Improving the Operational Picture through Icon Enrichment

Yushan Pan, Monan Yao, Carl-Erik Kopseng

Project background

Our project is under supervision of the ICT department of SINTEF. The objective of our project is to enhance the current icon design implemented in the BRIDGE Master system. The BRIDGE project is an EU-funded social-technical system that supports interoperability in large-scale emergency management. The objective of the Master system is to provide support for decision-making and multi-agency collaboration in large-scale emergency situations. The key is to ensure interoperability, harmonization and cooperation among stakeholders on the technical and organizational level (Bridge, 2010).

In the current Master system, the icons only represent a limited amount of attributes, mainly the category of the icon. In order to find more information about the entity, users need to interact with the icon to open further panels containing additional information. This results in inefficient operation and increases the emergency response time.

To solve this problem, we propose a new icon design that incorporates more attributes into the icon. We started interviewing operational leaders in the police and fire departments. Based on their input and suggestions, the most important attributes were selected and used in our proposed design.

Literature review

For most industrial field there are icon standards, such as ISO graphical symbols (ISO Booklet, 2013). However, not all researchers and designers follow these standards due to the cost of ISO manual (Wang et al 2007). An example of not following a standard is icon classification based on functions and forms (Rogers 1989).

Lidwell describes icons that use images that bear little or no relationship to the action, object, or concept (Lidwell et al 2003). Icon design is often influenced by marketing. Absence of ergonomics during the icon design phase will often result in poor usability (Russo and Moraes, 2003). Kascak describes in a paper on the Patient Monitoring System, that current icons design for UIs are not always designed to reflect the needs, experiences and limitation of the end-user (Kascak et al 2013). In system design, a good user interface should not only have structured layout, but also a reasonable icon design.

Several studies reported that certain icons are difficult to read, metaphors and icons difficult to be interpreted that can make it hard to operate a system (Gargiulo et al 2010).

Usability testing is import for icon design of operational systems (Eisenstein et al 2001). Usability testing should be taken into account when designing the MASTER system's icons. A good icon design is intuitive, easy to use and inexpensive to maintain (Dourish 1997). In order to deliver a good icon design,

human factors should be considered early in the system design process, and user testing should be conducted throughout the whole design process involving participants from the end-user population (Leonardi et al 2008).

Research goal

Our research goal is to present the most relevant information to the end-users in the icons, so that users are not required to needlessly interact with the system to get the information they need.

In order to achieve the research goal, several sub-research goals were introduced.

1. Understand the users' needs in the Master system
2. Understand weaknesses of the old design.
3. Categorize and select the most important attributes an icon should have.

Design process

We chose to use user-centered design for our project. The reason is that we have to know what users really need from this system. Unfortunately the design process was less than ideal. The core problem was recruiting users. The target group of the Master system is operational leaders from the police, fire and ambulance departments, and it proved difficult to get in touch with these users without help from SINTEF. In the beginning we therefore had to make a deal with what resources we could come up with ourselves.

Interview with a police officer

A policewoman was contacted. Consent was signed by both design team and interviewee (Appendix 1). We wanted to learn about how the police handle emergence issues. We also wanted to understand which icons from the old icon set are confusing (see Figure 1).

Figure 1, Testing of previous icon design

Interview findings

Among the main findings from this session was that the system for creating patrol ids results in ids with a high semantic content. For instance, for the Trøndelag district *Sierra-05* is the operational task leader in the command car.

Several issues arose during our interview. The interviewee had troubles finding the correct icons. It seemed that when icons were difficult to distinguish when victims, incidents, and responses were all bundled together. As previously mentioned, we had problems getting hold of users in our target group. Therefore we had to use our personal networks to try and find relevant users. Although we were lucky to get hold of a police officer, a problem with our interviewee was her lack of operational leader experience.

Interview with an instructor at the police academy

After we had firsthand information of weakness of previous icons, we interviewed Bjørn Danielsen, an instructor at the Police Academy. A lot of the training operational leaders in the police receive is given through his instruction, and he has first-hand experience in operational leadership. The aim of the interview was to verify the information we collected in the first interview and collect information from the viewpoint of an operational leader.

At the start of the session the consent form was signed (see Appendix 1). After getting the consent to use audio and pictures we set up basic recording equipment. We then set up the master system to display on the projector.

Interview findings

The interviewee was very much active during our session, and would frequently connect his computer to show us various documents or instructional videos. This resulted in much valuable input, such as ongoing discussions on the use of different vest colors than what is currently used, that a policeman has two IDs, the need for different type of vehicle types to be presented, and so forth (see Appendix 5). Figure 2 presents selected information during the interview.

Figure 2, Examples of selected images during the interview

Based on the information we collected a new set of icons for the police were designed (see appendix 2). We already knew that cooperation with the fire and ambulance departments is required during an emergency situation, but an important finding was that our interviewee had little interest in the details of other departments, with the exception of who the leader is. Therefore we would need to focus on ways of clearly discerning the leaders when improving on the icons.

Interview with the Chief Fire Officer

The interview with Chief Fire Officer Håvard Bakken followed the same steps as the previous interview with regards to setup and consent signing.

Interview findings

As in the case of the interview with the police instructor, we quickly discovered that the fire department also employs rich semantic codes when communicating, but an important difference was that the interviewee did not express any desire to expand the system beyond basic identification. All enhanced icons for fire mission leader can be found in appendix 2.

Further interview were not conducted

After the interview with the Chief of Fire we decided not to interview anyone from the ambulance service, as there was too little time left.

Experiment Design

The issue of testing our finding is a difficult one, as actual users of the system is hard to come by. The issues we have worked with involves working with professionals in the target group to extract what the most essential attributes of each entity are and if a new icon would need to be created. A lot of semantics from the professional domain has therefore been incorporated into the designs. This aspect makes it hard to test the usefulness of the icons without access to users in the target group.

At one point in the process we were advised not to use students for any summative testing, as the results would not be valid for the intended user group. If access to actual users was limited and we only had access to users outside the target group, then formative testing of non-semantic issues, such as contrast and size, could provide more useful feedback. There was internal dissent in the group on whether we could do more formative testing, as this usually belongs in the start of a design phase (Lazar et al, p.260), rather than at the end of the process. As such, we ended up trying to find ways of doing summative testing to better align with the intended project structure outlined by the course.

One thing we found interesting to test was *whether our design principles were easy to remember.* If given an introduction to the guidelines we used to create the icons, would a user remember what information the different colors and shapes intended to convey?

Evaluation Plan

To test whether our design principles were easy to remember we set up two experiments (see sections below for details on each). We ran the experiments over the course of three days, planning to use approximately half an hour in total per participant. This estimate later proved to be somewhat optimistic.

Participants

Our target group should be experts who will use Master systems in their future work. However, this group is difficult to get in touch with due to various reasons. First, recruiting people from this group is time-consuming; we would have to wait a considerable time. The time should also need to fit our entire group member as well. Secondly, recruiting a sufficient number of users might prove difficult. Five users have been claimed to a magic number for usability testing (Virzi, 1992) and that five users will find approximately 80% of usability issues, but other researchers disagree with that statement. Studies have found that five users are not sufficient to discover and identify a majority of usability flaws (Lindgraad and Chattratichart, 2007; Spool and Schroeder, 2001).

After some reconsideration we concluded that we needed to recruit more participants than five. In order to reduce issues of memorization, usability evaluation should not involving the users that already have been interviewed and/or been partaking in co-design (Nielsen, 1993). Moreover, considering we would have to spend a considerable amount of time to run the evaluation with a few professionals (probably less than 3), it is inefficient for usability evaluation (Lindgaard and Chattratichart, 2007; Spool and Schroeder, 2001). The participants are recruited from university students.

Procedure

The experiment includes two phases. The first phase is an introduction to the design guidelines (see Appendix 3) for the proposed icons. The purpose of this introduction is to quickly educate a non-experienced participant, so that the participant would be able to interpret the meaning of the proposed icons. The introduction lasts 5-10 minutes and only one participant is involved each session. During this time, the participant learns about the coded meanings for each shape, color and visual attributes. It should also be noted that the introduction does not explicitly present all the icons and details in design. Instead, only principles and patterns of the design are shown to the participant. Figure 3 illustrates the design principles of resources icons which are able to represent up to 7 attributes. The participant is shown how to read the attributes from colors, codes and outline. In addition a few example icons are also shown to the participants.

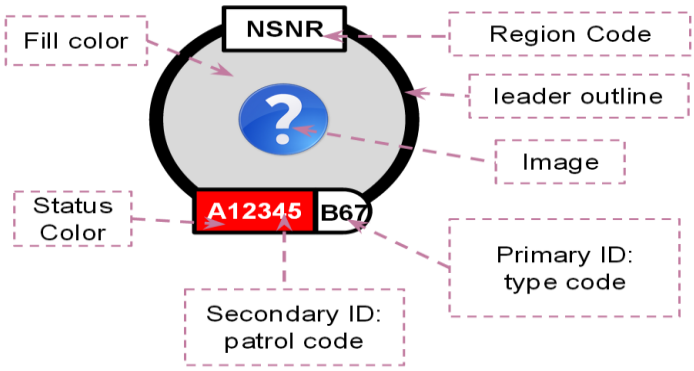

Figure 3, Design principles of resource icons

Immediately after the introduction, the participant is invited to the second phase of the experiment: the testing phase. The testing phase includes two tasks.

Task 1: List as many attributes as you can find for each icon

In the first experiment a subject is presented with sheet of six icons, where the experimenter is asking the subject to list as many attributes as can be found for one icon at a time. When the subject seems unable to list any more attributes, the process continues with the next icon until all icons have been covered, at which point the process continues with the next set of icons. There were 5 sets of icons in total, grouped by category. The icons were drawn from the categories that have proved to be the most important during the interviews of our professional users. These icons also happen to be the most information rich in our icon set.

The first task emphasizes on how well a participant interprets the icons in an idealized situation. The icons are presented somewhat enlarged and listed horizontally on a line. The purpose of such setting is to make it as easy as possible for the participants to observe the icons. Several sets of icons from different categories are shown to the participant. The participant is expected to list as many attributes as possible for each individual icon. Based on the given description of attributes, the participant will receive a score for each icon. For example, if one icon has 4 attributes and the participant lists 3 correct attributes, the participant receives 3 points for the icon. If the main icon category is wrong, such as saying a police car is a civilian car, this result in a score of 0. The score is given by the experimenter.

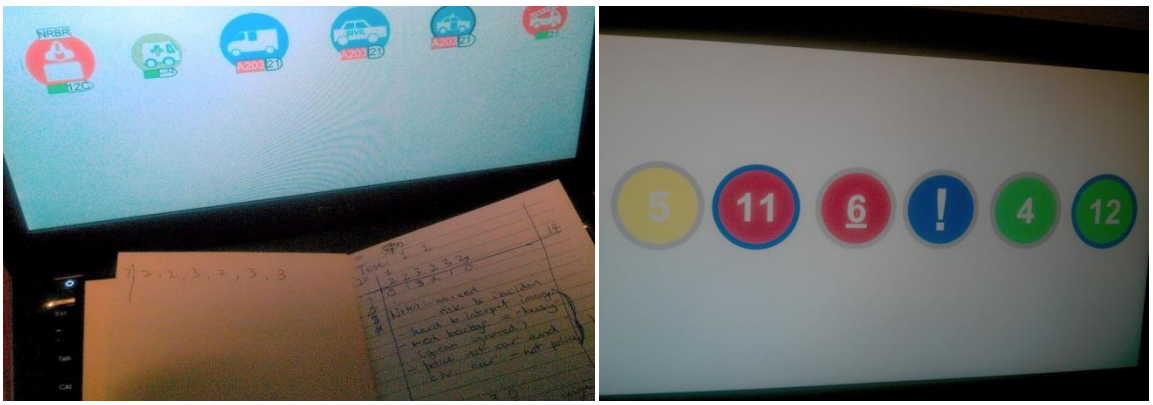

Figure 4, Task 1: Describe as many attributes as possible

Purpose of the test

We would like to see if people can recognize the icons and the meaning of their attributes.

Task 2: Answer questions based on two created scenarios

Task 2 is a within-group experiment, consisting of two scenarios. In the first scenario icons are presented on a map, whereas in the second scenario the icons are presented on a blank white background. For each of the scenarios, the participant is asked 11 questions that revolve around finding the number of icons matching the description in the question. One such question was "how many people are vitally injured and requires immediate medical attention?" The participant should discover all the victim icons that are filled with red color. The icons were picked from the set of icons presented in task 1.

Designing icons without regard to the context they are used in can be misleading due to the fact that readability can be severely affected by the background the icon is laying on. We therefore wanted to assess whether our icons are affected by this.

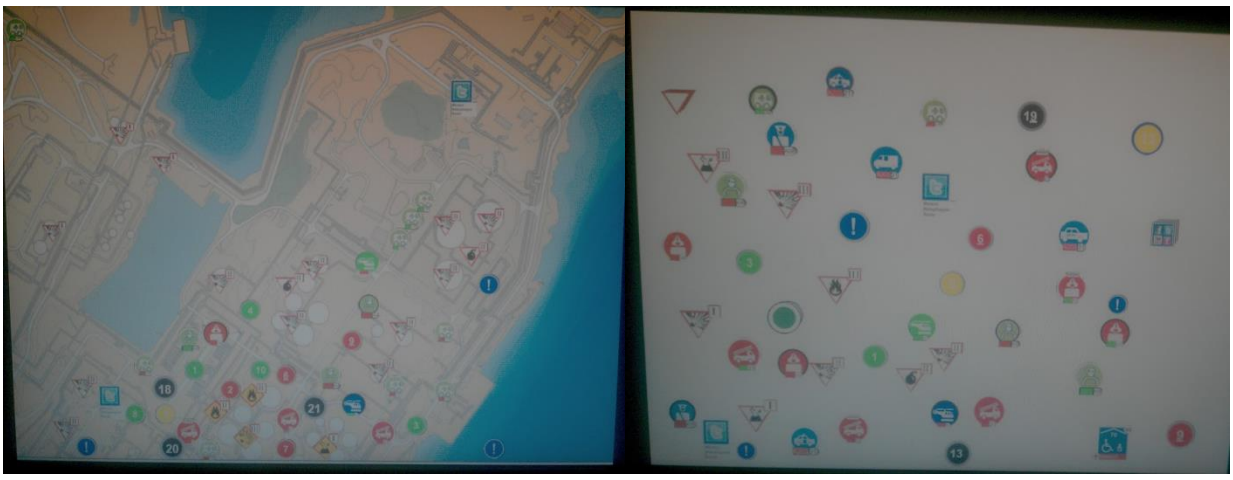

Figure 5: Two scenarios in the task 2

Variables

In our second task our independent variable is whether the icons are shown with a map background. The dependent variable is the number of icons found.

Hypotheses Formulation

H_0: There is no difference between the number of icons found with a map background and the number of icons found with a blank background.

Experiment results

Data analysis

The sample size has not been estimated, so we treat all data as valid, because we do not know whether our data set is too small. As the first step of analyzing the data, several graphical statistics are presented to visualize the data with the purpose of illustrating data distribution. A score of 1 or more means the respondent got at least the main icon type right.

Data visualization of results from task 1

Task 1-1

- Icon 1: The maximum number of attributes is 2, the median value is 1, the mode value is 2, and the variance is 0.77
- Icon 2: The maximum number of attributes is 2, the median value is 2, the mode value is 2, and the variance is 0.68
- Icon 3: The maximum number of attributes is 3, the median value is 2, the mode value is 3, and the variance is 1.04
- Icon 4: The maximum number of attributes is 2, the median value is 2, the mode value is 2, and the variance is 0.22
- Icon 5: The maximum number of attributes is 3, the median value is 2, the mode value is 3, and the variance is 0.79
- Icon 6: The maximum number of attributes is 3, the median value is 1, the mode value is 0, and the variance is 1.40

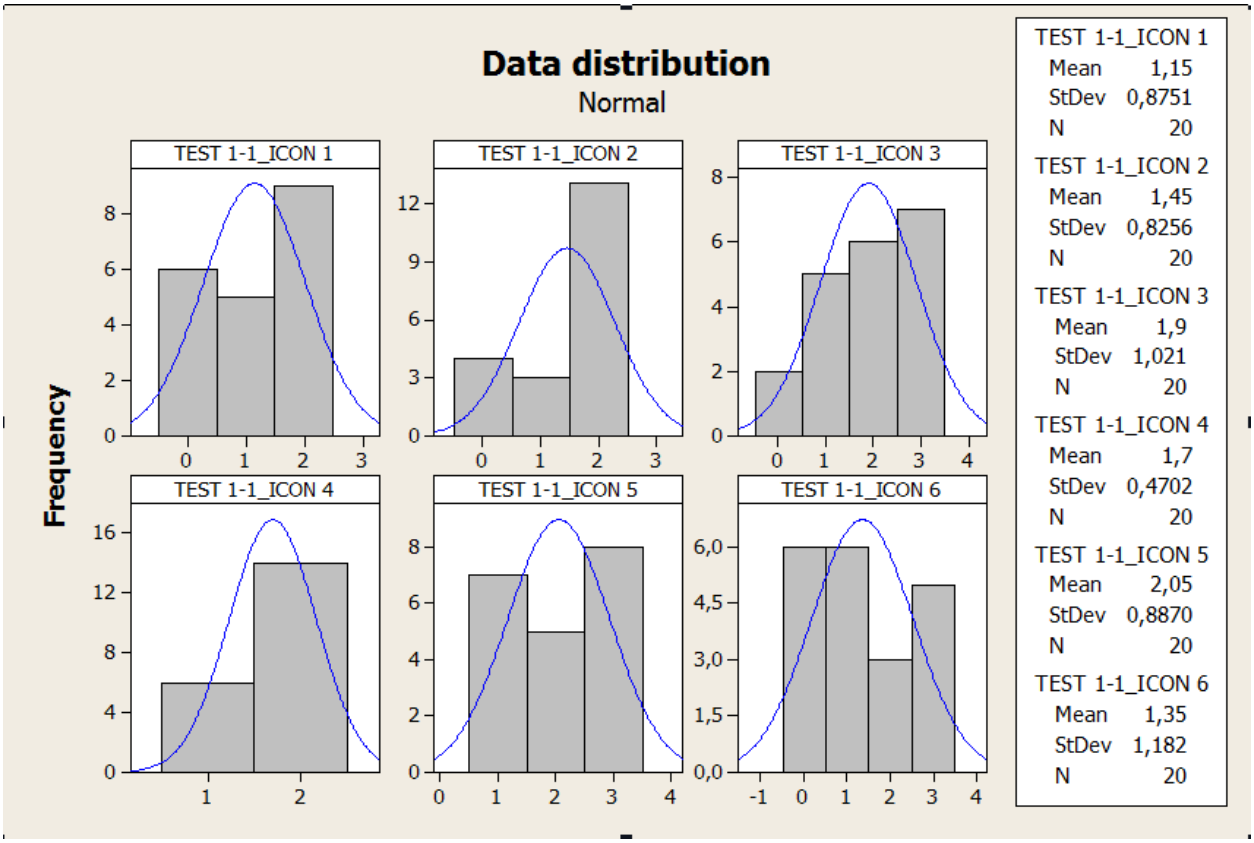

Figure 6, Score distribution for the test 1-1 icons

Task 1-2

- Icon 1: The maximum number of attributes is 4, the median value is 2, the mode value is 2, and the variance is 1.73
- Icon 2: The maximum number of attributes is 3, the median value is 3, the mode value is 3, and the variance is 0.30
- Icon 3: The maximum number of attributes is 3, the median value is 2, the mode value is 2, and the variance is 1.00
- Icon 4: The maximum number of attributes is 3, the median value is 2, the mode value is 3, and the variance is 1.62
- Icon 5: The maximum number of attributes is 4, the median value is 4, the mode value is 4, and the variance is 1.46
- Icon 6: The maximum number of attributes is 3, the median value is 2, the mode value is 3, and the variance is 1.31

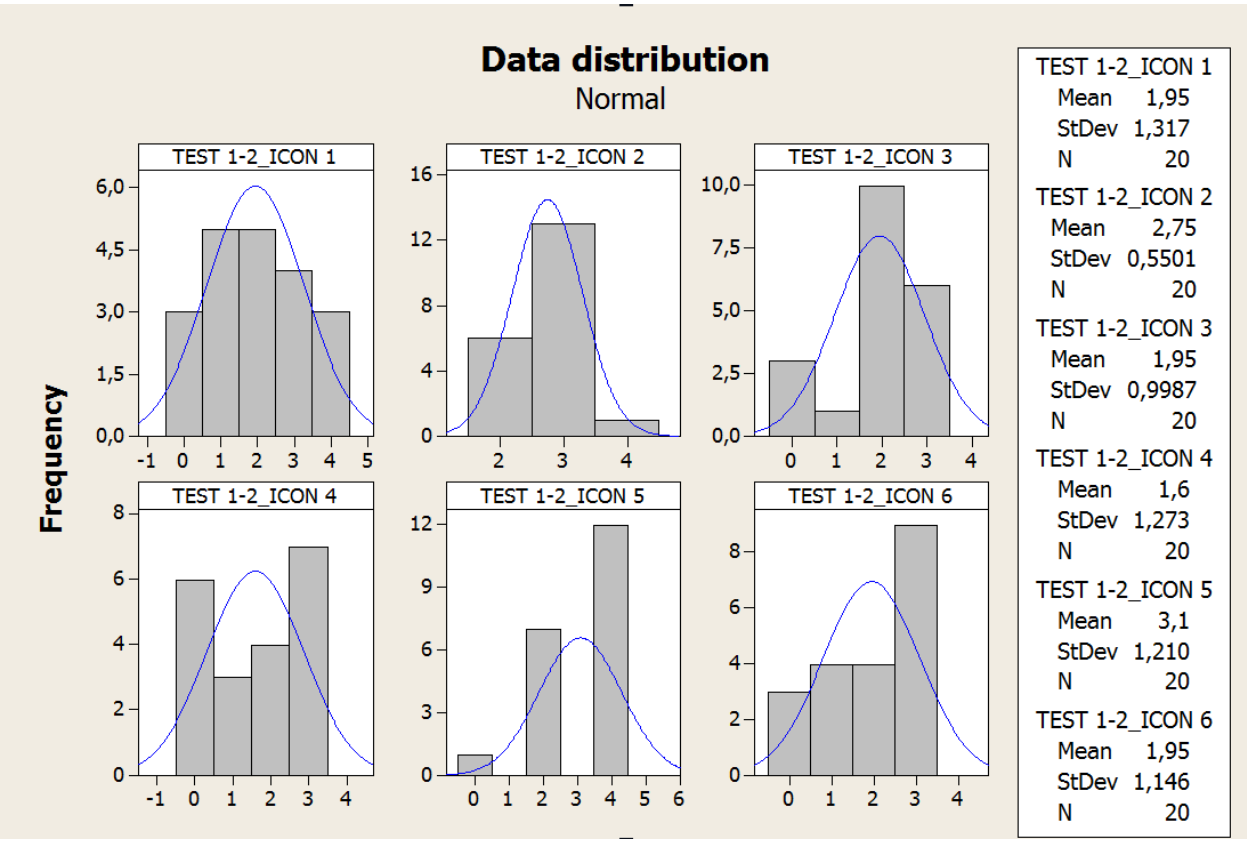

Figure 7, Score distribution for the test 1-2 icons

Task 1-3

- Icon 1: The maximum number of attributes is 3, the median value is 3, the mode value is 3, and the variance is 1.21
- Icon 2: The maximum number of attributes is 4, the median value is 3, the mode value is 4, and the variance is 1.00
- Icon 3: The maximum number of attributes is 4, the median value is 2,the mode value is 2, and the variance is 2.04
- Icon 4: The maximum number of attributes is 3 ,the median value is 2,the mode value is 2, and the variance is 0.26
- Icon 5: The maximum number of attributes is 4, the median value is 2.5 ,the mode value is 2, and the variance is 1.84
- Icon 6: The maximum number of attributes is 3, the median value is 2,the mode value is 2, and the variance is 0.68

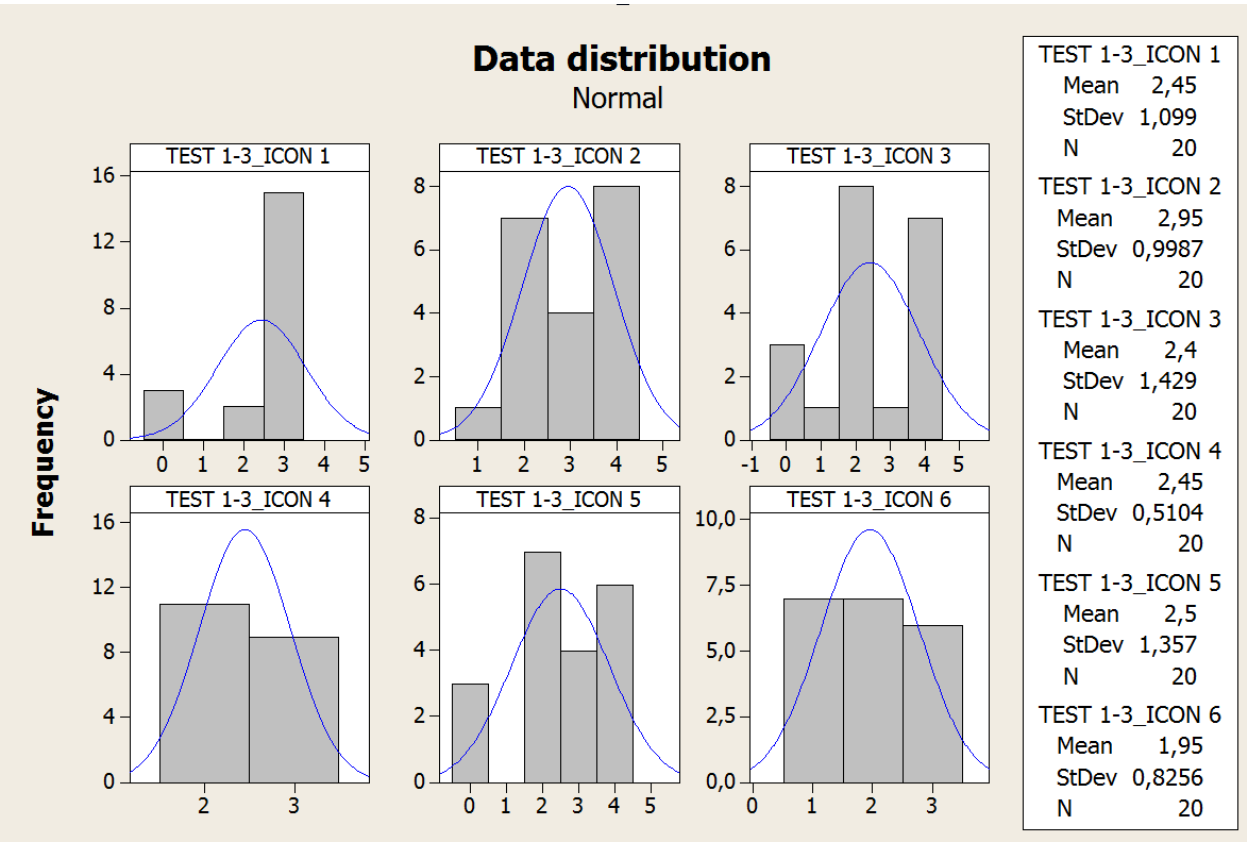

Figure 8, Score distribution for the test 1-3 icons

Task 1-4

- Icon 1: The maximum number of attributes is 2, the median value is 2,the mode value is 2, and the variance is 0.99
- Icon 2: The maximum number of attributes is 3, the median value is 2,the mode value is 1, and the variance is 1.12
- Icon 3: The maximum number of attributes is 2, the median value is 2,the mode value is 2, and the variance is 1.20
- Icon 4: The maximum number of attributes is 1,the median value is 1,the mode value is 1, and the variance is 0.58
- Icon 5: The maximum number of attributes is 2, the median value is 2,the mode value is 2, and the variance is 1.20
- Icon 6: The maximum number of attributes is 3, the median value is 2,the mode value is 2, and the variance is 0.75

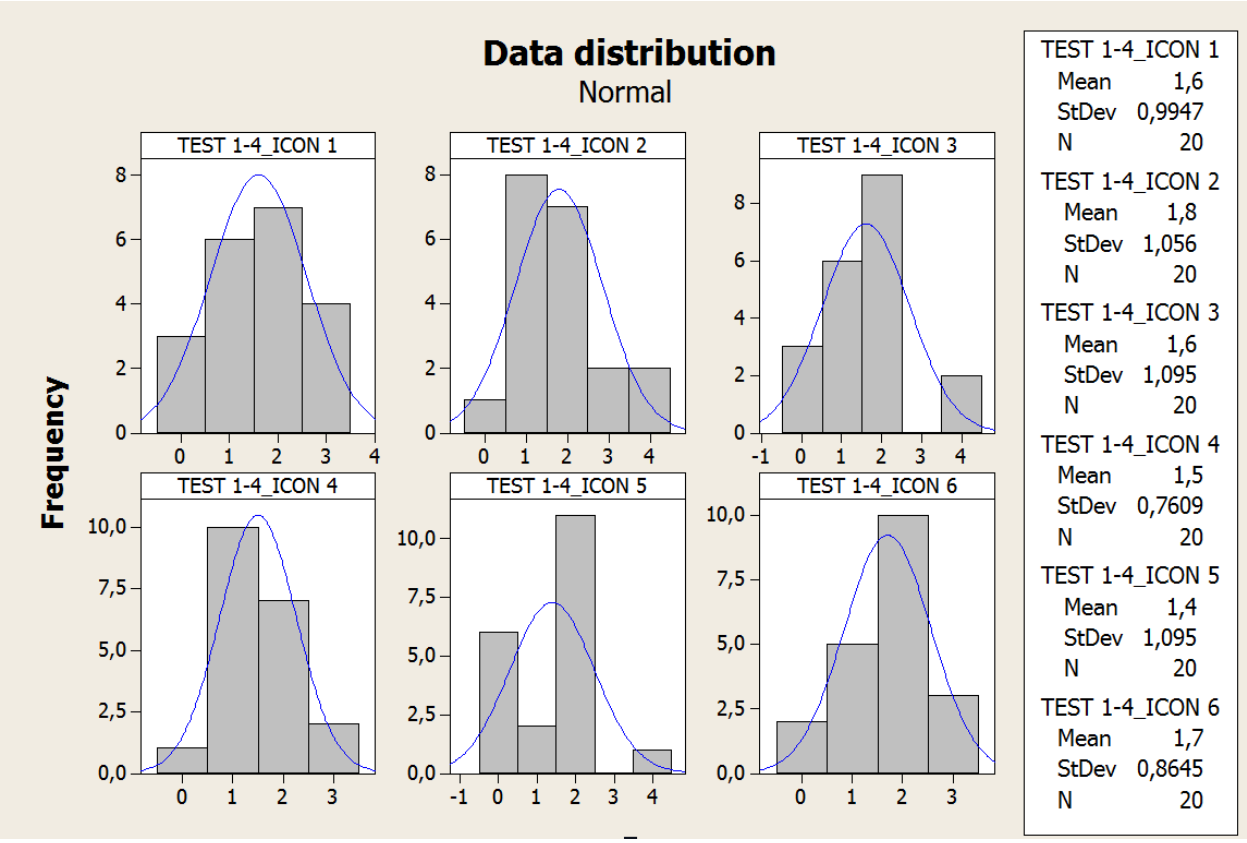

Figure 9, Score distribution for the test 1-4 icon

Task 1-5

- Icon 1: The maximum number of attributes is 2, the median value is 2,the mode value is 2, and the variance is 0.80
- Icon 2: The maximum number of attributes is 2, the median value is 2,the mode value is 2, and the variance is 0.82
- Icon 3: The maximum number of attributes is 2, the median value is 2,the mode value is 2, and the variance is 0.58
- Icon 4: The maximum number of attributes is 2, the median value is 2,the mode value is 2, and the variance is 0.60
- Icon 5: The maximum number of attributes is 1, the median value is 1 ,the mode value is 1, and the variance is 0.32

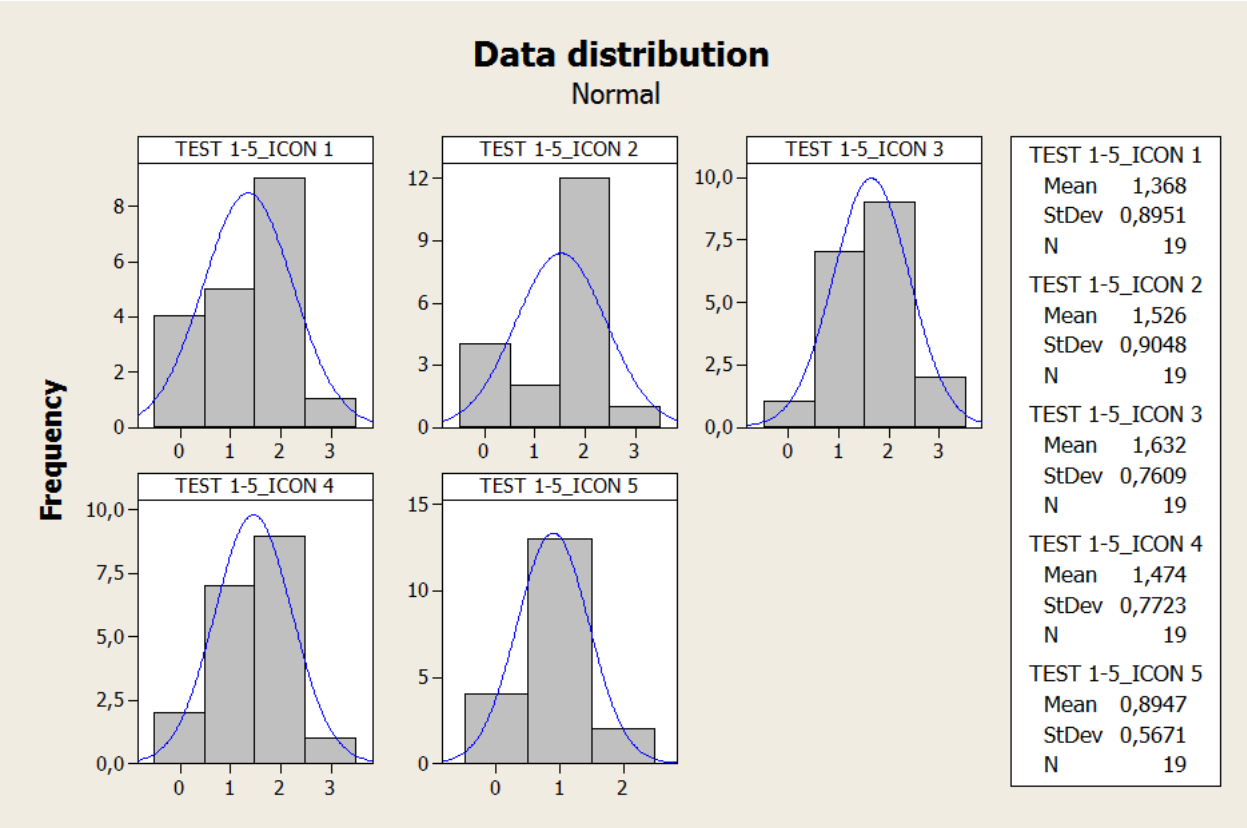

Figure 10, Score distribution for the test 1-5 icon

Data visualization of results from task 2

Significant differences can be found where the users are asked to find leaders (p = .001), aggregated icons (p = .000), explosion and bomb risks (p = .000), fire incidents (p = .000), and chemical incidents (p = .001). Hypothesis is rejected. There are differences between the number of icons found with a map background and the number of icons found with a blank background.

Figures 11 and 12 present the data distribution. We are specifically interested in the icons where there is a significant difference between the two scenarios.

Numbers of accounted icons with map background

- ***Find the number of leaders:*** the median value is 3, the mode value is 3, and the variance value is 3.04
- ***Find the number of aggregated icons:*** the median value is 3, the mode value is 3, and the variance value is 1.20
- ***Find the number of explosion and bomb risks:*** the median value is 5, the mode value is 5, and the variance value is 3.33.
- ***Find the number of fire incidents:*** the median value is 0, the mode value is 0, and the variance value is 0.68

- ***Find the number of chemical incidents:*** the median value is 0, the mode value is 0, and the variance value is 0.47

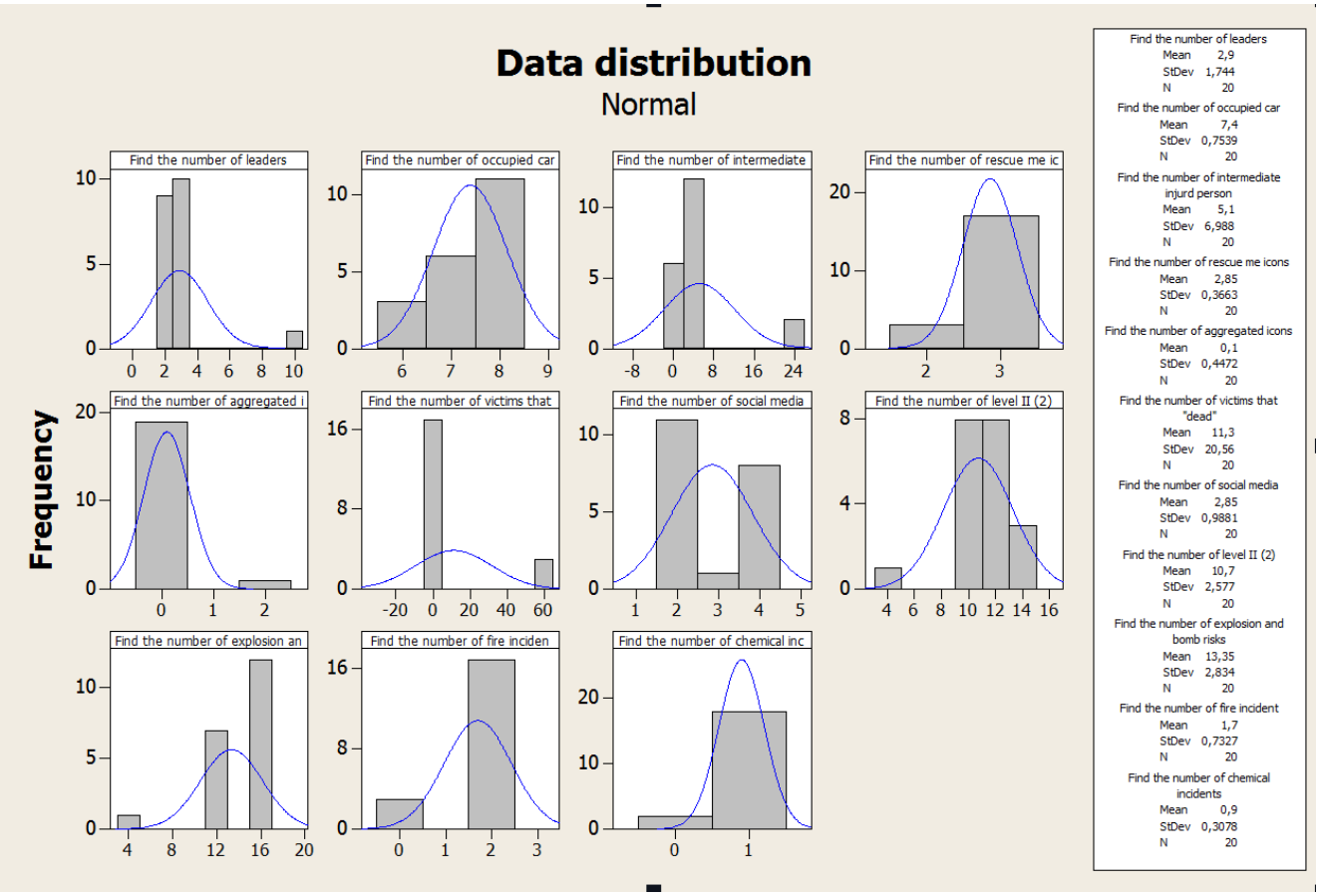

Figure 11, Numbers of accounted icons with blank background

- ***Find the number of leaders:*** the median value is 6, the mode value is 6, and the variance value is 2.21
- ***Find the number of aggregated icons:*** the median value is 0, the mode value is 0, and the variance value is 0.20
- ***Find the number of explosion and bomb risks:*** the median value is 15, the mode value is 15, and the variance value is 8.03.
- ***Find the number of fire incidents:*** the median value is 2, the mode value is 2, and the variance value is 0.73
- ***Find the number of chemical incidents:*** the median value is 1, the mode value is 1, and the variance value is 0.31

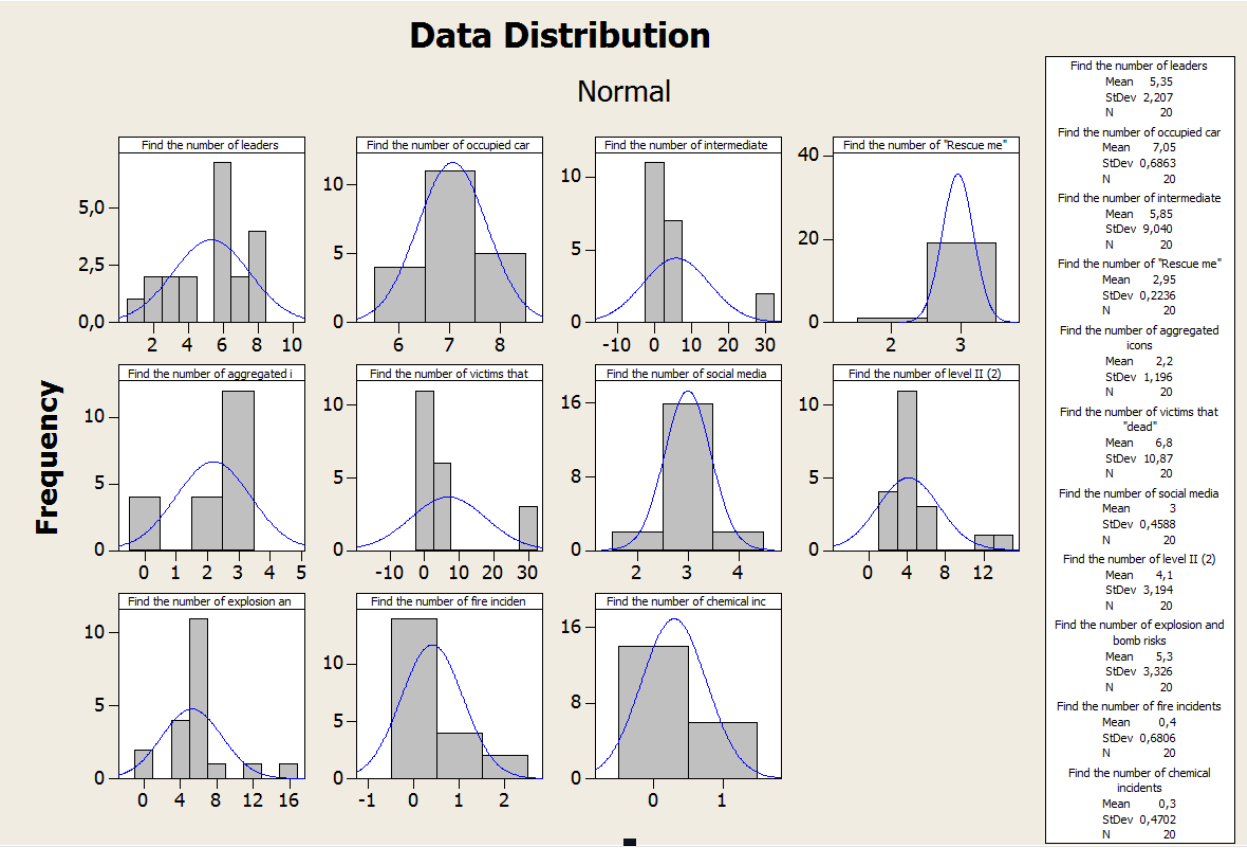

Figure 12, Numbers of accounted icons with map background

Findings and discussion

In numeric terms this means if a user gets at least one point, the main icon category has been recognized.

In test 1-1, 30% got the first icon wrong (car accident risk). We believe this is due to people mixing up the risk and incident categories, because from icon two most people had a consistent score of at least 1 point. About 30% also got the last icon wrong (explosion incident). We are unsure on why this happened, but we observed that most people expressed that they did not understand the symbol.

In test 1-2, which was mainly concerned with resource icons, most people scored well, but the icon for civilian police car had a large error rate. Some subjects expressed that they associated blue with information signs, and the text "SIVIL" might have contributed to further confuse them. On a positive note, 60% found all the 4 possible attributes of the police car with leader function.

In test 1-3, almost everyone had a score of at least one on icons 2,4 and 6. The error rate for the police helicopter and police man were the highest, although not alarming at 15%.

In test 1-4, somewhat surprisingly, the icon for a victim with minor injury had an error rate of 30%. When performing the test some respondents expressed that they were sure that the grey had some kind of meaning, and they were therefore trying to put meaning into elements that had none. This can indicate

that our design choice of adding thick grey outline to the victims' category might have been somewhat haphazardly chosen. The intention of the outline was originally meant merely to increase readability, but our results suggest that we should have followed the style of the other icons in having just a thin black border.

In test 1-5, 20% got the aggregated incident and aggregated victims' icon wrong. For the incident icon the cause was often the same as in test 1-5: they stated that it was a risk icon. For the aggregated victims icons we believe that the victims' category has an issue with "overloading" the same info segment. By that we mean that the same representation (the number shown in the icon) is used for both showing the patient number and as the number of people in that category when used in aggregation. That effect can also be spotted in Task 2, where some respondents report wildly inaccurate numbers due to summing id's instead of counting the individual icons.

In task 2, we found significant differences when examining the results for the questions concerning leaders, aggregation, explosion and bomb risks, fire incidents and chemical incidents, but *this was not necessarily due to the map background.* Almost everyone got the chemical incident accident right when using a map background, but 35% got it wrong when there was a blank background. This was probably caused by the phenomena we observed earlier; that people mix up the risks and incidents category, and since the blank background had *no incident icons*, this might have caused the high error rate. The same applies for the high error rate of the fire incident icon and the aggregated icons. See the section on *threats to validity* for further discussion. For the leader icons most people can find the correct number, but there is a large off-by-one error in both scenarios. We believe this is due to the helicopter icon with leader function, since the black outline has a very low contrast against the dark blue and is hard to see unless you know it is there. A lighter blue might prove to have better performance.

Threats to validity

Wohlin presents 4 categories for experiment validity, as discussed below.

Conclusion validity

It concerns the relationship between the technique used and the outcome in terms of scores for tests. One important question is whether the sample size is big enough to justify the conclusions drawn. The main effect claimed about understanding was a significant advantage for new designed icons.

Denoting the Type I error probability by α and the Type II error probability by β, the following relationship holds:

$$N = \frac{4(\mu_{\frac{\partial}{2}} + \mu_{\beta})}{ES^2}$$

If we use α = 0.05 and β = 0.20, we get $N = \frac{32}{ES^2}$ (Claes Wohin 2000) as a required sample size. We got a large Cohen's d = .053 (large effect size). The effect size is 1.23 yields N = 22 would be needed to observe the efficiency in the experiment. 2 additional participants are needed for an ideally experiment results. In an addition we do not have sufficient observations in all cases but we note that these effect sizes must be used with caution, because the data are not in general normally distributed.

Internal validity

Internal validity assesses whether the observed outcomes were due to the treatment or to other factors. Because of a within group method used in the experiment, the outcome might have been impacted by task. But in order to avoid the threat of selection bias, two different tasks were conducted in the same group. The potential problems of learning effects cannot be avoided because in task 1 the design guideline was introduced to the participants due to the problem of participants' selection (discussed above). Due to the briefness of the testing we ignore possible effects of boredom and fatigue.

Construct validity

Construct validity concerns whether it is legitimate to infer from the measures made in the experiment to the theoretical constructs that one was trying to observe. Some icon types were only present in one of the scenarios. Since the users were being asked to find them in both scenarios, this exaggerated the statistical differences. This is a design flaw of our test.

External validity

External validity is concerned with the question of whether it is possible to generalize from the experimental setting to other situations. The use of students instead of practitioners is a notable threat as we discussed above. The participants had learnt about how we redesign the icons through the design guideline.

Concluding remarks

Most of the design faults in Task 2 could be remedied by moving to a between-group design and having the set of icons. Further work should focus on the problem points noted in the finding sections, such as the grey outline of the victims' icons, consistent color use throughout a category, and careful consideration to contrast and readability when picking colors.

References

1. Bridge at a Glance, http://www.bridgeproject.eu/en/about-bridge, Access date: 07.11.2013
2. Virizi, R. (1992) Refining the best phase of usability evaluation: How many subjects is enough? Human factors, 34(4): 457-468
3. Nielsen, J. Usability Engineering Elsevier Inc., 1993
4. Davis, F.D., Bagozzi, R.P., Warshaw, P.R., "User Acceptance of Computer Technology: A Comparison of Two Theoretical Models", Management Science, 35, 982-1003, 1989
5. Wang, H., Hung, S., and Liao, C. A Survey of Icon Taxonomy Used in the Interface Design. Proceedings of the ECCE 2007 Conference, 28-31 August 2007, London, UK
6. Lidwell, W., Holden, K., and Butler, J. (2003). Universal Principles of Design. Massachusetts. Rockport Publishers
7. Beatriza R., Moraes A. The Lack of Usability in Design Icons An Affective Case Study About Juicy Salif. DPPI'13, JUNE 23-26, 2003, Pittsburgh, Pennsylvania, USA
8. Kascak, L., Rebola, C., Braunstein, R., and Sanford, J. A. Icon Design for User Interface of Remote Patient Monitoring Mobile Devices. SIGDOC'13, September 30- October 1, 2013, Greenville, North Carolina, USA
9. Gargiulo, G., Bifulco, P., Cesarelli, M., Jin, C., McEwan, A. and van Schaik, A. Wearable dry sensors with Bluetooth connection for use in remote patient monitoring systems. Studies In Health Technology and Informatics, 161, 2010, 57-65
10. Eisenstein, J., Vanderdonckt, J., and Puerta, A. Applying model-based techniques to the development of UIs for mobile computers. In Proceedings of the 6th international conference on Intelligent user interfaces, New Mexico, USA 2010.
11. Dourish, P. Accounting for system behaviour: Representation, reflection and resourceful action. Computers and design in context, MIT press, Cambridge, MA, USA, 1997, 145-170
12. Leonardi, C., Mennecozzi, C., Not, E., Pianesi, F. and Zancanaro, M. Designing a familiar technology for elderly people. Gerontechnology, 7, 2, 2008, 151.
13. Spool, J. and Schroeder, W. (2001) Testing web sites: Five users is nowhere enough. Proceedings of the ACM Conference on Human Factors in Computing Systems, 285-286
14. Claes Wohin, P. R. Experimentation in software engineering: An introduction. 2000.
15. ISO Graphical Symbols Booklet, http://www.iso.org/iso/graphical-symbols_booklet.pdf, accessed on Nov 1 2013
16. Lazar, J., Feng, J., Hochheiser, H. Research Methods in Human-Computer Interaction. Wiley, Glagow, 2010