Usability Evaluation of an Ecological Interface for Process Control Health Monitoring

Connor Upton[1], Gavin Doherty[1] & Myra O'Regan[2]
[1]Department of Computer Science, [2]Department of Statistics,
Trinity College Dublin, Dublin 2, Ireland
{connor.upton@cs.tcd.ie, gavin.doherty@cs.tcd.ie, moregan@tcd.ie}

Abstract. Generating graphical displays for complex systems is a difficult task. The Ecological Interface Design framework provides guidelines for work domain analysis and visual design; however the visual design principles are quite abstract and sometimes difficult to apply. In a companion article, we propose a visual design methodology that extends ecological interface design and provides a structured approach to generating graphical forms. In this paper, an experiment compares an existing chart to a new design generated through application of the methodology. The aim is to determine whether the redesign affects the usability of the chart and to explore possible causes for these differences. The results suggest that the new design provides better support in terms of efficiency, accuracy and satisfaction for a range of key tasks. A discussion investigates general issues relating to the validation of ecological interface designs.

1. Introduction

Advancement in automation technology has increased the importance of graphic displays for monitoring complex systems. While improvements in charting applications reduce the technical difficulties associated with generating dynamic displays, understanding what information to display and what visual format to display it in remain difficult tasks. The Ecological Interface Design (EID) framework provides methods of work domain analysis and visual design guidelines for control displays in complex systems [1][2]. While the analysis phase of the framework provides distinct artifacts for revealing what information to include, the design guidelines are quite abstract making it difficult to know what visual format is most suitable.

In a companion paper [3] a methodology is proposed that builds on the EID framework. This methodology combines work domain analysis with task analysis to reveal not only the information requirements, but also how they are used. This additional knowledge can inform a structured approach to the visual design of graphic forms based on data scale analysis, data transformations and visual scale matching. One of the key advantages of this approach is that it can be used with non-physical process systems where EID has traditionally been difficult to apply. A case study applied this methodology to a health monitoring system used in the semiconductor manufacturing industry, resulting in a new ecological design [3]. A control task

analysis suggests that the new display should provide better support for many of the cognitive tasks carried out by process engineers during monitoring. Here a usability evaluation aims to reveal performance differences between the original and the redesigned display.

1.1 Scope of the Study

It has been noted that evaluation of ecological designs can be problematic [4]. The variability of real world scenarios is difficult to simulate in a laboratory environment and in many cases the EID approach can radically change usage models, hindering comparative analysis techniques. In this case, an evaluation has been carried out on a portion of the redesign, namely the On-Target Indicator (OTI) chart. There were a number of reasons for this. Firstly, the proposed methodology guides the design of individual interface components that make up the overall ecological display. As such, it is appropriate to test the usability of these outputs. Secondly, the original OTI chart and the new design use the same underlying data and share well-defined, measurable tasks which allow a comparative study to be carried out. Finally, a detailed study of a single graphic form permits us to explore the differences that can be attributed to visual presentation. This level of exploration cannot be carried out with an integrated multidisplay interface as it becomes difficult to differentiate between the effects of the different graphic forms.

1.2 The Displays

The original OTI chart (fig. 1) takes the form of a modified control chart, with parameters on the horizontal axis, values on the vertical axis and tools (machines) encoded by way of icons. Control charts are widely used in industrial settings and play an important role in statistical process control. During a task analysis, a number of problems were noted with the original design (see [3]). One of the main issues was that the display allowed the key users (process engineers) to identify problems with particular parameters, but did not provide adequate support for diagnosis of these problems. It also did not support the identification of specific tool performance, another desirable feature. This chart was originally selected from a range of templates provided by a charting application.

One of the key ideas behind EID is to embed a model of the work domain within the visual design of the display. This externalised system model supports the user when dealing with unanticipated events. A work domain analysis of the on-target indicator chart revealed that the information it displayed related to two perspectives of the work domain; the functionality of the monitoring system and the physical organisation of the equipment. While the original design highlighted the former quality, the visual encoding of the tools made specific equipment issues difficult to discern. The redesigned OTI chart (fig. 2) is an ecological display that captures both perspectives, providing equal support for off-target parameter and tool detection and diagnosis of equipment issues. Through the proposed visual design methodology, data transformations were carried out. This reduced the quantitative data associated with the sensor readings to a set of ordinal ranges. Following this visual scale matching was carried out to generate a design space of potential solutions from which the

redesign was chosen. This experiment studies whether the data transformations and subsequent scale matching have had an impact on the usability of the chart.

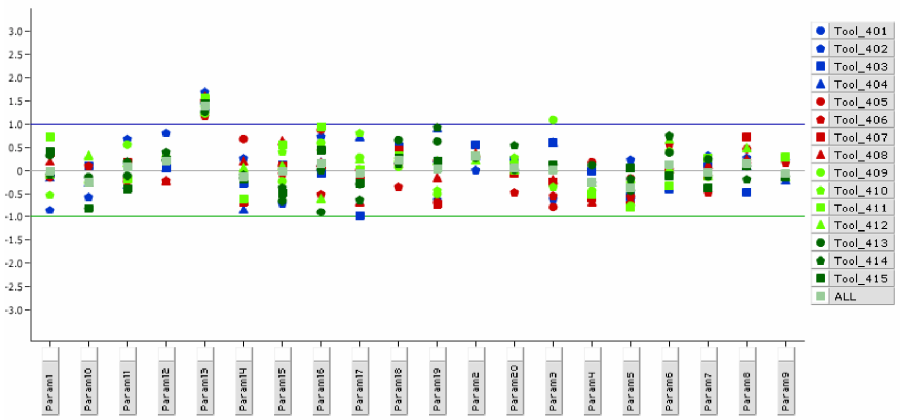

Figure 1. The original OTI chart (Chart A)

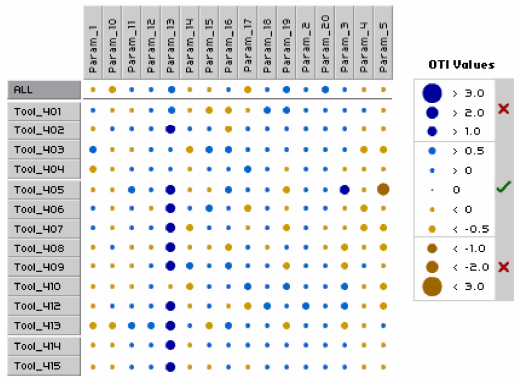

Figure 2. The redesigned OTI chart (Chart B)

2. Method

2.1 Participants

A total of 20 participants, 14 males and 6 females, took part in the study. Their ages ranged from 22 to 40 years of age. 10 were postgraduate students from the computer science department of Trinity College Dublin and 10 were industry employees. None were considered to be domain experts as they had no knowledge of the process control health monitoring or the displays involved, however all were experienced computer users. Despite lacking domain expertise this group was considered suitable due to the perceptual nature of the experiment. The participants carried out the study during regular working hours but were not compensated in any other way for their time. Access to expert users was difficult, however the experiment was repeated on a much smaller group of four process engineers providing anecdotal evidence presented in the discussion section of this paper.

2.2 Experimental Platform

2.2.1 System Data

During the interface design process a number of OTI charts, using real system data, were studied to identify key features indicating abnormal behavior. Twenty mock datasets were generated with specific features encoded in each. Each dataset involved 300 sensors, consisting of 20 parameters on 15 tools. Process engineers validated these data sets as being representative of the scale and complexity involved in real-world monitoring.

2.2.2 Interfaces

Two displays were studied in the experiment. The original OTI chart takes the form of a modified control chart. This was labeled chart A in the experiment. The redesigned OTI chart is a more ecological display incorporating the sensor values and their equal relationship to the physical system (tools) and the functionality of the monitoring system (parameters). This was labeled chart B in the experiment.

2.2.3 Materials

A custom web application was developed using Macromedia Flash software and a MySQL database to carry out the experiment. This application both presented the information to the participants and logged their performance. The study was carried out on desktop computers running Windows XP. The graphics were presented on 17" LCD Monitors with a 32bit colour setting.

2.3 Tasks

Four primary tasks were selected from the range of activities associated with the OTI charts.

1. *Select off-target sensors.* This involves identifying individual sensor readings that lie outside of the control limits. These need to be brought back into control to keep the process stable.

2. *Select off-target tools.* This involves identifying tools that contain sensors that lie outside of the control limits. A task analysis [3] showed that users often need to see the performance of a specific tool based on information from outside sources (e.g. machine technicians). While the action here reverses this process it provides a good indication of whether the relationship between tools and sensors is made explicit in the display.
3. *Select parameters that are off-target but matched.* This involves (a) identifying parameter sensors (labeled ALL) that lie outside of control limits and then (b) identifying whether this parameter is matched. Matched parameters exhibit tight clustering of their tool readings. Unmatched parameters have a highlighted label. An off-target but matched parameter indicates that its control-limit were set incorrectly and need to be adjusted.
4. *Select tools with three or more off-target sensors.* This state indicates a "dog" tool, one that exhibits erratic behavior. This involves identifying individual sensor readings on the same tool that lie outside of the control limits. This tool must be taken down for maintenance.

In each case, the participant was required to identify features relating to their task by selecting the appropriate interface elements i.e. sensor icons, tool labels, parameter labels. A chart can contain from 0 up to 3 features. Once all features are selected a submit button must be pressed to mark completion of the task.

2.4 Design

This is a within-subject design. The four tasks were presented in a random order and the chart type order was alternated for each user and for each task. The tasks were repeated four times for each chart (8 in total) to capture the four different number of features (0-3). The order of the number of features was also randomized for each user. An increased number of features is thought to increase the complexity of the task. As a result, some interaction between the independent factors was expected. Separate models were used for measuring efficiency, accuracy and satisfaction and the analyses were carried out separately for each task.

2.5 Performance Measures

Efficiency relates to the amount of time taken to complete a task. This is measured as the time between the initial presentation of a chart and the selection of the submit button once the task is complete. Accuracy relates to the number of errors incurred. An error is the incorrect selection of an interface element or failing to select an element that corresponds to a feature. Satisfaction is a subjective judgment of the displays. Once the participant had completed the task with both displays they were required to select which one provided better support or if they were equal. All of these performance measures were recorded by the application during the experiment.

2.6 Training & Supplementary Materials

Each participant was presented with a short animation giving an overview of the work domain, the tasks and the chart types, including interaction techniques for each chart. Following this, they registered their name and were presented with the tasks in a random order. Each task was preceded by a description accompanied by two animated

demonstrations of how to complete the task with either chart. At this stage the participants were asked to explain the task and their interaction strategies. If correct they were allowed to proceed, if not they were asked to re-read the instructions and were tested again to see whether they fully understood the task. The original design was labeled Chart A and the redesigned ecological display Chart B.

2.7 Hypotheses

Task 1, select off-target sensors, involves detecting ordinal differences between objects i.e. is a sensor greater or less than the control limit. Based on the basic tasks model of graphic efficacy [5] chart A, the original design, should give better performance results as it encodes the sensor values and control limits using position along a common scale. This encoding is shown to be the best for quantitative perceptual tasks.

Task 2, select off-target tools, involves detecting ordinal differences between objects, then identifying nominal relationships between objects. While chart B may prove slower for the initial ordinal task, its matrix layout provides better support for the nominative association between icon and label. This layout also removes the risk of data occlusion, where icons of similar value lie on top of each other. Together these should result in faster completion times and less errors for chart B.

Task 3, select parameters that are off-target but matched, involves identifying nominal relationships between labels and icons (i.e. finding the "ALL" reading), detecting ordinal differences between objects (position of "ALL" reading), then identifying a nominal state (matched status). The layout of Chart B separates the parameter reading from the sensor readings. It also presents the parameter reading beside the label where the matched status is encoded. Based on the proximity control principle [6] this should result in better performance for chart B.

Task 4, select tools with three or more off-target sensors, involves identifying nominal relationships between objects, then detecting ordinal differences between objects. The task constitutes a global question and involves understanding the data from the quantitative and two nominal variables. As chart B follows Bertin's rules for graphic construction [7], its visual form should make the target area pop out of the graphic form and result in better performance.

3. Results

The analyses were carried out separately for each task. For the efficiency (log of time) and accuracy (number of errors) measurements, generalised linear models were employed incorporating the repeated measures aspect of the design. As the satisfaction measurement was taken at the end of each task block, it had a smaller number of observations making a significance test unsuitable. Instead a confidence interval for the proportions is reported.

3.1 Task 1: Select off-target sensors

3.1.1 Efficiency
An Analysis of Variance (ANOVA) shows effects for chart type $F(1, 57) = 9.9918$, $p < 0.01$ and number of features $F(3, 57) = 16.954$, $p < 0.001$ but also a chart type by number of features interaction $(3, 57) = 8.5018$, $p < 0.001$. A Fisher LSD post hoc test on this interaction shows no significant difference between the charts ($p=0.594$) where no features exist, but mean performance time improvements for chart B were significant with 1 & 2 features ($p<0.0001$ & $p<0.0005$ respectively) and present but not significant ($p>0.056$) with 3 features.

3.1.2 Accuracy
An ANOVA shows strong interaction between chart type and number of features. A post hoc test was carried out with the following results. Chart A results in more errors than chart B in all cases where a feature exists. This difference is significant for 1 and 2 features ($p < 0.001$ and $p = 0.016$ respectively) but not significant for 3 features.

3.1.3 Satisfaction
14 out of 20 participants chose the redesigned chart compared to 3 out of 20 each for both the original chart and no preference. A 95% confidence interval for preference of Chart B over the other two options ranges between 55% and 91%.*

3.2 Task 2: Select off-target tools

3.2.1 Efficiency
An ANOVA shows effects for chart type $F(1, 57) = 32.9$, $p < 0.001$ and number of features $F(3, 57) = 35.327$, $p < 0.001$ but again a chart type by number of features interaction $F(3, 57)= 7.5698$, $p < 0.001$. The mean performance time was better for chart B in all cases where a feature existed. A fisher LSD post hoc test on the interaction shows this difference to be significant for 1 and 2 features (both $p<0.001$) and for 3 features ($p<0.01$).

3.2.2 Accuracy
An ANOVA again shows a strong interaction between the two factors. As no errors were incurred when no features were present, this level was not included in the analysis. A post-hoc test was carried out on the other results and showed that chart A resulted in more errors than chart B in all cases and that this difference is significant for 1 feature ($p < 0.001$) and 2 features ($p =0.010$) but not significant for 3 features.

3.2.3 Satisfaction
16 out of 20 users chose the redesigned chart compared to 3 out of 20 for the original chart and 1 out of 20 expressing no preference. This time the 95% confidence interval for chart B over the other two options ranges between 67% and 97 %.*

---

* generated using wilsons standard error

3.3 Task 3: Select parameters that are off-target but matched

3.3.1 Efficiency

An ANOVA shows no-interaction between chart types and number of features $F(3, 57)= 1.0498$, $p < 0.3777$. However, a strongly significant main effect is reported for chart type $F(1, 57) = 12.1$, $p < 0.005$ with chart B giving significantly faster performance times than chart A, and a weaker effect for number of features $F(3, 57) = 3.1829$, $p < 0.05$.

3.3.2 Accuracy

An ANOVA showed a weak interaction between factors. The post hoc test showed a significant difference ($p <0.001$) in favor of chart B where no feature exists. Although the number of errors was greater for chart A than chart B for 1 & 2 features no significant difference between chart types was shown. For 3 features the number of errors incurred was matched.

3.3.3 Satisfaction

16 out of 20 users chose the redesigned chart compared to 3 out of 20 for the original chart and 1 out of 20 expressing no preference. This time the 95 % confidence interval for chart B over the other two options ranges between 67% and 97 %.*

3.4 Task 4: Select tools with three or more off-target sensors

3.4.1 Efficiency

An ANOVA shows effects for chart type $F(1, 57) = 24.2$, $p < 0.001$ and number of features $F(3, 57) = 29.8$, $p < 0.001$ and again a chart type by number of features interaction $F(3, 57)= 5.5$, $p < 0.005$. Mean performance time was faster for chart B in all occasions and a fisher LSD post hoc test on the interaction shows this difference to be significant for no features ($p<0.001$), one feature ($p<0.05$) and two features ($p<0.001$) but not significant for 3 features ($p=0.53$).

3.4.2 Accuracy

An ANOVA showed no interaction between number of features and chart type. This task demonstrates a main effect for chart type with chart B having significantly fewer errors than A ($p< 0.001$) and a feature effect with 2 ($p<0.05$) and 3 features ($p<0.01$) having significantly more errors than 1 feature. In this analysis 0 features was omitted.

3.4.3 Satisfaction

16 out of 20 users chose the redesigned chart compared to 3 out of 20 for the original chart and 1 out of 20 expressing no preference. This time the 95 % confidence interval for chart B over the other two options ranges between 67% and 97 %.*

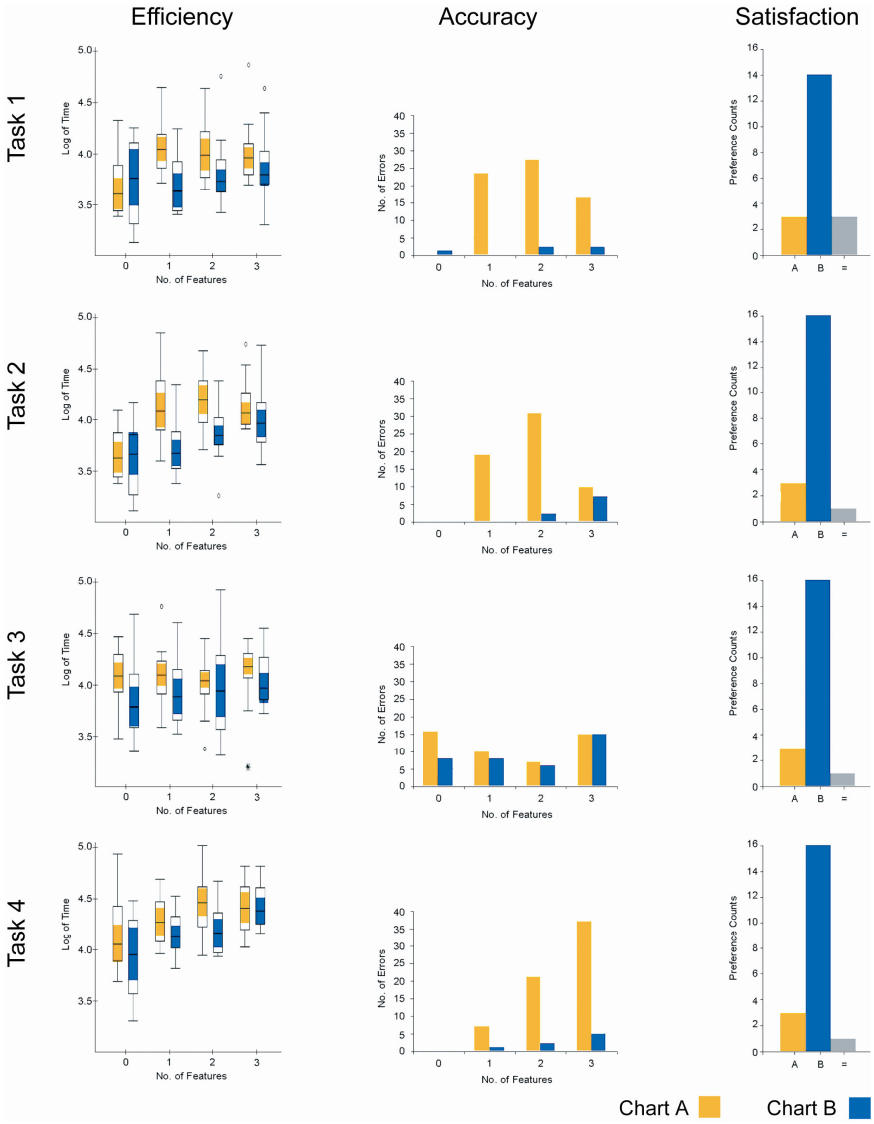

Figure 3. Results for each of four tasks and three performance measures

4. Discussion

For most tasks both number of features and chart type have an effect on user performance. An interaction between these two factors is also present making it difficult to report main effects. We provide a general discussion of the results below.

4.1 Task 1: Select off-target sensors

Chart B gave faster performance times in all cases except where no feature was present; in this case chart A was faster. In general, chart B resulted in fewer errors than chart A and gave a higher rating for satisfaction. It was originally expected that chart A would outperform chart B for this task. The results show that this is the case only when no features are present i.e. when the system is in control. While chart A's use of position on a shared scale should improve detection of a feature, the large number of icons may create visual noise that reduces performance. Chart B's encoding method causes off-target sensors to increase in scale and saturation. This improves the salience of these features. Their presentation within a matrix display eliminates the potential for data occlusion which may have resulted in the improvements in accuracy.

4.2 Task 2: Select off-target tools

Again chart B was faster in all cases where a feature existed but this time the differences are greater. Chart B resulted in fewer errors than chart A and again gave a higher rating for satisfaction. This was the expected result and is attributed to the matrix presentation. This layout makes it easier to relate the sensors to their tools as they are located on a shared spatial axis.

4.3 Task 3: Select parameters that are off-target but matched

As predicted the results show a significant improvement in efficiency for chart B and better accuracy in all cases except where three features exist. In this case equal numbers of errors are committed. This was the most complex task as is evident from the high number of errors committed with both charts. We attribute the improvements in chart B to the graphic encoding that makes it easier to detect the "ALL" (parameter mean) icon and to integrate it with the matched parameter status.

4.4 Task 4: Select tools with three or more off-target sensors

It was predicted that chart B would give a better performance due to the spatial encoding of the tools. This eliminates the need to temporarily store values in short term memory and allows the user to assess a tool by scanning the chart vertically. The results show that this is the case with a strong chart effect for accuracy and general improvements for efficiency.

4.5 The Number of Features Effect

At the outset of the experiment an effect was expected for number of features. The strong interaction between the two main factors was not expected as it was assumed

that an increase in features would increase difficulty incrementally for both charts. The results clearly show that this is not the case. If we look at number of errors we can see that this assumption only holds for task 4. For chart A with tasks 1 and 2 the number of errors increased from one to two features, but dropped off with three features. This is an interesting result requiring further exploration. It is possible that with three features present, the additional noise in the display causes the user to change their task performance strategy. While the current study can only identify different responses, future investigations of the displays using methods such as eye-tracking may provide useful information on viewing and task perforamce strategies.

4.6 Supplementary Study

While it was difficult to access a reasonable number of expert users, four process engineers agreed to carry out the experiment. The small study was carried out as a validation exercise to tesk the acceptability of the new design to the target users. We expected a certain amount of bias towards chart A due to their familiarity with the display. In fact, when presented with the new design (chart B) one engineer stated, "I don't like it and I don't think it will work". While the numbers were not sufficient to generate a statistical model, we observed some interesting results. There was a similar pattern of behavior between this test group and the main group for efficiency. In all tasks chart B gave faster mean response times than chart A where a feature existed. There was too much variation in the errors figures to draw significant conclusions, but the satisfaction measurement showed chart B was preferred for tasks 1 and 4, chart A and B were considered equal for task 2 and chart A was preferred for task 3. This is an encouraging result considering the engineers were more familiar with chart A.

5. General Discussion

Many psychophysical theories e.g. [5][6] give general guidelines for representing data based on specific cognitive tasks. The original OTI chart was constructed in-line with these guidelines using position to support quantitative judgments between datum. However, the results suggest that the new design is at least equal, and in many cases better, for carrying out the required tasks. This raises the question whether traditional approaches to cognitive graphics processing are too narrow for interactive displays? Many of these approaches rank visual variables in terms of their ability to support a specific task, but cognitive tasks rarely occur in isolation when working with dynamic charts and often a range of tasks can occur in quick succession. Also, while earlier theories have tended to focus on quantitative relationships, ordinal and nominal relationships play an important role in understanding complex systems. The proposed methodology suggests that the visual encoding of information requirements should be defined by both their position within a work domain model and the tasks for which they are used and the results are supportive of this.

5.2 Evaluation Issues in EID

The same characteristics that make it difficult to apply simple graphics guidelines also make it difficult to evaluate visual displays for complex systems. While carrying out this experiment a number of specific evaluation challenges were identified.

Firstly there is a difficulty in accurately representing work scenarios. While this experiment measures performance for a range of tasks associated with the OTI chart, this is just part of a larger health monitoring system that is used by process engineers. The engineers have access to a much wider set of resources including tacit knowledge and information from co-workers. These factors are beyond the scope of this experiment which can only show what an individual can understand through the displays. A similar issue relates to data. Original data is often unavailable for use in experiments for confidentiality reasons. Even when it is accessible the format is often unusable. In our case users had to identify stable and unstable system states. However, the frequency and severity of problems is unpredictable so it would be unreasonable to expect participants to monitor real world data. As a result mock datasets had to be generated.

Secondly there is a trade-off between representing the real-world and the practical limitations associated with experimental evaluation. The number of features factor was introduced to make the study more representative of a real-world monitoring scenario, but the interaction between number of features and chart type makes it difficult to generate statistically significant results for the main effects. A smaller range in the number of features factor would make it easier to obtain significant results but would reduce the validity of the case study. In light of this it is better to think of the experiment in terms of exploration and validation of potential design solutions rather than purely and evaluation study.

Finally, there is an issue as to whether the metrics of efficiency, accuracy and satisfaction provide the best means for evaluating an ecological interface. While these metrics tend to be pervasive in usability testing, the results can only inform us in general terms about the differences between displays. Knowing that one design performs better than another is obviously very helpful when choosing a system to implement, however usability metrics do not reveal the actual strategies that users employ when working with graphics. In sections 4.1-5 we attribute possible causes for the performance differences between displays. Alternative measurement techniques such as eye-tracking could help to accurately identify these causes and increase our understanding of how graphic forms are used during decision making.

6. Conclusions

This aim of this experiment was to study whether a redesign of a chart following the proposed design methodology would affect its usability. The results suggest that the new design provides better support in terms of efficiency, accuracy and satisfaction for a range of key tasks. While the experimental design resulted in strong interactions, post-hoc analyses suggest that chart type is responsible for the improvements in the performance metrics, providing evidence that the design methodology can result in a more usable design.

References

[1] Burns, C. M., Hajdukiewicz, J. R., 2004. Ecological Interface Design. CRC Press, Boca Raton, FL.

[2] Vicente, K. J., Rasmussen, J., 1992. Ecological interface design: theoretical foundations. IEEE Transactions on Systems, Man and Cybernetics 22, 589-606.

[3] Upton, C., Doherty, G. Extending Ecological Interface Design Principles: A Manufacturing Case Study, submitted for publication

[4] Vicente, K. J., 1999. Cognitive Work Analysis: Toward Safe, Productive, and Healthy Computer-Based Work. Erlbaum and Associates, Mahwah, NJ.

[5] Cleveland, W. S., 1985. The elements of graphing data. Wadsworth, Monterey, CA.

[6] Wickens, C. D., Carswell, C. M., 1995. The Proximity Compatibility Principle: Its psychological foundation and relevance to display design, Human Factors 37, 473-479.

[7] Bertin, J., 1983. Semiology of Graphics. The University of Wisconsin Press, Madison.