphrase change page loses the menu.	1
		Low	Help information opens in a separate window possibly confusing the user.	3
		Low	Help information does not give you a clear way out.	1
		Medium	Help information doesn't give you help on task in the system it instead gives you general help.	3
		Low	On the help page they should have a clickable list of the campuses first.	1
		Low	On the home page the headings in the center don't look clickable but are.	1
		Low	The menu item is still clickable when you are on that page.	1
		Low	The error message on pages that you don't have access to is terse non descriptive.	2
		High	If a user already has an account you can attempt to create a new account until several screens in.	2

Low	On the create new accounts page the users name is listed as n/a instead of their name.	2
Low	On the account creation page the services should be listed by most frequently used.	2
Low	An option to be able to delete some of the questions but not all of the questions would be helpful.	1
Medium	On the create new account page you can't create a new e-mail account.	2
Low	should be manage your e-mail account.	2
Medium	The label "View your accounts" seems to indicate you can see all the accounts you have and can create.	1
Low	Instructions at the top of the main page are superfluous.	2
High	The title of the link and the title of the destination page for several links don't match.	1
Low	On the "set up passprase questions" page the menu itmes are no longer in bold.	1
Low	The warning messages are not consistent across the windows.	1
Low	The account management logo does not link back to the home.	1
Low	Use the CAS login button on the login page to keep consistent with other systems.	1
Medium	User is not given a confirmation screen when the attempt to create a new account.	1
Medium	User should have to confirm that they understand the implications of forwarding their e-mail.	1

Low	The system tells me that my e-mail is forwarded but I don't understand why since I didn't do this.	1
Medium	Understanding what accounts you have created and what they are doing is difficult.	1
Medium	Links to common tasks should be front and center on the main page.	2
Medium	On the view your accounts page you don't have an option to delete.	1
Low	Change network ID username to user name	1
Low	Error messages explain the nature of the problem but not the cause.	1
Positive	The system is very minimalistic in design. The only things that are present are the items that help the user move through the screen.	1
Positive	Arrow next to menu indicates where the user is located.	1
Positive	The create new account page shows how many steps are required to activate the account.	1
Positive	The use of continue on the login screen is good.	1
Positive	On just about all of the pages the menu has a link to the home page.	1
Positive	On the password maintenance screen there is an emergency exit	1

Table 2.

Comparison between the severity of UT & HE

Severity	UT	HE
High	2	5
High/Medium/Low	0	1
High/Medium	0	0
Medium	0	15
Medium/Low	0	1
Low	8	22
Total	10	44