Exploring workload and performance through the use of visual analytics

Joanne Kitchin & Chris Baber

University of Birmingham, UK

ABSTRACT

In Visual Analytics, the output of automated analysis is presented to users in an interactive visualisation. By responding to this, the user can modify the parameters of the computer visualisation. This raises questions about the design of the visualisation and the appropriate level of interaction for users. This paper focuses on the impact of visualisation on user performance. A simple air target detection task (in which automated support identified possible threat aircraft) was combined with a secondary task (in which target letters had to be detected against a background). Four visual analytic displays were used to complete a target detection task over two studies. The first study explored how the displays affected workload, attentional demand and performance, and the second how workload, attentional demand and performance are affected by task load (using the same displays). Results show that the use of visual analytic displays maintains response time and primary task performance when task load increases. This suggests that the demand on attention is easier to manage when visual analytic displays are used.

KEYWORDS

Visual analytics, visualisation, workload, performance

Introduction

Visual Analytics (VA) combines automated data analysis with interactive visualisation (Keilman et al., 2009; Thomas and Cook, 2006). While this is often presented as a ‘new field’ (Scholtz,2006), for an ergonomist it feels very much like a familiar domain, with operators facing the output of complex data gathering and analysis; and then interpreting this output to make appropriate control actions (Sheridan,1987). One reason why VA has grown in popularity could be the trend towards ‘Big Data’ (Drury, 2015) and the need to triage the volume of data that operators face. As Simon (1971) pointed out, information “consumes attention…Hence, wealth of information creates poverty of attention and a need to allocate that attention efficiently among the overabundance of information sources…” [Simon, 1971, p41]. A second explanation for the growth of VA, could be the recognition that ‘thinking’ arises, not as something solely done ‘in the head’ but, through interaction with information sources (including other people) in the environment (Hutchins, 1995). So, the challenge is to ensure that information enables the user to reason appropriately, without being overwhelmed or distracted. The aim of the two studies reported in this paper are to: 1) explore the effects on workload, attentional demand and performance of different visualisations of a computer-supported task; and 2) determine which display type (if any) results in low attentional demand and workload but maintains and/or improves performance.

STUDY 1

Method

The experiment comprised of four conditions in a repeated measures design, counter-balanced using Latin Squares. Sixteen participants were recruited from the School of Engineering (fourteen male and two female), with a mean age of 26 (± 9 years). The experimental task was explained to each participant via written participant information pack and verbally, they were also given the opportunity to ask for more clarification of the task if needed. Each participant provided written consent and was made aware that they could withdraw (either in person or their data) at any time during the experiment. All participants had normal visual acuity, including corrected vision through spectacles. The University of Birmingham ethical review process approved the experiment.

Interface design

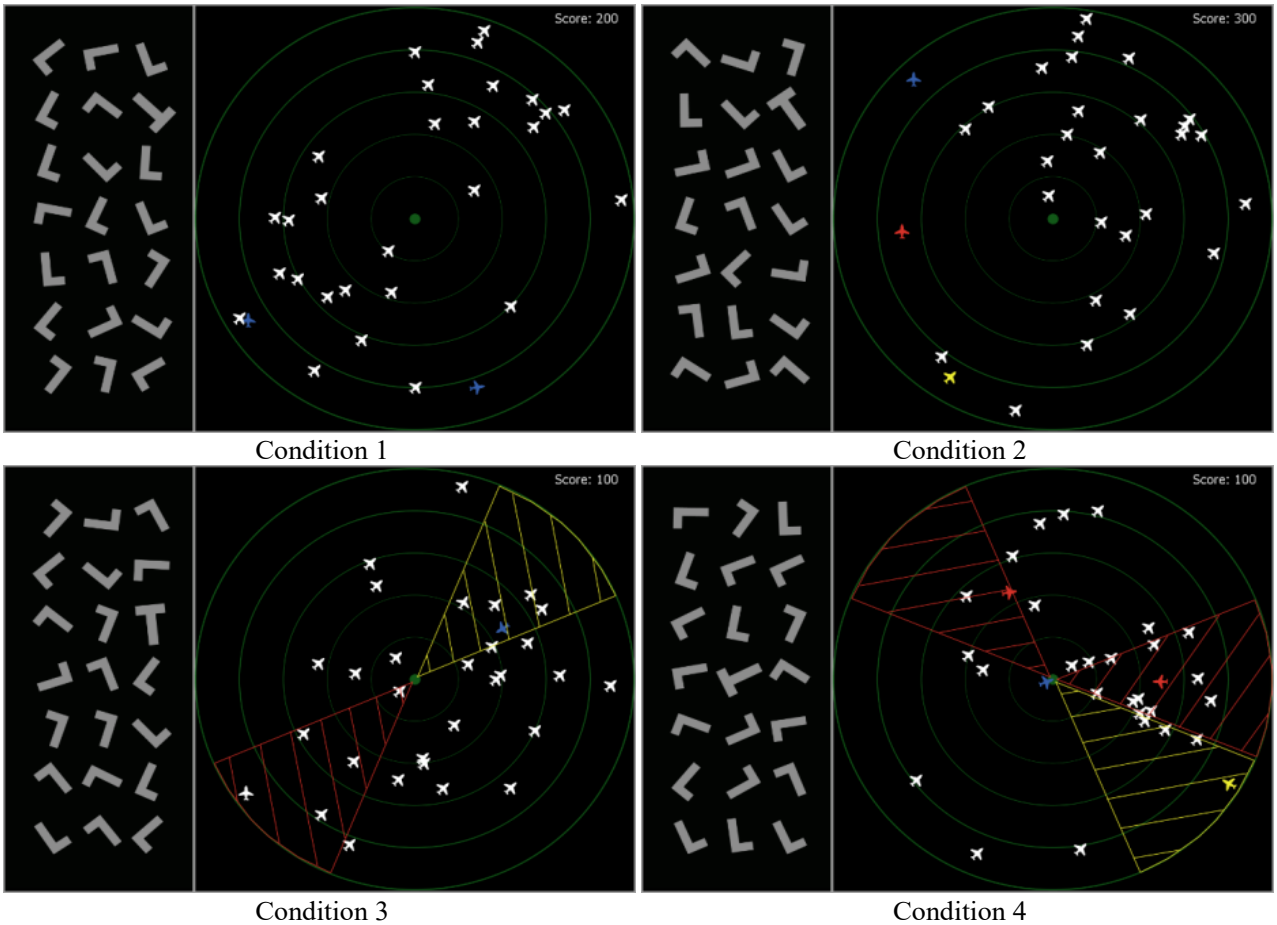

Figure 1: Example views of the displays used in each condition of the experiment.

Figure 1 shows examples of the interface in each condition. On the left of each display, there is an 'odd one out task' which was comprised of a three by grid of 'L' shapes in random static orientations. Hidden among the 'L' shapes was a single 'T' shape. This task is a variation on the target detection task of Neiser's (1964) visual search paradigm. The interface enabled the participant to click on the 'T' shape, resulting in an addition to the current score. The grid would re-order, placing a new 'T' shape in a different random position. The view on the right of each

interface was called the ‘radar view’. The ‘radar view’ consisted of aircraft tracks, which moved across the screen at set speeds and headings and were coloured white.

At predetermined intervals ‘odd aircraft’ entered the radar view. The ‘odd aircraft’ were going faster than the norm, or had a different heading to the norm, or travelled in a zig-zag manner (or any combination of these). There were seventeen “odd aircraft” in each condition.

In condition 1, all aircraft were white. Participants needed to check the behaviour of each aircraft and decide if it was an ‘odd aircraft’. If a participant clicked on what they decided was an ‘odd aircraft’, it would turn blue to signify to the participant that they had correctly identified a ‘threat aircraft’. If they selected an aircraft which was not odd, it would turn grey.

In condition 2, ‘odd aircraft’ were automatically assigned a colour: red or yellow. If the aircraft was red, it signified that this was a ‘threat’, if clicked on, it turned blue. If the aircraft was yellow, it signified was an ‘unknown threat’, when clicked it changed to either grey (to signify it was not a ‘threat aircraft’) or red (to signify it was a ‘threat aircraft’ and required additional action).

In condition 3 the colour of the aircraft remained white, as in condition 1. However, the regions in the radar were highlighted, red or yellow. This signified to the participant that there is either a ‘threat aircraft’ or an ‘unknown threat aircraft’ in that region.

In condition 4, ‘odd aircraft’ were presented by colour, as described in condition 2, and with the coloured regions as described in condition 3.

Task

The experimental task came in two parts, the primary task and the secondary task. In the primary task the participants were asked to search for the ‘odd aircraft’ in the ‘radar view’, when the participant believed that one had been found they were instructed to click on that aircraft. The objective of the primary task was to identify (click on) all ‘threat aircraft’ and the objective of the secondary task was to find (click on) as many ‘T’ shapes as they could when they felt they had time. Participants were encouraged to achieve the highest score, ‘threat aircraft’ were worth 100 points and ‘T’ shapes were worth five points.

The secondary task involved searching for the ‘T’ shape within the ‘L’ shapes in the ‘odd one out task’. Participants were told that this could be done as many times as they liked whenever they felt that they had the time and that the grid would refresh after every selection.

Experimental Setup

The experiment was carried out in one of the postgraduate offices in the Department of Electrical, Electronic and Systems Engineering, University of Birmingham, which allowed for strict control of the environmental conditions. The equipment used was a computer and mouse, the computer ran Microsoft Windows 7 and the experiment was accessed via the university network.

Data Collection

After each condition and at the end of the experiment participants completed a NASA-TLX Weighting Scale questionnaire (Hart & Staveland, 1986, Hart & Staveland, 1988) to record subjective participant workload. During each condition, the follow data were also recorded:

- The number of identified ‘threat aircraft’ (primary task).
- The time from ‘odd aircraft’ creation to ‘odd aircraft’ aircraft identification (primary task).

- The number of detected ‘T’s (secondary task).
- The number of times the participant switched between the ‘radar view’ and the ‘odd one out view’.

Data analysis

After reviewing the NASA TLX data, normal distribution curves, boxplots and z-scores of skewness and kurtosis it was determined that the data did not violate the assumptions for parametric data. Therefore, a Repeated Measures General Linear Model was used to analyse the workload results between the conditions in this experiment. All other data collected did, however, violate these assumptions and therefore a Friedman ANOVA was used to analyse results between conditions for the remaining data, using a Wilcoxon Signed Ranks Test for post hoc analysis with an adjusted significance level of 0.008 via a Bonferroni correction where applicable.

The maximum number of ‘threat aircraft’, 17, was predetermined and the same in each condition. This was used to normalise the primary task data into a percentage score. The maximum number of correctly identified ‘T’s in the secondary task was determined by the highest secondary task score achieved over all conditions, 131, and this was used to normalise the secondary task data into a percentage score.

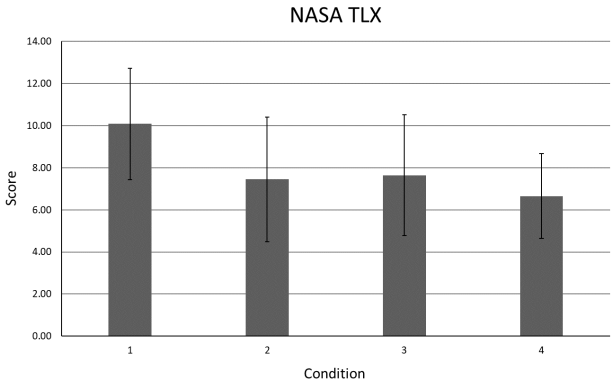

Figure 2: Results of the NASA TLX questionnaire

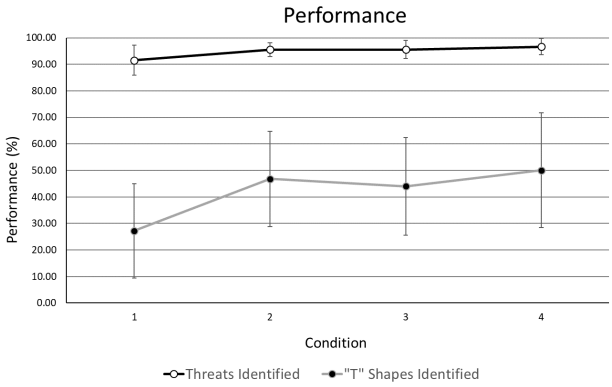

Figure 3: Primary and secondary task results

Results

Workload

Figure 2 shows the results of the NASA TLX scores over the four conditions. The score for condition 1 was statistically significantly higher than all other conditions ($P < 0.05$). The mean score was 10.08 with a standard deviation of 2.36. The largest decrease in score compared to condition 1 was in condition 4 ($P < 0.001$), with a mean score of 6.66 and standard deviation of 2.01. There was no statistically significant difference in score in all other comparisons ($P \geq 0.513$). The range of scores was lowest in condition 4, 6.5, compared to all other conditions (condition 1 = 9.3, condition 2 = 8.8, & condition 3 = 10.0).

Performance

Figure 3 shows the performance results for both the primary and secondary tasks. An analysis of Threats Identified (Primary Task) and ‘T’ Shapes Identified (Secondary Task) showed that Threats Identified was statistically significantly higher for each condition ($P < 0.05$) compared to ‘T’ Shapes Identified.

Condition 1 was statistically significantly lower than all other conditions ($P < 0.008$) in terms of the number of 'T' Shapes Identified over the four conditions. There was no statistically significant difference in score in all other comparisons ($P \geq 0.083$). The range in performance was noticeable in each condition with the maximum scores being 100, 123, 110 and 131, and the lowest scores being 10, 29, 10 and 20 for conditions A, B, C and D respectively.

A comparison of the 'threats' identified over the four conditions showed that the score for condition 1 was statistically significantly lower than condition 4 ($P < 0.008$). There was no statistically significant difference in score in all other comparisons ($P \geq 0.018$). The ranges of scores for the primary task was not so noteworthy with the maximum being 17 (highest achievable result) for all conditions and the minimum being 14, 16, 15 and 16 for conditions A, B, C and D respectively.

Reaction Time

The results for the aircraft life (i.e. the time from aircraft creation to aircraft detection) of the aircraft are shown in Figure 4. A comparison of the Aircraft Life over the four conditions showed that there was no statistically significant difference in all other comparisons ($P \geq 0.017$). Looking at the range of detection times between participants does show similarities between condition 1 and condition 4 (6.11 seconds and 5.87 seconds respectively) and condition 2 and condition 3 (2.94 seconds and 3.91 seconds respectively).

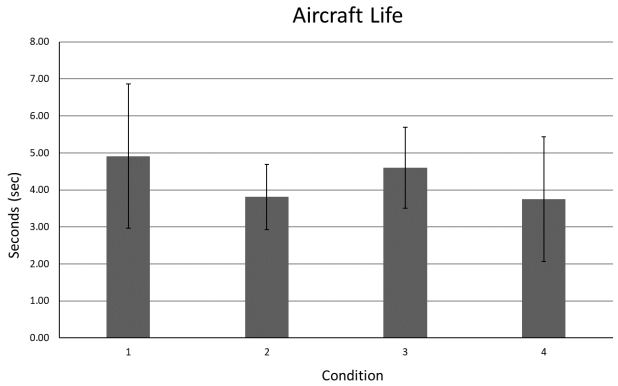

Figure 4: Aircraft life results: time from aircraft creation to aircraft detection

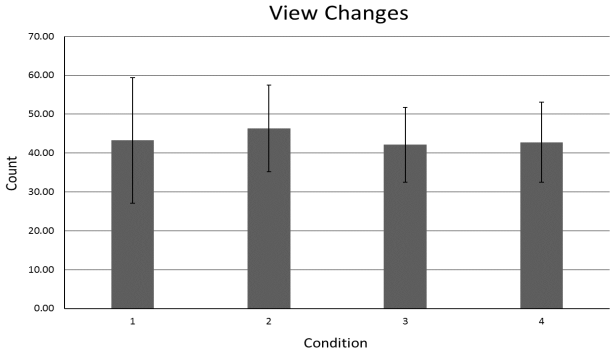

Figure 5: View change results: number of times participants switch from the primary task view to the secondary task view

Attentional demand

Analysis of the View Changes over the four conditions showed that there was no statistically significant difference in all other comparisons ($P \geq 0.053$). The range of view changes was also similar condition 2 and C (39 and 36 respectively) however, it was much higher in condition 1 (56), with condition 4 also showing an increase (46). Figure 5 shows a graph of these results.

Discussion

The results indicate that condition 1 (the control condition) had the highest rated workload compared to the other three conditions; the greatest improvement on this was seen in condition 4 with a difference of 3.42. No significant difference in reported workload was seen between the conditions B, C and D; however, after speaking to participants it was agreed that the task was 'easiest' in condition 4. This is supported by the performance results: participants achieved the highest scores in both the primary and secondary task in condition 4 with a significant difference

between condition 1 and condition 4. Although there were no other significant differences in the primary score, compared to condition 1 the secondary score was higher in all other conditions. This suggests two things: firstly, all changes in display provided a performance improvement compared to the control; and secondly the lowest perceived workload along with highest performance was in condition 4. The results also show that in all conditions, the primary score was significantly higher than the secondary score, suggesting that participants prioritised the primary task over the secondary task regardless of reported workload.

With the performance increase observed, it was expected that the time between aircraft creation and aircraft detection (i.e. aircraft life) would decrease over the four conditions, however, although a decrease is observed in all conditions compared to condition 1, notably in condition 2 and D, no statistical differences were found. As the primary task performance was high in all conditions ($\geq$ 91.54%) it is suggested that the detection time for the participants was near optimal and it would have been difficult to improve on this.

The average workload score in condition 1 suggested that even during the control condition perceived workload was low. This was a leading factor in creating a second study where the workload in two of the conditions is increased. After reviewing the results, only the displays in the control condition 1 and condition 4 would be presented in the second study and that the importance of both tasks would need to be emphasised.

STUDY 2

Method

Experimental Design, Experimental Setup, Data Collection & Participants

The experimental design, experimental setup and data collection was the same as STUDY 1. Sixteen participants were recruited from the School of Engineering (fifteen male and one female) and with a mean age of 19 ($\pm$ 0.62 years). Participants were encouraged to treat both tasks with equal importance.

Interface Design

Condition 1 and condition 2 used the display from condition 1 in STUDY 1. Condition 3 and condition 4 used the display from condition 4 in STUDY 1. Both the ‘radar view’ and ‘odd one out view’ provided a target score the participants were encouraged to reach. For the secondary task in the ‘odd one out view’ the target was 100 in each condition. For conditions 1 and 3 the target score for the primary task in the “radar view” was 17 and in condition 2 and 4 the target score was 52.

Results

Workload

A comparison of the NASA TLX scores over the four conditions can be seen in Figure 7. The results show that the score for condition 1 was statistically significantly lower than condition 2 ($P < 0.05$) and statistically significantly higher than condition 3 ($P < 0.05$). The there was no significant difference between condition 1 and condition 4 ($P = 0.485$). Condition 2 was statistically significantly higher than both condition 3 and condition 4 ($P < 0.05$) and Condition 3 was statistically significantly lower than condition 4 ($P < 0.05$).

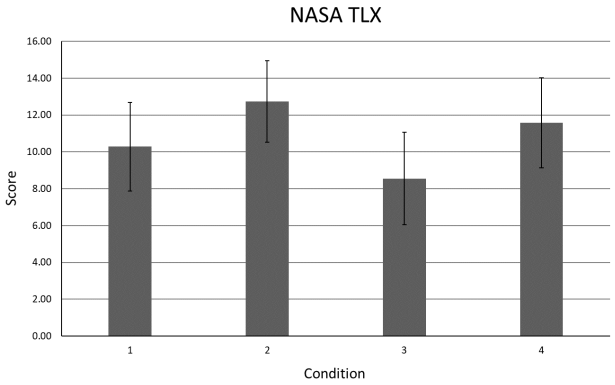

Figure 7: Results of the NASA TLX questionnaire

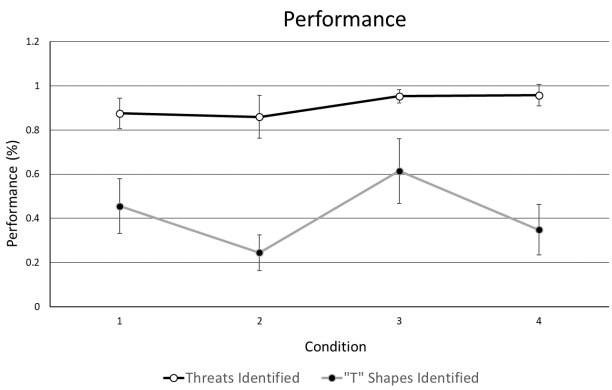

Figure 8: Primary and secondary task results

Performance

Figure 8 shows the performance results for both the primary and secondary tasks. Analysis of the threats identified (Primary Task) and 'T' shapes identified (Secondary Task) showed that threats identified was statistically significantly higher for each condition ($P < 0.05$) compared to 'T' Shapes Identified.

The amount of Threats Identified in condition 3 and condition 4 were statistically significantly higher than condition 1 and condition 2 ($P < 0.008$) but there was no statistically significant difference in score between condition 1 and condition 2 ($P = 0.535$) or condition 3 and condition 4 ($P = 0.175$).

The amount of 'T' Shapes Identified in condition 1 was statistically significantly higher than condition 2 and condition 4 and statistically significantly lower than condition 3 ($P < 0.008$). Condition 2 was statistically significantly lower than condition 3 and condition 4 ($P < 0.008$). Condition 3 was statistically significantly higher than condition 4 ($P < 0.008$).

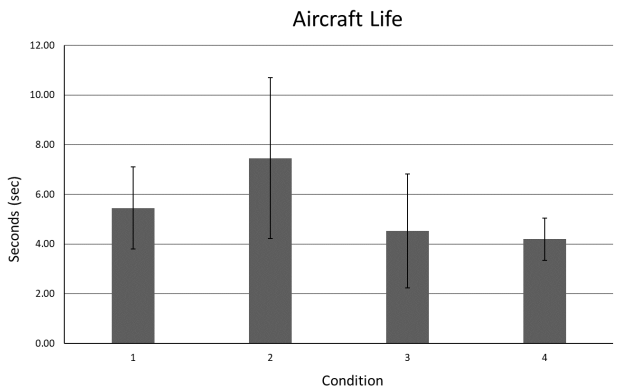

Figure 9: Aircraft life results: time from aircraft creation to aircraft detection

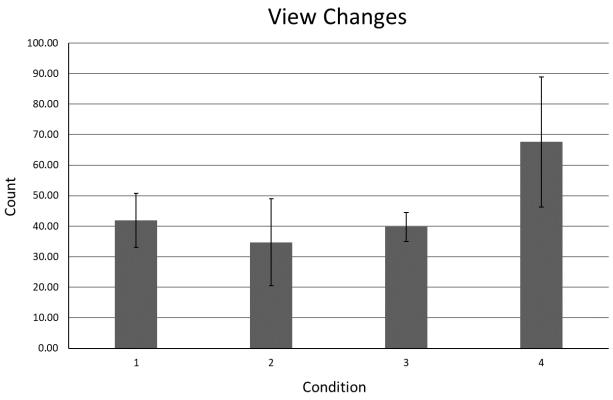

Figure 10: View change results: number of times participants switch from the primary task view to the secondary task view

Reaction time

Figure 9 shows the results for the aircraft life of the aircraft, i.e. the time from aircraft creation to aircraft detection. A comparison of the Aircraft Life over the four conditions showed that condition 4 was statistically significantly lower than condition 1 and condition 2 ($P < 0.008$ Primary and secondary task results) but not condition 3 ($P = 0.679$). Condition 2 was statistically significantly higher than condition 3 ($P < 0.008$). There were no statistically significant differences between condition 1 and condition 2, condition 1 and condition 3, and condition 3 and condition 4 ($P \geq 0.01$).

Attentional demand

Analysis of the View Changes over the four conditions showed that condition 4 was statistically significantly higher than in all other conditions ($P < 0.008$). There were no statistically significant differences between any of the other conditions ($P \geq 0.028$). Figure 10 shows a graph of these results.

Discussion

As expected the reported workload was higher when the task load was increased, however, this increase in workload was not equal across the displays used. There was a lower workload score when using the VA display in the low task load condition. Although the workload score was visually lower in the high task load condition the result was not significant. Statistically there was no difference between the VA displays during the high task load condition 3 compared to either of the control display conditions. When responding to the task demands, a high view change result could be explained in the high task load conditions because the participant needs to switch between the two views to obtain a high score in both tasks. The view count is high when using the VA display, but not when using the control display. In fact, there was no difference between view change between either of the control display conditions with either task load or the VA condition with a low task load. The increase in view changes indicated that participants were able to 'handle' the two tasks better when the VA display was used. This is supported by the increase in both correct primary and secondary task scores when changing from the control to the VA display with both task loads.

The aircraft life results also support the idea that participants could handle the two tasks better while using the VA displays. Aircraft life was not different between either of the VA conditions, implying that participants responded to aircraft in the same amount of time regardless of task load or even reported workload, which contrasts to the control where participants responded much slower to the aircraft in the high task load condition. Also there was no difference in primary task performance while using the VA display over the two task loads, but there was an increase compared to the control display under the same task load. However, a decrease in secondary task performance was observed for both display conditions when the task load was increased. This suggests that participants prioritised the primary task over the secondary task as seen in STUDY 1, even though it was stressed that both tasks had equal importance. Still, an increase in secondary task score did increase when the VA display was used, so even though task load reduced performance in this part of the task overall, using the VA display did improve performance.

Summary

- The VA display maintains response time when task load increases.

- The VA display improves overall primary task performance and maintains the performance when task load increases.
- The VA display improves overall secondary task performance however does not maintain the same level of performance when task load increases.
- The primary task has priority regardless of the display used or task load.

References

Drury, C. (2015) Human Factors / Ergonomics implications of Big Data analytics: Chartered Institute of Ergonomics and Human Factors annual lecture, *Ergonomics*, 58, 659-673

Hart, S. and Staveland, L. (1986) *NASA Task Load Index (TLX) V1. 0 User's Manual*.

Hart, S.G. and Staveland, L.E. (1988) Development of Nasa-Tlx (Task Load Index): Results Of Empirical And Theoretical Research, *Advances In Psychology, 52*, 139-183

Kielman, J., Thomas, J. and May, R. (2009) Foundations and frontiers in visual analytics, *Information Visualization, 8*, 239-246

Neiser, U. (1964) Visual search, *Scientific American,* 210, 94-107

Scholtz, J. (2006) Beyond usability: Evaluation aspects of visual analytic environments, *2006 IEEE Symposium On Visual Analytics Science and Technology*, New York: IEEE, 145-150)

Sheridan, T. (1987) Supervisory control, In G. Salvendy (ed.) *Handbook of HF,* NY: Wiley, 1244-1268.

Simon, H. (1971) Designing organizations for an information rich world. In M. Greenberg (Ed.) *Computers, Communications, and the Public Interest*, Johns Hopkins Press, 40–41

Thomas, J.J. and Cook, K.A. (2006) A visual analytics research agenda, *IEEE Computer Graphics and Applications, 26*, 10-13