Effect of domain knowledge on system usability evaluations

A thematic analysis of subjective usability experiences through real case scenario

Oscar Bjurling, Josefin Carlbring, Amanda Jaber, and Jacob Weilandt

Linköping University, 2016-06-03

Abstract

In this study, we investigated three groups of employees from two different departments at the software company IFS. We wished to assess which group would be best suited for conducting future usability tests of the company application, depending on their level of domain knowledge in the system project module. Analysis of data collected through think-aloud protocols, self-reported cognitive strain, mouse dynamics measurements, usability issues identified and follow-up interviews revealed that participants with and without domain knowledge were significantly different in several of these regards. We conclude from the results that employees without domain knowledge are attentive enough to identify the most serious issues. We suggest, however, that an iterative test process be implemented, addressing serious problems before re-testing the system with employees that have domain knowledge. In this way, problems in both usability and functionality are more likely to be found.

Keywords: Usability Test, Domain Knowledge, Think-aloud test, Thematic analysis

Introduction

When companies upgrade and expand their systems it can be hard to integrate the new with the old. It is important to remember to consider how older parts can remain usable, as well as testing the new system as a whole on a usability level. After all, the user of the system is human, so it is essential that the system is adapted for the human, both the beginner and the expert. By conducting a base study where you look at usability problems one could learn not only what problems there are in the system, but also what kind of user is the most representative and rewarding for future usability tests. This way you can ensure that the results will be helpful to the company in driving the system to a better usability solution.

The purpose of this study was to assess which group of employees at software company IFS would be best suited for evaluating the usability of the company application in the future. To supporting our analysis, we wanted to know in what ways the groups differed. We expected participants without domain knowledge to report higher cognitive loads, perform more interactions with the system, move the mouse further, and finally identify more usability issues than their domain expert colleagues.

Theory

Thematic analysis (TA) is a qualitative data analysis method used to identify and analyze themes within the data, providing answers to the given research questions. Themes are identified through clustering, data coding, theme development and revision (Braun & Clarke, 2006).

Think-aloud protocols are, according to Arvola (2014), a technique which involves asking the participant to think out loud while they are solving their given task. At the same time, everything that they say and do are recorded. Analysis of the collected data can provide insights into the participants' thought processes.

NASA-TLX (NASA-Task Load Index) forms are, according to Grimes and Valacich (2015), tools for assessing the subjectively perceived workload of a task and the workload it entails.

Usability – Nielsen (2012) defines usability as a combination of five components; learnability, efficiency, memorability, errors, and satisfaction. We chose to use his definition in this study.

Utility - according to Nielsen (2012), utility is whether or not the system provides the appropriate functions to successfully navigate in

the system and complete tasks. Usability and utility are closely related, and the same methods can be used to measure them both.

Expertise - or *domain knowledge*, is, according to Sternberg (2009), often characterized by deeper domain-specific knowledge and problem solving abilities. Domain experts develop advanced mental representations - schemas - of problems that allows them to apply their problem solving skills to novel problems in their domain. If these problems contradict existing knowledge and schemas, experts are more flexible in revising their strategies than be novices. Domain experts also process their knowledge information more effectively by automating sub-processes of domain-specific tasks.

Cognitive load - Cognitive load often refers to the amount of *working memory* and mental effort required to perform a task (Wickens, Hollands, Banbury, & Parasuraman, 2013). A general agreement is that working memory is a limited resource that is quickly depleted when novel, complex or difficult tasks are undertaken. Cognitive load is highly dependent on the level of expertise of an individual. A fundamental assumption is that schemas developed and stored in long-term memory allows for more effective cognitive processing. The automatization of sub-processes may also contribute to lowering the amount of working memory needed to complete a task. In this way, expertise and cognitive load are intrinsically linked (Kalyuga, 2009).

Cognitive load and mouse movement behavior - Cognitive load can be measured in several ways, some of which are rather intrusive. One un-intrusive way of measuring cognitive load is by studying *mouse dynamics* which includes mouse travel distance, average mouse movement speed, and number of mouse clicks during task performance. These measurements are recorded during task completion and does not impact the performance of the participant in any way (Grimes & Valacich, 2015). In their study, Grimes and Valacich (2015) found statistically significant correlations between mouse dynamics measurements and cognitive load, as reported by participants using a NASA-TLX questionnaire. The strongest correlations with reported cognitive load were found with measures of mouse travel distance and mouse travel speed.

Method

Participants

The present study focused on three groups of employees at IFS:

1. Employees of IFS R&D (Research & Development) - Have no domain knowledge.
2. Employees of IFS R&D - Have domain knowledge.
3. Employees of IFS Scandinavia - Mixed levels of domain knowledge.

The study involved 34 participants. Group sizes were roughly equal, with 11 participants in groups 1 and 2, and 12 participants in Group 3. Participants were selected based mainly on their availability, although some basic requirements regarding their level of experience had to be met. We were initially provided with an email list by IFS of potential employees that matched our requirements. Each participant was subsequently contacted by mail to discuss their participation.

Design

Our study was an example of a between-group design which means that each participant was a member of one group only. The results from each group were then compared to the others to examine differences and the effect of the independent variable - group membership. Our dependent variables were TLX scores, input interactions and mouse travel distance. We also quantified the number of usability problems found by each participant. By assessing the nature of a complaint we counted how many problems of each priority level were identified. The number of problems found in total by each participant were put into our last dependent variable.

Procedure

A user study was conducted using a think-aloud test model with semi-structured interviews. The user study took about 30 minutes per participant including a brief introduction. During the test, the screen was filmed and voices were recorded. The test took place in a room at IFS where

disturbing elements could be minimized. There were two test leaders with different roles. One researcher, with whom the participants were asked to constantly communicate, led the test and performed the interview. The other researcher took notes during the test. The same test leaders were used for all test sessions.

A test scenario was designed in IFS Application. The scenario involved creating a new project using an existing project structure as a template. This could be done in several ways using existing functionality with the only requirement being that a new project head had to be created first, otherwise copying the existing project template would not work. The scenario was split into three separate subtasks.

The participants were initially sat down in front of the computer and asked to read and sign the test information and consent form. Participant were then verbally informed about the test components - the think-aloud protocol, the test scenario itself, and the TLX questionnaire. They were given an opportunity to ask questions before the test. Participants then received three task to complete. After each completed task they were asked to fill in the corresponding TLX questionnaire. Following the test scenario, each participant was interviewed in a semi-structured way. They answered a list of basic questions about their experience levels, professional background and the test scenario itself. These questions were focused on usability and task structure.

Results

A one-way multivariate analysis of variance (MANOVA) was conducted to assess the effect of group membership on four dependent variables - mean overall TLX score, number of input interactions, mouse distance travelled, and number of usability problems identified - as a whole. Following a nonsignificant Box's *M* test, Pillai's trace results of the MANOVA indicated significant differences between the groups on the dependent measures at the $p < .025$ level, $V = .57$, $F(8, 50) = 2.49$, $p = .023$, $\eta_p^2 = .285$, observed power = .778. Follow-up univariate analyses of variance (ANOVAs) for each dependent variable were done at the $p < .025$ level using the Bonferroni method to control for Type 1 errors generated by multiple comparisons. ANOVA results were significant for TLX scores, $F(2, 27) = 4.49$, $p = .021$, $\eta_p^2 = 0.25$, observed power = .605, and a Tukey post hoc test revealed significant differences in mean scores between Group 1 ($M = 32.78$, $SD = 8.72$) and Group 2 ($M = 16.22$, $SD = 13.19$). ANOVA results were also significant for mouse travel distance, $F(2, 27) = 4.43$, $p = .022$, $\eta_p^2 = .247$, observed power = .598, where post hoc tests again revealed differences in means between Group 1 ($M = 20.57$, $SD = 11.59$) and Group 2 ($M = 9.58$, $SD = 4.60$). In post hoc tests following both significant ANOVAs, Group 3 failed to differ enough from Group 1 or Group 2 to reach statistical significance.

Nonsignificant ANOVA results were shown for number of interactions, $F(2, 27) = 1.84$, $p = .178$, $\eta_p^2 = 0.12$, observed power = .244, and number of problems found, $F(2, 27) = 0.88$, $p = .424$, $\eta_p^2 = 0.062$, observed power = .117. No post hoc tests were therefore investigated.

Results of the MANOVA showed that there were indeed significant differences between the groups of participants across multiple variables, and follow-up ANOVAs revealed where these differences were found. Group 1, on average, reported significantly higher TLX scores and mouse distance travelled than Group 2, but there were no significant differences between them in the input interactions and problems found variables. Group 3 did not differ significantly from the other groups in any of the four variables.

Analyses of correlations between variables revealed that for TLX scores, there were positive correlations with mouse travel distance, $r(28) = .39$, $p = .033$, number of interactions, $r(28) = .468$, $p = .009$ and usability problems identified, $r(28) = .582$, $p = .001$. Apart from their correlations with TLX scores, mouse travel distance and number of interactions were also positively correlated with each other, $r(28) = .826$, $p = .000$.

As mentioned above, the number of usability problems found was positively correlated with TLX scores but did not show any significant correlation, positive or negative, to neither interactions, $r(28) = .244$, $p = .194$, nor mouse travel distance, $r(28) = .051$, $p = .787$.

To summarize the correlation results, TLX scores showed positive correlations to all other variables. Interactions and mouse travel distance were positively correlated to each other and to TLX scores. Lastly, the number of problems found was positively correlated to TLX scores but did not show any correlation of any kind to interactions or mouse travel distance.

Analysis

From the thematic analysis of the qualitative data, three themes were abstracted, the gap between reality and expectation, Navigation in the system, and Leaps of faith.

The gap between reality and expectation

This theme is defined by the group's initial expectations of how the copy function should work and also how the groups differed in their ability to accept and adapt to how the function actually worked. The biggest issue that all participants, no matter the group, voiced concern over the fact that the process of copying a project is unnatural and unintuitive in the sense that it did not match the participants' expectations of how such a process should work. However, Groups 1 and 2 approached this issue differently with Group 1 expressing a lot more faith in the system and therefore getting more caught up when realizing their view did not match reality and Group 2 merely pointing out the backwardness of the workflow and then falling back on previous experience to be able to do a workaround and proceed with the task. When Group 1 got stuck in the system they quickly got emotional and audibly frustrated, often cursing and expressing confusion, and in the interview they pointed out that this was unacceptable since they probably wouldn't have been able to proceed with the task in a real life scenario without asking for help and mentioning several times that the order of things goes against the natural expectation. Whilst performing the tasks participants from group 1 would often redo the same thing over and over again certain that they were doing something wrong, showing little understanding for the fact that their expectations do not match up with the reality of how the function actually works. This also shows in the statistics where Group 1 showed a significantly higher mental load when performing the task compared to the other groups, meaning that they had to focus and think a lot harder during the test.
As previously mentioned, Group 2 also considered this as an issue and most of them also considered it to be one of the bigger issues with the function, but they also showed more acceptance towards it, often displaying an attitude of "it could be better but it is easy once you have learned it". Although they expressed that it did not work as they expected and most of them made the same errors as Group 1 during the test, they did not seem anywhere near as flustered as Group 1 on the issue. The gap between reality and expectations were not as big for them since they often knew how to work around the issue and were more accepting of the fact that this is just way it works. Again, they did not like the way it worked but they didn't mind it either.

Navigation in the system

This theme is defined by the process by which the different groups moved differently through the system. The choice of navigation method differed between the groups depending on whether or not they had domain knowledge. Those with domain knowledge, Group 2, often relied on using shortcut command keys to navigate through the system because they were more familiar with the system as a whole and therefore knew how to navigate more effectively than their counterparts in Group 1. But this caused them to often do something that was not intended, they went outside the system and worked around the intended workflow we had put before them. Therefore they often showed confusion over things that were not relevant to the task at hand e.g. getting lost in parts of the system that they weren't even supposed to be in, and therefore losing focus of providing actual feedback. The reason why this is relevant is that even though Group 2 always were successful in finishing their task, they went outside the intended system to do it. Their first impulse was to do what they knew how to do, because they are coloured from previous experiences and for the most part this led to them largely ignoring the "manual" way, which leads to them bypassing important problems.
Group 1 rather chose to do it the manual way

e.g. using RMB (right mouse button) to find menus, because according to them that is the most common way to do it. And it is here that the two groups differ in a major way. If you would like to know if your system has an intuitive interface or not, you need testers that go through the system the way a regular user would. Sometimes during the test participants from Group 1 also went outside the intended workflow to find a solution to the task but not with the same reasoning as the participants in Group 2. Participants in Group 2, as previously mentioned, went outside the system with clear goals and intentions because they knew what they were looking for. They were aware of a workaround. Group 1 participants did it exploratory, because they thought they couldn't find the solution they were looking for in the part of the system where we had placed them, even though they could.
There is a big difference in these two ways of reasoning and we would argue that the latter is more useful for testing. If your tester thinks that they have to move outside the intended workflow to find the way something works, your system is probably not intuitive enough in the way it introduces the required task, an observation you probably wouldn't have found with a tester who know what to do and therefore work around your problems in an unintended way.

Leaps of faith

This theme can be seen throughout the entire task and abstracted from the interviews as a lot of insecurities were expressed about where, how and when things happened. All the groups showed signs of what we have chosen to call leaps of faith in the sense that they took a lot of chances and had to guess a lot when they performed the task. In the interviews and during the test the participants showed this by often saying "let's see what happens" right before they committed to a choice. But there were some differences between participant with domain knowledge and those without as to why a leap of faith occurred. Group 1 often criticized the system for having too little information or guidance about what they need to do. Group 1 also criticized the system for presenting the user with too many choices with too little information about what these choices mean. With so many choices and so little information about them the participants in Group 1 found it hard to know whether or not the information was relevant and therefore did not know if it could be skipped. Group 2 however, took the same leaps of faith as Group 1 but for different reasons. Group 2 who had domain knowledge often read through all the checkboxes and information, because they often knew what it meant or at least had some notion of what it meant. Group 2 showed their knowledge of the domain in the feedback where they rather than criticizing the lack of information, criticized the way in which information was displayed. They pointed out that they felt like the system asked for too much information that was not relevant in the given situation, taking on a bigger perspective in their critique than Group 1. Group 2 reacted to the information being displayed as confusing, they felt that it was out of place which made them uncertain and hesitant about what they were currently doing and what the task was about. This caused them to take chances, moving forward under the attitude "Well, I suppose I can move forward now".

Conclusions

Based on the results of the statistical tests and thematic analysis, we believe initial usability tests with employees without domain knowledge, regardless of their department, would identify the most serious issues and assess the overall intuitiveness of the system. After addressing the identified usability issues further tests can be run with employees that have domain knowledge. These employees are likely to comment on any functional problems in the new and revised system, and evaluate it from a broader system perspective. Because of this, we recommend an iterative way of testing usability where initial tests with domain novices reveal the most serious usability problems which are then addressed before the next iteration of tests are conducted with domain experts, if possible. In this way, problems of all levels of severity are more likely to be found.

References

Arvola, M. (2014). *Interaktionsdesign och UX - om att skapa en god användarupplevelse* (1st ed.). Lund: Studentlitteratur AB.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(May 2015), 77–101. http://doi.org/10.1191/1478088706qp063oa

Grimes, M., & Valacich, J. (2015). Mind Over Mouse: The Effect of Cognitive Load on Mouse Movement Behavior. *ICIS 2015 Proceedings*, (Furnham 1986), 1–13. Retrieved from http://aisel.aisnet.org/icis2015/proceedings/HCI/9

Kalyuga, S. (2009). Knowledge elaboration: A cognitive load perspective. *Learning and Instruction*, *19*(5), 402–410. http://doi.org/10.1016/j.learninstruc.2009.02.003

Nielsen, J. (2012). Usability 101: Introduction to Usability. Retrieved May 5, 2016, from https://www.nngroup.com/articles/usability-101-introduction-to-usability/

Sternberg, R. (2009). *Cognitive Psychology* (5th ed.). Belmont: Wadsworth.

Wickens, C., Hollands, J., Banbury, S., & Parasuraman, R. (2013). *Engineering Psychology and Human Performance* (4th ed.). Upper Saddle River: Pearson.