An Eye Tracking Study on the Effects of Dark and Light Themes on User Performance and Workload

Tobias Julien Ettling*
tobias.ettling@gmx.de
University of St. Gallen
St. Gallen, Switzerland

Dominik Steinmann
dominik.steinmann@gmail.com
University of St. Gallen
St. Gallen, Switzerland

Kenan Bektaş
kenan.bektas@unisg.ch
University of St. Gallen, HOCH Health Ostschweiz
St. Gallen, Switzerland

Amine Abbad-Andaloussi
amine.abbad-andaloussi@unisg.ch
University of St. Gallen
St. Gallen, Switzerland

Abstract

The visual theme of a dashboard, whether light or dark, is a prominent design choice with potential implications for user experience. This research investigates the effect of visual theme on user performance and workload during decision-making tasks on dashboards. In a within-subjects experiment, we measured the effect of dark and light themes and task complexity (easy, medium and hard), on task completion time, accuracy, confidence, fixation counts, pupil dilation, and workload. The dark mode improves accuracy, confidence, and average fixation count for medium task complexity levels, suggesting its utility in specific scenarios. In dark mode, the relative pupil dilation was higher, but the perceived workload was lower than in light mode. These findings highlight the need to study the interrelation between objective workload measurements and subjective questionnaires. This study advances empirical foundation for theme selection in data-driven interfaces of varying complexity.

CCS Concepts

• **Computing methodologies** → **Perception**; • **Human-centered computing** → **Empirical studies in visualization**; *Visualization.*

Keywords

eye tracking, dark theme, dark mode, light theme, light mode, display polarity, cognitive load

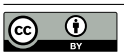
ACM Reference Format:
Tobias Julien Ettling, Dominik Steinmann, Kenan Bektaş, and Amine Abbad-Andaloussi. 2025. An Eye Tracking Study on the Effects of Dark and Light Themes on User Performance and Workload. In *2025 Symposium on Eye Tracking Research and Applications (ETRA '25), May 26–29, 2025, Tokyo, Japan.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3715669.3725879

1 Introduction

Dark theme, dark mode, night mode, or negative display polarity, is a user interface (UI) design approach, where a dark background and dim UI-elements are used with sufficient visual contrast. Dark mode has become increasingly prevalent in software products, particularly since 2018 according to Google trends, coinciding with a growing awareness of potential benefits of reducing blue-light exposure and eye strain, especially in dim-light environments [Yang et al. 2023]. Historically, dark mode has existed since early cathode ray tube (CRT) displays were used [Erickson et al. 2021; Ma et al. 2023]. At that time, a light text on a dark background (i.e., dark mode) was the standard until the advent of the desktop metaphor, where text editing tools with "What you see is what you get (WYSIWYG)" capabilities became predominant. Recently, dark mode has been introduced by many major software companies such as Apple, Google, and Microsoft, as an alternative to traditional light mode interfaces [Yang et al. 2023].

While some studies have explored the effects of dark mode on visual fatigue and readability [Cosima Piepenbrock and Buchner 2013; Xie et al. 2021], there is limited research examining its impact on user performance and workload when interacting with dashboards. For example, Qiao et al. explored the effects of light and dark mode with interactive maps and found that dark mode could improve performance at night, however, it was less favored by users in practice [Qiao and Wu 2023]. Dashboards, with their dense information and visual complexity, present a unique challenge for interface design [Yigitbasioglu and Velcu 2012], and the choice between light and dark mode could significantly influence user experience and performance.

In the current work, we aim to address this gap by investigating the effects of light and dark mode on user performance and workload during mentally challenging decision-making tasks on dashboards. In an eye-tracking experiment, we studied objective measures of user behavior, such as task completion time, error rates, and eye movement patterns. Thus, one crucial aspect of this research involves examining how eye movement patterns differ between dark and light mode dashboards. By analyzing metrics such as fixation duration and pupil size, we can gain insights into users' visual attention and cognitive processing when interacting in different modes, as previously done by others (e.g., [Qiao and Wu 2023; Sethi and Ziat 2023]). While objective measures such as eye tracking provide valuable information, it is equally important to consider users' subjective preferences for dark and light mode dashboards and their perceived complexity of tasks using the NASA TLX questionnaire [Hart and Staveland 1988] and the Single Ease Question (SEQ). By combining objective and subjective measures,



This work is licensed under a Creative Commons Attribution 4.0 International License.
ETRA '25, Tokyo, Japan
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1487-0/2025/05
https://doi.org/10.1145/3715669.3725879

we aim to gain a comprehensive understanding of how the dark and light modes influence user interaction and perceived cognitive load when working with dashboards.

2 Related Work

The emergence of dark mode as a prominent design trend has prompted research into its effects on various aspects of user experience, including visual comfort, readability and cognitive performance [Yang et al. 2023]. While research has been conducted on the effects on text readability [Cosima Piepenbrock and Buchner 2013; Sethi and Ziat 2023] and maps [Xie et al. 2021], its impact within the context of complex dashboard designs remains limited. According to Xie et al., one of the primary motivations for adopting dark mode is its potential to reduce eye strain and improve visual comfort, particularly in low-light environments in a navigation context [Xie et al. 2021]. By minimizing glare and reducing the amount of light emitted from the screen, dark mode can alleviate fatigue and discomfort associated with prolonged screen usage [Xie et al. 2021]. Furthermore, while dark mode can improve readability of text for some users with visual impairments [Aleman et al. 2018], it may also reduce the readability for others [Cosima Piepenbrock and Buchner 2013; Sethi and Ziat 2023]. Light mode generally offers better visual acuity due to the higher contrast between text and the background. Studies have shown that light mode can enhance reading comprehension and proofreading performance, particularly for small font sizes [Cosima Piepenbrock and Buchner 2013]. Ultimately, the choice between light and dark mode often comes down to personal preference and individual needs. Some users may find dark mode more comfortable and aesthetically pleasing, while others may prefer light mode for its clarity and readability [Sethi and Ziat 2023].

Eye-tracking technology has emerged as a valuable tool for evaluating user interface designs [Holmqvist et al. 2011], including the impact of light and dark modes on user behavior and visual attention [Bergstrom and Schall 2014; Dadzie et al. 2016; Qiao and Wu 2023]. Eye-tracking metrics such as fixations, saccades, and pupil dilation can provide insights into how users interact with different interfaces and how their visual attention is affected [Bergstrom and Schall 2014; Hostettler et al. 2023; Zagermann et al. 2016]. Studies have shown that eye tracking can reveal differences in visual search patterns, attention allocation, and cognitive processing [Jacob 1995]. Sethi and Ziat investigated the impact of display polarity (light-on-dark vs. dark-on-light) on cognitive load, using eye-tracking and subjective measures [Sethi and Ziat 2023]. Their findings revealed that older adults experience higher cognitive load with negative polarity (dark mode) in bright environments, while younger adults exhibit higher cognitive load with dark mode in dim environments. The authors concluded that older adults tend to favor positive polarity to minimize mental fatigue, whereas younger adults often prefer negative polarity in dim environments for aesthetic reasons [Sethi and Ziat 2023]. Qiao and Wu conducted a user experiment with eye-tracking to evaluate the usability of dark and light maps in different lighting environments (day and night) [Qiao and Wu 2023]. They found that the light-during-the-day performance was generally the best, followed by the dark-at-night performance. Conversely, the dark-during-the-day performance was the worst, followed by the light-at-night performance. They emphasized the importance of aligning map design with the lighting environment for better communication. They also observed a low preference for dark maps in practice, suggesting a need to shape user attitude towards dark mode in public-facing web maps. Similar to [Qiao and Wu 2023], in our study, we focus on specific effectiveness, efficiency, and cognitive load metrics:

- **Effectiveness metrics:** *Task Accuracy* is measured based on the correctly answered task questions [Qiao and Wu 2023; While and Sarvghad 2024], *Task Confidence* refers to the participants subjective Likert scale rating of their confidence after completing a task, and *Fixation Count* is the total number of fixations during the task [Arana-De las Casas et al. 2023; Qiao and Wu 2023].
- **Efficiency metrics:** *Task Completion time (TCT)* refers to the total time taken by a participant to complete a task, *Time to first fixation (TFF)* is the duration from the beginning of the task to the first fixation in a area of interest (AOI).
- **Cognitive load metrics:** *Questionnaires* were used to assess subjective perceived workload, NASA-TLX and Single Ease Question (SEQ) [Hart 2006], *Pupil Dilation* (i.e., the relative change in pupil size) [Dong and Liao 2016; Hostettler et al. 2023; Qiao and Wu 2023; Sethi and Ziat 2023], and *Total Fixation Duration* on the areas of interest (AOIs) [Qiao and Wu 2023].

3 Methodology

We combined objective measures of user behavior and subjective assessments of workload to investigate the effects of dark and light themes on dashboard interfaces. Data was collected using eye tracking, task performance metrics, and subjective workload questionnaires. As the user task (see section 3.3) stayed the same in each trial, three types of dashboards were created to induce cognitive workload in three different levels: *High*, *Medium*, and *Low*. Each dashboard has been created in light and dark mode, which makes a total of 6 dashboards. Questionaires where staged before (demographics and theme preference) and after the experiment (NASA-TLX ranking), as well as during the experiment after each trial (task answers and confidence, SEQ and NASA-TLX).

Table 1: Summary of the six trials in Step 5 (Figure 2), each linked to a light or dark mode dashboard. Complexity (Easy, Medium, Hard) is based on the probability of belonging to G1 or G2. Indicators relate to AOIs in Figure 1: Neutral attributes are indecisive, Supportive favor the dominant group, and Contradictory favor the opposite group.

		Group Probability (%)			Indicators		
Complexity	ID	G1	G2	Theme	Neutral	Contradictory	Supportive
Easy	1	100	0	Light	-	-	All
	2	0	100	Dark	-	-	All
Medium	3	10	90	Light	Height	Age	Job, Hobby
	4	20	80	Dark	Job	Height	Age, Hobby
Hard	5	60	40	Light	-	Job, Height	Age, Hobby
	6	60	40	Dark	-	Job, Height	Age, Hobby

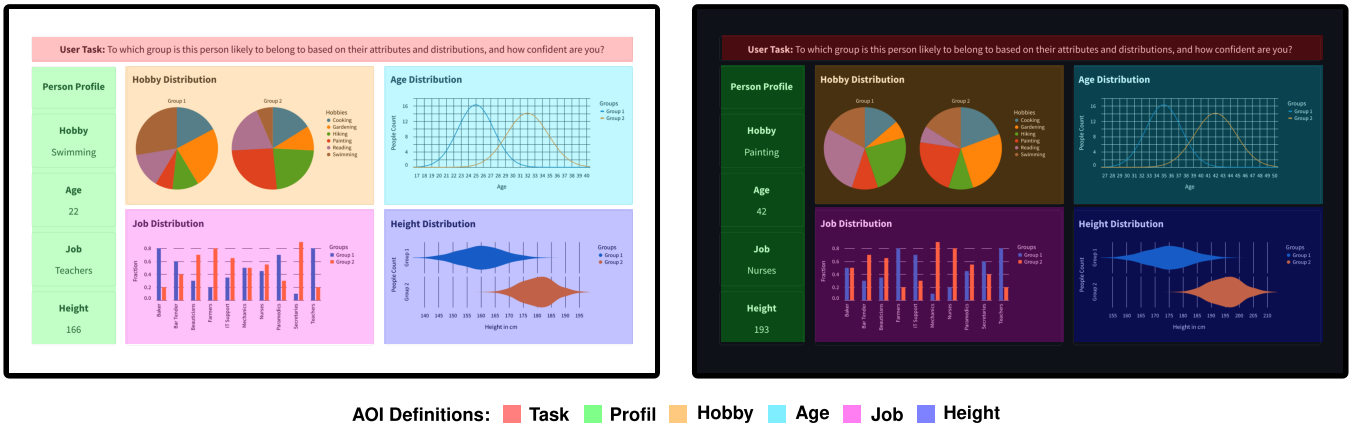


Figure 1: Visualization of the light and dark mode dashboard variants, highlighting the Areas of Interest (AOIs) used for eye-tracking analysis and user interaction assessment.

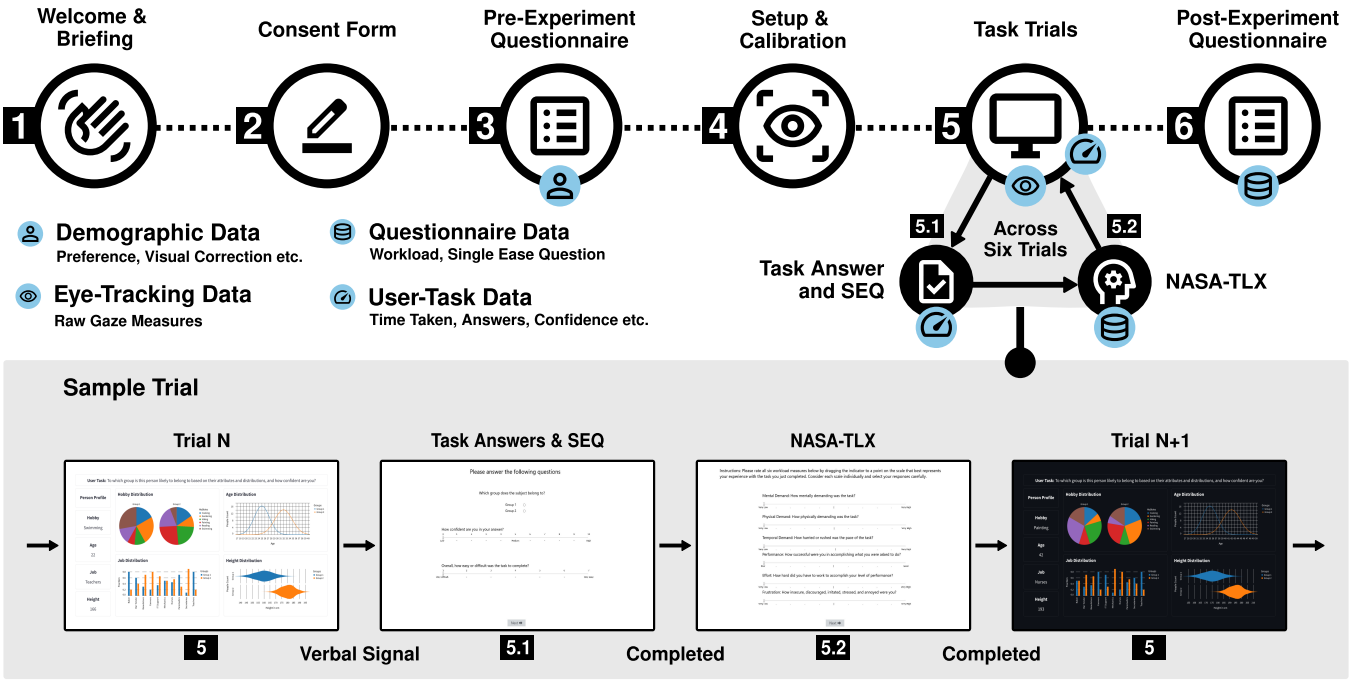


Figure 2: Overview of the experimental procedure, detailing the six-step workflow from briefing to post-experiment questions. Step 5 is repeated for for each one of the six dashboard variations. Blue circles indicate the type of data collected at each step.

3.1 Participants

For the experiment, we recruited 17 volunteers. Participants ages ranged from 18 to 38 years (Mean = 26.9, SD = 6.3). The gender distribution included 13 participants identifying as male, 2 female, and 2 participants identifying as other. Regarding their typical interface preferences, 10 participants reported preferring dark themes, 4 preferred light themes, and 3 had no preference. The initial eye-tracking calibration quality was rated 'Excellent' for 10, 'Good' for 6, and 'Poor' for 1 participant using the iMotions software's assessment.

Data from participants were excluded based on pre-defined criteria: One was excluded due to exhibiting consistently rapid task completion times across trials, suggesting a lack of sufficient engagement with the decision-making process, and another participant who achieved 'Poor' calibration quality was excluded, as reliable gaze data could not be ensured. Furthermore, two additional participants provided poor pupil size evaluation, which have been excluded in pupil related analysis. Following these data exclusion criteria, analysis of task performance and subjective questionnaires were conducted using data from N=15. Due to specific issues with

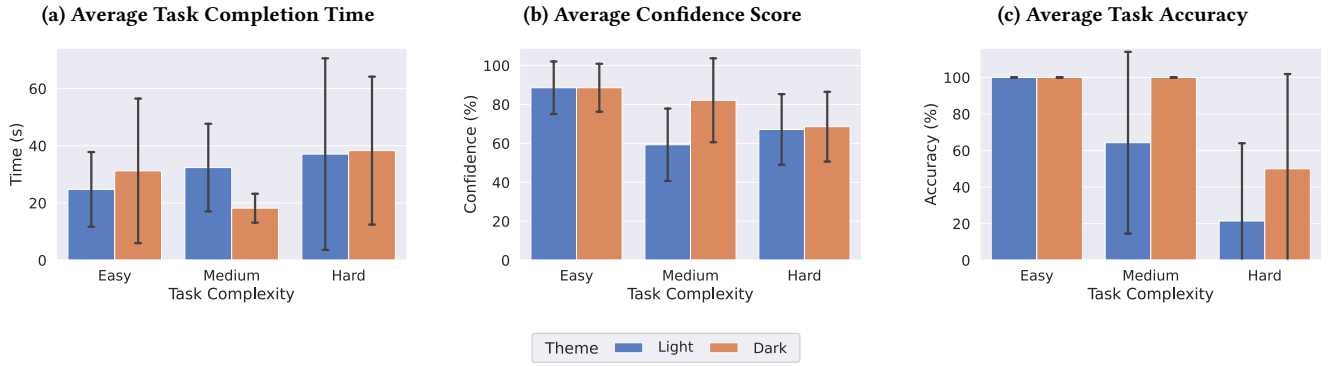


Figure 3: Displays (a) the average Task Completion Time, measured as the duration of each trial, (b) the confidence score provided by each participant immediately after completing the task, and (c) the accuracy of task answers given after each trial. We compute the average across all participants and trials. Error bars represent the ± standard deviation and color indicates the theme, light (blue) or dark (orange) theme.

the pupil recordings for two participants, analysis involving pupil dilation metrics were restricted to a subset of N=13.

3.2 Apparatus

We used a screen-based Tobii Pro X3-120 eye tracker (120 Hz) to record participants' eye movements on a 27-inch monitor with 1920×1080 resolution. Eye tracking data was analyzed to measure fixations, pupil dilation, gaze duration, and time to first fixation on areas of interest (AOIs) [Blascheck et al. 2014; Holmqvist et al. 2011] that are defined within the dashboard interfaces (Figure 1). The luminance of the lab was 114 LUX (i.e., a dimly lit work environment with closed window blinds), which was chosen due to eye-tracking hardware requirements for reliable data acquisition.

3.3 User Task

The user task was designed to simulate a realistic analytical decision-making scenario that requires the integration of information from multiple visualizations, a common activity when using dashboards [Few 2013]. The dashboard (see Figure 1) comprised six key components: (i) a textual user task description outlining the objective, (ii) a subject profile featuring plausible demographic-style attributes: two numerical (age and height) and two categorical (job and hobby) variables, providing concrete data points for evaluation, (iii) a pie chart illustrating hobby distribution (categorical), (iv) a line chart depicting age distribution (numerical, continuous), (v) a grouped bar chart showing job distribution (categorical), and (vi) a horizontal violin chart comparing height distributions (numerical, continuous) between two potential groups (G1 and G2).

The selection of these specific dashboard elements and chart types was motivated by several factors. First, we aimed to include a mix of visualization types commonly used in dashboards [Yigitbasioglu and Velcu 2012], representing different ways of encoding data. Second, by incorporating charts suited for both categorical data (pie chart for proportions, grouped bar chart for discrete comparisons) and numerical data (line chart for trends/distribution shape, violin plot for density comparison), we ensured the task required users to process and interpret diverse visual information formats. This heterogeneity reflects typical dashboard complexity and allowed us to assess potential theme effects across different visual encodings, which may vary in perceptual difficulty. Third, the overall layout required users to actively shift attention and integrate information from spatially distinct areas (profile text, individual charts), simulating the visual search and synthesis inherent in dashboard usage.

The core task required users to analyze the subject's profile attributes (age, height, job, hobby) in the context of the distributions shown in the four charts and decide whether the subject was more likely to belong to Group 1 (G1) or Group 2 (G2). This classification task simulates a common analytical goal where evidence must be gathered and weighed from multiple visual sources to reach a conclusion.

Crucially, to investigate whether the effects of light versus dark themes might depend on the cognitive demands of the task, we systematically varied the task complexity. We synthetically generated the underlying data for the distributions (using fixed means and standard deviations) and carefully adjusted subject profile attributes across trials. This manipulation altered the alignment and strength of the evidence provided by the charts, creating three distinct levels of complexity: Easy, Medium, and Hard (see Table 1 for probability breakdowns and indicator roles). For example, *Easy* trials presented unambiguous evidence favoring one group across all indicators. *Medium* trials introduced some ambiguity or slightly conflicting information. *Hard* trials presented substantially conflicting evidence between indicators, requiring more careful deliberation and integration to resolve, thus inducing higher cognitive load.

To mitigate potential learning effects across the six trials, we slightly adjusted the value ranges for the numerical attribute distributions (age and height) and shuffled the specific labels used for the categorical distributions (job and hobby) between trials, while ensuring that the underlying shape, proportions, and relative difficulties of the distributions remained consistent according to the task complexity.

3.4 Procedure

The experimental procedure is illustrated in Figure 2. Each Participant progressed through the following steps:

(1) Welcome and Briefing: Participants were welcomed, and the general purpose of the study (without revealing specific hypotheses) was explained.

(2) Informed Consent: Participants read and signed an informed consent form detailing their rights, the voluntary nature of participation, and data confidentiality.

(3) Pre-Experiment Questions: Participants completed a questionnaire collecting demographic information (age, gender), visual correction use (glasses/contacts), and prior experience and preferences regarding interface themes (light/dark mode usage, switching frequency).

(4) Setup and Calibration: Participants were seated comfortably in front of the monitor equipped with the eye tracker. They were instructed to maintain a relatively stable head position. The Tobii Pro eye-tracking system was then calibrated for each participant. Calibration was repeated if the initial attempt was unsuccessful, ensuring adequate tracking quality before proceeding.

(5) Task Trials (Repeated): Once calibration was successful, participants were informed that the main task trials would begin. This step was repeated six times, once for each unique combination of dashboard theme (light/dark) and task complexity (Easy/Medium/Hard), presented in a randomized order across participants to minimize order effects and potential biases. The order of dashboards and themes can be found in Figure 5. Each trial consisted of:

(5.1) Dashboard Presentation and Task Execution: A dashboard visualization was displayed. Participants analyzed the information to determine whether the subject profile belonged to Group 1 or Group 2, based on the criteria described in Section 3.3. Eye movements were recorded from the moment the dashboard appeared until the participant signaled task completion (verbally or with a gesture, as instructed).

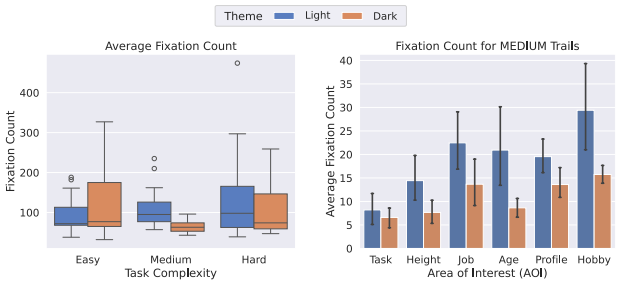


Figure 4: Fixation count grouped by theme and task complexity and averaged across all participants and trials (Left) as well as Areas of Interest for medium Trials (Right). The boxplot (Left) shows the IQR with fences at 1.5 times the IQR, while the barplot (Right) displays standard deviation error bars representing variability.

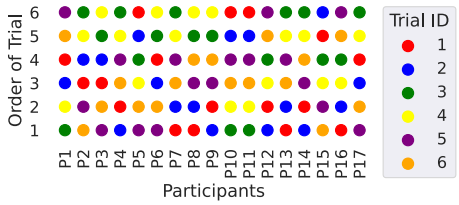


Figure 5: Overview of dashboard theme order for each participant. Colors indicate trial ID (see Table 1), while the y-axis indicates the order (1 - first to 6 - last) across all participants.

(5.2) Task Answers and SEQ: Immediately after signaling completion, the dashboard was replaced by a survey prompt on the screen. Participants verbally reported their answer (Group 1 or Group 2) and their confidence in that answer on a scale from 0 (not confident at all) to 10 (completely confident), followed by a Single Ease Question (SEQ) regarding the task they just finished.

(5.3) NASA-TLX: Following the task response, participants verbally completed the NASA-TLX questionnaire items presented on the screen.

(6) Post-Experiment Questions: After completing all six dashboard trials, participants answered final questions, primarily involving ranking the contribution of each NASA-TLX dimension to their overall perceived workload during the experiment.

Throughout the experiment, we collected four main types of data: (i) demographic and preference data (from questionnaires), (ii) gaze data (eye-tracking metrics), (iii) task performance data (completion time, accuracy, confidence), and (iv) subjective workload and usability data (NASA-TLX, SEQ). The independent variables manipulated across trials were the **theme mode** (light vs. dark) and the **task complexity** (Easy vs. Medium vs. Hard).

4 Results

This section presents the results of our study on how theme mode $T \in \{light, dark\}$ and task complexity $C \in \{easy, medium, hard\}$ affect participants' gaze behavior (fixations, pupil dilation), perceived

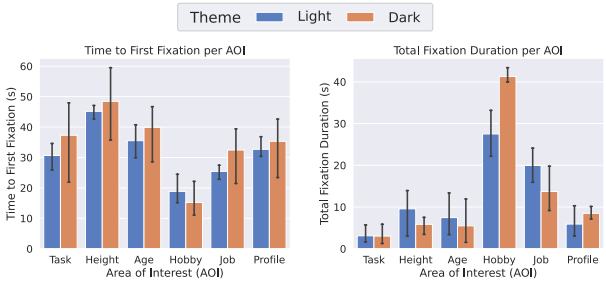


Figure 6: Time to first fixation (Left) and the total fixation duration (Right) on each area of interest (AOI) averaged across participants and trials. Error bars show the ± standard deviation.

workload (NASA-TLX, SEQ), and task performance (accuracy, completion time). We first analyze task completion measures (time, confidence, and accuracy), followed by eye-tracking measures, and concluding with questionnaire data. As noted in Section 3.1, we used N=13 for pupil analysis (due to bad data recording) and N=15 for all other measures. For statistical analysis, we first assessed normality using the Shapiro-Wilk test (W) and homogeneity of variance with Levene's test (F_L). If normality was violated the Mann-Whitney U test (U) was used; otherwise, an ANOVA test was conducted.

Task Completion Time (TCT) is defined as the duration of a trial, measured from the moment the dashboard appears until the participant completes the trial. In Figure 3a you can see the average trial duration across complexity levels (C) and theme modes (T). Overall, no significant ($U : p > 0.05$) effect of T and C on TCT was found. Generally TCT increased with difficulty for the light theme, but no clear trend emerged for the dark theme.

Task Accuracy was calculated as the average across all participants (see Figure 3c). The Shapiro-Wilk test indicated a violation of normality ($W : p < 0.05$), and Levene's test showed a significant violation of homogeneity of variance ($F_L : W = 4.238, p = 0.003$). Results did not show significant differences for the easy ($U = 84.5, p = 1.000$) and hard ($U = 65.0, p = 0.237$) conditions. However, the Mann-Whitney U test revealed a significant effect of theme on accuracy for medium task complexity ($U = 117.0, p = 0.016$).

Task Confidence was reported by participants after giving the task answer in each trial. Figure 3b shows the average confidence score grouped by theme (T) and complexity levels (C). The confidence scores for easy (light & dark) and medium (light) C violated normality ($W : p < 0.05$). For hard (light & dark) and medium (dark) C normality was given ($W : p > 0.05$). Levene's test indicated a significant violation of the assumption of homogeneity of variance ($F_L : p = 0.077$). The Mann-Whitney U test revealed no significant differences for easy ($U = 86.0, p = 0.955$) and hard ($U = 87.5, p = 0.896$), but a significant difference for medium ($U = 132.0, p = 0.015$), indicating that theme affected confidence for in medium task complexity.

Fixation Count was measured for each trial and averaged across participants (see Figure 4). We examined fixation count across task complexity (C) and theme mode (T). Normality was violated in most cases, except for medium C and dark T. Levene's test showed no significant variance violation ($F_L : W = 1.221, p = 0.304$). The Mann-Whitney U test revealed no significant effect of T for easy ($U = 112.0, p = 1.0$) and hard ($U = 102.0, p = 0.678$) complexity, but a significant effect was found for medium ($U = 191.0, p = 0.001$), indicating that T impacted fixation counts at medium complexity. The chart on the right in Figure 4 shows that for medium complexity trials, the fixation count was consistently higher for the light theme across all areas of interest (AOIs).

Time to first fixation (TFF) is defined as the duration from the start of the trial to the participant's first fixation on an Area of Interest (AOI), see Figure 1. The left plot in Figure 6 shows the average TFF. We conducted a one-way ANOVA to assess the effect of theme (T) on TFF. The Shapiro-Wilk test confirmed normality for both themes (dark: $W = 0.950, p = 0.421$; light: $W = 0.962, p = 0.647$). Levene's test showed no significant violation of homogeneity of variance ($W = 2.415, p = 0.129$). The result indicated no significant effect of theme on TFF, $F(1, 34) = 0.502, p = 0.484$.

Total Fixation Duration (TFD) is the sum of the duration of all fixations made on areas of interest (AOIs) during a trial (see Figure 6 right). We analyzed the effect of theme (T) and complexity (C) on the TFD. After normalizing the data, the Shapiro-Wilk test confirmed normality across all conditions ($W : p > 0.05$), and Levene's test indicated no significant variance violations ($W = 0.351, p = 0.843$). Two-way ANOVA results showed no significant main effects of C ($F(2, 30) = 0.006, p = 0.994$) or T ($F(1, 30) = 0.055, p = 0.816$), nor a significant interaction effect ($F(2, 30) = 0.088, p = 0.916$).

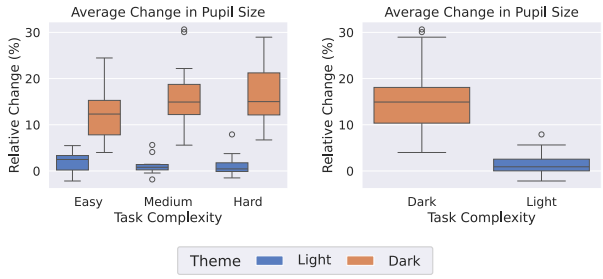


Figure 7: Average relative change in pupil size (%) from baseline to post-task phase for each participant and trial. (Left) Grouped by task complexity across themes. (Right) Grouped by theme. Boxplots show the IQR with fences at 1.5 times the IQR, and outliers as circles beyond the fences.

Pupil Dilation was measured as the relative percentage change in pupil size from baseline to post-task. As noted in Seection 3.1, two additional participants were excluded due to poor pupil measurements, leaving N=13 total participants. Pupil size change was calculated by first filtering outliers (5th–95th percentiles) and removing the first second of recordings to avoid display transition effects. We also evaluated the data after removing the first 2s and 3s of recordings, but the trends remained consistent. Baseline pupil size was averaged over the next 5 seconds, and post-task size was averaged from then until the end of each trial. The relative percentage difference between baseline and post-task values determined pupil change, repeated for each participant and stimulus. We analyzed pupil change across complexity levels (C) and theme modes (T), see Figure 7. Normality (Shapiro-Wilk test) was given for almost all cases ($W : p > 0.05$), except for hard task complexity with light theme ($W = 0.840, p = 0.021$). Levene's test indicated significant variance differences ($F_L : W = 10.018, p = 0.0$), so we conducted the Mann-Whitney U test, which revealed that theme had a significant ($U : p = 0.000$) effect on pupil dilation across all C.

Questionnaires included the Single Ease Question (SEQ) and NASA-TLX, along with dimension ranking. Looking at the average ranking for each dimension of NASA-TLX, we found that most participants reported mental demand as the largest contributor to their perceived workload, followed by performance and effort. In contrast, physical demand and frustration were ranked the lowest on average. In Figure 8, we observe that the workload increased on average from easy to hard tasks. We analyzed workload scores across task complexity (three levels) and themes (two levels). The Shapiro-Wilk test confirmed normality for all conditions ($W : p > 0.05$), and Levene's test indicated no violation of variance homogeneity

(F_L : $p = 0.932$). A two-way ANOVA was performed to examine the effects of task complexity and theme mode on workload scores. The analysis revealed a significant main effect of task complexity, $F(2, 84) = 4.62, p = 0.012$, and a significant main effect of theme mode, $F(1, 84) = 4.46, p = 0.038$. However, the interaction effect between task complexity and theme mode was not significant, $F(2, 84) = 1.19, p = 0.311$. For the Single Ease Question (SEQ) (see Figure 9), participants reported the highest scores for the easy level, followed by medium and hard. A higher SEQ score indicates a task was perceived as easier, while a lower score suggests greater difficulty. On average the dark themed trials were perceived easier than the light themed trials. The Shapiro-Wilk test indicated violations of normality in the easy ($p = 0.031$) and medium ($p = 0.001$) task complexity with dark theme. Levene's test revealed a significant violation of variance homogeneity ($p = 0.025$), necessitating the use of the Mann-Whitney U test. No significant differences were found between theme modes across easy ($U = 110.500, p = 0.949$), medium ($U = 137.0, p = 0.308$), or hard ($U = 83.0, p = 0.218$) tasks.

5 Discussion

This section discusses the study's findings, focusing on the effectiveness, efficiency, and cognitive load associated with task performance across different theme modes (Dark and Light) and task complexity levels (Easy, Medium, Hard).

5.1 Effectiveness

Task Accuracy: As expected, accuracy generally decreased with increasing task complexity. For the easy tasks, both themes performed similarly. However, for hard and medium complexity tasks, the dark theme consistently outperformed the light theme in terms of task accuracy. This aligns with the findings of [While and Sarvghad 2024], who investigated the effect of theme mode among different age groups. Specifically, for the young adult group, they observed an improved performance in dark mode.

Task Confidence: Similar to the accuracy results, participants' confidence slightly decreased with increasing task complexity. Theme mode significantly affected confidence only at medium complexity, with the dark theme leading to higher confidence. Thus, dashboard

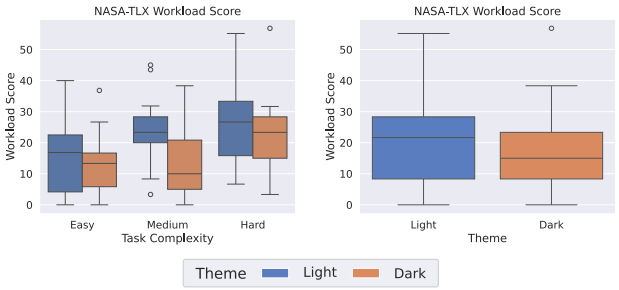


Figure 8: NASA-TLX workload score by complexity level and theme mode, averaged across trials and participants. Boxplots show the IQR with fences at 1.5 times the IQR, and outliers are displayed as individual circles beyond the fences.

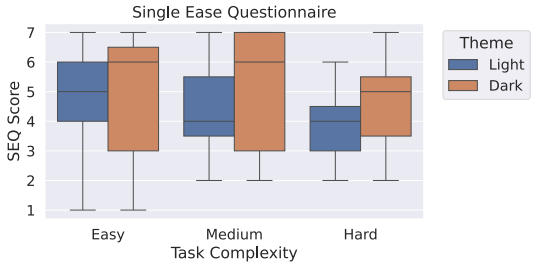


Figure 9: Average score for the Single Ease Question (SEQ) by complexity level and theme mode. Boxplot shows the IQR with fences at 1.5 times the IQR.

theme might have an effect on participants' confidence depending on the complexity of the task.

Fixation Count: Fixation count was significantly higher for the light theme (again) only at the medium complexity level, aligning with previous findings from [Qiao and Wu 2023]. This suggests that dark mode may facilitate more efficient processing, requiring fewer fixations while still achieving higher task accuracy (see Figure 3c) in medium task complexity on dark mode trials.

5.2 Efficiency

Task Compleation Time (TCT): No significant differences were found between theme modes or task complexity levels. However, for medium complexity trials, these findings align with [Qiao and Wu 2023], who observed shorter TCT in dark mode at night compared to light mode, suggesting that dark mode may enhance task efficiency under specific conditions.

Time to first fixation (TFF): We observed (see Figure 6) that the time to first fixation (TFF) was shortest for the Job and Height AOIs, likely due to their central placement on the screen. In light mode, we observe a trend of faster TFF, which aligns with the findings of [Qiao and Wu 2023], who reported faster TFF in light mode, suggesting it may facilitate quicker visual attention allocation. However, theme mode did not significantly affect time to first fixation in our study.

5.3 Cognitive Load

Questionnaires: NASA-TLX results show that task complexity significantly affected perceived workload, with higher workload reported for hard trials and lower workload for easy trials. Participants also found the dark theme significantly less mentally demanding than the light theme, consistent with [Sethi and Ziat 2023], who found lower workload in dark mode for text-based tasks. The interaction between task complexity and theme mode was not significant. Most participants perceived lower complexity in dark mode than on light themed dashboards, however the effect of theme on the SEQ score was not significant.

Pupil Dilation: In our study, pupil dilation, which is often associated with cognitive load [Dong and Liao 2016], was significantly affected by theme mode in all task complexities. The relative change in pupil size was constantly higher in the dark mode compared to the light mode, suggesting a greater impact of the dark mode on cognitive load. However, these results do not support the perceived

workload (NASA-TLX) scores. We carried out experimental sessions in a relatively dimly lit room with constant ambient illumination condition. We consider that the effect of the dark mode on pupil dilation was therefore more prominent than that in the light mode. Thus, the difference between measured (i.e., pupil dilation) and perceived workload measurements can be attributed to the interaction between ambient illumination and the mode of visualization. However, further investigations would be necessary to reveal the main cause of these findings.

5.4 Limitations and Future Work

Building upon these findings, several promising research directions can be listed to enhance the current understanding of the impact of light and dark themes on the design of dashboards.

Firstly, replicating this study with a larger and potentially more diverse participant pool is recommended. While our findings provide initial insights, a larger sample size would enhance statistical power and increase confidence in the generalizability of the results across user groups and preferences. Furthermore, future research should focus on in-depth analysis of eye-tracking metrics beyond fixation counts and pupil change. Analyzing scan paths could reveal crucial details about users' sequential information gathering strategies. Comparing how users navigate between different AOIs in light vs. dark theme, especially during complex tasks, might uncover systematic differences in visual search patterns and cognitive processing. Additionally, examining dwell times could complement fixation counts by indicating the duration of attention within each AOI. Finally, the scope of the investigation could be broadened by exploring different dashboard designs (e.g., varying density, chart types, interactivity), various task types (e.g., monitoring, exploration, specific data extraction), and varying ambient lighting conditions, to study how these factors interact with theme choice, extending related work by [Qiao and Wu 2023; Sethi and Ziat 2023]. Studying specific user populations, such as older adults or users with visual impairments, would also be valuable for accessibility considerations.

Pursuing these research directions would contribute significantly to developing nuanced, evidence-based guidelines for dashboard theme selection, ultimately helping designers create interfaces that are both effective and comfortable for a wider range of users.

6 Conclusion

This study investigated the effects of light and dark themes on user performance, perceived workload, and eye-tracking metrics during decision-making tasks on dashboards of varying complexity. Our findings reveal a nuanced picture: dark mode consistently resulted in significantly lower perceived workload, as measured by the NASA-TLX, compared to light mode across all task complexities. While task completion time did not differ significantly between themes, dark mode demonstrated clear advantages under medium task complexity, leading to significantly higher task accuracy, greater user confidence, and fewer fixations. This suggests potentially more efficient visual processing in dark mode under moderate challenge. No significant differences in accuracy or fixation count were found for easy or hard tasks, although accuracy trends favored dark mode. Furthermore, pupil dilation relative to baseline showed significant differences between the two modes among all complexity levels.

The primary contribution of this work is to provide empirical evidence using eye-tracking and subjective measures that highlight the interactions between visual theme, task complexity, performance, and user experience in a dashboard context. Although dark mode did not universally improve objective metrics like speed or accuracy among all conditions in our study, its consistent benefit in reducing subjective workload and its performance advantage under medium complexity offer valuable data for designers. These results underscore the importance of considering both performance metrics and user comfort. However, these conclusions should be considered in light of the limitations of our study. The findings are based on a specific analytical task and a dashboard layout. The study was carried out with volunteers and within a controlled and dimly lit laboratory environment. These factors may influence the direct generalizability of the results to different real-world scenarios, user groups, or ambient lighting conditions. These limitations, along with detailed findings, motivate further investigation.

Acknowledgments

We thank our colleagues Jonas Krogedal and Francesco De Blasi for their involvement during the course project that provided the foundations for this paper. The study is partially supported through a grant from the Research Committee of the HOCH Health Ostschweiz (Grant number 24/25).

References

A. Aleman, M. Wang, and F. Schaeffel. 2018. Reading and Myopia: Contrast Polarity Matters. *Scientific Reports 8* (2018). doi:10.1038/s41598-018-28904-x

Nancy Ivette Arana-De las Casas, Jorge De la Riva-Rodríguez, Aide Aracely Maldonado-Macías, and David Sáenz-Zamarrón. 2023. Cognitive Analyses for Interface Design Using Dual N-Back Tasks for Mental Workload (MWL) Evaluation. *International Journal of Environmental Research and Public Health* 20, 2 (2023). doi:10.3390/ijerph20021184

Jennifer Romano Bergstrom and Andrew Schall. 2014. *Eye tracking in user experience design.* Elsevier. 55 pages.

Tanja Blascheck, Kuno Kurzhals, Michael Raschke, Michael Burch, Daniel Weiskopf, and Thomas Ertl. 2014. State-of-the-art of visualization for eye tracking data.. In *Eurovis (stars).* 29. doi:10.2312/eurovisstar.20141173

Iris Mund Cosima Piepenbrock, Susanne Mayr and Axel Buchner. 2013. Positive display polarity is advantageous for both younger and older adults. *Ergonomics* 56, 7 (2013), 1116–1124. doi:10.1080/00140139.2013.790485 PMID: 23654206.

Aba-Sah Dadzie, Emmanuel Pietriga, Bo Fu, Natalya F. Noy, and Margaret-Anne Storey. 2016. Eye tracking the user experience – An evaluation of ontology visualization techniques. *Semantic Web* 8, 1 (2016), 23–41. doi:10.3233/SW-140163

W. Dong and H. Liao. 2016. Eye tracking to explore the impacts of photorealistic 3d representations in pedstrian navigation performance. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLI-B2 (2016), 641–645. doi:10.5194/isprs-archives-XLI-B2-641-2016

Austin Erickson, Kangsoo Kim, Alexis Lambert, Gerd Bruder, Michael P. Browne, and Gregory F. Welch. 2021. An Extended Analysis on the Benefits of Dark Mode User Interfaces in Optical See-Through Head-Mounted Displays. *ACM Trans. Appl. Percept.* 18, 3, Article 12 (May 2021), 22 pages. doi:10.1145/3456874

Stephen Few. 2013. *Information Dashboard Design: Displaying Data for At-a-glance Monitoring.* Analytics Press.

Sandra G. Hart. 2006. Nasa-Task Load Index (NASA-TLX); 20 Years Later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 50, 9 (2006), 904–908. doi:10.1177/154193120605000909

Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Advances in Psychology*, Peter A. Hancock and Najmedin Meshkati (Eds.). Human Mental Workload, Vol. 52. North-Holland, 139–183. doi:10.1016/S0166-4115(08)62386-9

Kenneth Holmqvist, Marcus Nyström, Richard Andersson, Richard Dewhurst, Halszka Jarodzka, and Joost Van de Weijer. 2011. *Eye tracking: A comprehensive guide to methods and measures.* Oxford.

Damian Hostettler, Kenan Bektaş, and Simon Mayer. 2023. Pupillometry for Measuring User Response to Movement of an Industrial Robot. In *Proceedings of the 2023 Symposium on Eye Tracking Research and Applications* (Tubingen, Germany) *(ETRA '23)*. Association for Computing Machinery, New York, NY, USA, Article 52, 2 pages. doi:10.1145/3588015.3590123

Robert J. K. Jacob. 1995. *Eye tracking in advanced interface design.* Oxford University Press, Inc., USA, 258–288.

Suyu Ma, Chunyang Chen, Hourieh Khalajzadeh, and John Grundy. 2023. A First Look at Dark Mode in Real-world Android Apps. *ACM Trans. Softw. Eng. Methodol.* 33, 1, Article 3 (Nov. 2023), 26 pages. doi:10.1145/3604607

Lige Qiao and Mingguang Wu. 2023. Light mode and dark mode: Which one is suitable when using public-facing web maps? An experimental evaluation using eye-tracking. *Transactions in GIS* 27 (03 2023). doi:10.1111/tgis.13038

Tara Sethi and Mounia Ziat. 2023. Dark mode vogue: Do light-on-dark displays have measurable benefits to users? *Ergonomics* 66, 12 (2023), 1814–1828. doi:10.1080/00140139.2022.2160879

Zack While and Ali Sarvghad. 2024. Dark Mode or Light Mode? Exploring the Impact of Contrast Polarity on Visualization Performance Between Age Groups. In *2024 IEEE Visualization and Visual Analytics (VIS)*. IEEE, 211–215. doi:10.1109/vis55277.2024.00050

Xiaojiao Xie, Fanghao Song, Yan Liu, Shurui Wang, and Dong Yu. 2021. Study on the Effects of Display Color Mode and Luminance Contrast on Visual Fatigue. *IEEE Access* 9 (2021), 35915–35923. doi:10.1109/ACCESS.2021.3061770

Aiyi Yang, Chu Hiang Goh, and Lim Jing Yi. 2023. The Research into Dark Mode: A Systematic Review Using a Two-Stage Approach and the S-O-R Framework. *International Journal of Communication Networks and Information Security* (2023), 358–395. doi:10.22399/ijcesen.1415

Ogan M. Yigitbasioglu and Oana Velcu. 2012. A review of dashboards in performance management: Implications for design and research. *International Journal of Accounting Information Systems* 13, 1 (2012), 41–59. doi:10.1016/j.accinf.2011.08.002

Johannes Zagermann, Ulrike Pfeil, and Harald Reiterer. 2016. Measuring Cognitive Load using Eye Tracking Technology in Visual Computing. In *Proceedings of the Sixth Workshop on Beyond Time and Errors on Novel Evaluation Methods for Visualization* (Baltimore, MD, USA) *(BELIV '16)*. Association for Computing Machinery, New York, NY, USA, 78–85. doi:10.1145/2993901.2993908