Berkeley
UNIVERSITY OF CALIFORNIA

The Effect of Background/Foreground Color on User Productivity

Jacob Sycoff
UC Berkeley MIDS
jsycoff@berkeley.edu

Matt Wong
UC Berkeley MIDS
mbearw@berkeley.edu

Daphne Yang
UC Berkeley MIDS
daphne.yang@berkeley.edu

Abstract

The effect of background contrast polarity has been debated for decades and with the wave of "dark mode" app settings debuting in many mobile applications, arguments have been made for dark mode's impact on productivity or lack thereof. Conflicting literature (Cushman 1984 and Wang 2003) indicates that the causal effect of background polarity is not well understood. We conducted a field experiment through the UC Berkeley X-Lab and piloted through Amazon's MTurks platform to evaluate the causal effect of background contrast polarity on our operationalized definition of productivity. Our treatment consisted of 4 total versions of a survey. The survey consisted of questions pertaining to reading comprehension (through SAT reading comprehension questions), recall (through matrix memorization questions), visual acuity (through finding the difference questions), and logic (through pattern recognition questions). Our data suggests that there is no statistically significant effect of background contrast polarity on user productivity.

Introduction

1.1 Dark vs Light Mode in Popular Culture

In the past two years, the use of "dark mode" themes has become increasingly popular and many social media platforms and apps have begun to shift to include a toggle for "dark mode". According to *phoneArena.com*, "dark mode" was actually created as the by-product of technology in the early beginnings of personal computing. However, it was not until 2019 that the popularization of "dark mode" themes began to take off. 'Dark Mode" loosely refers to display settings and

color schemes that are primarily darker colored backgrounds accented by vibrant pops of color in text and borders, also referred to as negative contrast polarity. Use of Dark mode became a common alternative to the traditional white/grey backgrounds in web pages and apps. Personal accounts from dark mode enthusiasts claim that dark mode helps to reduce eye fatigue and increase productivity and studies from companies like Twitter claim that dark mode helps to maintain user engagement for longer periods of time.

1.2. Existing Literature

Currently, there is conflicting information and claims around the benefit or drawback of using variations in background contrast polarity. Cushman (1986) found that subjects exposed to content with negative contrast polarity (light character text on dark backgrounds) performed better in reading tasks with less visual fatigue, making a case for the use of dark mode. However, Wang (2003) found that subjects performed better in some comprehension tasks when content was shown with negative contrast polarity (dark text on a light background). Existing literature evaluates the impact of background contrast polarity on visual fatigue and acuity; however, our research aims to measure the causal effect of background contrast polarity on an operationalized definition of "productivity".

1.3 Definition of Productivity

We operationalized productivity as the summative quantity of accuracy scores across 3 tasks (reading comprehension, visual acuity/error detection, pattern recognition).

$$\text{productivity} = \sum^{t} \frac{n \text{ correct answers}}{k \text{ total questions in task, } t}$$

Additionally, the inconsistency in the existing literature surrounding a user's "productivity" in the differently themed environments informs our research into the effects of background contrast polarity on user productivity.

1.4 Research Question and Hypothesis

The exact research question will be: Is there a significant effect of contrast and polarity on user productivity?

We believe that with conflicting evidence in the literature, there is no increase in productivity as a result of background contrast polarity.

Methods

2.1 Participants

In this experiment, a total of 270 individuals were recruited to participate in the study. This sample included UC Berkeley community members including graduate students, undergraduate students, and faculty members as well as Amazon MTurks workers. After applying our inclusion and exclusion criteria to the total group of participants, our study sample consisted of 189 UC Berkeley students (graduate students, undergraduate students, and faculty members). Our total study sample had a mean age of 25 (± 3.85 yrs). The participants were recruited as a part of a larger omnibus research survey for the graduate Experiments and Causal Inference Final Semester project.

Our final study sample consisted of the following participant sample:

Table 1: Summary of Participant Groups

Treatment/ Group	Light Mode	Dark Mode	Low Contrast	Neon Mode
Graduate	3	4	8	3
Undergraduate	43	42	36	40
Staff	3	1	0	2
Faculty	1	0	0	0
Total Participants	**50**	**47**	**44**	**45**

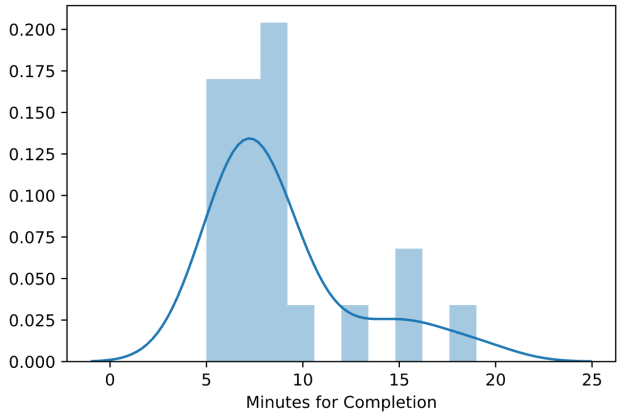
NOTE: This table shows the number of participants in each group, broken down by role at UC Berkeley

Our final study sample consisted mostly of undergraduates and although it is not representative of the general population, students spend a majority of their time in class and in settings where productivity is incredibly important. Therefore, we posit that this sample can still be used to determine the effect of background on productivity.

2.2 Inclusion/Exclusion Criteria

We collected data for all 270 individuals who completed our survey. We then applied our inclusion criteria to only analyze data from qualifying participants. We chose to undergo this data collection method because our survey was included as part of a larger omnibus survey used to collect more participant data through the UC Berkeley X-Lab. Our inclusion criteria for our sample was based on the results of our pilot study, conducted through Amazon's MTurks and powered by Qualtrics survey.

Figure 1: Pilot Study - Survey Completion Time



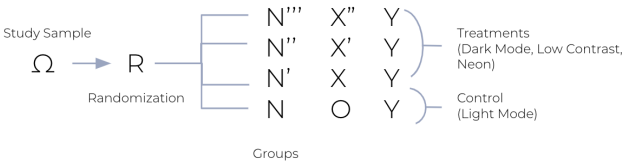
Note: Distribution of minutes to complete the initial pilot study

Pilot data showed variance in survey completion time (as shown in **Figure 1**) as well as low accuracy scores. Therefore, we defined our inclusion criteria as including all respondents with UC Berkeley affiliation.

2.3 Potential Outcomes Notation

We began designing our experiment by using potential outcomes notation (as shown in **Figure 2**). We started with our study sample, Ω, and randomly selected individuals into one of four groups, denoted in the figure as N*. From these groups, we assigned three groups to be treatment groups and one to remain as the control. We then collected their outcomes, Y, in the form of accuracy scores.

Figure 2: Experimental Design as Potential Outcomes Notation



Note: Potential Outcomes Notations for the study design. (R:Randomization, N: Group, X: Treatment, O: Control, Y: Outcome)

2.4 Statistical Power

To conduct this experiment, we started by conducting a power calculation to understand how our budget constraints might lead to a decrease in the statistical power of findings. To determine our statistical power, we calculated the minimum sample size needed in each of the 4 groups to obtain a power of 80% with a moderate effect size of 0.25 and a standard confidence level of 95%. To achieve this level, we determined that we would need at least 45 individuals in each group.

Figure 3: Output of our Power Analysis

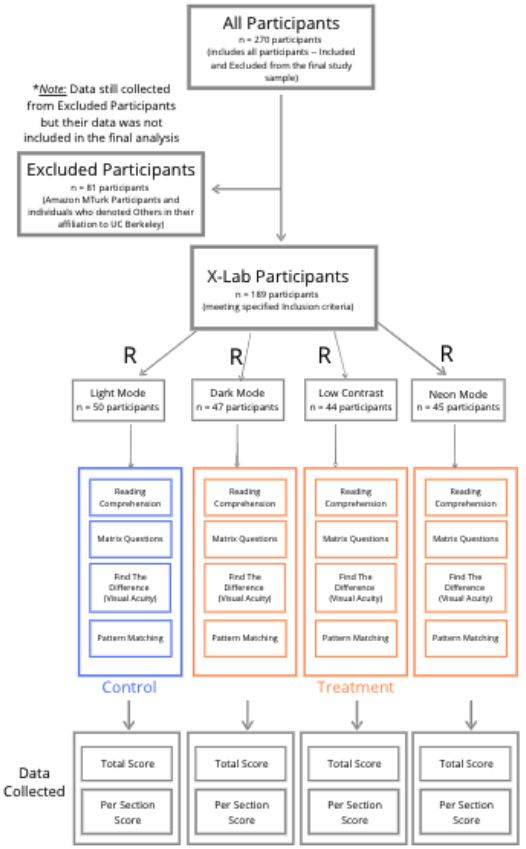
```
     Balanced one-way
analysis of variance power
calculation

              k = 4
              n = 44.59927
              f = 0.25
      sig.level = 0.05
          power = 0.8

NOTE: n is number in each
group
```

NOTE: The output of our power analysis indicates each group needed a minimum size of 45 participants each. Actual sample sizes with the exception of one group met the minimum size for power

2.5 Experimental Procedure

To conduct this experiment, we split our population into four treatment groups: Light Mode (Control Group), Dark Mode (Treatment 1), Low Contrast (Treatment 2), and Neon mode (Treatment 3). We included Low Contrast and Neon Mode as additional groups in our experimental design because we wanted to ensure that our experiment was detecting the effect of dark mode and light mode-- not just the effect of any change to the background (also known as the Hawthorne Effect). We then subject each group to a series of tasks to test for productivity.

Figure 4: CONSORT Documentation and Flowchart of Proposed Experimental Design



NOTE: The CONSORT diagram of the experimental design including counts of individuals in each group and the counts of people who were excluded from the study

2.5.1 Task Sections

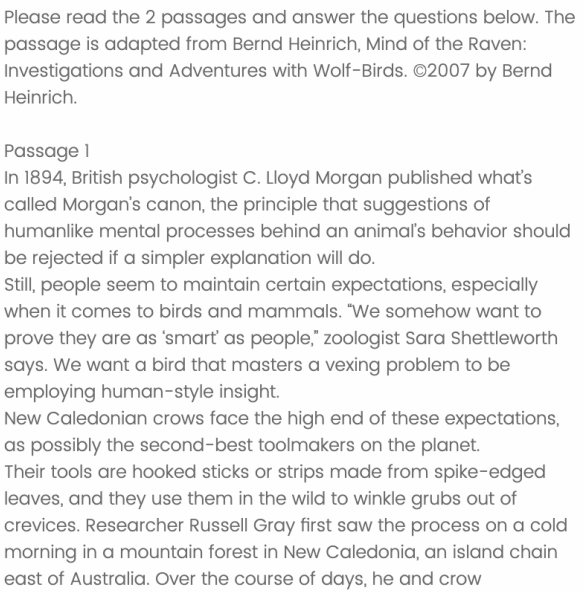
Each of the groups were subject to slight variations in treatment, but all received the same ordering of sections and received the same questions. However, the ordering of the multiple choice selections was randomized.

Reading Comprehension Section

Each group received 4 multiple choice questions for reading comprehension that were sourced from current PSAT testing

material (light mode sample below in **Figure 5**). This section was intended to test for any effect of different backgrounds on comprehension tasks. Questions asked in this section required more than a surface understanding of the text. The data collected from this section was the number of questions correct for each participant.

Figure 5: Example Light Mode Reading Comprehension Question


Please read the 2 passages and answer the questions below. The passage is adapted from Bernd Heinrich, Mind of the Raven: Investigations and Adventures with Wolf-Birds. ©2007 by Bernd Heinrich.

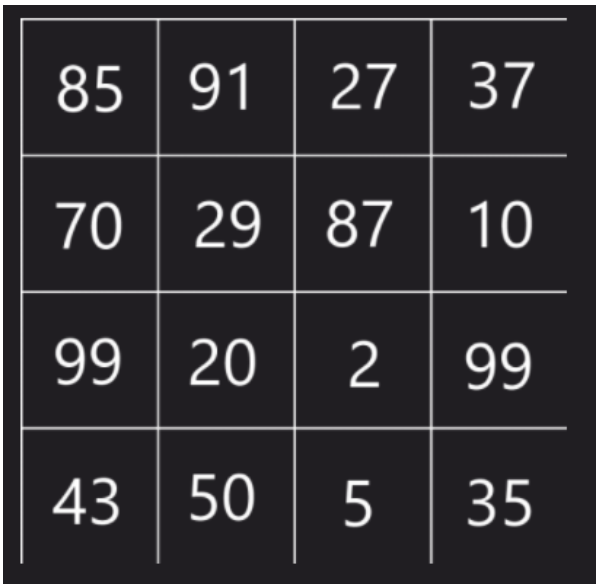
Passage 1
In 1894, British psychologist C. Lloyd Morgan published what's called Morgan's canon, the principle that suggestions of humanlike mental processes behind an animal's behavior should be rejected if a simpler explanation will do.
Still, people seem to maintain certain expectations, especially when it comes to birds and mammals. "We somehow want to prove they are as 'smart' as people," zoologist Sara Shettleworth says. We want a bird that masters a vexing problem to be employing human-style insight.
New Caledonian crows face the high end of these expectations, as possibly the second-best toolmakers on the planet.
Their tools are hooked sticks or strips made from spike-edged leaves, and they use them in the wild to winkle grubs out of crevices. Researcher Russell Gray first saw the process on a cold morning in a mountain forest in New Caledonia, an island chain east of Australia. Over the course of days, he and crow

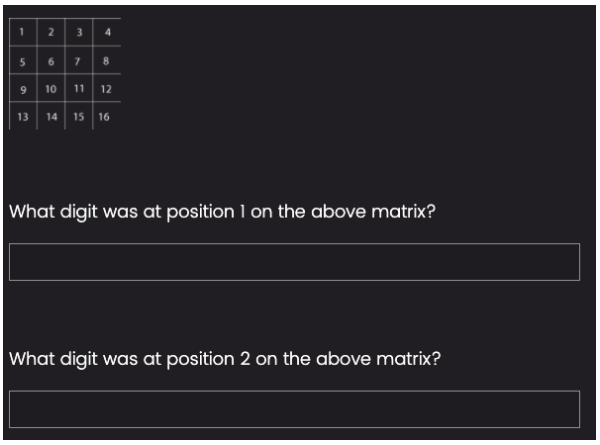
NOTE: The data collected from the reading comprehension section consists of the average score across all 4 questions for each participant.

Matrix Memorization Section

After the completion of this section, participants were then asked to look at a matrix of randomly generated numbers for one minute and were then asked to complete a series of fill in the blank questions to recreate the matrix to the best of their ability (dark mode sample below in **Figure 6**). This section was intended to test for any effect of different backgrounds on recall tasks. The data collected from this section was the number of correct digits remembered for each participant.

Figure 6: Example Dark Mode Matrix Memorization Question





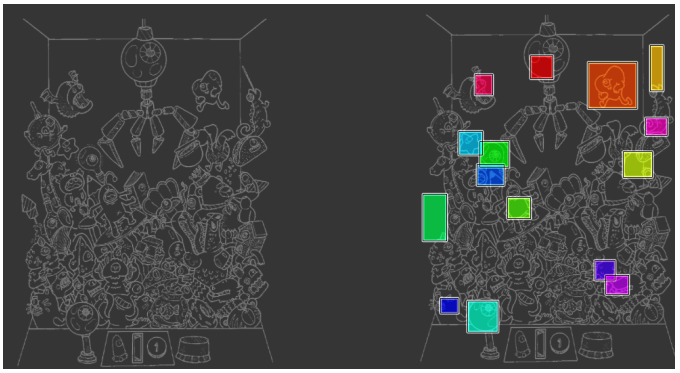
NOTE: The data collected from the matrix memorization section consists of the average score across all 16 fill-in-the-blank questions for each participant.

Find the Difference Section

Next, each participant was provided two juxtaposed images with fifteen differences

between the left image and the right image. The participants were asked to first denote the number of differences spotted in the pair and then asked to list the differences that they had noted (low contrast example with increased contrast below in **Figure 7**). This section was intended to test for any effect of different backgrounds on visual acuity/scrutiny tasks. The data collected from this section was the number of differences noted, and although we did not take the free response list of answers into consideration, we did include this question in the survey to ensure participants were not randomly guessing values.

Figure 7: Example Low Contrast *Find the Difference* with Highlighted valid regions

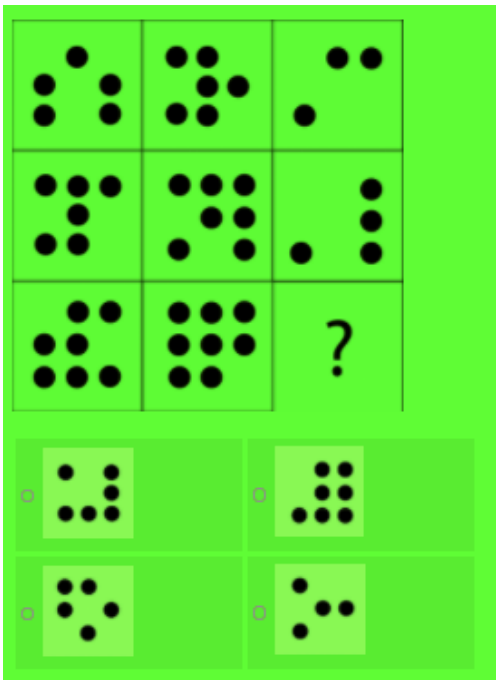


NOTE: The data collected from the find the difference section consists of the proportion of differences spotted for each participant.

Pattern Completion Section

Finally, the participants were asked to complete a pattern completion challenge that inquired about which pattern came next in a sequence (neon example in **Figure 8**). This section was intended to test for any effect of different backgrounds on pattern recognition and logic tasks. The data collected was a binary mapping to whether the participant had answered the question correctly or not.

Figure 8: Example Neon Pattern Completion Question



NOTE: The data collected from the pattern matching section consists of a dummy variable measuring whether the answer is correct for each participant.

Liar Detection

As presented in our MTurks pilot study data, there was evidence that participants were selecting random answers for responses instead of properly performing the tasks. We therefore implemented *liar questions* to assess a participant's completion of the tasks. These questions were not designed to be difficult but rather to be simple checks that participants were genuinely completing the

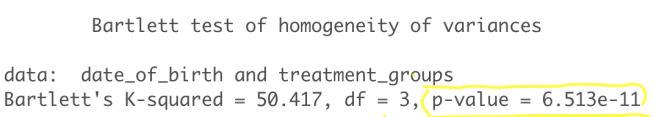
survey instead of just going through the motions in order to receive their incentives. We chose to not exclude participants that failed the *liar questions* because we believed that these questions may be indicative of intangibles important to our study. For example, a participant in low contrast may feel more inclined to go through the motions because of the higher level of difficulty than in an individual in other background settings.

2.6 Proper Randomization

Randomization for this experiment was conducted primarily through the Qualtrics randomization settings. In our experiment, each participant was randomly shown one of the four versions of our survey questions and was required to complete the full survey in their assigned setting.

Of note, our study design intended to have equally sized groups in each of the four treatments; however, by removing individuals who denoted themselves as "Other" or no affiliation to UC Berkeley from our sample, we were not able to ensure equally sized treatments and control groups. Additionally, we conducted covariate balance checks to ensure that our randomization was not compromised and that our groups were similar. The covariate balance check was conducted through the Bartlett Test of Homogeneity of Variance. Unfortunately, our groups failed the covariate balance check on year of birth (results shown in **Figure 9**). This meant that our groups were not randomized correctly and that the groups differed beyond just their exposure to a treatment/control. While we acknowledge that there was failure in the randomization process, we were not able to find a remedy for this problem ad-hoc and proceeded with our analysis.

Figure 9: Result of the Covariate Balance Check



NOTE: The low p-value indicates that the distribution of variances across the different groups is not equal and therefore, the groups differ in ways beyond just their exposure to the treatment.

Results

3.1 Exploratory Data Analysis

The Preliminary results, shown in Table 2, suggest several differences between groups for specific survey questions. Most of these, however, would prove insignificant upon statistical analysis.

Table 2: Preliminary Results

	Total Score	Reading Comprehension Score	Difference Score	Memory Matrix Score	Pattern Score	Honesty Score
Overall	0.47	0.56	0.32	0.51	0.49	0.72
Light Mode	0.49	0.53	0.36	0.54	0.55	0.74
Dark Mode	0.47	0.52	0.38	0.46	0.52	0.65
Low Contrast Mode	0.45	0.61	0.23	0.54	0.43	0.78
Neon Mode	0.47	0.59	0.32	0.52	0.43	0.70

NOTE: This table shows the final calculated scores for each section of the survey for each group in the experiment.

3.2 Regression Analysis

We begin our analysis by creating three models in an iterative fashion. First, we regressed the Overall Score (the sum of the scores from each question) on each of the treatment groups. Next, we wanted to see if there was an effect of simply being in any non light mode group, so a second model was created. Our results are shown in **Figure 10**. We hypothesized that the time of day when taking the survey would influence our treatment effect, so a third model was created that included the interaction with taking the test during the day or at night, shown in **Figure 11** in the following page.

Figure 10: Regression Table for Model 1 and 2

	Dependent variable:	
	Overall Score	
	(1)	(2)
Dark Mode	-0.022	
	(0.038)	
Low Constrast	-0.036	
	(0.039)	
Neon	-0.023	
	(0.039)	
Any Treatment		-0.027
		(0.031)
Intercept	0.488^{***}	0.488^{***}
	(0.027)	(0.027)
Observations	187	187
R^2	0.005	0.004
Adjusted R^2	-0.012	-0.001
Residual Std. Error	0.189 (df = 183)	0.188 (df = 185)
F Statistic	0.293 (df = 3; 183)	0.740 (df = 1; 185)
Note:		*p<0.1; **p<0.05; ***p<0.01

3.3 T-Test

To understand our treatment effect at the individual question level, we decided to run a Difference in Means T-test comparing our treatment groups to the control. In addition to the four main survey questions, we included one of the *liar questions* in our T-test analysis. Out of the fifteen tests performed, only one provided evidence to reject the null hypothesis with Benjamini-Hochberg adjusted P-value threshold: *Find the Difference* in Low Contrast Mode vs Light Mode.

Figure 11: Linear Regression w/ time of day interaction

	Dependent variable:
	Score
Light Mode	0.018
	(0.049)
Low Contrast	0.023
	(0.047)
Neon	0.030
	(0.047)
Night Test	0.090
	(0.058)
Light Mode & Night Test	-0.007
	(0.079)
Low Contrast & Night Test	-0.111
	(0.086)
Neon & Night Test	-0.092
	(0.084)
Constant	0.436***
	(0.034)
Observations	187
R^2	0.031
Adjusted R^2	-0.007
Residual Std. Error	0.188 (df = 179)
F Statistic	0.820 (df = 7; 179)
Note:	*p<0.1; **p<0.05; ***p<0.01

3.4 Multiple Comparisons Problem

By performing fifteen T-tests simultaneously we run into problems controlling our False Discovery Rate. In order to account for this, we use the Benjamini-Hochberg Procedure to adjust our P-value threshold to reject the null hypothesis as accurately as possible. This is done by ordering the T-test P-values from smallest to greatest and selecting the P-value that satisfies the equation below:

$$max(p_i < \frac{i\alpha}{M})$$

where *i* is the index of the ordered P-value, α is the previous P-value threshold (.05 in our case) and *M* is the total number of tests performed. With our correction, our new P-value threshold was **.0002914**. This is a marked decrease from our previous threshold and should be sufficient in controlling the FDR.

Discussion

4.1 Main Points

Based purely on the lack of statistically significant effects in any of our regression models and our T-tests of this study, we can say that there is no evidence that Dark mode (nor Low Contrast or Neon) provide a significant boost in productivity when compared to Light mode. In addition, we can say that Low Contrast is significantly worse than Light mode when performing *Find the Difference* tasks. Due to the specificity of the task however, we don't find this result to be too worthy of excitement. It does, however, suggest a degree of validity to our experiment, as the most obvious effect is able to be detected with our statistical methods

4.2 Experiment Improvements

Our future iteration of this experiment would see some key changes. Foremostly, we would aim to have more control over randomization and explore an in-person rather than at-home survey. In addition, we would add time of day into our groups, with an emphasis on

balanced group counts. Improvement on our operationalization of productivity with an improved *Find the Difference* section could allow for a more accurate measure of visual acuity. Incorporation of question timing might additionally shed light on another dimension of productivity and our treatment.

Our implementation of *liar questions* allowed for the detection of false answers and future iterations of this experiment may consider adding more.

4.3 Applications

Looking forward, this study does have practical use. The Dark mode & Night Test interaction in our third model had a slightly positive coefficient with a P-value of .124. While not statistically significant, this is a marker that a future experiment investigating Dark mode's effect at night would be a potentially useful endeavor.

Conclusion

As we spend an increasing amount of time interacting with computers, the optimization of productivity remains a topic of great import. We've seen in this study that it *is* possible to affect productivity in a negative way with a specific work environment. This provides hope that the reverse can also be uncovered.

References

Wang, An-Hsiang. *Effects of VDT leading-display design on visual performance of users in handling static and dynamic display information dual-tasks,* International Journal of Industrial Ergonomics, Volume 32, Issue 2, 2003, Pages 93-104.

Cushman, W. H. 1986. Reading from microfiche, a VDT, and a printed page: Subjective fatigue and performance. Hum. Factors 28,63-73.