Patrik Ståhl | Tomas Eklund | Franck Tétard | Barbro Back

A Cooperative Usability Evaluation of the Domino Software

Turku Centre for Computer Science

A Cooperative Usability Evaluation of the Domino Software

Patrik Ståhl
> Department of Information Technologies, Åbo Akademi University
> `patrik.stahl@abo.fi`

Tomas Eklund
> Department of Information Technologies, Åbo Akademi University
> `tomas.eklund@abo.fi`

Franck Tétard
> Department of Information Technologies, Åbo Akademi University
> `franck.tetard@abo.fi`

Barbro Back
> Department of Information Technologies, Åbo Akademi University
> `barbro.back@abo.fi`

Abstract

This report contains the results and findings of the usability evaluation performed on the Domino software at the Institute for Advanced Management Science Research (IAMSR) at Åbo Akademi University, Turku, Finland. The evaluation sessions were held between 28.11.2005 and 9.12.2005. The report consists of five sections:

An executive summary that briefly summarizes the results and lists the recommendations for the usability problems found.

A description of the usability evaluation performed. This section outlines the goals of the evaluation and describes the methodology, methods and procedures used. It also contains information about what kind of data was collected, what tools were used to perform the analysis, a description of the participants and a description of the model used to rate the severity of the usability problems found.

The evaluation results contains the results and summaries of all the data collected during the evaluation.

Findings and recommendations contains detailed information about the usability problems and bugs found in the software and recommendations for solving them.

The final section, conclusions, contains concluding thoughts on the findings of the evaluation.

TUCS Laboratory
Data Mining and Knowledge Management Laboratory

1 Executive summary

In may 2005 the Institute for Advanced Management Systems Research (IAMSR) at Åbo Akademi University requested a usability evaluation on a software tool, Domino, developed at the same department. The tool was developed as a part of the Domino project, financed by the Academy of Finland. The purpose of the request was to obtain a report on the usability issues with the current prototype of the software. The planning of the evaluation began in September 2005 and lasted until November 2005 when the first pilot test began.

After careful consideration of different methodologies and methods, five participants were chosen to take part in the evaluation, all students at Åbo Akademi University participating in a course involving the Self Organizing Map (SOM).

The evaluations were all conducted in a usability lab, equipped with tools to record the interaction between the participants and the software, as well as the dialog between the moderator[1] and the participant. A set of tasks and scenarios were created to test the most important and functioning parts of the software. The interaction between the participants and the software as well as the post-test questionnaires and interviews revealed both strengths and weaknesses in the current prototype. Listed in Table 1 is a list of the usability problems found in the evaluation.

---

[1]The term moderator refers to the person in charge of the evaluation.

Problems	Recommendations
Participants had trouble with the window handling.	Introduce a more obvious way to manage windows within the application.
Participants were confused by the simulation dialog window.	Rethink the simulation dialog window.
Participants had trouble finding a way to identify the ID of different clusters.	Make it easier to identify the different clusters.
Participants were confused by the cluster colors and mixed them up with SOM colors.	Make a bigger distinction between the colors of the clustered map and the SOM colors.
Participants were confused by the import labels dialog window.	Some confirmation, visual or other, that labels have successfully been imported is required
Participants had trouble identifying SOM colors because of the node selection color.	Find a different way to show that a node is selected.
Participants had trouble finding the simulation tool from the menus.	Relocate "Simulation" in the menus.
Participants were confused by the clustering dialog window.	Rethink the buttons in the clustering dialog window.

Table 1: *Usability problems found and recommended action.*

In addition to these usability problems a number of bugs were also found in the current prototype. These bugs also needs to be addressed in the next iteration of the software.

2 Description of the usability evaluation

The following section contains a description of the usability evaluation and how it was shaped. It describes what data was collected and how. The goals of the evaluation are also described along with the evaluation procedures, the testing environment and the participants.

2.1 Test goals

The purpose of the evaluation was to acquire a list of usability problems found in the current prototype of the software so that the developers could set the goals for the next phase in the development of Domino.

The evaluation focused on specific areas of particular interest to the designers in this particular state of development. More specifically the following features of Domino were examined:

- *Clustering a map*. Study how easy or difficult it currently is to cluster a map in Domino.

- *Importing labels*. Although a small task it can be central when comparing different nodes to each other and should, therefore, be fairly simple.

- *Simulating*. A central tool of Domino. Study how easy or difficult users find the current implementation of the simulation tool to use. The underlying simulation algorithm is by no means final but an interface has none the less been designed and should be tested.

- *Interpreting maps*. Central to Domino. If the users are familiar with the SOM but cannot make sense out of the maps in Domino due to its interface, then Domino will have no purpose.

2.2 Testing environment and equipment

The test facility used was the HCI laboratory located in the University of Turku's premises at the Department of Information Technology. The laboratory consists of two rooms, divided by a one-way mirror. It is equipped with hardware and software for recording the user's interaction with the tested

device. The hardware consists of a digital video camera, a scan raster, a mixer and mixer monitor and a PC for analyzing and reporting. The examination room also has a PC, on which the user performs the evaluation.

2.3   The evaluation procedure

The methodology used in the usability evaluation of Domino was based on the methodology described by Rubin [6] for exploratory tests. The process for evaluating the interface is very much informal, and collaboration between the user and the moderator is very high. This means that the focus will be not be on gathering quantitative data to be able to measure *how well* users did in the evaluation, but rather on collecting qualitative data to be able to understand *why* users performed the way they did.

The method used for evaluating the usability of Domino was cooperative evaluation [3]. The method is described by its creator Monk et al. [3] as a method suited for testing prototypes. The rules for moderator intervention are not as strict as in a classical usability test, making it easier to steer the users attention away from unfinished areas of the software and also to guide the user back on track again, should the user get lost.

In a cooperative evaluation the users perform a set of tasks, selected by the designer, using the prototype begin tested. While the users perform the tasks they are asked to think aloud and and ask questions. The moderator also asks the users questions while they are performing the tasks. Through out the whole session the moderator looks for unexpected behavior and comments on the interface by the user [3].

The number of participants for the evaluation was chosen based on the discount usability engineering approach by Nielsen [5]. This meant choosing a number of between three and five participants. Ten users were interested in participating and from these ten, five were chosen based on their previous experience with financial analysis and financial ratios. This was in turn measured with help of a screening questionnaire asking the participants whether they had passed either of the two financial courses offered at the Department of Information Systems and if they had any previous working experience within the area. They were also asked to rate their experience level with a number of other things, which can be seen from screening questionnaire, included in Appendix A.

2.3.1 Pre-evaluation procedures

Along with the email saying that the participants had been chosen for the evaluation, a brief introduction to the SOM and Domino was also sent. The document can be found in Appendix A. The participants were all asked to read the document before coming to the evaluation. On the scene they were again asked if they had read the document and if not, they were given a few moments to read it.

2.3.2 Evaluation procedures

The evaluation began with a small presentation of the evaluation and the testing facilities. This was followed by a small walkthrough of the evaluation procedure. After walking the user through what was going to happen the user got to familiarize him- or herself with the test scenarios, which can be viewed in the Appendix A.

Effort was made to perform the above procedures in a relaxed manner in order for the participants to not feel stressed by the fact that they were in a test situation in which they were being observed.

After having familiarized themselves with the test scenarios the participants were asked to take a seat at the computer where they performed the actual tasks. During this interaction with Domino the moderator sat right next to the participant, asking questions in order to get the participant to voice their thoughts on what he or she was doing at the moment. While performing the tasks the users were also required to fill out a report. This report can be viewed in the Appendix A.

During the session two types of observations were captured using the equipment in the lab. To capture comments of the participants sound was recorded with a video camera, set to record sound only. The camera was, at all times, pointed away from the participant, in order to lessen the feeling of being observed. The second type of observation captured was the interaction between the participants and the software. Video of this interaction was captured through a scan raster.

2.3.3 Post-evaluation procedures

On completing the scenarios of the evaluation, the users were given a post-test questionnaire (see Appendix A) to fill in, to capture their thoughts and feelings

about the software while they were still fresh. The questionnaire was then followed by a short interview, consisting of five questions.

1. What did you like best/least about the software?

2. What kind of tasks do you think the software could be used for?

3. Are there any analyzing-tools you would like to see added?

4. What one thing would you tell the designers to change?

5. Anything else you would like to comment about the software?

The interview aimed at capturing anything the questionnaire could have missed and gave the participants a chance to freely verbalize any thoughts they might have on the software.

2.4   The participants

The participants for the test was chosen based on three criteria.

- *Availability*. Since the evaluation was performed at a university, students were easy to find and were chosen as the target population for the test.

- *Background*. From the students, business and engineering students were chosen in accordance with a state of the art survey performed by Eklund et al. [2], which found that common end users of this kind of software in large Finnish companies have a degree in either business or engineering.

- *SOM knowledge*. The last criteria was that the test population should have some knowledge of SOM and if possible some knowledge of financial key ratios and financial analysis.

Students were recruited from a course that contained an introduction to the SOM as well as a few exercises. A screening questionnaire was designed to easily pick out the most qualified participants of the ten willing to participate. The questionnaire consisted of three categories of questions: general, education and background. The questionnaire used can be found in Appendix A. Based on the answers from this questionnaire five participants were chosen.

2.5 Pilot testing

Before performing the evaluation with actual participants, it was decided that one or a few pilot runs of the evaluation should be conducted. Participants for these pilots were recruited from "down the hall"[2]. Before the pilot testing began it was decided that the whole evaluation session should not take much more than one hour, in order to have the participants full attention for the whole session Barnum [1, p. 161]. In case of sessions running longer than this, alterations should be made to shorten the sessions.

First pilot test
For the first pilot the pair testing method was chosen. This meant that two users were recruited to perform the scenarios together with a test moderator present, in case they got stuck and were unable to solve the problem and move forward on their own.

As it turned out, the whole test configuration needed some modifications, starting with the scenarios and tasks. The participants were confused with the tasks and did not know what to do, and the test moderator waited too long to intervene and give the participants the guidance they needed. The dialog between the two participants was rich but sometimes led to over-analyzing of tasks, mostly because of glitches and unclarity in the scenarios and tasks.

As can be seen from Table 2 the first pilot lasted well over one hour in total and a substantial part of that time was spent trying to solve the first task. This task was in no way meant to be difficult or time consuming and was, therefore, revised.

Second pilot test
In the second pilot, cooperative evaluation was used instead of pair testing as the evaluation method. The second pilot was more successful in the participant understanding the tasks. There was, however, still a problem with the first task being slightly unclear, and therefore, took too long to complete (see Table 2). The dialog between the moderator and the participant was very natural and provided good information. As a result this evaluation method was also chosen for the third and final pilot.

At the end of the evaluation session a discussion began concerning the method to use and so the post-test interview was skipped because of lack of time. A total time is therefore not available for the second pilot.

Third pilot test

---

[2]This expression suggests recruiting e.g. colleagues or other people that are easily available and not necessarily representative of the end user population

| | Pilot 1 | Pilot 2 | Pilot 3 |
	(Pair testing)	(Coop. eval.)	(Coop. eval.)
Task 1	39:38	25:21	10:42
Task 2	00:58	01:37	00:58
Task 3	01:03	00:35	00:30
Task 4	03:06	11:18	10:05
Task 5	14:07	04:32	10:07
Questionnaire and interview	14:54	-	21:28
Time total	1 h 12 min 46 sec	N/A	53 min 50 sec

Table 2: *Evaluation Pilots*.

The third pilot went well overall. A few minor things were detected in the scenarios that still needed correcting or adjusting. After this pilot it was decided to move on and start calling in the chosen participants to start the real evaluations.

2.6   Types of data collected

The data collected from the evaluation was mostly qualitative data in the form of the video recorded during the sessions and the interview conducted after the participants had completed the scenarios. In addition to this, some quantitative data was also obtained from the different questionnaires the participants were asked to fill out.

In accordance with Monk et al. [3, p.35] two things were sought for in the video collected: deviations and comments. Deviations meaning deviations from the baseline described in the next section and comments that the participant made while performing the tasks.

A deviation from the baseline does not necessarily equal a usability problem but can just mean that the user found a slightly different way to complete the task or that the user made a simple mistake. However, a deviation can also mean that the user really had problems completing a task because of some underlying usability problem and in these cases the usability problems are recorded. The comments collected further clarify the intentions and thoughts of the participants at a certain point.

2.6.1 The baseline and participant deviations

In order to examine how the users deviated from the way that the developers had intended the software be used, we asked a developer to perform the same evaluation tasks that the users had performed. The data collected in that session was then used as a baseline in analyzing the data collected from the participants. The session was recorded in the same way as the others and then analyzed to create a chart of how the designer went about solving the tasks. This chart can be seen as Figure I in Appendix B.

The boxes to the far right, with bolder borders, represent the task as a whole. The boxes to the left of each of these boxes represent subtasks, i.e. smaller tasks that are required to be performed in order to complete the current task. Each evaluation was compared to this chart and a separate chart was created for each individual participant (see Figures II - VI in Appendix B). Deviations from the baseline was marked with blue borders and blue text. The task between task two and three, "Tile Windows", was not given a task number since it was not required by the user to perform this task. It was still included since it was what the designer did, when performing the tasks.

An aggregated chart was also made showing all the places where the users deviated (see Figure VII). Not all deviations proved to be usability problems though, and another visualization of the collected deviations was made with usability problem areas marked with red (see Figure VIII).

2.7 Severity Ratings

The severity rating method used to rate the usability issues in this report is based on the model presented by Nielsen [4]. Figure 1 illustrates this method, which combines the two most important dimensions of a usability problem: how often the problem occurs and to which degree it hinders the user from performing the task at hand.

Nielsen [4] recommends having multiple usability experts rate the usability problems and then aggregating the results for the best result. In this case, however, the severity of the problems were rated by one person only, the first author of this report.

		Users experiencing the problem	
		Few	Many
Impact of problem on the users who experience it	Small	Low severity	Medium severity
	Large	Medium severity	High severity

Figure 1: *Severity rating combining two dimensions of usability problems Nielsen [4, p. 104].*

3 Evaluation results

The following section presents the results of all the data collected in the evaluation. This includes data from questionnaires, the post-test interview and the report that was required by the participants to fill in during the actual evaluation. This section also lists all deviations and comments gathered from the audio and video.

3.1 Participant results

The results from the questionnaire revealed that the chosen participants had the following characteristics:

	$\bar{x}$	σ	**Median**
Age	24.2	1.92	24
Credits	105.6	52.87	101

Table 3: *Age and credits of participants.*

The large standard deviation in credits for the participants is a result of participant number 5, who had fairly little credits in comparison to the others. All participants, apart from the last one, were in the final stages of their studies with a total amount of credits exceeding 100.

The larger part of the participants were male. All were majoring in information systems except one who was majoring in "Communications and autom-

	Number of participants	%
Male	4/5	80
Majoring in Information Systems	4/5	80
Other degree(s)	0/5	0
Had passed Investment Planning	4/5	80
Had passed Financial Planning	2/5	40
Had passed either IP or FP	5/5	100
Had passed both IP and FP	1/5	20
Color blind	0/5	0

Table 4: *Additional information about the participants*.

atization". None of the participants had any previous degrees. Most of the participants had passed the course Investment Planning, with the exception of one. Less than half of the population had passed Financial Planning and only one participant had passed both courses.

The last fact in Table 4 was not in the original screening questionnaire. Arguably it should have been since a color blind participant would very likely have little use of a tool using colors to communicate information abouts the data. Hence an email was sent out to participants later on regarding the matter. Participants were simply asked to answer the question "Are you color blind?" with a simple "yes" or "no" in reply to the email.

Questions	$\bar{x}$	σ	Median	Mean
Windows operating system	4.2	1.1	5	5
OS X operating system	1.8	1.1	1	1
Linux operating system	2.2	0.84	2	2
Microsoft Word	4.2	0.84	4	4
Microsoft Excel	3.2	1.1	3	3
Databases	2.8	1.48	3	3
Financial benchmarking	2.4	1.14	2	2
Financial analysis	2.2	0.84	2	2
Self-organizing maps	1.2	0.45	1	1

Table 5: *Screening questionnaire summary*.

Table 5 shows a summary of the answers given by the chosen participants in the final part of the screening questionnaire, background. The individual answers can be viewed in Table 9, in Appendix C. The summary suggests that the participants were very familiar with the Windows operating system, Microsoft Word and Microsoft Excel. The participants felt that they had very

little experience with the SOM, suggesting that none of the participants had worked with the SOM before the course.

3.2   Task completion times and error rates

Figure 2 shows all of the participants' task completion times in proportion to the total completion time, meaning the total time it took for the participants to complete all five tasks. The biggest difference in how much time any participant spent on a task compared to the developer, referred to as the baseline in the Figure, is clearly the case of participant 2 and task 3. In this particular case the user experienced difficulties finding a way to identify a specific cluster.

All the participants, aside from participant 2, spent the most time on the first task. This is quite natural since this task also involves the first interactions with the interface. The fact that the baseline also spent the most time on this task suggests that was also the most time consuming. There was a notable difference in the total completion time between the designer and the participants due to the fact that the designer was already very familiar with the interface. This can be seen from Table 10 in Appendix C.

On the whole there were no large deviations from the baseline in this aspect, aside from the one mentioned, indicating that there are no really large usability glitches making it impossible for a user not familiar with the interface to work with the software in order to complete tasks similar to the ones used in this evaluation. One must of course take into consideration the fact that this evaluation involved a moderator or co-evaluator sitting beside the evaluator and guiding this person from venturing into the most unfinished parts of the interface.

As can be seen from Table 6 the participants had no real difficulties providing the right answers for the evaluation report. The lowest percent correct answers was given by participant 4 who still managed to get 60 % of all the tasks right. The reason for the wrong answers in this case stems from the fact that this user also had trouble identifying the ID of different clusters. The error that participant 1 made was also for the same reason.
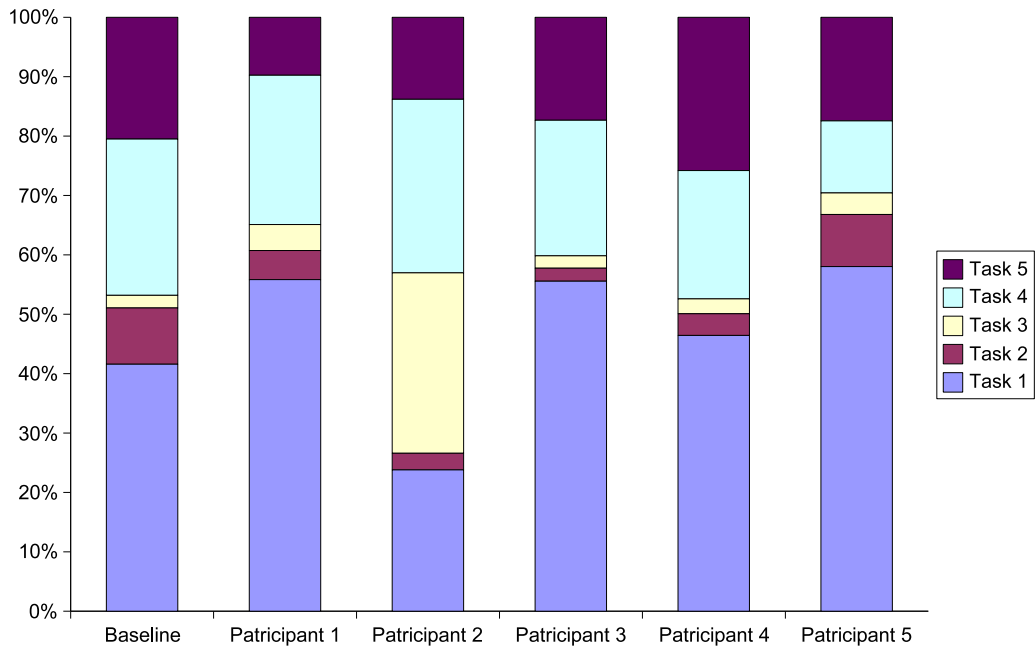
Figure 2: *Task completion times in proportion to total completion time.*

	Participant 1	Participant 2	Participant 3	Participant 4	Participant 5	% correct
Task 1	14/14	14/14	14/14	14/14	14/14	100
Task 2	1/1	1/1	1/1	0/1	1/1	80
Task 3	0/1	1/1	1/1	0/1	1/1	60
Task 4	14/14	14/14	14/14	14/14	14/14	100
Task 5	1/1	1/1	1/1	1/1	1/1	100
% correct	80	100	100	60	100	

Table 6: Task error rates

3.3 Deviations and comments

The numbers of each deviation and comment refer to the numbers in the Figures in Appendix B. The Figures provide a visual representation of the tasks and where the users deviated from the baseline.

3.3.1 Deviations

Participant 1
* *Tile windows*. The user skipped this step and had difficulties handling the different windows required to work with in order to complete the tasks. After clustering the map into five clusters the user was asked to again locate the component planes to be able to complete task 1. The user had problems finding the component planes. After a few minutes of struggling with trying to find the component planes the user finally found the right window.

3. Identify clusters 3 and 4. User got stuck and asked for help. User had trouble figuring out how to identify different clusters. The user got a hint or two and started trying to select different neurons on the clustered map and noticed changes in the "Selected nodes" tab. The user spent a long time trying to find some kind of cluster ID.

3.3. Identify ID of selected clusters from the clusters tab. When the user finally found the "Cluster ID" he mixed the IDs with the bold numbers to the far left. After identifying cluster 3 and 4 on the clustered map the user was asked to point the same clusters out on the component planes. The user confuses the clustered map colors with the SOM colors of the component planes.

4. Identify cluster characteristics. The user still had some problems in distinguishing between the SOM colors and the clustered map colors. It took some further explaining before the user understood the concept of clusters and neurons.

7.4. Click left "Select All". The user only chose the relevant companies instead of selecting all.

7.6. Click "Apply". The user did not notice any change in the interface after clicking "Apply" since he did not have the clustered map window open. The user tried clicking the right "Select All" and then "Apply" again. After clicking "Close" the user noticed a new "Labels" tab had appeared next to the "Clusters" tab.

8.1. Deselect selected neurons. The user had random neurons selected most of the time but this did not seem to bother him.

8.2. Select nodes K-C and G-P. The user only selected the K-C neuron before moving on.

8.3. Deselect selected neurons. The user analyzed the neurons with only one neuron selected.

Participant 2
* Tile windows. The user skipped this step and had difficulties handling the different windows required to work with in order to complete the tasks.

3. Identify clusters 3 and 4. The user had no trouble identifying cluster 3 but when asked to repeat the process in order to identify cluster 4 the user experienced problems.

3.1. Select nodes on clustered map and *3.2. Select "clusters" tab.* The user did these two in the opposite order. This made little difference and the user had little trouble finding out which cluster had which ID, the first time. Finding an ID for cluster 4 proved much more difficult for unknown reasons.

5.2. Select node from perceived best cluster on the clustered map and *5.3. Identify ID of selected cluster from the "clusters" tab.* The user remembered the ID of the cluster he wished to identify so he did not have to, nor did he feel like going through the process of identifying it again.

6.3. Identify ID of selected cluster from the "clusters" tab. The user kept selecting and deselecting nodes and had trouble answering which cluster was which. When probed a little about the nodes and their ID it turned out the user was quite lost. He suggested counting the number of nodes in the clusters and comparing the number to the "Nr of Nodes" column in the clusters tab.

When asked to bring up the clustered map again the user clustered the map again instead of just bringing the old clustered map window up. The user suggested somehow putting visual labels on the clusters. The evaluator had to hint to the user to try to deselect all other nodes than the ones in the relevant cluster.

When asked what the user thought the Cluster ID color meant he asked if it meant a poor value. The user did not realize the difference between the SOM colors and the clustering colors until informed by the evaluator.

7.4. Click left "Select All". The user marked and moved the labels one by one, instead of moving all the labels at the same time to the "Visible labels" area.

7.6. Click "Apply". The user only had the components planes window open and noticed no visible change after clicking "Apply". The user then tried selecting one of the companies and clicking "Apply" again. When nothing happened the

user once again tried clicking "Apply".

Because there was no indication of the labels getting imported when the user did not have the clustered map window open at the time, the user tried using "Customize labels" to get the labels to appear in the component planes. The user tried repeated times to "Apply" and "Close". After a while the user found the labels in the clustered map but wanted the labels to appear on the component planes as well.

8.1. Deselect selected nodes. The user had no nodes selected so there were no nodes to deselect.

8.3. Deselect selected nodes. The user did not deselect the nodes when analyzing them.

9.4. Click "Create a new label". The user directly clicked "Close" without first clicking "Create a new label". The user then checked both the component planes window and the clustered map window for any visible changes without finding any. The user then tried repeating the process. On the second attempt the user successfully managed the simulation.

Participant 3
* Tile windows. The user skipped this step and had difficulties handling the different windows required to work with in order to complete the tasks.

3. Identify clusters 3 and 4. The user explored the clustered map for some time, clicking around, having trouble trying to find some kind of ID.

3.1. Select nodes on clustered map and *3.2. Select "clusters" tab*. The user first selected the clusters tab and then kept trying to figure out how to identify the different clusters. The user continued to click around and even found his way into the simulation window. It took some guidance before the user could successfully identify the clusters.

7.4. Click left "Select All". The user only selected G-P and K-C.

7.6. Click "Apply". The user only had the component planes window open and after not seeing any change in the window after clicking "Apply" one, he tried clicking it a few more times.

8.3. Deselect selected notes. The user did not deselect the selected nodes before starting the analysis. After a while the user started deselecting the nodes while studying their color to then reselect them again before moving to the next ratio.

9.4. Click "Create a new label". After increasing the values by 100 % the user clicked "Close" instead of "Create a new label". The user then thought that the

colors for the G-P node had changed, or that they should have changed, in the component planes, instead of the label moving. He started comparing the previous analysis with the "new" color, and confirmed that the values or colors had not changed. He also noted that there was no change in the clustered map either. The user obviously seemed to understand the clustered map based on his comments.

Having both windows open while experimenting with the Simulation box the user noticed a change right away after clicking "Create a new label".

Participant 4
* *Tile windows*. The user skipped this step and had difficulties handling the different windows required to work with in order to complete the tasks.

2.4. Click "Close". The user seemed unsure what to do with the dialog box after having clicked "Apply" and just dragged it to the side leaving it open. After a while the user closed the dialog box by clicking the red X in the top right corner of the box.

7.4. Click left "Select All". The user only selected the relevant companies to import.

8.3. Deselect selected nodes. The user kept deselecting and selecting the two nodes while analyzing them.

9.1. Select "Simulation" from the "Map" menu. The user did not find it very obvious to choose "Simulation" from the "Map" menu right away, but instead explored the program for a while. The user looked at the "Map" menu but continued his explorations to come back later and found it the second time around.

Participant 5
* *Tile windows*. The user skipped this step and had difficulties handling the different windows required to work with in order to complete the tasks.

3. Identify clusters 3 and 4. The user found cluster 3 and 4 too similar in color to notice any difference between them at first glance. The user had some problems figuring out how to get the ID of the different clusters and tried both importing labels and simulation in the process of getting the ID.

9.1. Select "Simulation" from the "Map" menu. The user had some trouble finding "Simulation". The user knew what she wanted to do but not how to achieve this goal with the program.

9.4. Click "Create a a new label". The user clicked "Close" right after having changed the values. After clicking "Close" and concluding that nothing hap-

pened the user went back to the simulation dialog box again and said: "It does not save it".

3.3.2 Comments

Participant 1
* Tile windows. The user commented on the window handling and said it could be "a little clearer". The user mentioned Excel's solution with tabs at the bottom of each sheet to easily switch between sheets.

2.3. Click "Close". The user: "Is it ok to close this?"

3.3. Identify ID of selected clusters from the clusters tab. The user: "Why would "Cluster ID" start from 0? I think starting from 1 would be more logical."

9.4. Click "Create a new label". The user: "Did it do it already? Do I need to apply somehow?"

Participant 2
8.2. Select nodes K-C and G-P. The user pointed out that he guessed the values under "1" in the "Selected Nodes" tab represented Operating margin but that he was not sure and would like them labeled instead of numbered.

Participant 3
4. Identify cluster characteristics. The user was talking about the cluster color displayed next to the "Cluster ID" in the clusters tab. The user: "Is it the mean value of all the selected neurons?"

7.6. Click "Apply". When asked if something should have happened when the user clicked "Apply", the user answered: "Labels should have appeared in the component planes."

8. Identify neuron characteristics. When first asked to describe the characteristics for the nodes the user confused the previous analysis with the current one and said: "All values were better for K-C."

9.4. Click "Create a new label". The user: The company should have moved a little bit in that direction (indicating movement toward KC).
Going to the Simulation dialog box again the user commented: "I think we have to create a new label."

Participant 4
8.3. Deselect selected nodes. When asked why the user kept deselecting and selecting the two nodes the user replied: "It (referring to the shading used to

imply that a node is selected) can be distorting."

Participant 5

2. Cluster map. The user spent some time studying the clustered map. The user: "I can only see four clusters." The user: "How can I remove these shadows?!"

3. Identify clusters 3 and 4. After the user had found the ID of the different clusters, with some help, the user was asked if the ID was hard to find. The user replied: "Yes, because normally when I use software I go to these menus (referring to the menus in the top left corner of the window). Of course this is quite convenient but I think it would be better if you add something up here also. I did not think that I by simply doing this (referring to selecting nodes) could find out the id of different clusters."

8.3. Deselect selected nodes. When the user deselected the selected nodes and selected a node next to it instead the user was asked why. The user replied that it was distorting the underlying color.

9.1. Select "Simulation" from the "Map" menu. When asked what kind of word or what it should say in the menu for achieving the users goal, the user replied: "To assess the original database."

9.4. Click "Create a new label". The user: "Close? Can I automatically save it when I close it?"

4 Findings and recommendations

The following section presents the usability problems found, bugs and other issues found in more detail. The section also contains a list of in detail recommendations to each usability problem.

4.1 Findings

4.1.1 Usability problems

Table 7 lists all the usability problems found in the evaluation, percentage of affected participants and severity level of the problem. A total of eight different usability problems were found. The usability problems can also be viewed in Figure VIII in Appendix B. The tasks that involve the usability issue are marked with red.

	Usability Problems	Participants affected	Severity level
1.	Participants had trouble with the window handling.	100 %	High
2.	Participants were confused by the simulation dialog window.	80 %	High
3.	Participants had trouble finding a way to identify the ID of different clusters.	80 %	High
4.	Participants were confused by the cluster colors and mixed them up SOM colors.	40 %	Medium
5.	Participants were confused by the import labels dialog window.	60 %	Medium
6.	Participants had trouble identifying SOM colors because of the node selection color.	40 %	Medium
7.	Participants had trouble finding the simulation tool from the menus.	40 %	Low
8.	Participants were confused by the clustering dialog window.	20 %	Low

Table 7: *Collected findings, rated by severity.*

4.1.2 Bugs and other issues

Five bugs were also found while performing the evaluation.

All labels get imported
A few users found that found that while only selecting the two labels for the companies relevant to the task for import, all the labels were instead imported.

Labels get imported when closing the simulation dialog window
This problem occurs when simply opening the simulation dialog window, before importing any labels, and then just closing it by clicking "Close".

Importing labels can cause the clustered windows to get distorted
When importing labels with the clustered map window in windowed mode, the clustered map can become distorted with all borders "out of sync" (see Figure 3).

Blank clustered map
One user found a bug that made the clustered map go completely blank with only its' outer borders remaining. The user used the following actions to produce this bug:
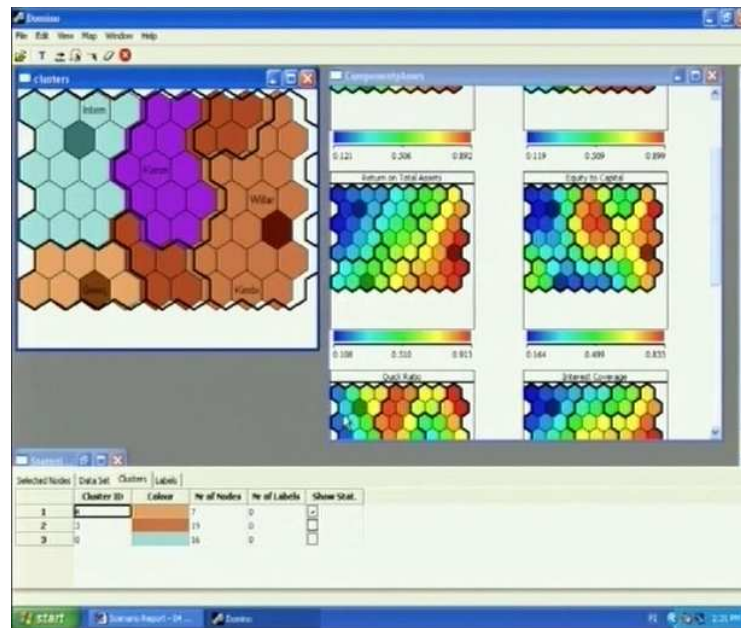
Figure 3: *Clustered map borders "out of sync"*.

1. Opened the map

2. Closed the U-matrix window

3. Minimized the component planes window

4. Clustered the map into 5 clusters

5. Minimized the clustered map window

6. Imported labels by selecting GP and KC from "Labels not visible" one at
   a time and moving them over to the right side using the "»" button and
   then clicking "Apply"

7. Minimized the component planes

8. Brought up the clustered map window by double clicking on it

The result is a blank clustered map, as is illustrated in Figure 4.

Inconsistent labeling
In the import labels dialog box there are two boxes labeled: "Labels not visible"
and "Visible labels". To be consistent in the language used this should be:
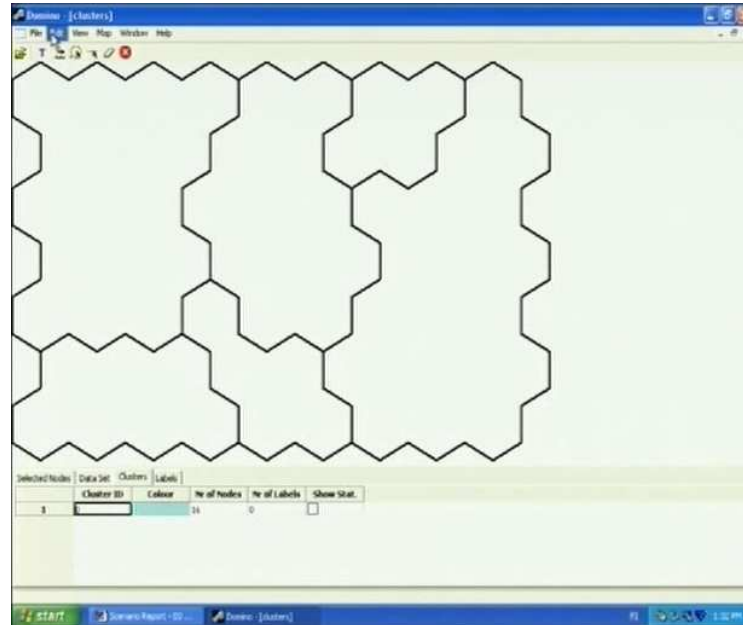"Labels not visible" and "Labels visible".

Figure 4: *Blank clustered map*.

4.2 Recommendations

Following is a list of recommendations to solve the usability problems found in the evaluation. It is recommended that at least the usability problems with a severity rating of medium be addressed in the next iteration of development for the software.

1. *Introduce a more obvious way to manage windows within the application.* In the current prototype working with many windows at the same time can become very cluttery and sometimes confusing. The fact that the windows do not always reach to a change affecting them does not help either. E.g. importing labels will not bring about any visual sign of any change if the clustered map window is minimized.

2. *Rethink the simulation dialog window.* The fact that only one participant

22

got the simulation process right the first time suggests that the logic is not perfectly clear. After having changed values in the dialog window 80 % of the participants clicked "Close". Changing the label of the "Create a new label" button to "Apply" and providing some visual feedback that a change was made might help.

3. *Make it easier to identify the different clusters.* 80 % of the participants had trouble finding a way to identify the different clusters. The fact that the users many times made it as far as to 3.3 (see Figure VIII), meaning that the Cluster ID row was fully visible to the user, without finding any way to identify the clusters suggests that it might be very difficult to locate an ID. If this is to be a central feature is need to be brought forward so it is easier to find. In addition to being difficult to find the cluster IDs also start from 0. Although numbering often starts from 0 in the world of programming languages this does not make sense in a user interface not specifically intended for programmers. Participant 1 commented on the matter while trying to identify cluster 0.

4. *Make a bigger distinction between the colors of the clustered map and the SOM colors.* Although only 40 % of the participants had some trouble with this, it can be potentially disasterous to the analysis of the clusters. A user could, by mistake, make an analysis based on the clustered map colors.

5. *Some confirmation, visual or other, that labels have successfully been imported is required.* In the current prototype if one is not working with the clustered map window when importing labels, there is no way to tell if the labels were imported or not, aside from that a "Labels" tab is added to the lower tabs. Many users had trouble with this dialog box because there is no immediate feedback when clicking "Apply". This led to some confusion in the evaluation.

6. *Find a different way to show that a node is selected.* Some of the participants found it hard to analyze the underlying value of a node based on its color, while having it selected. This is because the current way of showing that a node is selected is by darkening its color. One participant suggested making selected nodes with dots and another way could be to highlight or emphasize the borders.

7. *Relocate "Simulation" in the menus.* Some users had trouble finding the simulation tool from the menus. This suggests that its current position in the menus may not be the best. However only 40 % of the participants experienced this problem and none of these spent an exceptionally long time looking for it.

8. *Rethink the buttons in the clustering dialog window.* Some users found it confusing that they had to "Close" this window after having successfully clustered a map. Having the window just close upon applying the desired action could make more sense to users. A cancel button should also be available in the dialog window since the user should always be able to cancel his/her actions. Therefore changing the current buttons to "Apply" and "Cancel" would be recommended.

4.3 Other feedback

Information on how the participants felt about working with the software was collected with a post-test questionnaire and a post-test interview. The following sections list the a summary of the results gathered from these.

4.3.1 Post-test questionnaire summary

In table 8 a summary can be seen of the post-test questionnaire. A scale of 1 to 5 where 1 was used to indicate a total disagreement with the statement and 5 a high agreement with the statement.

Question	x̄	σ	Median	Mode
Using the software was easy	4.2	0.45	4	4
The concept of the self-organizing map was difficult to understand	2.4	1.14	2	2
Finding the desired menu choices was easy	4.4	0.89	5	5
I liked the look-and-feel of the software	3.8	0.84	4	3
The software has a clear purpose	4.4	0.89	5	5
The software felt complete	3.4	1.14	3	3
It was easy to compare the companies using the software tool	4.6	0.55	5	5
The visual representation of the map was clear	4	1	4	3
Clustering the map was hard	1.8	0.84	2	1
The component planes were clear and easy to understand	4.4	0.55	4	4
Comparing the component planes simultaneously was difficult	1.8	1.3	1	1
Importing labels was easy	4.8	0.45	5	5
Interpreting the map was easy	4.4	0.55	4	4
Simulating was hard	1.6	1.34	1	1
I felt that the tasks were difficult to complete using the software	1.6	0.89	1	1
The tasks would have been easier to complete using tabular data in Excel	1.6	0.89	1	1

Table 8: *Post-test questionnaire summary.*

According to the participant's answers:

- *The participants found the concept of the self-organizing map somewhat difficult to understand.* Given this, some good documentation on the matter could probably be included with a final version of the software.

- *The participants where not thrilled by the look-and-feel of the software.* This is to be expected since the software is still a prototype. Taking a look at this area of the software in a later stage of development would be recommended though.

- *The participants did not feel that the software felt complete.* Also to be expected since the software is still a prototype.

- *Not all participants felt that the visual representation was clear and easy to understand*. Part of this may be because of the similarities between the SOM colors and the clustered map colors used in the current prototype.

- *The participants felt that comparing the component planes simultaneously was difficult*. All participants had trouble with handling the windows withing the software.

- *Some participants thought simulating was difficult*. Some participants had trouble with the simulation dialog box and its buttons.

On a more positive note the answers also imply that:

- The participants thought the software was easy to use

- Navigating the menus was no problem

- The participants understood the purpose of the software

- The participants thought using the software to compare companies was easy

- The participants thought the component planes were easy to understand

4.3.2 Post-test interview summary

Following are the participants answers to the questions in the post-test interview conducted at the end of each session.

1. What did you like best/least about the software?

   Participant 1
   Best: The simplicity of the menus, not too many options.
   Worst: The look and surfing between the windows. Some sort of excel windowing system would be great.

   Participant 2
   Best: You can see easily which are the best companies. I like that you can pick how many clusters you want.
   Worst: The labels. In the U-Matrix (the user was referring to the clustered map) you can find the labels but not in the component planes

Participant 3
Best: When you got it, it was pretty easy to use.
Worst: Without a manual the concept was a little bit hard to get.

Participant 4
Best: The program was pretty easy to use. Was able to quickly to learn how it worked.
Worst: It was pretty hard to get an overview of the situation when performing the analysis because of the windowing system. The distortion of the nodes when selecting them.

Participant 5
Best: It is simple. If you select one node in one window it will be selected in the other.

2. What sort of task do you think the software could be used for?

Participant 1: To analyze companies. To compare just about anything that is comparable.
Participant 2: Well I study electronic communication. Maybe you can cluster electronic components by deffisures, differences in voltage, amperes...
Participant 3: To compare different companies against each other and to compare their financial ratios.
Participant 4: For simulations, what happens if a key ratio increases and so on. If you have a log of data you can easily categorize it.
Participant 5: For clustering any data that you want to cluster.

3. Is there any analyzing-tool you would like to add?

Participant 1: Diagrams to better visualize differences between companies.
Participant 2: Some way to label the clusters or quickly get their ID.
Participant 3: Some other way to get the values of the nodes, other than the SOM colors, and some tool to get a summary of a cluster.
Participant 5: The more choices you have the more difficult it is to make a choice.

4. If you could tell the designers to change one thing what would it be?

Participant 1: The visual part, the window, it looks kind of boring at the moment.

Participant 2: For me the most difficult thing was to find which cluster was which.

Participant 3: That it would not crash! (laugh) Perhaps you could make it more graphically appealing and add some functionality.

Participant 4: Mark a selected node with a dot or something else.

Participant 5: Just add a button to show the ID of the clusters.

5. Anything else you would like to add?

Participant 2: If you compare Nenet with this software maybe this was better because Nenet only has one color in the U-Matrix (user was referring to the clustered map). It is the same green all the time (Nenet). Easier to see with the different colors. Preferred Domino over Nenet for analyzing.

Participant 4: Was easy to use and importing labels was easy.

Participant 5: I liked it! We should have used it in the course.

5   Conclusions

The fact that the participants had never used the software before and had relatively little experience with the SOM suggests that Domino faired well overall in the evaluation. A few usability problems and bugs were found but that is only to be expected since it is a prototype. All participants were able to complete all tasks, most times without much help, and among the answers from the post-test questionnaire and interview, some good feedback can be found. There is, however, still room for improving Domino significantly by first tending to the usability problems and bugs found.

One thing that the participants liked about Domino was its simplicity. Since the current version is a prototype there will most likely be some tools and features added before the final release. However, it would be a good idea to, if possible, preserve the simplicity that the participants so liked. This means not adding too many tools and too many features.

Finally it is also the recommended [3; 4] to include a heuristic evaluation in any usability evaluation on coming iterations since this can make a very good addition to other evaluation methods.

References

[1] Barnum, C. M. (2002). *Usability testing and research*. Longman Publishers, NY.

[2] Eklund, T., Back, B., Vanharanta, H., and Visa, A. (2004). *Financial benchmarking in finnish companies - a state of the art survey*. Technical Report 618, Turku Centre for Computer Science, Turku.

[3] Monk, A., Wright, P., Haber, J., and Davenport, L. (1993). *Improving your human-computer interface - A practical technique*. Prentice Hall, New York, NY, USA.

[4] Nielsen, J. (1993). *Usability Engineering*. Academic Press, San Diego, CA, USA.

[5] Nielsen, J. (1994). *Guerrilla HCI: using discount usability engineering to penetrate the intimidation barrier*, pages 245 – 272. Academic Press, Inc., Orlando, FL, USA.

[6] Rubin, J. (1994). *Handbook of Usability Testing*. John Wiley & Sons, INC., USA.

Appendices

A   Evaluation Documents

Introduction to SOM and Domino

What is the self-organizing map (SOM)?

The self-organizing map is a two-layer unsupervised neural network. A neural network is an artificial intelligence tool that attempts to learn patterns and similarities in data by mimicking a simple model of the human nervous system. In unsupervised learning applications, the network is provided with data and is allowed to freely organize itself based upon the patterns and similarities that it identifies in the data. In the case of the SOM, the result of the training process is a two dimensional grid of neurons in which each neuron is "tuned" to a specific combination of attributes, and to which the surrounding neurons display some degree of similarity. The distances between neurons are used to divide the map into a number of clusters, so essentially the SOM is a visual clustering tool. The SOM has been used in thousands of different applications since its invention in the early 1980's, including engineering, process monitoring, medicine, financial and economic analysis, etc.

What is Domino?

Domino is a software tool based upon the SOM. The program is a prototype developed within a software laboratory at the Department of Computer Science at Åbo Akademi on assignment from the Institute for Advanced Management Systems (IAMSR) at Åbo Akademi. The purpose of the program will essentially be to support Business Intelligence managers in companies for tasks that involve financial benchmarking of companies, i.e. the comparison of different parties based upon the use of financial performance measures, such as financial ratios. The primary difference between Domino and existing commercial SOM packages is that Domino includes a number of analysis tools specifically to support benchmarking, features which are often lacking in existing program packages. However, Domino is still a prototype, and therefore, a uniform testing group is required to suggest improvements and identify problems, which is why the usability test will be conducted.

How is Domino used?

As was previously noted, the map is essentially a two-dimensional grid of neurons, in which each neuron is tuned to a specific combination of attributes. The attributes are interpreted using the component planes (see figure). The component planes show the distribution of values across the map. By selecting a specific neuron and comparing its color on each of the corresponding feature planes, the attributes of the neuron, and its associated data, can be identified. Warm colors, such as red and orange, specify high values, whereas cool colors, such as blue, specify low values. In this case, the map illustrates a number of animals grouped according to a list of binary (0 or 1) attributes, such as two-, or four-legged, wings, feathers, etc. The map is clearly divided into bird animals on the right side, and mammals on the left side. In addition, meat eating animals are positioned on the lower half of the map, while herbivores are on the upper half.

A SOM clustered into five clusters
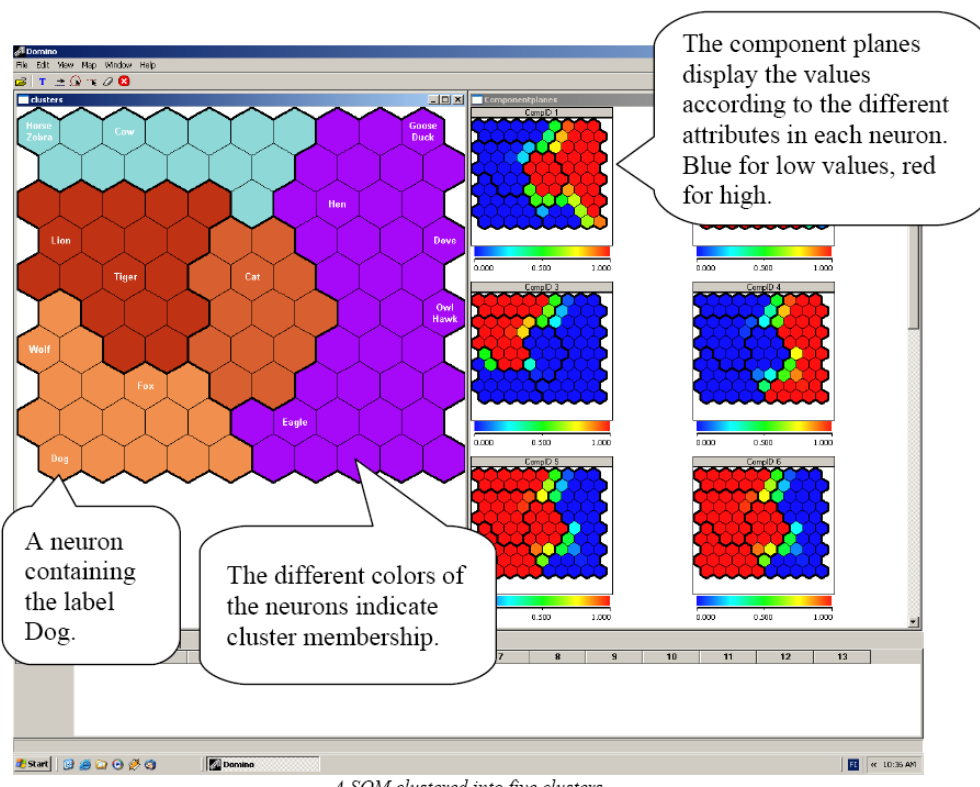
What does the test situation look like?

The software will be tested in a usability lab where the participants' interactions with the software can be monitored and recorded. The only thing that will be recorded is the dialog between the evaluator and the user and the interaction between the user and the software (what is happening on the screen).

What is required of the participant?

The only thing required of the participant is to show up on time for the test having read this document.

Domino Usability Test Screening Questionnaire

General

Email   _____

Age   _____

Gender   Male [ ]   Female [ ]

Education

Faculty   _____

Department   _____

Major / Minor(s)   _____

Total amount of credits (approx.)   _____ sv / sp / ects

Degree(s)   _____

Have passed   [ ] Investment Planning   [ ] Financial Planning

Background

Please rate your experience level with the following things.

	No experience				A lot of experience
Windows operating system	1	2	3	4	5
OS X operating system	1	2	3	4	5
Linux operating system	1	2	3	4	5
Microsoft Word	1	2	3	4	5
Microsoft Excel	1	2	3	4	5
Databases (e.g. MySQL, Oracle, Access)	1	2	3	4	5
Financial benchmarking	1	2	3	4	5
Financial analysis	1	2	3	4	5
Self-organizing maps	1	2	3	4	5

If you have previous work experience with financial analysis, please specify. Use the back.

Test Scenarios

Scenario 1

You have been given an assignment by your boss, who himself flew to Hawaii over the weekend on a business matter and told you that he would not be available for communication until Monday. Today is Sunday and it is time to get the assignment done.

The assignment is to write a report on your findings from creating and analyzing a graphical presentation of five companies' current financial situation, using a software tool called Domino.

There is a shortcut icon to the program on your Windows desktop labeled "Domino". Start the program. The map and all its' related files can be found in c:\evaluation\pnp_map. Load the map. Once the map loads you can close the U-matrix window right away, you will not need it.
.

Scenario 2:

Now that you have loaded the map it is time to get to work on the assignment. The first thing you need for your report is a description of two clusters of the map, when clustered into five clusters.

1. Describe the primary characteristics of cluster 3 and 4 according to the component planes.
2. Of all the clusters, which is best?
3. Of all the clusters, which is poorest?

Scenario 3:

Your boss also wanted the report to include an analysis of the company's top five competitors. To see where these companies are positioned on the map you might want to import some labels.

4. What are the primary differences between Kimberly-Clark and Georgia-Pacific?

Scenario 4:

Before your boss left for his business trip you had a discussion with him where you were speculating what would happen if Georgia-Pacific's profitability ratios (operating margin, return on equity, return on total assets) all were increased by 100%. Your boss told you to further investigate the matter and include your findings in the report. Finish up your assignment by simulating such a change and include the findings in the report.

5. What happens to Georgia-Pacific's position on the map when simulating a 100% increase in all profitability ratios?

Report for Domino analysis

Map clustering

1)

1 = Low, 2 = Medium, 3 = High

	Cluster 3	Cluster 4
Operating Margin		
Return on Equity		
Return on Total Assets		
Equity to Capital		
Quick Ratio		
Interest Coverage		
Receivables Turnover		

2)

3)

Company analysis

4)

1 = Low, 2 = Medium, 3 = High

	Georgia Pacific	Kimberly Clark
Operating Margin		
Return on Equity		
Return on Total Assets		
Equity to Capital		
Quick Ratio		
Interest Coverage		
Receivables Turnover		

Simulation

5)

Domino Usability Test Post-test Questionnaire

	strongly disagree				strongly agree
Using the software was easy	1	2	3	4	5
The concept of the self-organizing map was difficult to understand	1	2	3	4	5
Finding the desired menu choices was easy	1	2	3	4	5
I liked the look-and-feel of the software	1	2	3	4	5
The software has a clear purpose	1	2	3	4	5
The software felt complete	1	2	3	4	5
It was easy to compare the companies using the software tool	1	2	3	4	5
The visual representation of the map was clear	1	2	3	4	5
Clustering the map was hard	1	2	3	4	5
The component planes were clear and easy to understand	1	2	3	4	5
Comparing the component planes simultaneously was difficult	1	2	3	4	5
Importing labels was easy	1	2	3	4	5
Interpreting the map was easy	1	2	3	4	5
Simulating was hard	1	2	3	4	5
I felt that the tasks were difficult to complete using the software	1	2	3	4	5
The tasks would have been easier to complete using tabular data in Excel	1	2	3	4	5

B Evaluations
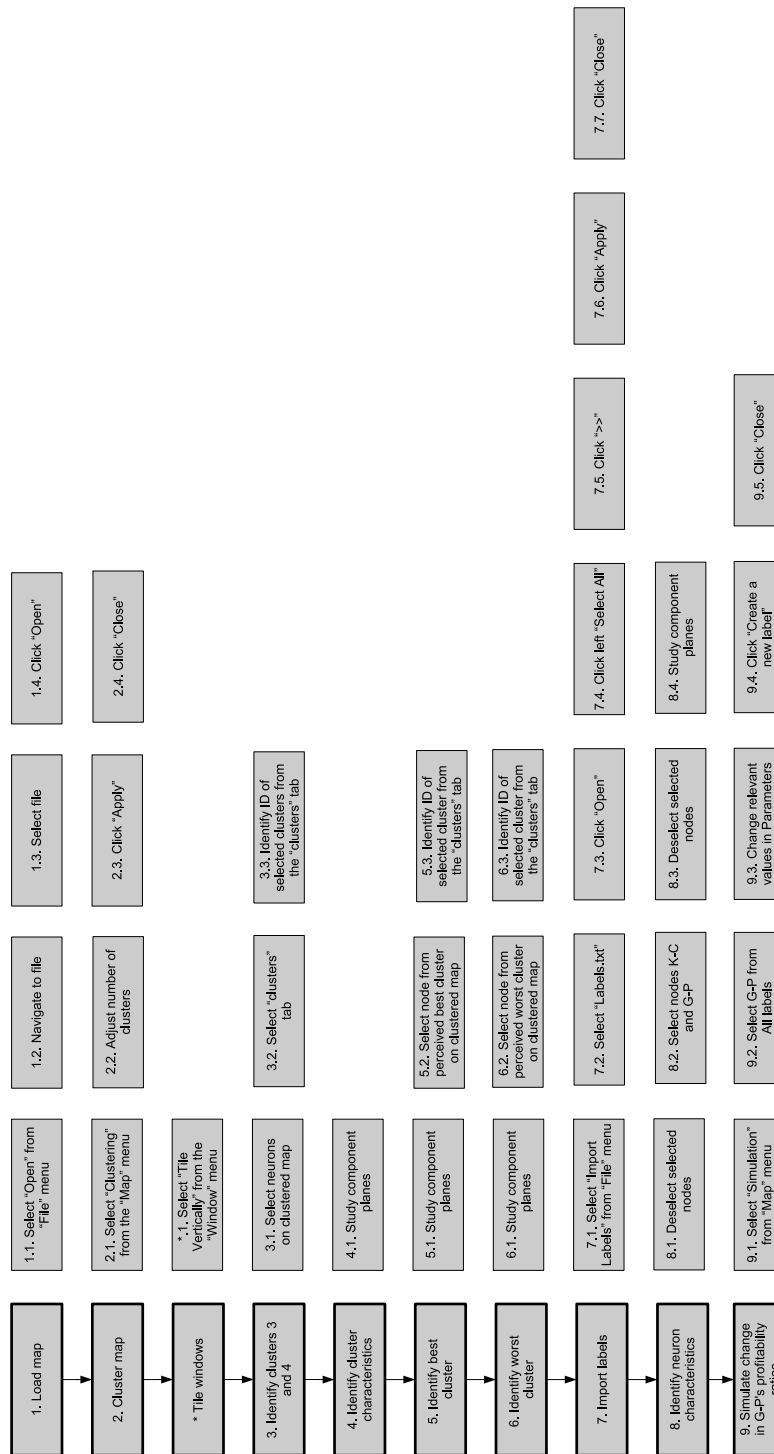
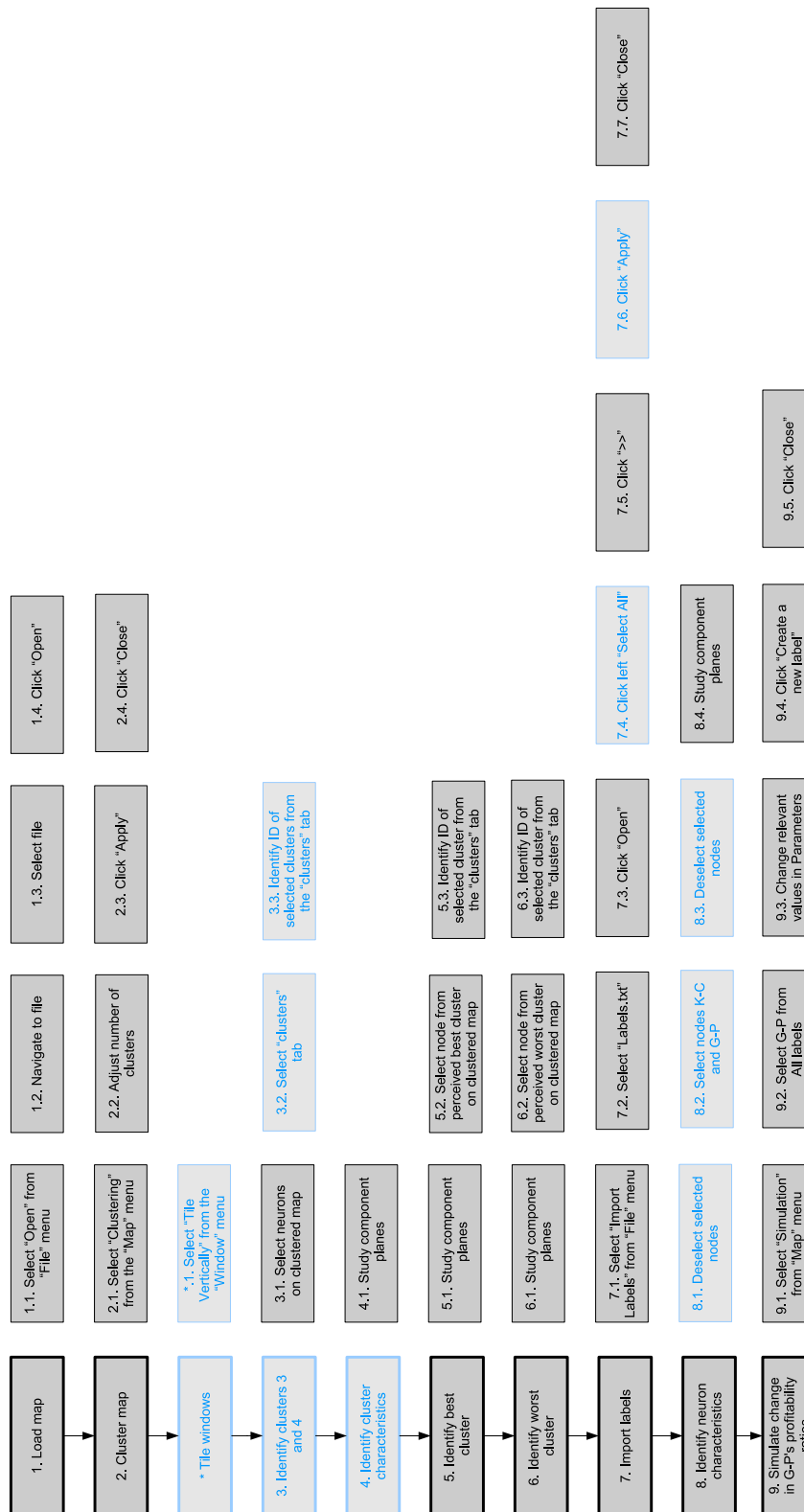Figure I: *Baseline - The evaluation tasks solved by the designer.*

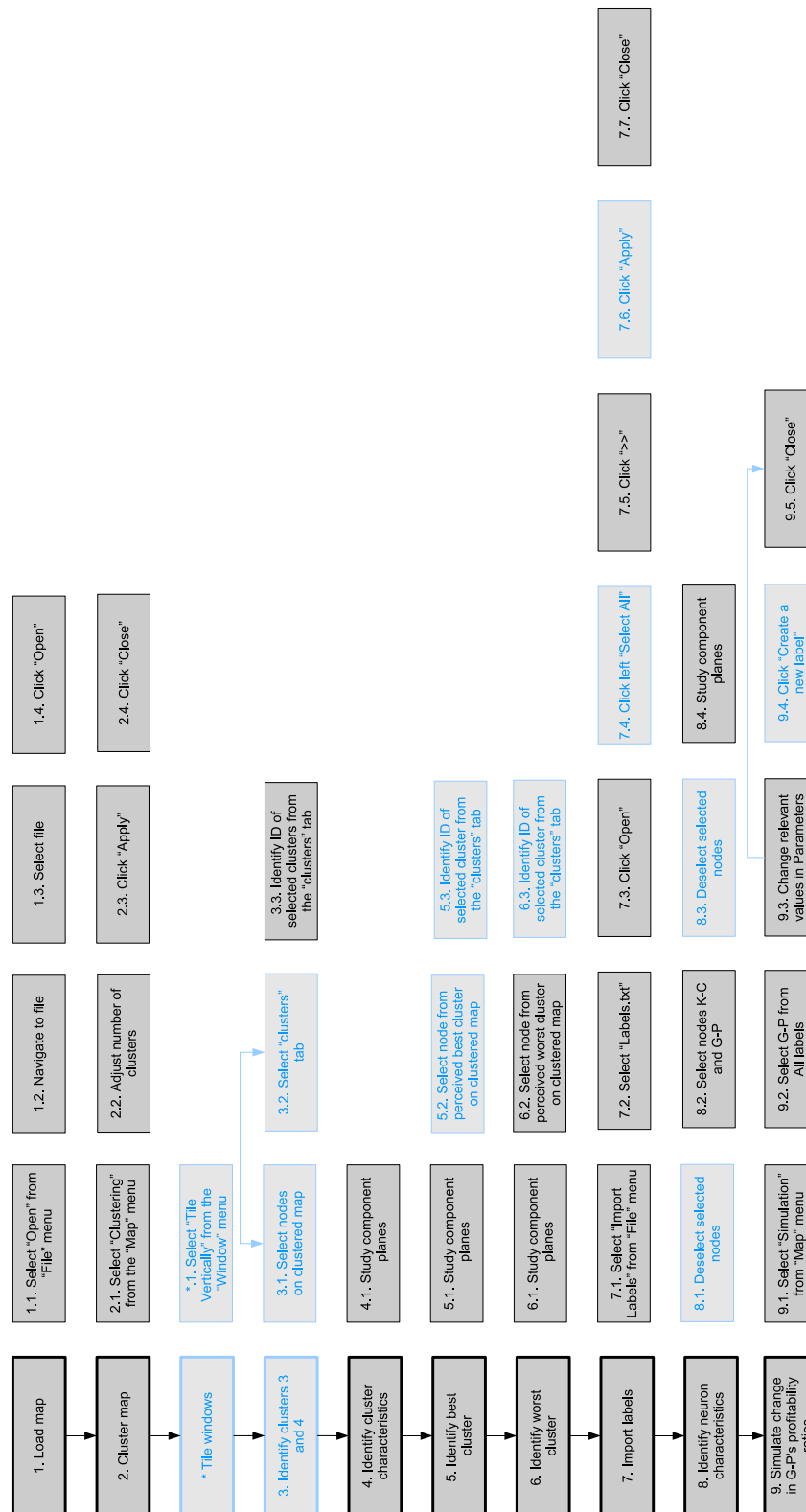Figure II: *Deviations of participant 1.*
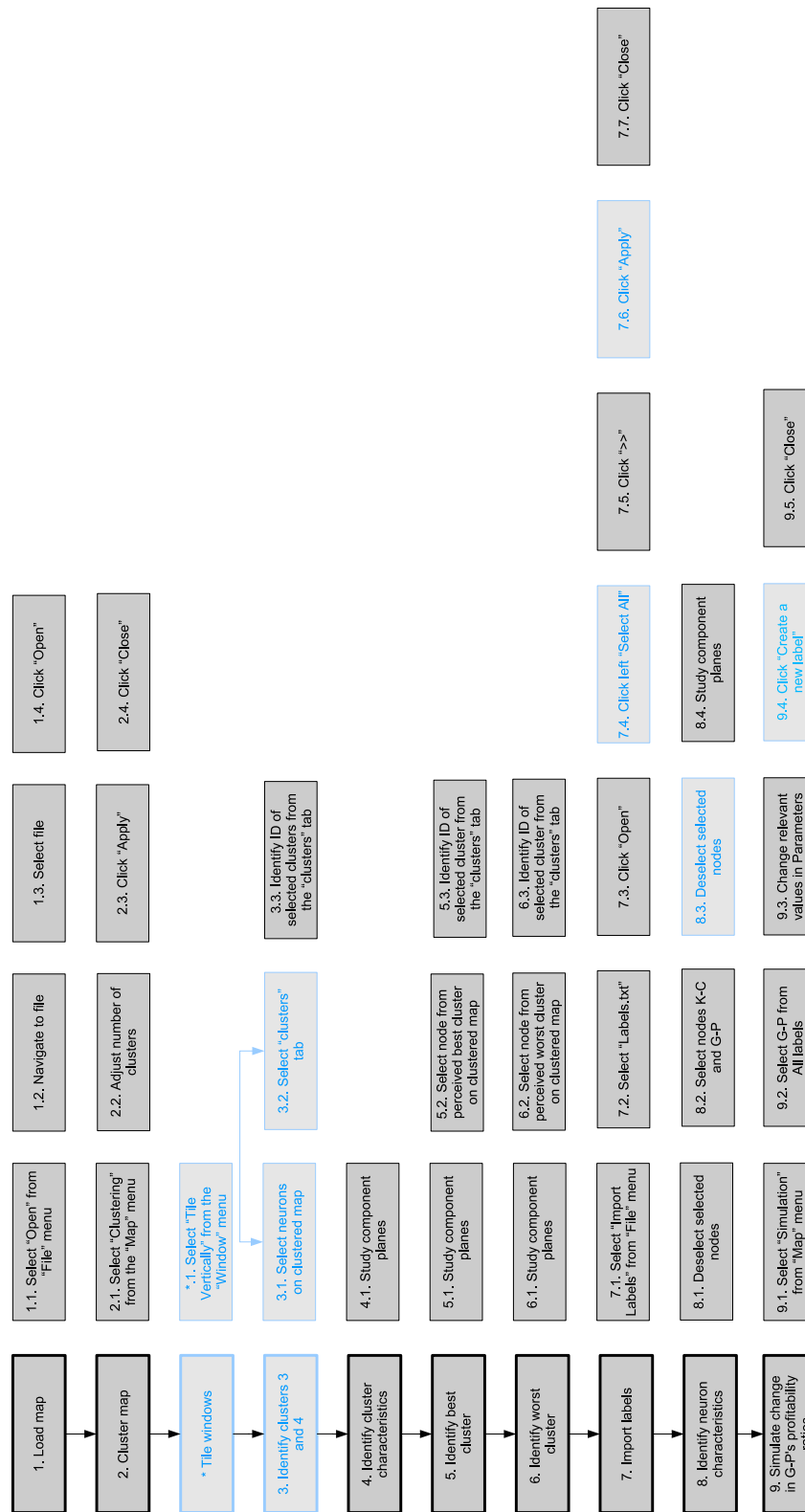
Figure III: *Deviations of participant 2.*

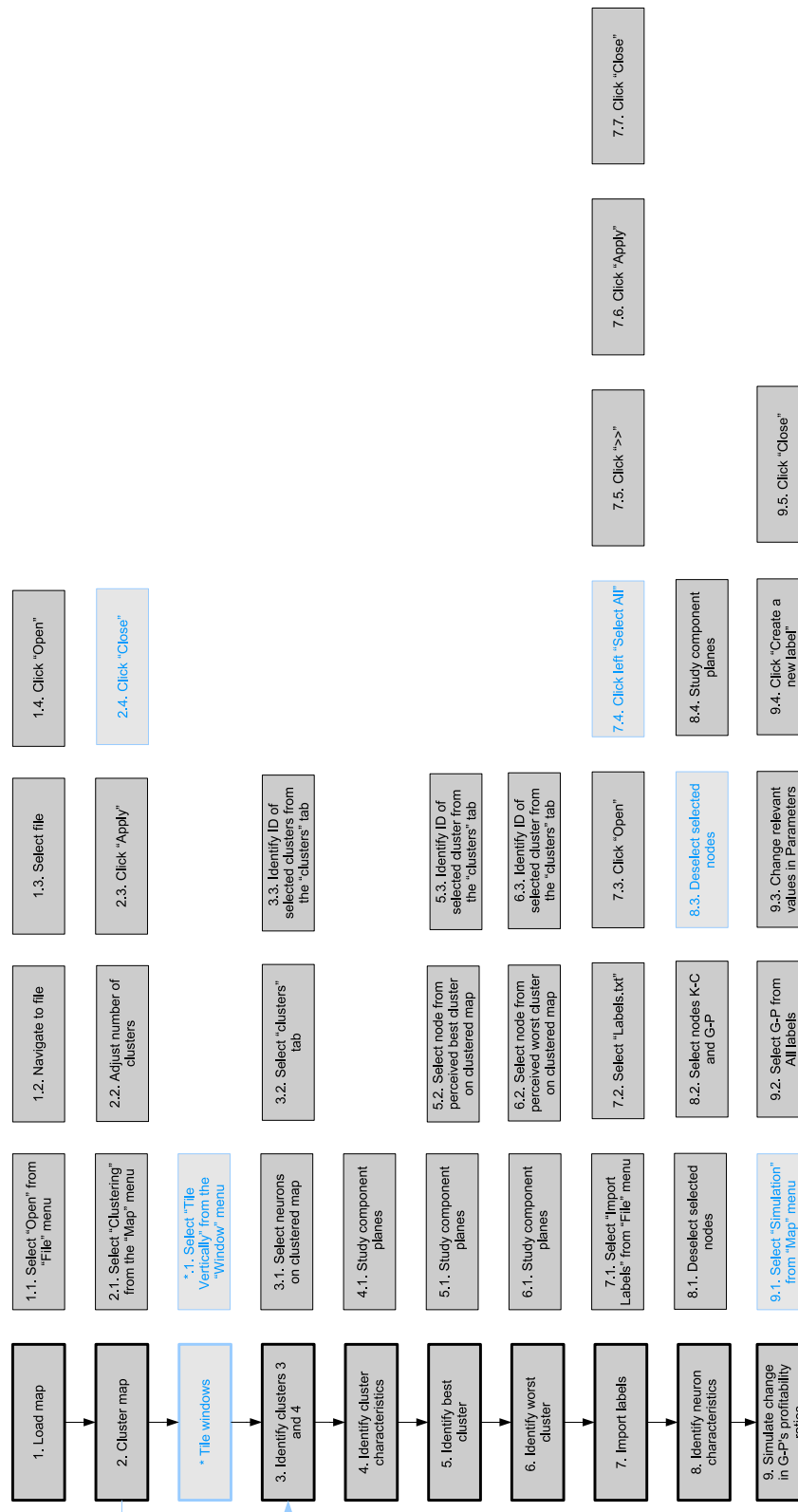Figure IV: *Deviations of participant 3.*
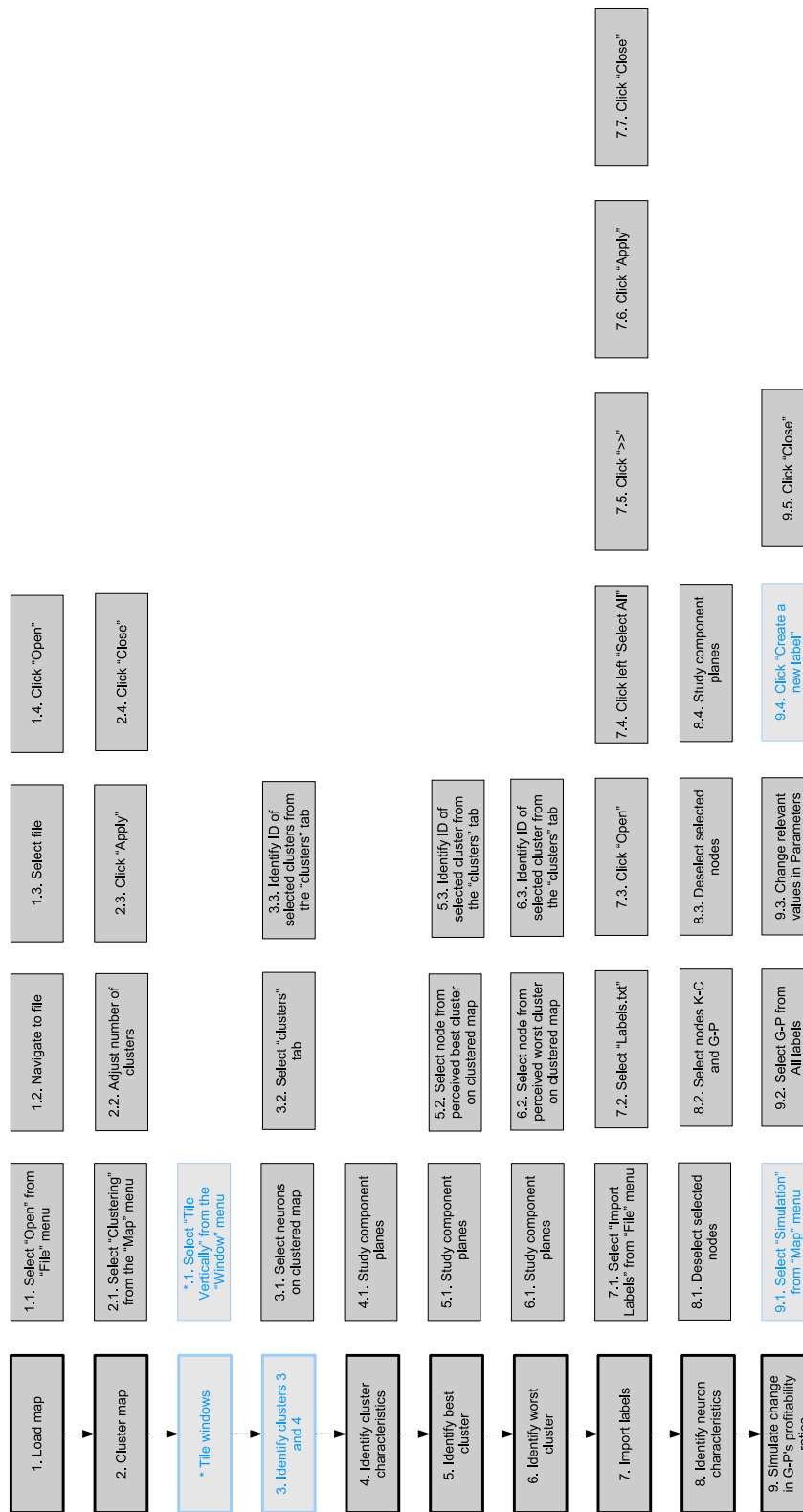
Figure V: *Deviations of participant 4.*

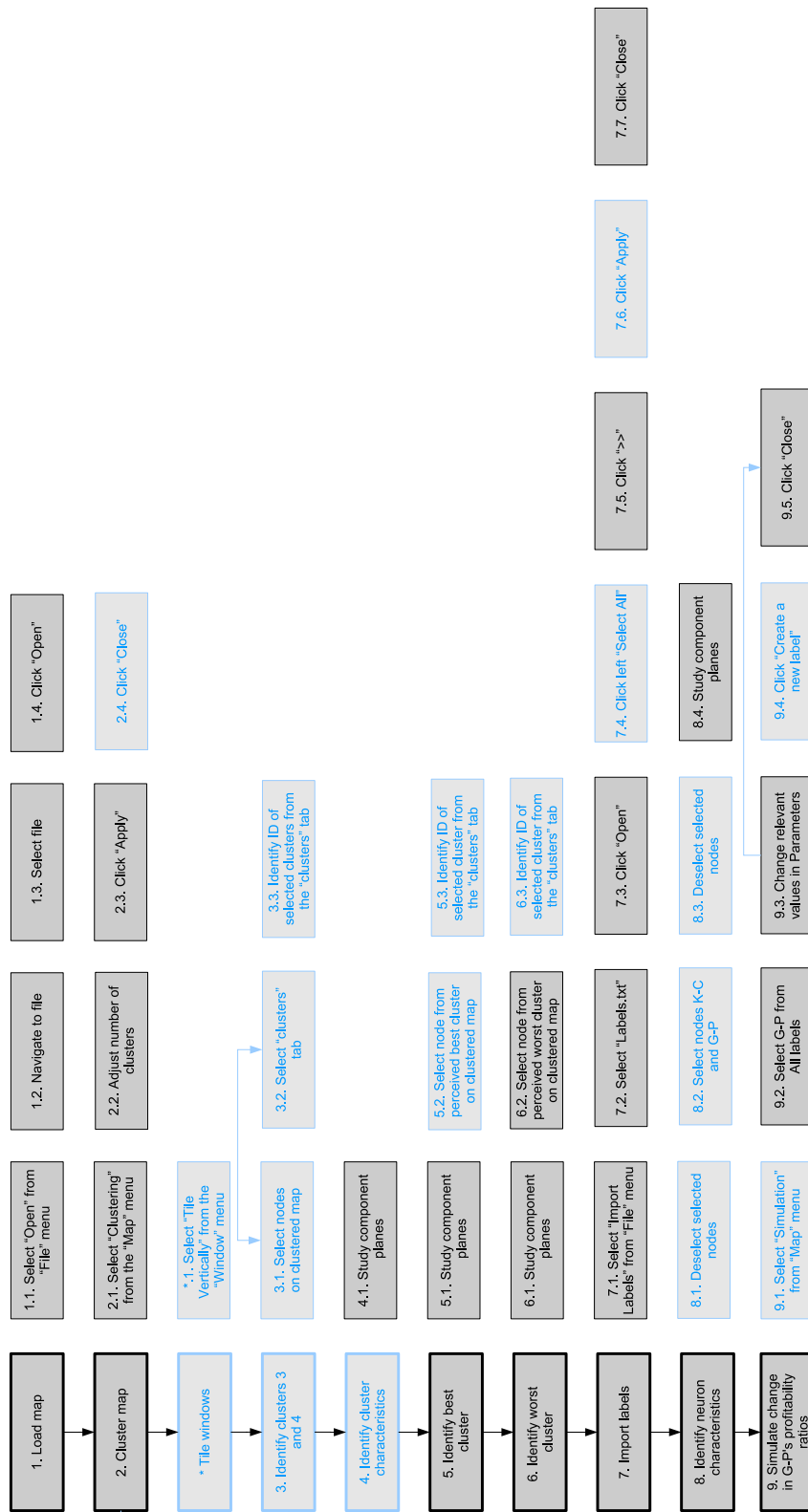Figure VI: *Deviations of participant 5.*

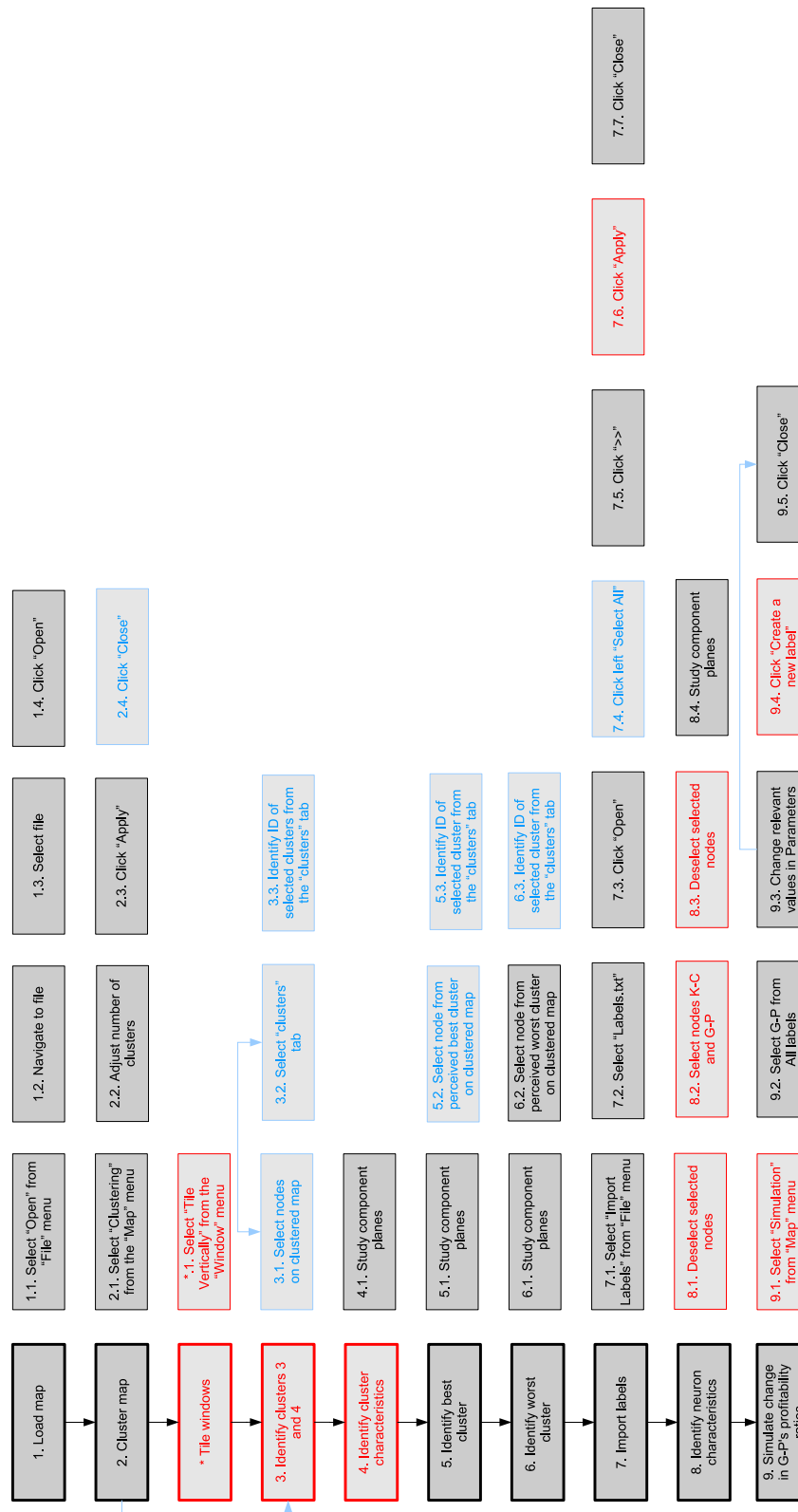Figure VII: *Aggregated deviations of all participant.*

Figure VIII: *Aggregated deviations of all participant with usability problems.*

C Additional Data

Questions	Participant 1	Participant 2	Participant 3	Participant 4	Participant 5
Windows operating system	3	5	5	5	3
OS X operating system	1	3	3	1	1
Linux operating system	2	3	2	3	1
Microsoft Word	4	5	4	5	3
Microsoft Excel	3	5	3	3	2
Databases	2	5	3	3	1
Financial benchmarking	2	3	2	4	1
Financial analysis	2	3	2	3	1
Self-organizing maps	1	2	1	1	1

Table 9: *Screening questionnaire answers.*

Tasks	Baseline	Participant 1	Participant 2	Participant 3	Participant 4	Participant 5
Task 1	03:35	24:53	08:37	18:13	12:43	25:54
Task 2	00:49	02:11	01:01	00:43	01:00	03:55
Task 2	00:11	01:57	11:00	00:41	00:41	01:37
Task 4	02:16	11:12	10:35	07:29	05:55	05:25
Task 5	01:46	04:21	05:00	05:41	07:04	07:47

Table 10: *Task completion times in minutes and seconds.*

Question	Participant 1	Participant 2	Participant 3	Participant 4	Participant 5
Using the software was easy	4	4	5	4	4
The concept of the self-organizing map was difficult to understand	2	1	2	3	4
Finding the desired menu choices was easy	5	5	5	4	3
I liked the look-and-feel of the software	3	4	4	3	5
The software has a clear purpose	3	5	4	5	5
The software felt complete	2	4	3	3	5
It was easy to compare the companies using the software tool	4	5	5	4	5
The visual representation of the map was clear	3	5	4	3	5
Clustering the map was hard	3	1	2	2	1
The component planes were clear and easy to understand	4	4	5	4	5
Comparing the component planes simultaneously was difficult	1	1	2	4	1
Importing labels was easy	4	5	5	5	5
Interpreting the map was easy	4	5	4	4	5
Simulating was hard	1	1	1	4	1
I felt that the tasks were difficult to complete using the software	3	1	1	1	2
The tasks would have been easier to complete using tabular data in Excel	3	1	1	2	1

Table 11: *Post-test questionnaire answers.*

University of Turku

- Department of Information Technology
- Department of Mathematics

Åbo Akademi University

- Department of Computer Science
- Institute for Advanced Management Systems Research

Turku School of Economics and Business Administration

- Institute of Information Systems Sciences