DOT/FAA/TC-06/12

Federal Aviation Administration
William J. Hughes Technical Center
Atlantic City International Airport, NJ 08405

Optimal Design of Event Lists (ODELs) Phase 1: Does List Format Facilitate Visual Search for Information?

Vicki Ahlstrom, Human Factors Team – Atlantic City, ATO-P
Bonnie Kudrick, L-3 Communications, Titan Corporation

December 2006

Technical Report

U.S. Department of Transportation
Federal Aviation Administration

NOTICE

This document is disseminated under the sponsorship of the U.S. Department of Transportation in the interest of information exchange. The United States Government assumes no liability for the contents or use thereof. The United States Government does not endorse products or manufacturers. Trade or manufacturers' names appear herein solely because they are considered essential to the objective of this report. This document does not constitute Federal Aviation Administration (FAA) certification policy. Consult your local FAA aircraft certification office as to its use.

This report is available at the FAA, William J. Hughes Technical Center's full-text Technical Reports Web site: http://actlibrary.tc.faa.gov in Adobe® Acrobat® portable document format (PDF).

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
DOT/FAA/TC-06/12		

4. Title and Subtitle	5. Report Date
Optimal Design of Event Lists (ODELs) Phase 1: Does List Format Facilitate Visual Search for Information?	December 2006
	6. Performing Organization Code AJP-6110

7. Author(s)	8. Performing Organization Report No.
Vicki Ahlstrom, Human Factors Team – Atlantic City, ATO-P Bonnie Kudrick, L-3 Communications, Titan Corporation	DOT/FAA/TC-06/12

9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
Federal Aviation Administration Human Factors Team – Atlantic City, ATO-P William J. Hughes Technical Center Atlantic City International Airport, NJ 08405	**11. Contract or Grant No.**

12. Sponsoring Agency Name and Address	13. Type of Report and Period Covered
Federal Aviation Administration Human Factors Research and Engineering Group 800 Independence Avenue, S.W. Washington, DC 20591	Technical Report
	14. Sponsoring Agency Code ATO-P

15. Supplementary Notes

16. Abstract

This report documents the first in a series of studies on the optimal design of event lists (ODELs) for Technical Operations use. The ODELs study described in this report examines whether event list format has an impact on user performance when searching for information. The stimuli consisted of four different list formats: delineated, non-delineated, ledger shading, and white text on a blue background. These formats represented list formats currently in existence in the operational environment. Researchers measured task completion time, accuracy, and eye-scanning metrics such as number of fixations, fixation duration, blink frequency, pupil diameter, and number of reversals. Additionally, researchers collected subjective ratings of difficulty and preference rankings for each of the four conditions. The results indicated that the list design did not have a significant impact on task completion time or the number or duration of fixations. However, list design did appear to impact the error rate, subjective ratings of difficulty, and user preference. Participants made fewer errors in the ledger shading and delineated conditions, rated them as less difficult, and ranked them as most preferred.

17. Key Words	18. Distribution Statement
Table Format Technical Operations Visual Scanning	This document is available to the public through the National Technical Information Service, Springfield, Virginia, 22161. A copy is retained for reference at the William J. Hughes Technical Center Library.

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	22	

Form DOT F 1700.7 (8-72) **Reproduction of completed page authorized**

Table of Contents

Page

Acknowledgments v

Executive Summary vii

1. INTRODUCTION 1

1.1 Purpose 1
1.2 Metrics Defined 1
1.3 Task Design 2

2. METHOD 3

2.1 Participants 3
2.2 Apparatus 3
2.3 Procedure 5

3. RESULTS 5

3.1 Task Completion Time 5
3.2 Accuracy 6
3.3 Number of Fixations 7
3.4 Fixation Duration 9
3.5 Blink Frequency 9
3.6 Pupil Diameter 9
3.7 Number of Reversals 10
3.8 Subjective Ratings 10

4. CONCLUSION 11

References 13

Acronyms 14

List of Illustrations

Figures Page

Figure 1. The four conditions tested in this study: Delineation (DL), Ledger shading (LS), No delineation (ND), and White text on a blue background (WB). 4

Figure 2. Total number of errors for each of the four conditions. 6

Figure 3. Number of participants who made errors in each condition. 7

Figure 4. Relationship between number of fixations and response time. 8

Figure 5. Mean difficulty rating for each of the four conditions with a higher score corresponding to increasing difficulty. 10

Figure 6. Participant rankings for each of the four conditions from most preferred (1) to least preferred (4). 11

Tables Page

Table 1. Task Completion Time 6
Table 2. Number of Fixations 8
Table 3. Mean Fixation Duration 9
Table 4. Blink Frequency 9
Table 5. Pupil Diameters 10

Acknowledgments

The Federal Aviation Administration's Human Factors Division sponsored this work with the support of Beverly Clark and the Technical Operations Safety and Operations Support Office. Shantanu Pai wrote the experimental program. Ferne Friedman-Berg helped with the statistical analysis. John Dilks and Wallace Daczkowski set up the laboratory equipment, and Albert Macias helped extract the oculometer data. Mark Hale and Todd Truitt helped train Bonnie Kudrick on how to calibrate the oculometer properly. In addition, I would like to thank Ulf Ahlstrom who provided helpful discussions on data analysis. Finally, a note of thanks to the many specialists at the Technical Operations sites we visited who readily shared their insight and expertise.

Executive Summary

Maintenance specialists in the Federal Aviation Administration (FAA) routinely engage in visual searches for information on system events. Often, system information appears in the form of an event list. This study describes the first of a set of experiments aimed at identifying event list qualities conducive to rapid visual search.

We constructed an experiment based on four different event list formats currently used in FAA maintenance operations. The four formats were delineation (DL), no delineation (ND), ledger shading (LS), and white text on a blue background (WB). We asked participants to locate an information item in an event list, which was formatted in one of the four formats, as quickly and accurately as possible. We recorded reaction time and errors, and we used an oculometer to collect data on eye tracking during the trials such as fixation duration and number of fixations.

Results indicated that there were no significant differences in response time between the conditions. However, we found significant differences in accuracy (errors) between the conditions: participants made less errors in the LS condition than the other conditions, and fewer participants made errors in the LS and DL conditions than the ND condition. Participants ranked the LS and DL conditions significantly higher than the ND and WB conditions. These results led us to three major conclusions:

1. Event list format does not have a significant impact on visual search time.
2. Event lists should have some sort of demarcation differentiating the rows of data to facilitate accurate search.
3. Event lists that have rows that are clearly differentiated (conditions LS and DL) are perceived as easier to use and are preferred to those without clear differentiation (conditions ND and WB).

This study does have some limitations. Color coding is often used in event lists. It is unclear how color coding could be used together with ledger shading. We did not investigate the impact of color coding. We intend to investigate the impact of color coding on event lists in a future study.

1. INTRODUCTION

Technical Operations (TO) is the part of the Federal Aviation Administration (FAA) that is responsible for maintaining the ground equipment such as radars, communications, power, navigation aids, and environmental systems. As the National Airspace System (NAS) is becoming increasingly complex, TO Specialists must monitor and troubleshoot many pieces of equipment using computer interfaces. These computer interfaces take many formats, but one of the more common forms of information display is an electronic table known as an event list. Event lists exist throughout the TO environment.

TO Specialists spend vast amounts of time looking at and interacting with information presented in event lists, thus, it is important that the lists be well designed. Poorly designed interfaces such as lists can have a negative impact on user performance and can increase operating costs (Bednall, 1992). Optimizing lists could reduce time to complete a task, reduce error rate, and improve user satisfaction. For the FAA, optimizing performance could also mean an increase in equipment availability and, thus, fewer flight delays.

In spite of their extensive use, there is little systematic data on how event lists should be designed. Many alternatives are available including white text on a blue background separated by blank space, black text on a white background separated by lines (delineated), and alternating gray and white lines (ledger shading). Presumably, a design that lets the user find the desired item faster is better than one leading to slower performance. Additionally, a preferred format would minimize errors. Finally, the preferred format would not increase user workload.

1.1 Purpose

The purpose of this study is to identify characteristics that would facilitate the visual search for information in event lists for the TO environment. Thus, this study compares user performance with several different list designs to identify optimal design characteristics for event lists.

1.2 Metrics Defined

We utilized a combination of measures in this study. In addition to traditional measures of task completion time and accuracy, eye-movement measures are emerging as an effective way of evaluating user interfaces (Goldberg & Kotval, 1999). Literature has identified several different components of eye movement data as correlating with cognitive demands of a task. These eye activity measures could be potentially informative in determining the optimal format for information in event lists. Some of these potentially informative measures include number of fixations, fixation duration, blink frequency, pupil diameter, and number of reversals.

A user makes a series of eye movements when searching for information on a computer screen. The eye movements across a computer screen are made up of saccades and fixations. Saccades are the rapid movement of the eye from one location to another. Fixations are when the eye is held relatively stable in a position for a minimum duration to gather information.

According to Rayner (1995), more difficult processing, such as reading a difficult piece of text, leads to longer fixation durations. Mean fixation duration has been used as an indicator to reflect task difficulty for reading. Thus, longer fixation durations are thought to be an indicator of the difficulty the user has in extracting information from the display.

The number of fixations and the number of saccades are closely related; the number of saccades can be defined as the number of fixations minus one. The number of fixations overall is considered to be negatively correlated with search efficiency (Goldberg & Kotval, 1998). The number of saccades reveals the magnitude of search with more saccades suggesting more search is necessary to locate the target. Card (1982) found that the amount of time required to search a list of items in a menu was proportionate to the number of saccadic eye movements made by the user.

Several studies have used pupil diameter as a measure of workload (Iqbal, Adamczyk, Zheng, & Bailey, 2005; Iqbal, Zheng, & Bailey, 2004; Van Orden, Limbert, Makeig, & Jung, 2001). Studies have shown that pupil diameter covaries with performance measures that are used as an indicator of fatigue or drowsiness. Other studies found that pupil diameter decreased during times of increased tracking error (Van Orden et al.).

In other studies, researchers used blink duration and blink frequency as a measure of workload. Both blink duration and blink frequency have a tendency to decrease as workload increases (Van Orden et al., 2001), although this is not always the case (Ahlstrom & Friedman-Berg, 2005).

The eye tracking data also allowed us to look at patterns of how the individuals scanned the event lists. The most efficient scan path would consist of a short scan directly to the item of information being searched. Thus, an event list with characteristics that promote a more efficient search strategy would result in fewer reversals of the scan path. A scan path reversal can be described as a change in direction of more than 90 degrees from the preceding saccade.

1.3 Task Design

In preparation for this study, we visited TO locations. At each site, we collected observations about the characteristics of the tables that were present on operational systems. We also conducted structured interviews with personnel. These interviews provided information on how personnel use tables, including specific tasks. We used these data to structure the tables and tasks for the experiments.

Based on feedback from subject matter experts at operational locations and observational data collected, we found that the TO Specialists use tables of information for two general tasks: (a) searching for specific items of information and (b) monitoring events or systems for change. The current study focuses on the first of these tasks, searching for specific items of information.

We used familiarization visits to the TO facilities to collect information on table characteristics of current systems and the ambient environment in which these tables of information are presented. We found that the majority of the tables used by TO Specialists were organized in chronological order with the most recent event appearing at the bottom and older events, scrolling out of view, at the top of the table.

Operational facilities that are collocated with air traffic control facilities are often dimly lit. We measured ambient light levels at an operational facility, so that we could replicate the lighting in the laboratory for the experiment.

2. METHOD

This section describes the methods used in this study. Subsections describe the participants, apparatus, and procedures.

2.1 Participants

Thirty people participated in the study. We used a convenience sample, obtaining our participants from those available at the FAA's William J. Hughes Technical Center (WJHTC). We chose this sample of individuals instead of TO Specialists for several reasons; one reason was that of economy. We used local personnel to eliminate the travel cost necessary to bring in participants. We were also concerned that TO personnel may have developed biases after many years of working with a particular system. Participants ranged in age from 20 to 63. All had normal or corrected to normal vision. We pre-screened participants, precluding individuals wearing bifocals, in order to facilitate accurate oculometer data collection. Technical difficulties (e.g., the participant bumping the oculometer) caused us to exclude data from 10 participants. Five females and 15 males provided valid data.

2.2 Apparatus

We conducted this study at the Research, Development, and Human Factors Laboratory at the FAA's WJHTC. We wrote a custom data collection program in Visual Basic to display the find statements and table of information as well as collect reaction time data. The experiment used two computers: one for the oculometer and one for the experiment. The computer for the experiment had two monitors: one to display the find statements and the other to display the table of information. Prior to each session, we synchronized the two computers. Screen Pro recording software captured events on the screen. To mimic the lighting in TO environments, we set the overall ambient lighting level to +.38 foot-lamberts (fl).

The participants wore an Applied Science Laboratories Model 5000 oculometer comprised of eye and head tracking components. Prior to each session, a researcher calibrated the system using a nine-dot calibration grid (Willems, Heiney, & Sollenberger, 2005). The oculometer captured eye and head movements while recording x, y, and z point-of-gaze coordinates at the rate of 60 Hz (Ahlstrom & Friedman-Berg, 2005). Metrics derived from the oculometer data included pupil size, blink rate, fixation duration, number of fixations, number of reversals, and scan path.

In order to maintain congruency with tasks described by TO Specialists at the field sites visited, we set up the experiment to mimic a monitor and control interface for a building. The premise of the experiment was to mimic a monitoring and control function task without the technical expertise required for a TO environment. Rather than use names of actual NAS equipment, we chose items for the table that might be familiar to the participants such as lights, phones, and computers.

The table listed the date and time of each event as well as the item, description, level, and location. There were five levels used in the table: informational, normal, non-functioning, alert, and alarm. Locations included areas inside the building such as the Briefing Room, Experiment Room 2, General Purpose Lab, Reception desk, Conference Room A, Annex, Library, Section 4A, and rear of building (describing door location). The six column headings were Date, Time, Item, Description, Level, and Location. Because we were not using TO Specialists, we were

concerned that lack of familiarity with acronyms and terminology used for TO systems would impact the task. Therefore, instead of using actual NAS events, we created events related to the functioning of a building such as the monitoring of phones, doors, lighting, heating, and air conditioning.

There were 30 different find statements. The statements were presented on the leftmost screen (one statement at a time). The characters in the find statements were Courier New, Bold, Size 12 font and were approximately 2.0 mm (.039 in.) high.

Researchers presented the target information in a table of information on the rightmost screen. As illustrated in Figure 1, we tested four different table formats: Delineation (DL), Ledger shading (LS), No delineation (ND), and White text on a blue background (WB). In the DL condition, each cell was outlined in black. In the LS condition, rows of information alternated between white (61.0 fl) or light gray (46.5 fl) shading. In the ND and WB conditions, there were no visible lines between the rows. In the ND condition, information was presented in black text against a white background. The WB condition was exactly the same as the ND condition except the text was white on a blue background (4.56 fl, RGB values: R = 0, G = 0, and B = 255). The luminance of the black text, used in all of the conditions except the WB, was measured at .22 fl. We based the columns used in the information table on categories of information observed at field sites, including the date, time, item, description, status level, and location.

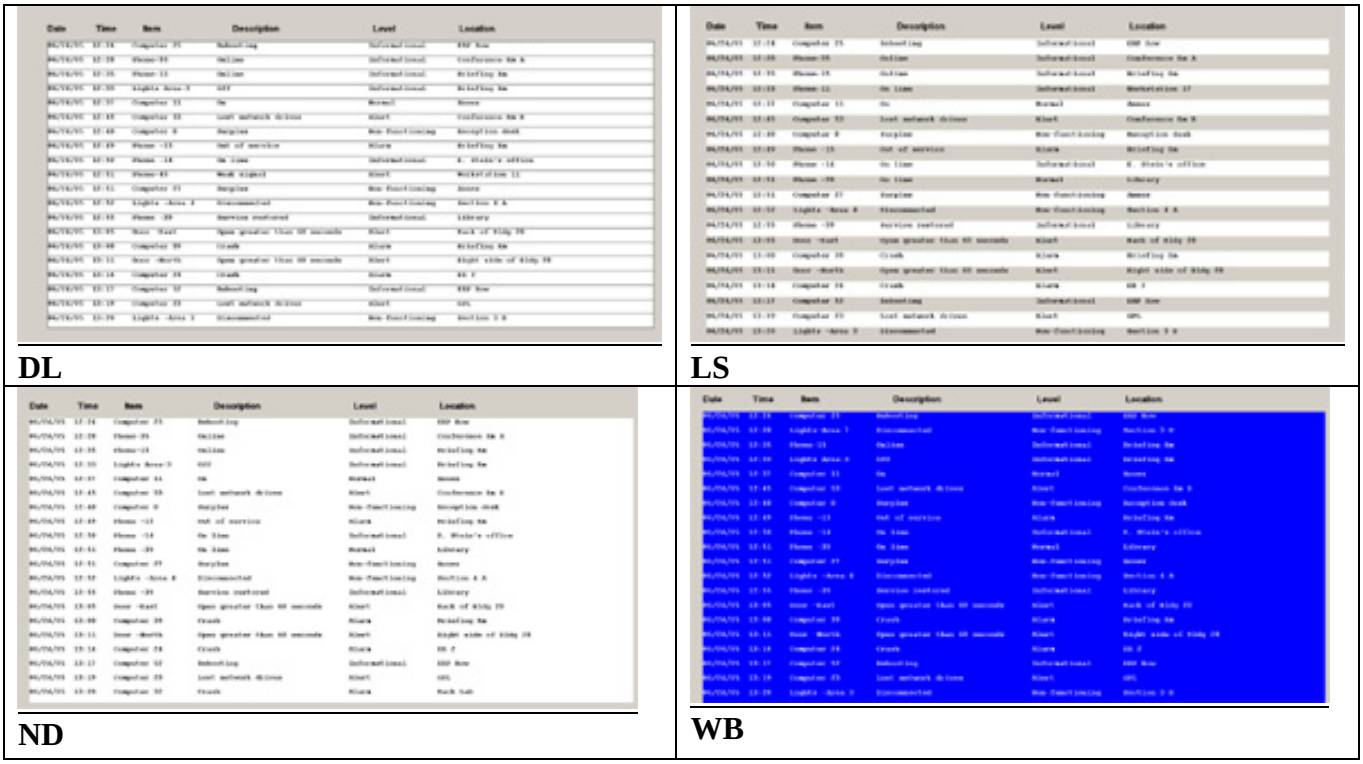

Figure 1. The four conditions tested in this study: Delineation (DL), Ledger shading (LS), No delineation (ND), and White text on a blue background (WB).

2.3 Procedure

For each condition, we collected reaction time data, accuracy data, fixation data, and pupil data. Finally, we collected subjective data including participant preference rankings and ratings for each of the conditions.

Each participant read and signed a written statement of informed consent. A researcher read the instructions aloud to each participant. Each participant provided background information by filling out a paper questionnaire. Information on the questionnaire included gender, job title, number of years in current position, and any current vision problems. We excluded individuals who reported wearing bifocals from continuing in the study due to technical limitations of the equipment used for data collection.

Each participant completed one practice session to familiarize themselves with the experimental procedures. The researcher encouraged participants to ask questions after the practice session. The practice session consisted of all four conditions with 10 tasks per condition. After the practice trial, each participant completed two experimental sessions. Total experiment time was approximately 1 hour.

In each trial, we presented participants with a piece of information to find on the leftmost screen of the two-screen display. The participants read each statement then searched in the table of information presented on the rightmost screen for the information item. If desired, the participants could look back at the find statement at any time by shifting the focus of their gaze to the leftmost screen. When the participants found the item, they used the mouse to click on the item. The search was self-terminating; it ended when the participants found the correct statement. A feedback statement appeared at the bottom of the screen indicating whether the answer was correct or incorrect. Once the participants answered correctly, a new find statement appeared on the leftmost screen. If the participants selected an incorrect answer, feedback appeared at the bottom of the screen that the answer was incorrect. This process continued until the participants selected the correct answer.

Each participant participated in all four conditions. The order of the conditions was randomly assigned. There were 30 trials per condition. At the conclusion of the experimental session, participants ranked the conditions, rated the display formats, and provided comments on any aspect of the study.

3. RESULTS

Technical difficulties (e.g., the participant bumping the oculometer) caused us to exclude data from 10 participants. Five females and 15 males provided valid data. For each condition, we collected reaction time data, accuracy data, fixation data, and pupil data. Finally, we collected subjective data including participant preference ratings and rankings for each of the conditions.

3.1 Task Completion Time

We measured task completion time for each condition. We defined task completion time as the difference from the time the question appeared until the time the participant selected the correct answer. Table 1 shows the mean task completion time for each of the four conditions. Although

task completion time was slightly faster for the LS condition, there was no significant difference across the four conditions. We used a one-way Analysis of Variance (ANOVA) to compute task completion times for all conditions. We found no significant difference between groups, $F(3, 57)$, $p > .05$.

Table 1. Task Completion Time

Condition	Mean	(*SD*)
DL	6.95	1.33
LS	7.07	1.95
ND	7.16	1.44
WB	7.10	1.54

$n = 20$

3.2 Accuracy

The researchers defined accuracy as completion of task without any errors. They defined an error as clicking on an item that was not the target item. Figure 2 shows the number of errors made for each condition. The overall accuracy rate was high (95%). The highest accuracy (fewest errors overall) was found in the LS condition (97.2%), followed by the DL condition (95.2%), the WB condition (94.7%), and then the ND condition (94.3%). A chi-square test for the number of errors indicated that the LS condition differed significantly compared to the ND condition, $\chi^2 = 9.01$, ($df = 1$), $p < .01$; the WB condition, $\chi^2 = 7.43$, ($df = 1$), $p < .01$; and the DL condition, $\chi^2 = 5.22$, ($df = 1$), $p < .01$. The other conditions did not reach statistical significance with the chi-square test.

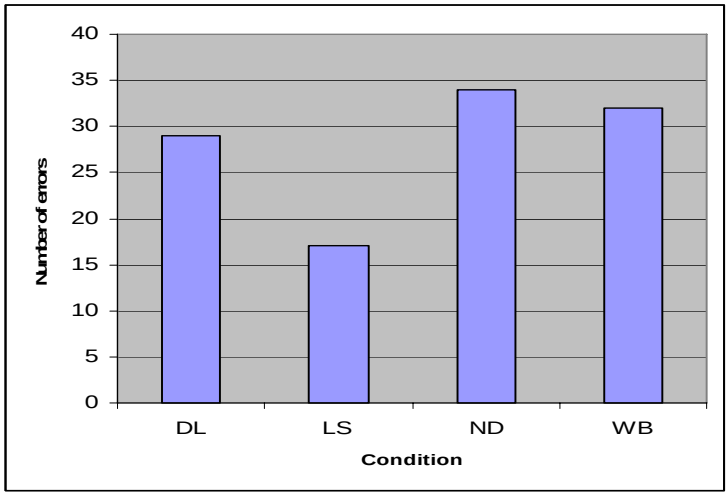

Figure 2. Total number of errors for each of the four conditions.

In addition to analyzing the number of errors overall, we looked at the number of individuals making errors (see Figure 3). This allowed us to look at the root of the errors more carefully. Overall, errors could be biased if a single participant made more errors in one particular condition. In order to see whether any condition increased the number of errors across participants, we looked at the number of participants who made errors per condition. A chi-square test for the number of participants making errors indicated that the DL condition differed significantly compared to the ND condition, $\chi^2 = 9.81$ (df =1), $p < .05$; and the LS condition differed significantly compared to the ND condition, $\chi^2 = 14.12$, (df = 1), $p < .01$. Chi-square tests also indicated a statistically significant difference between conditions LS and WB, $\chi^2 = 5.05$, (df = 1), $p < .05$. We did not find statistically significant differences at the $p < .05$ level between DL and WB, LS and DL, or ND and WB.

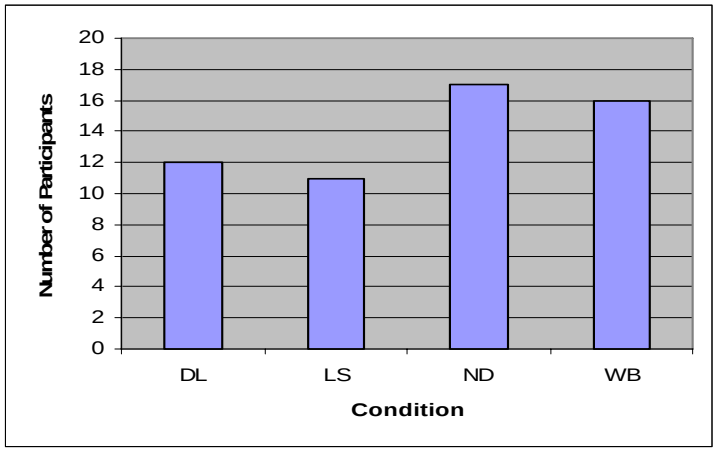

Figure 3. Number of participants who made errors in each condition.

User feedback indicated that there were two different categories of errors. In some cases, respondents clicked on an item one row away from the target, presumably by mistake. These types of errors were less frequently seen in the LS (2 adjacent errors) and DL (3 adjacent errors) conditions. These types of errors occurred more frequently in the ND condition (8 adjacent errors) and the WB condition (15 adjacent errors). In other words, for the DL and LS conditions, the incorrect answer selected by the participant was one row away from the correct answer for 10% and 12% of the errors, respectively. For the ND condition, the incorrect answer selected by the participant was one row away from the correct answer 24% of the time. However, for the WB condition, the incorrect answer selected by the participant was one row away from the correct answer 47% of the time.

3.3 Number of Fixations

In general, number of fixations is considered a metric of search efficiency. Fewer fixations would indicate that less searching was necessary to find the target, thus, producing a more efficient search.

We used a general linear model to analyze the relationship between task completion time and number of fixations for the four conditions. There was a statistically significant linear relationship for three of the four conditions at the $p < .05$ level: DL, $F(1, 18) = 5.94$, $p < .05$;

LS, $F(1, 18) = 4.72$, $p < .05$; and ND, $F(1, 18) = 6.93$, $p < .05$. The relationship for condition WB did not reach statistical significance at the .05 level, $F(1, 18) = 4.22$, $p > .05$. Figure 4 illustrates the relationship between the number of fixations and the response time for the four conditions. Also, as the number of fixations increased, there was a tendency for the response time to increase as well. The increased fixations may have been related to a lowered efficiency of information processing.

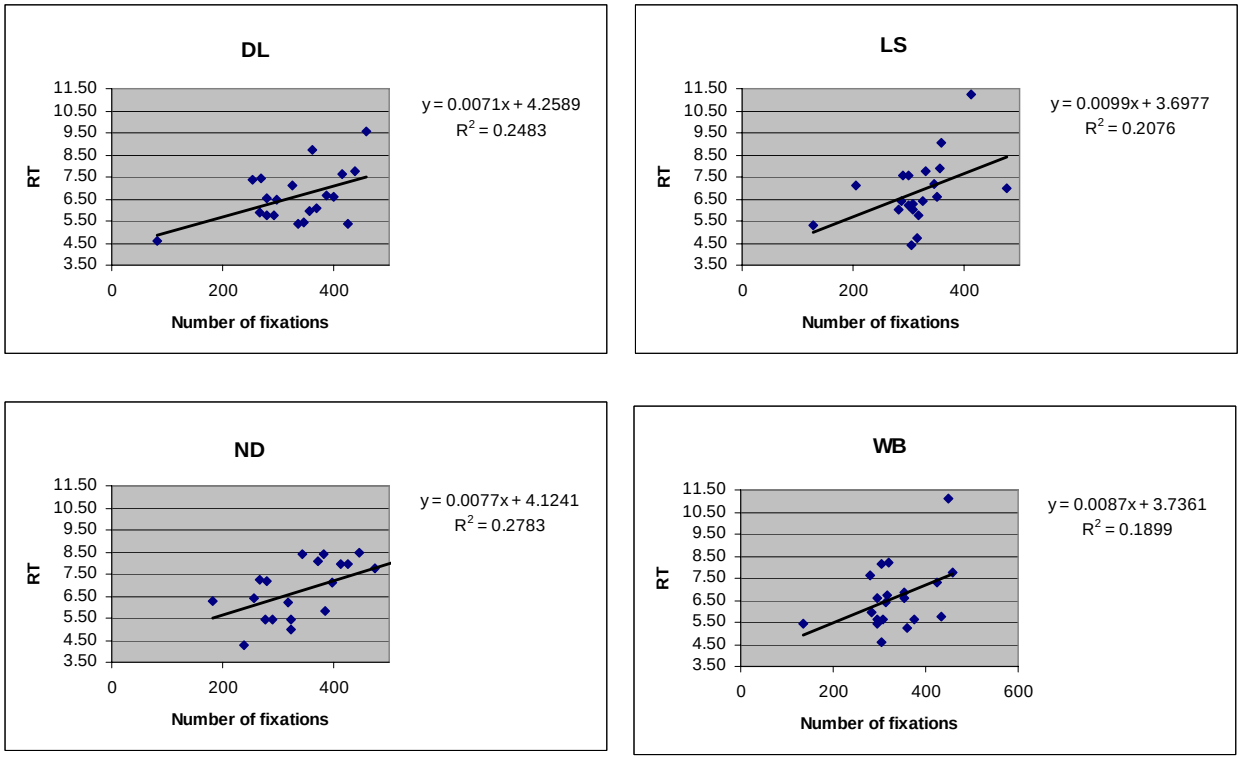

Figure 4. Relationship between number of fixations and response time.

As Table 2 shows, the LS condition had the fewest fixations, followed by the DL condition, the WB condition, and then the ND condition (with the most fixations). Although there appears to have been fewer mean fixations for DL and LS, we found no significant difference between groups when using an ANOVA to test the number of fixations for all conditions, $F(3, 57) =$ $p > .05$.

Table 2. Number of Fixations

Condition	Mean # Fixations	*SD*
LS	315.15	69.85
DL	332.10	85.86
WB	333.10	73.65
ND	345.85	87.15

n = 20

3.4 Fixation Duration

Mean fixation duration has been used as an indicator to reflect task difficulty for reading (Rayner, 1995). As text difficulty increases, the fixation duration tends to increase. As Table 3 shows, the LS condition had the longest mean fixation duration, followed by WB, DL, and ND. We used an ANOVA to compute mean fixation duration for all conditions. We found no significant difference between groups, $F(3, 57)$, $p > .05$.

Table 3. Mean Fixation Duration

Condition	Mean	*SD*
DL	.550	.099
LS	.566	.104
ND	.539	.105
WB	.562	.099

$n = 20$

3.5 Blink Frequency

In other studies, researchers used blink frequency as a measure of workload. Blink frequency tends to decrease as workload increases (Van Orden et al., 2001), although this is not always the case (Ahlstrom & Friedman-Berg, 2005). Table 4 shows the mean blink rate for each of the four conditions. We used an ANOVA to compute mean blink rate for each of the four conditions. We found no significant difference between groups, $F(3, 57)$, $p > .05$.

Table 4. Blink Frequency

Condition	Mean	*SD*
DL	72.15	46.93
LS	70.70	46.50
ND	75.45	49.71
WB	67.90	54.30

$n = 20$

3.6 Pupil Diameter

We compared the pupil diameter (in millimeters) across the four conditions. As Table 5 shows, DL and ND had the smallest pupil diameters, followed by LS and WB. The ANOVA test we used indicated statistical significance, $F(3, 57) = 19.228$, $p < .001$. A post hoc comparison using the Tukey honestly significant difference procedure showed that there were significant pairwise differences between all conditions ($p < .01$) except for DL and ND. In general, larger pupil size is correlated with higher workload. However, pupil size is also influenced by luminance, with the pupil size increasing for darker stimuli and decreasing for brighter stimuli. Because the luminance of the stimuli within our study decreased in the same direction as the increase in pupil size in our data, we suspect that the increase in pupil size is due to luminance changes rather than workload.

Table 5. Pupil Diameters

Condition	Mean	*SD*
DL	2.053	.255
LS	2.075	.263
ND	2.053	.252
WB	2.120	.271

n = 20

3.7 Number of Reversals

The eye tracking data allowed us to look at how the individuals scanned the tables of information. The most efficient search is one that goes directly to the target. It is less efficient if a person scans down a table of information, skips the target, and then has to reverse the scan direction while backtracking over items already scanned. Differences in the characteristics of the table could impact the efficiency of scanning behavior, lead to more reversals, and less efficient scanning behavior.

The LS and DL conditions were similar in the total number of reversals (summed across participants) with 960 and 966 reversals. Although the ND condition appeared to have had more reversals (1,008) than any of the other conditions, the WB condition had the fewest number of reversals of all of the conditions (911). An ANOVA indicated that the effect between the conditions in the number of reversals was not statistically significant, $F(3, 57)$, $p > .05$.

3.8 Subjective Ratings

Participants completed a debriefing questionnaire, which included their subjective ratings of each condition on a 5-point Likert scale. The Likert scale consisted of ratings from 1 to 5 (1 = very easy, 5 = very difficult). Figure 5 shows the average ratings for the four conditions. Overall, participants gave the WB and ND conditions the worst ratings, indicating that these were the most difficult. A Wilcoxon test showed that only the differences between the LS and ND conditions ($z = 2.097$, $n = 20$) and the DL and ND conditions ($z = 2.411$, $n = 20$) reached statistical significance at the $p < .05$ level.

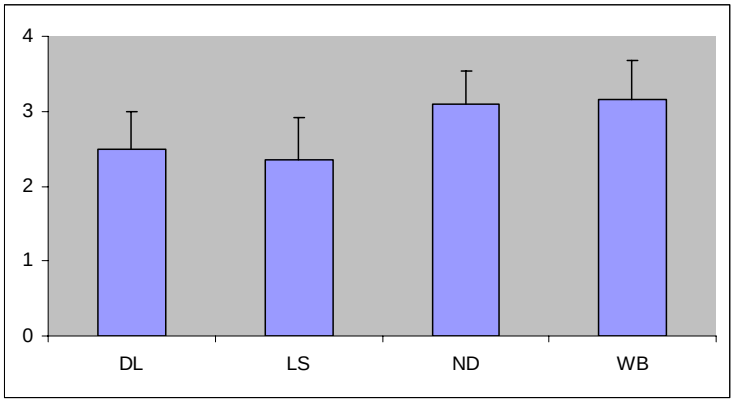

Figure 5. Mean difficulty rating for each of the four conditions with a higher score corresponding to increasing difficulty.

Participants also ranked each of the four conditions in order of preference. The rankings ranged from 1 (representing the most preferred) to 4 (representing the least preferred). Figure 6 shows the participant rankings for each of the four conditions. More participants ranked the LS condition as number 1 than any other condition. Although only 3 people ranked the DL condition as number 1, 11 people ranked it as number 2. None of the participants ranked DL as number 4. More participants ranked the WB condition as least preferred than any other condition. Thirteen people ranked it number 4 (least preferred). In contrast, 5 respondents ranked it number 1 (as their most preferred choice of the four conditions); second only to the LS condition. The ranking distribution indicates that participants had very strong preferences for the WB condition – either loving it or hating it – although more than twice as many hated it than loved it.

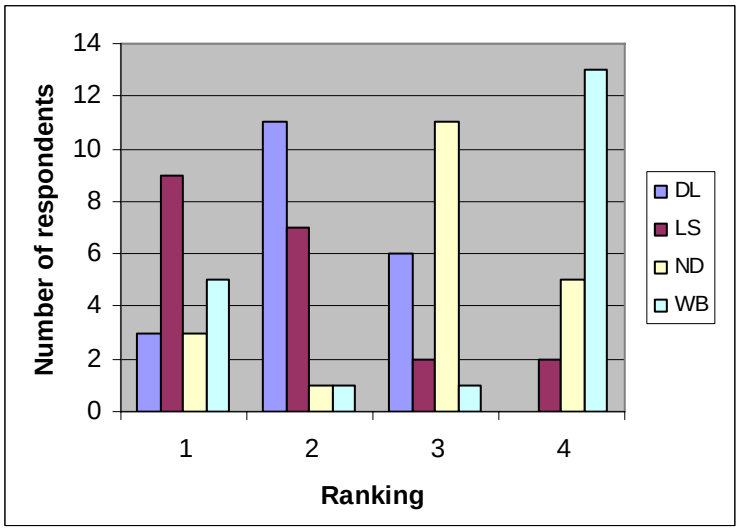

Figure 6. Participant rankings for each of the four conditions from most preferred (1) to least preferred (4).

4. CONCLUSION

This study looked at the impact of table format on search performance. We used different measures to capture different aspects of task performance. Overall, our results show that table characteristics had little impact on response time for the tasks we tested or on eye metrics, which are generally seen as measures of workload. We did find significant differences in pupil size across the four conditions. Pupil size, however, is affected by workload and luminance. Our results show that pupil size increases with the conditions that have lower luminance value more than the conditions that have higher measured luminance. It is likely that the pupil size results we found are due to differences in luminance rather than differences in workload. In our study, the differences in stimulus luminance could have confounded the use of pupil diameter as a measure of workload. Additionally, many of the studies that found pupil diameter to covary with workload and performance involved monotonous tasks lasting 2 or more hours or fatigued participants (Van Orden et al., 2001). Our study did not involve fatigued participants or tasks of long duration.

The primary difference between the conditions was the error rate. When every other row was shaded (LS), the errors decreased compared to the condition where there was no clear separation between the rows (ND). Thus, the results indicate that a person's response can be influenced by the design of event lists. Based on the data collected in this study, it would appear that there is a finite chance that a small investment in list design can reduce selection errors. This involves some sort of demarcation using lines or shading, which seems to make observer visual search more precise.

Considering that 24% of the incorrect responses in the ND condition and 47% of the incorrect responses in the WB condition were one row away from the correct answer compared to only 10% for the DL condition, where the rows were clearly delineated, one could speculate that if delineation was added to the blue condition, it may be possible that the overall error rate could decrease to less than that found in the DL condition. Further research is needed to determine whether a darker background with delineated rows will decrease errors to a level comparable to a lighter background with delineated rows.

Subjective ratings indicate that participants perceived ledger shading and delineated event lists as easier and preferred them to the non-delineated event lists. In the TO environment, however, event lists often have color coding. Further research is needed to determine how to appropriately combine ledger shading with color coding.

The WB condition provided us with some unexpected results. Although the majority of participants ranked this condition as the least preferred condition, there were many participants who ranked this condition as the most preferred. This condition had a relatively high number of incorrect responses that were only one row away from the correct response; we speculate that this condition may have been problematic for some users, but not for others. It seems clear from the data that further research is needed to fully comprehend the outcomes of this study as it applies to reverse polarity displays such as the white with blue background. For example, based on the results of this study, what would the outcome have been if the event lists we studied were designed with white text on a blue background and delineation?

The results of this study have implications for event list use in operational environments. As none of the different formats had clear advantages in response time, the decision on which format to use may be made based on other factors. Accuracy is usually an important consideration for event lists such as these. We found that the lists that had differentiation between rows, whether by shading or lines, resulted in fewer participants making errors. The LS and DL conditions also received higher user preference ratings and rankings.

There are several questions related to the implementation of event lists in operational environments that we did not answer in this study. As discussed earlier, we do not know how to fully explain the differences in ratings, rankings, and errors that we found between the WB and DL conditions when the main format difference between the two conditions was a reverse of polarity.

We found some advantage to the use of LS and DL event lists. In operational environments, items in a list are often color coded to indicate changes in status. The use of color coding may have had an impact on task performance. We intend to examine these issues more closely in future studies.

References

Ahlstrom, U., & Friedman-Berg, F. (2005). *Subjective workload ratings and eye movement activity measures* (DOT/FAA/CT-05/32). Atlantic City International Airport, NJ: Federal Aviation Administration, William J. Hughes Technical Center.

Bednall, E. S. (1992). The effect of screen format on visual list search. *Ergonomics*, *35*(4), 369-383.

Card, S. K. (1982). User perceptual mechanisms in the search of computer command menus. In *Proceedings of Human Factors in Computer Systems* (pp. 190-196). New York: ACM.

Goldberg, J. H., & Kotval, X. P. (1998). Eye movement-based evaluation of the computer interface. In S. K. Kumar (Ed.), *Advances in occupational ergonomics and safety* (pp. 529-532). Amsterdam: IOS Press.

Goldberg, J. H., & Kotval, X. P. (1999). Computer interface evaluation using eye movements: Methods and constructs. *International Journal of Industrial Ergonomics, 24*, 631-645.

Iqbal, S. T., Adamczyk, P. D., Zheng, X. S., & Bailey, B. P. (2005). Towards an index of opportunity: Understanding changes in mental workload during task execution. In *Proceedings of the ACM Conference on Human Factors in Computing Systems* (pp. 311-320). Portland, OR.

Iqbal, S. T., Zheng, X. S., & Bailey, B. P. (2004). Task-evoked pupillary response to mental workload in human-computer interaction. In the *Extended Abstracts of the ACM Conference on Human Factors in Computing Systems* (pp. 1477-1480). Vienna, Austria.

Rayner, K. (1995). Eye movements and cognitive processes in reading, visual search, and scene perception. In J. M. Findley (Eds.), *Eye movement research* (pp. 3-22). Amsterdam: Elsevier Science B. V.

Van Orden, K. F., Limbert, W., Makeig, S., & Jung, T. P. (2001). Eye activity correlates of workload during a visuospatial memory task. *Human Factors*, *43*(1), 111-21.

Willems, B., Heiney, M., & Sollenberger, R. (2005). *Study of an ATC baseline for the evaluation of team configurations: Effects of allocating multisector control functions to a radar associate or airspace coordinator position* (DOT/FAA/CT-05/07). Atlantic City International Airport, NJ: Federal Aviation Administration, William J. Hughes Technical Center.

Acronyms

ANOVA	Analysis of Variance
DL	Delineation
FAA	Federal Aviation Administration
fl	Foot-lamberts
LS	Ledger shading
NAS	National Airspace System
ND	No delineation
TO	Technical Operations
WB	White text on a blue background
WJHTC	William J. Hughes Technical Center