**Reading in the Dark: A Matched-Pairs Experiment on Whether Dark Mode Slows Down On-Screen Reading**

Sriram Achaththekoot, Aiden Gie

Grade 11, Category: Grades 10-12

DuPont Manual High School, Louisville, KY

## 1. Introduction and Statistical Question

Dark mode, which shows light text on a dark background, is now a standard setting on almost every phone, and many of our classmates say it lets them read faster and with less eye strain. Research in human factors points the other way. For years, studies have reported a “positive polarity advantage,” meaning people usually do better with dark text on a light background. A 2007 set of experiments by Buchner and Baumgartner found this advantage for proofreading, and Piepenbrock, Mayr, Mund, and Buchner reported the same pattern in 2013 for both younger and older adults. Most of that work measured proofreading or visual sharpness, not the speed of ordinary, continuous reading on a phone, so we wanted to find out whether the advantage shows up in plain reading speed.

Our statistical question is: *For high school students at DuPont Manual High School, does reading a passage in dark mode take a different average amount of time than reading a similar passage in light mode, and if so, by how much?* The value we set out to estimate is the mean within-person difference in reading time, dark minus light. We wanted to know both whether that difference is large enough to detect and whether it is large enough to matter in everyday life.

We used a within-person, matched-pairs design for this project. Reading speed varies a lot from one person to the next. For reference, adults read silently at about 238 words per minute on average, but individual readers spread across a wide range (Brysbaert, 2019). If we compared two separate groups of people, that person-to-person variation would easily hide a small mode effect. By having each person read in both modes and comparing each reader only to himself or herself, we cancel out individual differences and focus on the display mode alone.

**2. Data Collection and Experimental Design**

We recruited 30 volunteers from our school, mostly classmates and students from a shared study hall. This is a convenience sample, not a random sample, so the safest group to generalize to is students similar to the ones who volunteered. The part of the study that does support a cause-and-effect claim is the random assignment described below, which was applied within each participant.

Every trial used the same 6.1-inch phone, with brightness fixed at 60% and adaptive brightness turned off. All sessions ran in the same windowless room under the same overhead lighting, so the screen and the surroundings stayed constant. We wrote two passages, A and B, each exactly 250 words, on unrelated everyday topics, and matched them for difficulty using the Flesch-Kincaid grade level (7.0 and 7.2) and a similar average sentence length. Either passage could be shown as black text on white (light mode) or white text on black (dark mode) at the same on-screen text size.

The response variable was reading time in seconds. We started timing when the reader tapped “start” and stopped when the reader said “done,” then read the time off a silent screen recording to the nearest 0.1 second.

Each participant read once in each mode. For every participant we flipped a coin twice. The first flip decided which mode came first, which spread out any practice or fatigue effect. The second flip decided whether passage A or B was shown in dark mode, which kept any small leftover difference in passage difficulty from lining up with one mode. Because a coin, not the experimenter, set these choices, a difference between the modes cannot be explained by one passage being harder or by people always reading the first passage faster. That is what lets us read the result as cause and effect.

A fast time means nothing if the reader is skimming. To guard against that, each reader answered three short comprehension questions after each passage. We decided ahead of time to re-run any trial that scored below 2 out of 3, and only one trial needed a redo. Comprehension stayed high in both modes (light 2.80 out of 3, dark 2.87 out of 3), so readers were reading in both conditions rather than racing through them.

Taking part was voluntary, and anyone could stop at any time. The task had no foreseeable risk, and we recorded no identifying information, since each reader was given an ID number only. The full data for all 30 participants appear in the Appendix.

**3. Displaying the Data**

Figure 1 shows the reading times for the two modes on the same vertical scale. The two distributions overlap almost completely. Person-to-person times run from about 44 to 95 seconds, a much wider spread than any gap between the modes. Looked at this way, ignoring the pairing, the modes appear nearly identical, and their medians are within 0.2 second of each other. That is the main reason an unpaired, two-group comparison would have inappropriate use here.

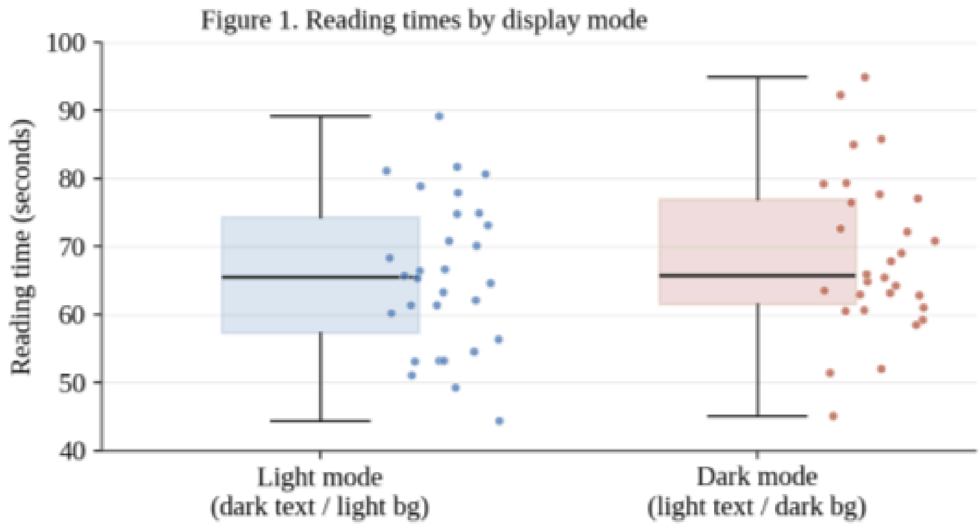
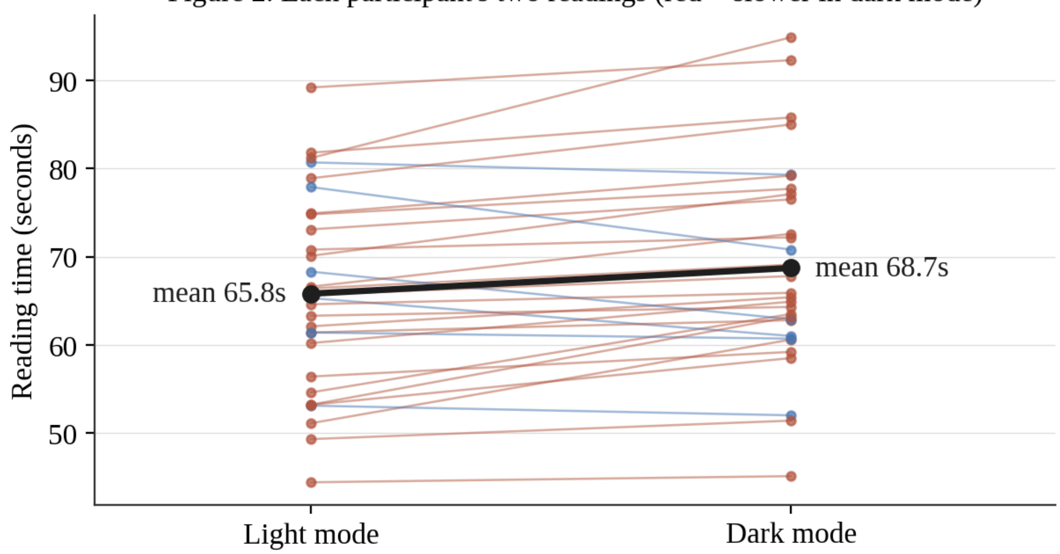



Figure 2 adds back the information the boxplots leave out by connecting each participant's two readings with a line. Now a pattern is visible: most lines slope upward from light to dark (shown in red), and the heavy line shows the group mean rising from 65.8 to 68.7 seconds.

Figure 2. Each participant's two readings (red = slower in dark mode)




**4. Analysis**

Table 1 summarizes both modes and the 30 paired differences. On average, dark mode took 2.92 seconds longer, with a standard deviation of the differences of 4.48 seconds, and 24 of the 30 participants (80%) were slower in dark mode. The two modes have almost the same median even though the mean difference is clearly positive. That is not a contradiction. A few readers were much slower in dark mode, which pulls the mean of the differences above zero while the two overall distributions still look alike.

*Table 1. Descriptive statistics for reading time by mode and for the within-person differences d = dark − light.*

| Variable | Mean (s) | SD (s) | Min (s) | Median (s) | Max (s) |
|---|---|---|---|---|---|
| Light mode | 65.8 | 11.1 | 44.4 | 65.5 | 89.2 |
| Dark mode | 68.7 | 11.7 | 45.1 | 65.7 | 94.9 |
| Differences (d) | 2.92 | 4.48 | −7.1 | 2.85 | 13.7 |

Let $\mu_d$ be the mean of the (dark − light) reading times for the population of readers our participants represent. We test $H_0$: $\mu_d = 0$, meaning mode makes no difference, against $H_a$: $\mu_d \neq 0$. We used a two-sided alternative because the earlier evidence for ordinary reading was mixed, so we did not assume in advance which mode would be faster.

First, independence and randomization: the order of the two conditions was set by a coin flip and participants worked on their own, so the 30 differences are independent. Second, the differences should be approximately normal. Figure 3a shows that they are fairly symmetric with no strong outliers, and the normal probability plot in Figure 3b stays close to a straight line, so this condition is reasonable. With 30 differences, the t-test also holds up well even when the distribution is not perfectly normal. Both conditions are met.

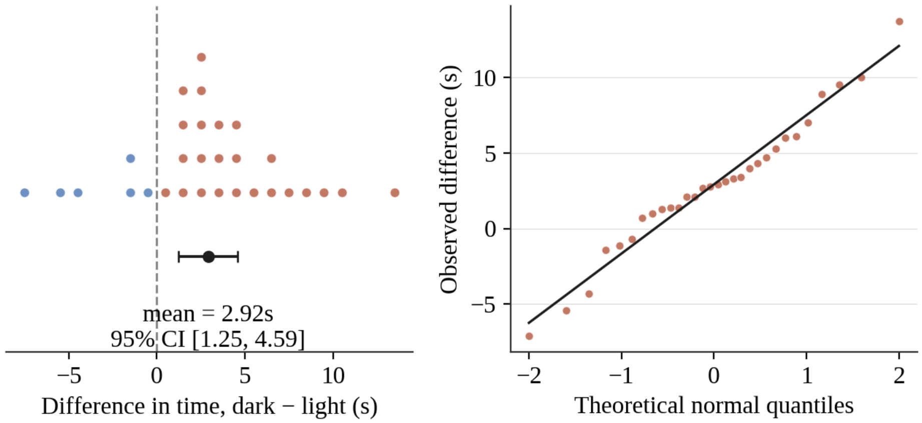



Figure 3a. Within-person differences

Figure 3b. Normal QQ plot of differences

The test statistic is $t = \bar{d} \div (s_d/\sqrt{n}) = 2.923 \div (4.476/\sqrt{30}) = 3.58$, with 29 degrees of freedom, and the two-sided p-value is 0.0012. Since this is well below 0.05, we reject $H_0$ and conclude that there is strong evidence display mode affects reading time. A 95% confidence interval for $\mu_d$ is $2.923 \pm 2.045 \times 0.817$, or about (1.25, 4.60) seconds. So we are 95% confident that, for readers like ours on this kind of short task, dark mode raises the mean reading time by roughly 1.3 to 4.6 seconds. The whole interval sits above zero, which agrees with the test.

The t-test assumes the differences are close to normal, so we also ran a check that does not. If display mode truly made no difference, then for each person the dark-minus-light value would be just as likely to be positive as negative. We modeled that idea by randomly flipping the sign of each of the 30 differences and recording the new mean, then repeating this 100,000 times. A mean as far from zero as our observed 2.92 seconds came up in only about 0.12% of those simulations, which is a p-value of about 0.0012. This nearly matches the t-test, so the conclusion does not depend on the normal assumption.

The result is clearly statistically significant, but the size is small in practical terms. An increase of 2.9 seconds is only about 4% of the 65.8-second light-mode average. We will look further in depth as to what this means in the conclusion.

**5. Conclusion**

Our data give strong and consistent evidence that dark mode slows down the reading of short passages for readers like our participants. The slowdown averaged about 2.9 seconds per 250-word passage, with a 95% confidence interval of 1.3 to 4.6 seconds, and a randomization test agreed with the t-test. This direction matches the positive polarity advantage from earlier research and extends it from proofreading to plain reading speed. Because we randomly assigned the order and the passage pairing within each person, we can attribute the slowdown to the display mode for these participants. Because the participants were volunteers and not a random sample, we are careful not to claim that the exact size carries over to all students. We also avoid saying the effect is proven, since statistics measure how strong the evidence is, and not certainty.

While there is a clear statistical effect, the practical application is narrow. Under five seconds on a one-minute read does not matter much for a quick text or a short caption, although over an hour of steady reading it could add up to a couple of minutes that some people might notice. Speed is also only one goal. Dark mode might still be the better choice for comfort in a dark room, for less glare, or for battery life, none of which we measured.

**6. Reflection**

The matched-pairs design was the best decision we made. Figure 1 shows that the differences between people were several times larger than the difference between modes, so a two-group study would probably have missed the effect. Pairing each reader with himself or herself revealed a clear signal with only 30 people. Counterbalancing both the order and the passage removed the two confounding variables we were most worried about, and the comprehension check kept a fast time from meaning a careless one. Reporting a confidence interval, and not just a p-value, is what let us separate statistical significance from practical importance.

The biggest weakness was the sample. Volunteers from one school are not a random sample, so the exact effect size may not hold for other ages, languages, or reading conditions. Timing from a spoken "done" adds a small reaction-time error. It should be about the same in both modes, but it adds noise and could be removed. We fixed brightness at one level and used a single phone and text size, but the polarity advantage is known to depend on brightness and lighting, so our number is tied to those settings. Two passages matched on readability are still not perfectly equal, and counterbalancing spreads that difference out without erasing it. Finally, 250 words is a short task, so any fatigue that builds up over long reading sessions could not show up.

A stronger version of this study would do four things. First, draw a random sample across grade levels instead of using volunteers. Second, replace the spoken "done" with an on-screen tap that is logged automatically, to cut timing noise. Third, add screen brightness and room lighting as a second factor in a two-factor experiment, to test whether dark mode's disadvantage shrinks or even reverses in a dark room. Fourth, use longer passages to capture effects that build up over time. Measuring comprehension accuracy and self-reported eye strain as well would let us weigh speed against comfort instead of treating speed as the whole story.

**Acknowledgments**

*We would like to thank our parent mentor, Muralee Achaththekoot, who suggested that we report a confidence interval alongside the hypothesis test and who reviewed a draft for clarity. We appreciate the opportunity to compete and display our statistical and research knowledge.*

**References**

Brysbaert, M. (2019). How many words do we read per minute? A review and meta-analysis of reading rate. Journal of Memory and Language, 109, 104047.

Buchner, A., & Baumgartner, N. (2007). Text–background polarity affects performance irrespective of ambient illumination and colour contrast. Ergonomics, 50(7), 1036–1063.

Piepenbrock, C., Mayr, S., Mund, I., & Buchner, A. (2013). Positive display polarity is advantageous for both younger and older adults. Ergonomics, 56(7), 1116–1124.

**Appendix: Raw Data (Reading Time in Seconds, All 30 Participants)**

*Diff = Dark − Light. Positive values mean the participant read more slowly in dark mode.*

| ID | Light | Dark | Diff | ID | Light | Dark | Diff |
|---|---|---|---|---|---|---|---|
| P01 | 74.8 | 77.7 | +2.9 | P16 | 70.8 | 72.2 | +1.4 |
| P02 | 73.1 | 76.5 | +3.4 | P17 | 44.4 | 45.1 | +0.7 |
| P03 | 54.6 | 63.5 | +8.9 | P18 | 62.1 | 65.4 | +3.3 |
| P04 | 51.1 | 60.6 | +9.5 | P19 | 81.8 | 85.8 | +4.0 |
| P05 | 66.4 | 69.1 | +2.7 | P20 | 56.4 | 59.2 | +2.8 |
| P06 | 80.7 | 79.3 | −1.4 | P21 | 61.4 | 60.7 | −0.7 |
| P07 | 81.2 | 94.9 | +13.7 | P22 | 65.7 | 67.8 | +2.1 |
| P08 | 74.9 | 79.2 | +4.3 | P23 | 49.3 | 51.4 | +2.1 |
| P09 | 70.1 | 77.1 | +7.0 | P24 | 60.2 | 64.9 | +4.7 |
| P10 | 89.2 | 92.3 | +3.1 | P25 | 68.3 | 62.9 | −5.4 |
| P11 | 78.9 | 85.0 | +6.1 | P26 | 66.6 | 72.6 | +6.0 |
| P12 | 65.3 | 61.0 | −4.3 | P27 | 53.2 | 58.5 | +5.3 |
| P13 | 53.1 | 52.0 | −1.1 | P28 | 64.6 | 65.9 | +1.3 |
| P14 | 61.4 | 62.8 | +1.4 | P29 | 77.9 | 70.8 | −7.1 |
| P15 | 63.3 | 64.3 | +1.0 | P30 | 53.2 | 63.2 | +10.0 |