Testing, testing, 1, 2, 3!

Exploring Different Conditions for Grading Test Strips Accurately

Emily Zhou and John Scott

MIDS 241, Fall 2022

Introduction

Lateral flow strips are a very quick, flexible, and inexpensive method of assaying – or determining valuable information about – a sample of interest. For decades, the most widely-available lateral flow assays were pregnancy strips, which measure the presence and concentration of hCG (human chorionic gonadotropin) in urine. In the past few years, the world has come into close contact with another type of lateral flow assay – the antigen test for COVID-19, where a person can tell if they have viral particles present in their mucous membranes. In general, when a person is asked to take a test at home, lateral flow is the preferred assay method because it is manufacturable en-masse at a low price, and is runnable by inexperienced users after reading a few minutes of simple instructions.

Figure 1: Left: A single cassette with one clearly visible line. This strip was imaged in the sunlight. Right: A simple way of assigning a value that represents the intensity of a test strip is to use a color scale such as the one shown above.

The assay method itself is simple, but a common occurrence during the "reading" (or grading/scoring) of the strip is ambiguity or confusion on the part of the user:

- *Is there a line present?*
- *If there is actually a line, how can we say how intense that line is?*
- *After all that, how do we relate the intensity of a test line to the concentration of the target of interest?*

Ways of reading test strips

Quantification methods exist on a spectrum from simplest and cheapest to most complicated and expensive. The simplest natural language descriptions (negative/positive) are common but inconsistent, and hard to relate to others in subsequent studies. This represents the low technological end of reading lateral flow strips. At the opposite end of the technological spectrum, an assay developer may want to use an electronic sensor and run an image processing algorithm to get an objective representation of the intensity. Between these two poles are intermediate ways of recording test strip values. One intermediate way of getting around this is to provide a printed color scale (see Fig ***??***), which shows example test lines numbered by their varying intensity. A user can write down the number of the example line that most closely matches the test strip.

Using the simplest language (negative/positive) of course costs nothing, but is extremely variable and creates data of low information. The reader + image processing approach produces data with the highest information quality, but the costs can be up to several thousand USD, and require other infrastructure (electricity,

computers, file storage, and trained users). The printed grading scale method may cost a few dollars per card, but relies on error-prone human judgment. Despite the limitations, a reader with a grading scales is still superior to a reader providing less refined estimates.

Setting the stage for the experiment

Our project investigates this intermediate method of having a human grader using a printed grading card. Under different lighting and background conditions, we assess the ability of subjects to accurately grade cassettes for their color given an image of a cassette alongside such a color scale. Given that the system of a human + color scale keeps the inconsistency of a human reader in the loop, this system is susceptible to many of the same psychological measurement issues as found in estimation tasks performed by humans. The psychological biases could include the position/orientation of the cassette vs the grading scale, recency bias (in which the previous cassette influences the way you measure a subsequent cassette), the level of audiovisual distraction while grading a cassette, and many other factors.

During an assay development project, thousands of images of strips may be taken, and while there is a concerted effort to make sure photographic conditions are consistent from experiment to experiment, conditions can vary from person to person within the same company, and from site to site between different companies. Aside from images taken during the development process, end users are often responsible for reading their own test cassettes. An assay may require that a person read in the natural light in a bathroom, a hospital room, a dentist's office, outside at a testing site, as well as other places. Additionally, testing outside may take place at different times of day, or under different weather conditions.

Given the absence of an industry-standard strip-imaging app, each developer who wants to understand imaging these strips in the field has to perform their own studies. While the field of assay development is heading towards digitization and more sophisticated means of quantifying test results, the current state is not a settled issue.

It is therefore important to understand how the imaging conditions can affect readings by users. The added benefit of taking the images is that further processing can be performed and the resulting values correlated with results from the users. The scope of this experiment is the effect of different photographic conditions on user-assigned test line scores; future work could focus on using quantified images to better understand both user bias as well as how to objectively quantify strips without incorporating a user's subjective estimation.

To make the imaging most closely resemble real-world lateral flow imaging, an iPhone 13 Pro Max using standard exposure settings was used for every image. Of course, a more carefully designed imaging setup could be achieved with controlled lighting and (for example) a manually-exposed high-end camera, but the purposely more amateur setup covers more genuine use-cases.

Causal Question & Treatments

Based on the above discussion, we posed the causal question: **Does lighting or background color affect how we read test strips?**

Given that we had to confine our project to a limited number of conditions that may impact user readings of these cassettes we decided on investigating:

1. **Lighting Condition**
    - **Ambient**: Indoor fluorescent-lit biological laboratory sp ace
    - **Shade**: Outdoors on a sunny day in the shade of a building
    - **Sunlight**: Outdoors on a sunny day in full sunlight
2. **Background**
    - **Black**: a black sheet of paper was used as the background to image the cassette and color scale
    - **White**: a white sheet of paper was instead used

For each of the six combination of these two factors, images were taken for ten different cassettes that spanned a range of test line intensities, for a total of sixty images. Each condition is named in the format `Lighting Background` (e.g. Shade Black). Fig ***???** shows two of the cassettes imaged under those conditions.

Figure 2: A single cassette imaged under six different conditions. From left to right, the conditions are: Ambient Black, Ambient White, Shade Black, Shade White, Sunlight Black, and finally Sunlight White. Since the cassettes are cropped out of their full images, the backgrounds are not visible; however, significant differences in exposure, image clarity, white balance, and shading are visible for the three lighting conditions. The test lines themselves look noticeably different in each condition.

Given that the cassettes used in every imaging condition are identical, there should theoretically be **zero** differences between groups, except possibly in the case where the camera's exposure were to cause the fainter test lines to disappear. Systematic bias or differences in reading variability may be signs of true subjective differences caused by the various conditions.

Experimental Design

To test our hypotheses, we used Qualtrics to deliver our treatments via online survey. We recruited participants from three sources: Prolific, our family and friends (Personal), and from MIDS students and instructors (MIDS).

Figs 3 and 4 shows how participants were sorted into the six treatment groups. One Prolific participant did not complete their survey successfully, which contributes to attrition.

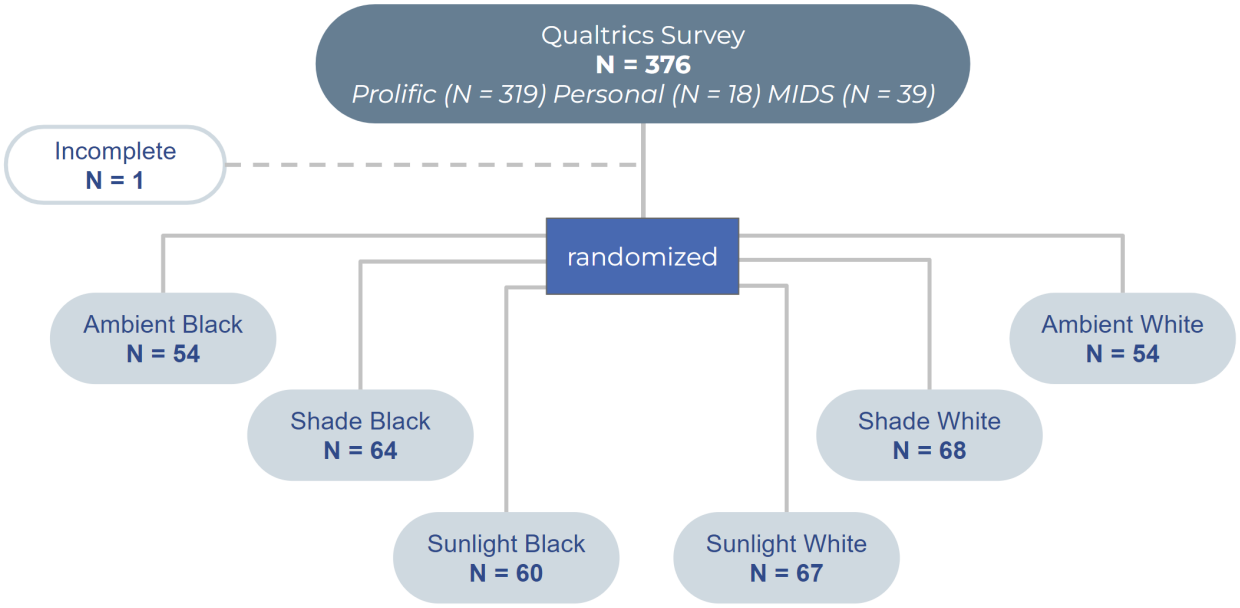

Figure 3: A flow diagram of participant recruitment, assignment to treatment groups, and attrition. Users were split fairly evenly between conditions. While the majority of subjects were sourced from Prolific, a significant number were participants from the authors' workplaces, friends, and fellow students.

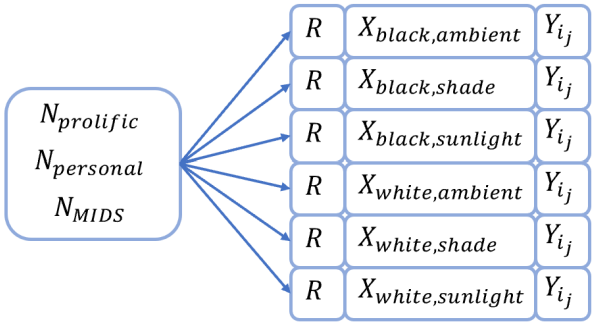

Figure 4: A ROXY diagram showing the experiment design. Y_{i_j} refers to a cassette score, i, for j cassettes. The score, i, is an integer from 0 to 10. The cassettes, j, were in the same order for all treatments. There were ten cassettes with a true grade of 1 through 10, not including the instruction image.

The Qualtrics survey began with some demographic questions (see the Appendix for more details). These formed the basis of our covariate data. Then, we showed participants one of six blocks of questions, randomizing which block would be shown with Qualtrics' built-in features. Each block began with an instruction image accompanied by an attention check question. Then, there was a page of 10 cassettes for each participant to grade (see Fig 5 for examples of the survey images). All questions were required before the survey was marked completed.

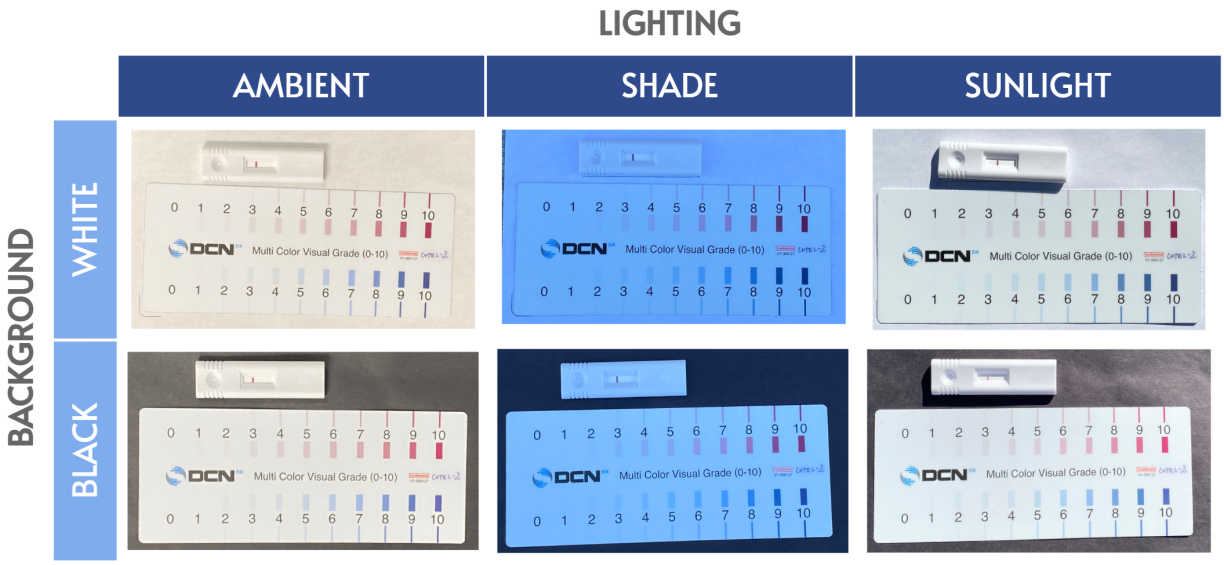

Figure 5: A table of example treatment photos included in the Qualtrics survey. An iPhone 13 Pro Max with normal photo settings was used to capture survey images. This image capture method was considered closest to the real use case of people imaging test strips.

Figure 6: A full series of cassettes with varying test line intensities from the Sunlight Black condition. These are the actual cassettes used in this experiment. The cassettes are cutout from their original images to show the intensity differences, but the full images looked as shown in the previous figure.

With our experiment design, the instrument for delivering the treatment was embedded in the survey images, like those above. As long as the participant took the survey, we were reasonably confident that they were treated as intended (as in, participants were compliant as long as they completed their survey). From the way the Qualtrics survey is designed, they would not receive the treatment for another group unless that survey-taker lived in the same household or worked in the same office as another survey-taker and took their survey for them. But this spillover risk is very low.

We had one case of attrition in the Prolific population. They did not complete the survey, which Prolific indicates is likely due to device failure or other external factors to the participant and the survey. This attrition is very low and does not greatly impact our analysis.

Exploring the data

Data processing

Our data is from Qualtrics surveys, which we loaded into R with some custom R scripts (see the Appendix for details). We created three surveys in order to keep track of the sources of the data. The covariate and test sections were the same, but for the personal and MIDS surveys, an email collection mechanism was added in order to conduct a raffle. The promised $20 was found to be a powerful motivator for students, friends, and professionals alike. Although the code in the Rmd version of this document begins here, all code is showed only at the end of this paper.

Since each participant generated 10 data points that should in the ideal case form a perfect line, we chose these summary statistics to aggregate this per-participant data.

1. **Root mean-squared error (RMSE)**, computed as follows: $\sqrt{\frac{1}{10}\sum_{i=1}^{10}(trueScore_i - userScore_i)^2}$
    - This describes the absolute distance from the truth, disregarding which direction the user is biased from the true cassette scores.
    - The closer to 0 the RMSE is, the more accurate the grades the user gave were.
2. **Linear model** for each user, generating: $userScore = \beta_0 + \beta_1 trueScore$
    - The slope, β_1, and intercept, β_0, are of interest to us. Perfect accuracy yields a linear model with $\beta_0 = 0$ and $\beta_1 = 1$.
    - The intercept in particular indicates the accuracy of user grades at the extremes of the cassette values. If the intercept is higher, this means that the values are generally overestimated for cassettes with true values near 0.

While the RMSE number measures deviation from the known values of the cassettes, intercept measures the ability of a user to correctly grade the low end of the curve, and the slope measures the overall linearity of their assigned scores.

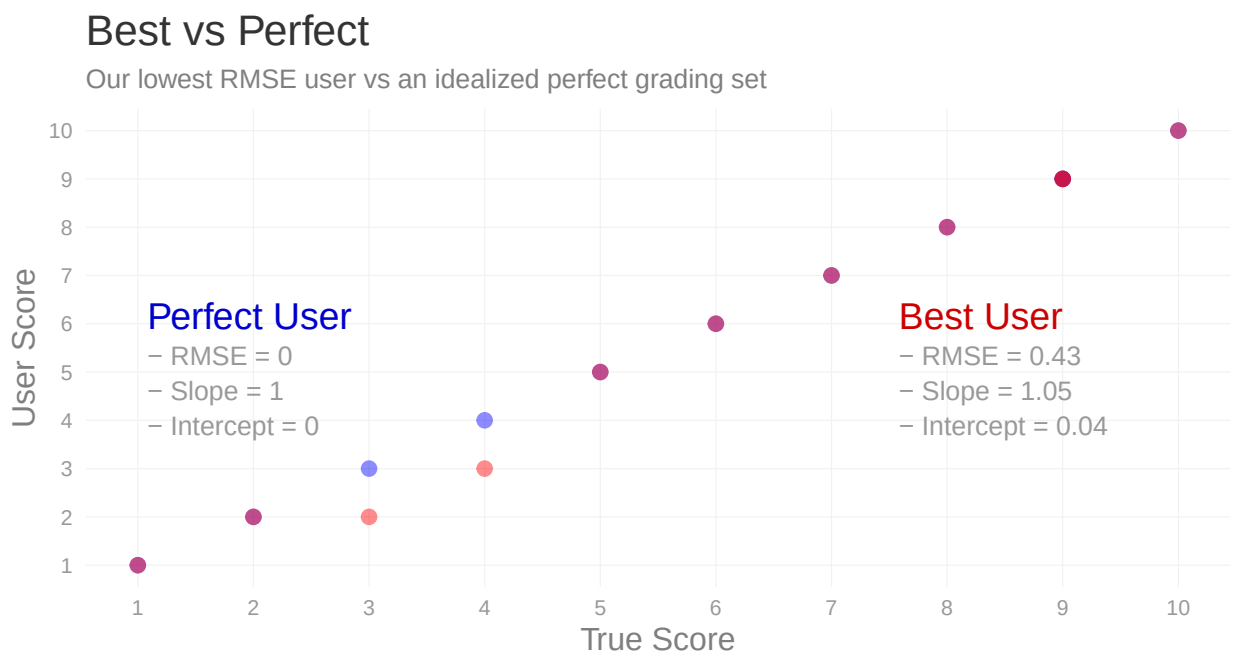

Figure 7: This figure shows two example subjects: one hypothetical perfect subject who gave all the correct answers (in blue), and one of the best-scoring subjects in red. The summary scores (RMSE, slope, and intercept) are shown as well.

The Attention Check

The attention check question was straightforward, asking:

> What grade is the test strip above? This question is an attention check.
>
> *For this question, you must select the grade that was given in the instructions above the image.*

The attention check would allow us to:

1. Train people what we are looking for with a simple example with the correct answer, and
2. Check whether people were paying attention to the instructions

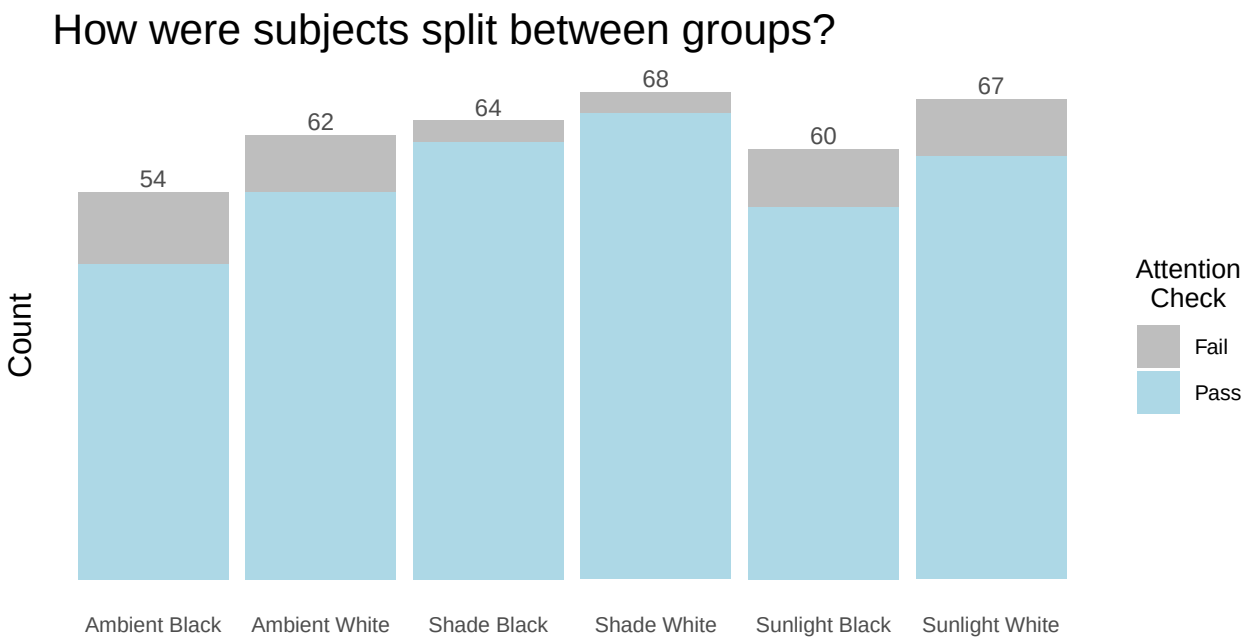

Figure 8: This shows how subjects were split among the different treatment groups, as well as the relative proportion of people who failed the attention test.

Failing the attention check was linked to poor performance, although insufficient room in this report is available to cover this specific topic. The poor performance was also not a concern for this particular project because the unusually bad graders were removed from calculations via a separate outlier removal process, described in a subsequent section.

Data Exploration and Outlier removal

The initial exploration through the data is composed of different sections: 1. All individual grades for each user and cassette, graphed at one time 2. Summary scores for each subject in aggregate 3. Discussion of how to remove the poorest performers 4. Outlier-removed data

Given that individual cassette gradings are easily shown as a scatter plot, it is relatively easy to take a look at the test data in aggregate. As shown in the **Best vs Perfect** graph in the previous section, the linear nature of the data collected allows for the collection of simple and meaningful summary data. Below is a look at all responses from all users from each treatment group. The groupings of points generally follow the ideal slope=1, intercept=0 format, but rather extreme dispersion is observed, with some points extremely far from each other.

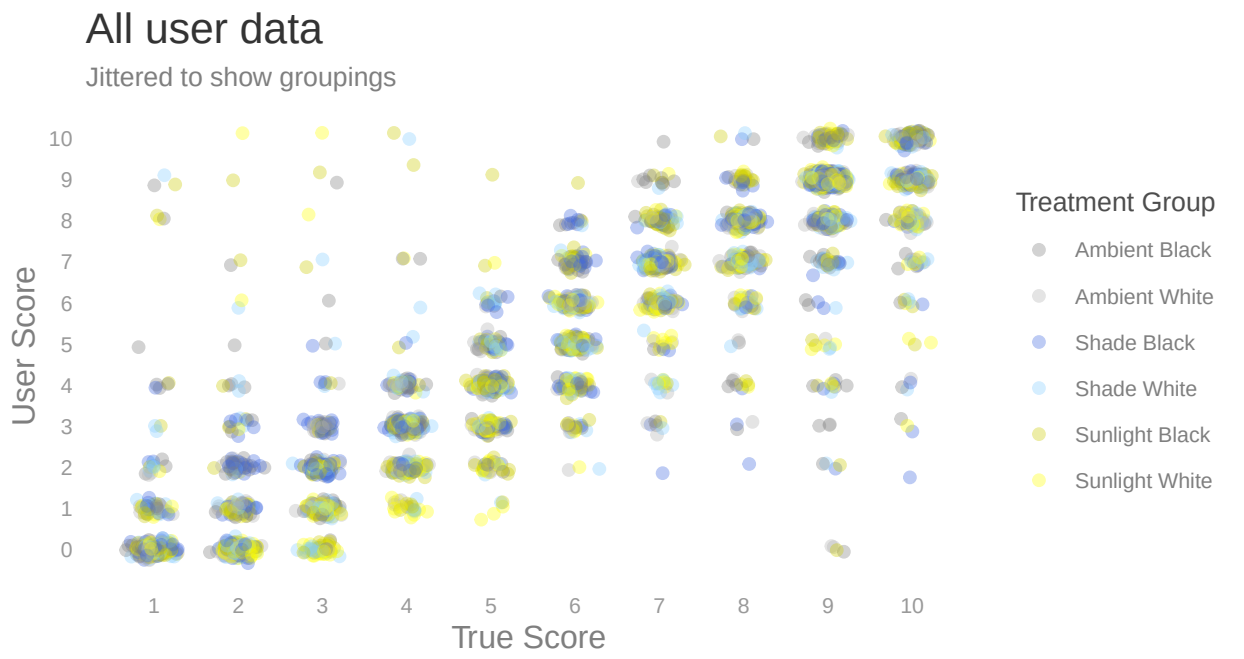

Figure 9: All subject data is graphed as a scatterplot at the same time. This view of the data provides a quick and simple overview of user performance. One trend that is visible is the concentration of yellow dots (representing the two Sunlight groups) weighted at the low end of the True Scores. In terms of (x,y) coordinates, some positions showing higher concentrations of yellow dots includes: (3,0), (4,1), and (5,2).

Upon first look at the data, without the use of jitter and alpha values for each point, the grading scatter plot looked like a single grid of points, causing the authors to doubt their code. Much more variation between individuals and between conditions was observed than expected, although it is clear that the general ideal linear trend is faithfully represented in the summation of guesses.

Figure 10: Cases of mistaken identity. Each row is an actual example of misperception taken from the subject data. The cassette on the left is the cassette that was shown to a subject; the cassette on the right shows the graded score that the subject assigned to that strip. The top row is an example from the Sunlight White condition, where a person was shown a cassette whose true value is a 1, but the user guessed that the cassette was an 8 on the provided scale. It is extraordinarily unlikely that this represents a good-faith attempt at scoring the cassette. Due to this, outlier removal is examined below.

Incredulity at some of the grading values is natural when some of the mistakes are as egregious as the ones shown above. Two ways of further analyzing these high-RMSE guesses include:

1. A programmatic tool to show images of what a person was shown vs a person's guess, and this would perhaps reveal systematic bias in grading estimations. Such a tool could rank the worst guesses by L1 distance from true score to estimated score, and present the results as a tiled image. It is possible that these are truly random.

2. Analysis of the randomness of the guesses by users, showing that (for example) the slope is close to zero or even negative, or that the intercept is far above the low end of the . Other statistical analyses to determine likelihood of subjects' guess to be random vs deliberate could be performed.

It is possible that, if all treatment groups suffered the same level of random guessing *and* the random guessing was not systematically biased for one treatment group differently than another, that minimal impact would be made to the graphs or analyses (variance being the only affected measure). Despite these potential approaches, the authors opted for simple outlier removal.

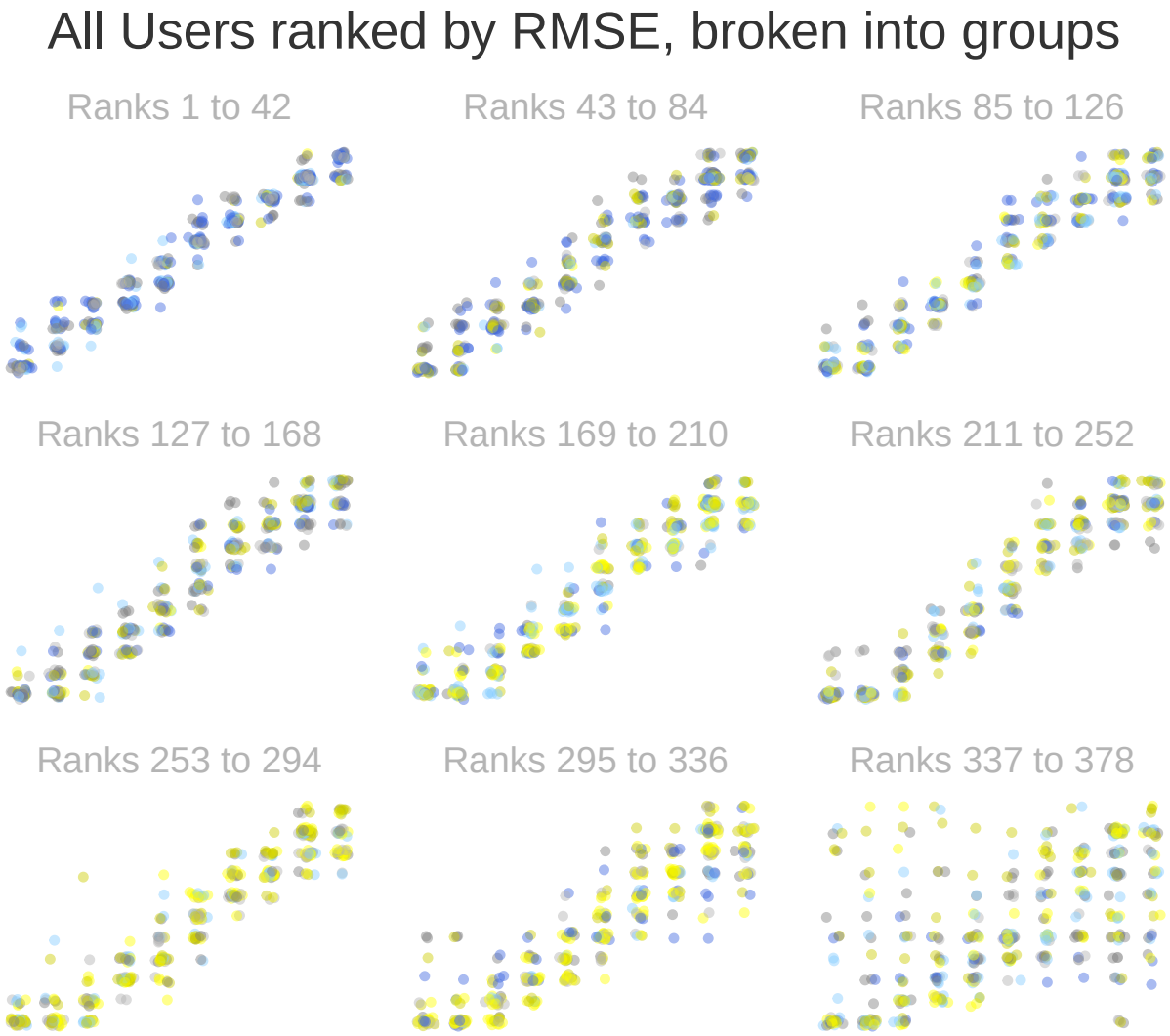

Figure 11: The figure above shows all users, ranked and sorted by RMSE and separated into nine groups. The general quality of the user responses is relatively consistent, with the exception of the final group, which show extreme spread and poor correlation with the true values.

One way of breaking that detailed user data into interesting and useful groupings is to first rank users by their RMSE scores, and then to group them with their similarly-ranked neighbors. Due to the coloration differences between treatment groups, trends are easy to observe. For instance, it is apparent in the first grouping (highest quality graders) that there is a heavy over-representation of the Shade groups (depicted as the blue dots) relative to the other lighting conditions. Likewise, some of the later groups show large

numbers subjects from the Sunlight treatment group. The increase in blue dots as users show better and better performance is consistent across the ranked groups.

A key finding from this graph is that the best performers tended to be from the shade groups, which was unexpected because of the deeply unrepresentative blue color of the images themselves. Upon further inspection (the Shade images can be seen in figures towards the top of this paper), it may be that the blue discoloration of the image led to the reds appearing more black/gray in color rather than red, and the exposure settings generated by the camera under these conditions led to more faithful representation of the low end of the color scale. This is perhaps worth a follow-up investigation but beyond the scope of this paper.

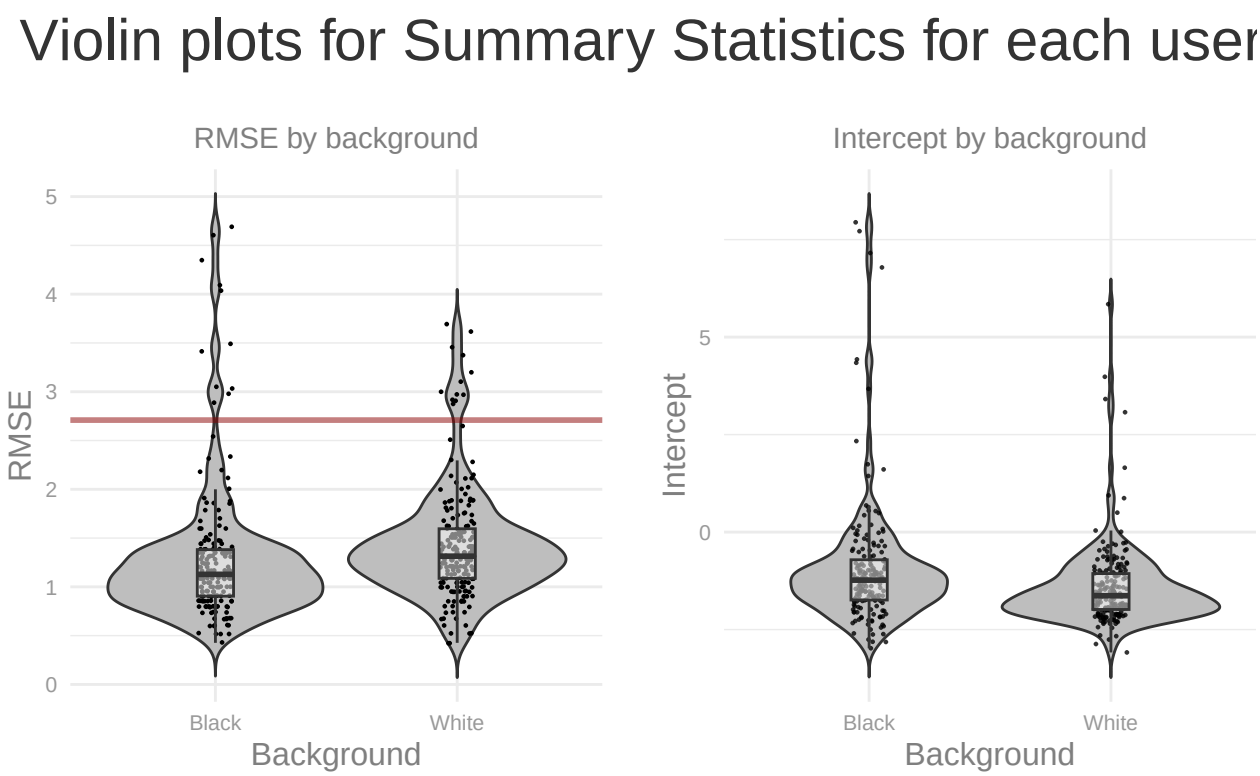

Figure 12: These violin plots show the distribution in subject score quality, grouped by background. While three summary statistics are collected for each subject, the two most relevant (RMSE and intercept) are shown in this figure. The reason these two are considered more relevant is because RMSE and slope both describe the overall quality or closeness to the perfect line. Intercept, on the other hand, is of particular interest because it deals with the ability of a person to distinguish the lowest test line signals, which is important in borderline cases, where the signal is nearest to the visual cutoff between negative and positive.

The violin plots show the distribution of errors and makes choosing a cutoff for removal of poor performers easier. In this case, having an RMSE value that is above two standard deviations is a suitable way of separating the main grouping of subjects who likely tried to complete the survey versus subjects who may have rushed or not tried to legitimately provide answers. Outliers in this case are extremely easy to showcase, meaning we can examine the specific cases and gauge how reasonable are the cases of alleged misperception.

Variance in guesses have been shown in three different ways in the images above:

1. Individual cassette estimations simultaneously for all users
2. Individual cassette estimations for subjects grouped by RMSE ranking
3. Violin plots (with box plots to show quartiles) on the key summary statistics

What was learned from these views on the data was that:

1. A high degree of variance in score estimations permeates all groups
2. Some of the guesses are likely farcical and perhaps worth deletion
3. The distribution of "good-faith" guesses could perhaps be identified via simple **Mean+ 2*(SD)** method, which eliminates the worst performers.

Figure 13: The application of the discussed **Mean + 2*SD** method on subject RMSE to filter out outliers results in the above chart.

After applying the outlier-detection method (which was not rigorously conceived or investigated). This represents a "first-passs" approach to removing outliers. Other methods could include correlating 1) the speed at which the survey was filled out, 2) screen size of the viewing device, 3) overall characteristics of the user scores, or 4) the attention check.

Despite this, the simplest of rules (treating the distribution as normal and excluding the extremes of one of the tails) seems to have gotten rid of the most incorrect estimators well enough.

Aggregate treatment group performance

Having separated what are believed to be many of the non-attempts at grading the cassettes from the data, it is time to summarize and examine how the groups performed in aggregate. While no single average treatment effect can be measured that covers all aspects of user performance, the averages and standard errors for each treatment condition can be compared at different points along the linear curve.

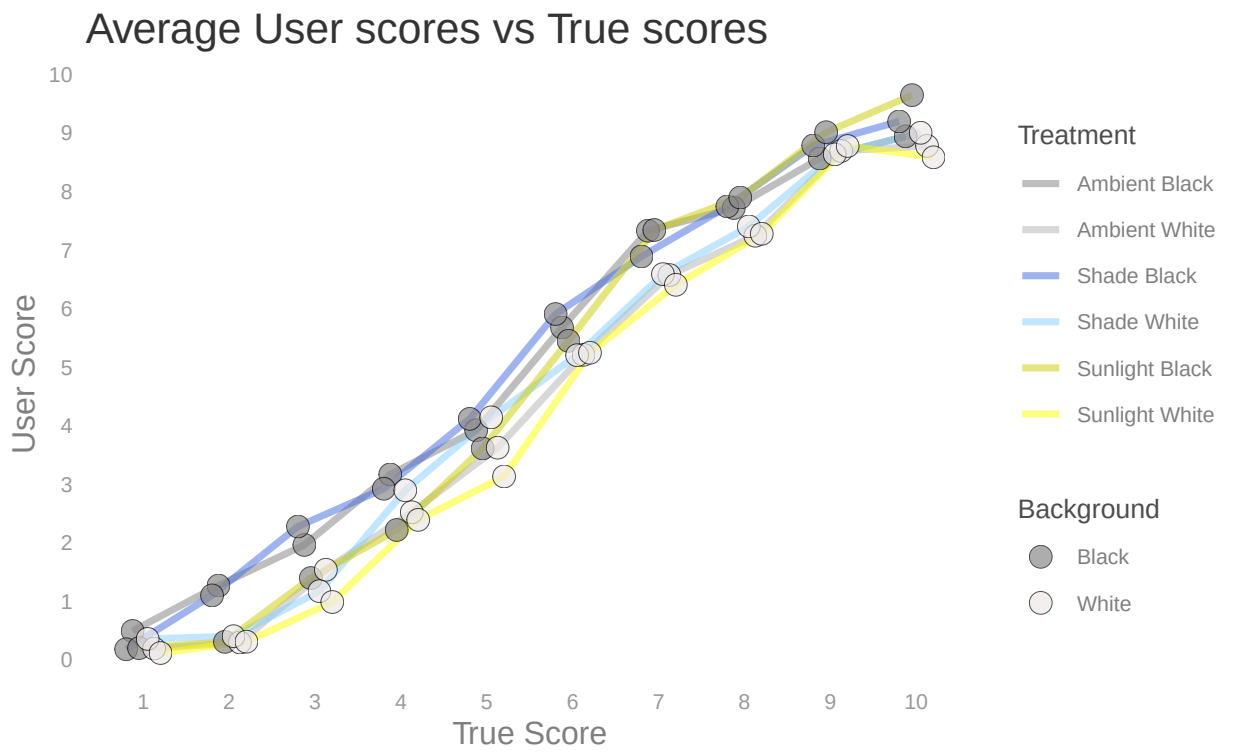

Figure 14: This summary figure flattens each treatment group into a single point per cassette. The color of the point corresponds to background condition, while the lines connecting subsequent points correspond to the lighting condition + the background.

The summary figure above shows the average user scores for each group and each true cassette value. Of particular note is the relatively consistent separation between the black and white background conditions, with the exception of the Sunlight Black condition, which tended to follow the remaining White background conditions. Some takeaways from this chart include:

1. The Black background conditions were associated with higher test line estimates
2. The Sunlight lighting conditions were associated with lower test line estimates
3. The Shade Black and Ambient Black conditions were able to, on average, distinguish between cassettes with true scores of 1 & 2, unlike every other group, which on average evaluated both levels 1 & 2 as 0s
4. *All conditions underestimated the test line signals up until a true value of six*

Further analysis is best done in statistical form, and is contained in subsequent sections.

Statistical analyses

Covariates in data

These were the covariates we focused on.

- `device`
- `gender`
- `colorblind`

- `age_group`
- `eyewear`
- `cassette_experience`

After the initial data exploration, we chose to focus on examining differences in grading accuracy due to `background` – either black or white – rather than between all six treatments.

Since all our covariates were categorical, we used chi-squared tests for comparison. This does have the drawback of possibly picking up a statistically significant test due to the number of tests being run.

This first test is a simple test of balance between the two treatment groups, white and black backgrounds. The p-value here is greater than 0.05, so we fail to reject the null hypothesis that the distribution is different from an equal distribution of values across the two backgrounds.

```
    Chi-squared test for given probabilities

data:  table(background)
X-squared = 0.92045, df = 1, p-value = 0.3374
```

Second, a set of p-values for checking the covariate balance.

Covariate	Chi-sq p-value
Device	0.692
Age group	0.132
Gender	0.516
Colorblind	0.072
Eyewear	0.789
Cassette Experience	0.717

In the above table, we see there is a possibility of imbalance for the `age_group` and `colorblind` covariates. Below are joint frequency tables for both. These imbalances are probably due to low sample size (N = 375) in certain covariate categories. For example in the `age_group` category, the 65+ age group only had 11 participants to be spread across the treatment groups. For the `colorblind` category, there was a group of 5 participants who were not sure, but all happened to be assigned to the white background treatment.

	18 - 24	25 - 34	35 - 44	45 - 54	55 - 64	65+
Black	39	67	27	18	12	4
White	41	60	47	25	6	6

	No	Not sure	Yes
Black	162	0	5
White	177	5	3

Treatment Effects

Modeling approach

We decided on modeling two different key variables: **RMSE** and **Intercept**. As stated previously, we believe these two to be the best ways of quantifying individual and group performance.

For each of these modeled outcomes, we decided on three different ways of modeling treatments. Descriptions of as well as defaults are listed below.

1. Background-only (*default Black*)
2. Background + Lighting (two levels, *default Ambient + Black*)
3. Treatment Group (each as a separate entity, *default Ambient Black*)

For each of the above ways of modeling treatments, we also included the following as covariates:

1. Attention Pass <- whether the person has passed the attention test (*default FAIL*)
2. Experience Levels <- experience with lateral flow tests (*default PROFESSIONAL*)
    - Little
    - Moderate
3. Device <- type of deviced used to view the survey (*default LAPTOP/DESKTOP*)
    - Phone
    - Tablet

With the three types of models, two outcome variables, and exclusion/inclusion of covariates, there are a total of twelve models to compare.

Modeling and excluded covariates

In fitting these models, we found that consistent with expectation, gender and colorblindness did not tend to provide strong explanatory power for the outcome (the cassettes are meant to be colorblind friendly and one's gender expression should not affect your ability to grade cassettes). The age of the user and what eyewear they were using also did not appear to have significant explanatory power for the outcomes.

Finally, we decided to include the source of the data as a fixed effect (the `source` covariate indicating whether the participants were sourced through Prolific, personal network, or MIDS). We also noted that the device used did tend to explain more variability in the outcome. In both outcomes, using a phone had a negative impact on the outcome due to the smaller size of the images affecting accurate grading. We also included the amount of cassette experience that the user had previously as a control. Additionally, whether the user passed the attention check was a strong predictor of their outcome for both the intercept and RMSE.

In our data exploration and subsequent investigation we did not note any significant heterogenous treatment effects based on certain covariates. We also did not expect HTEs from the covariates we gathered on the users.

Models

Table 4: Subject Grading Error (RMSE) based on Treatment

	Outcome: RMSE		
		rmse	
	BG only	2-level	1-level
	(1)	(2)	(3)
Background white (not black)	0.181***	0.179***	
	(0.042)	(0.041)	
Shade		−0.105**	
		(0.053)	
Sunlight		0.145***	
		(0.049)	
Ambient White			0.109
			(0.078)
Shade Black			−0.177**
			(0.079)
Shade White			0.068
			(0.084)
Sunlight Black			0.111
			(0.075)
Sunlight White			0.283***
			(0.078)
Constant	1.145***	1.135***	1.174***
	(0.030)	(0.046)	(0.063)
Observations	352	352	352
R^2	0.051	0.119	0.123
Residual Std. Error	0.392 (df = 350)	0.379 (df = 348)	0.379 (df = 346)
Note:			*p<0.1; **p<0.05; ***p<0.01

RMSE models without Covariates This simple comparison showed that, for RMSE, there was no explanatory power in separating treatments into individual groups. Accounting for either only background or background + lighting had nearly the same residual error. For Background, both White and Black backgrounds showed significance; for background + lighting, the two backgrounds had significance as well as one of the lighting conditions (shade). For the flat treatment groups model, only Shade Black mattered in comparison.

To summarize the the findings: the white background had higher amounts of error and the shade led to decreased error.

RMSE models with Covariates

Table 5: Subject Grading Error (RMSE) based on Treatment with covariates

	Outcome: RMSE		
		rmse	
	BG only	2-level	1-level, All Groups
	(1)	(2)	(3)
Background white (not black)	0.185***	0.183***	
	(0.041)	(0.040)	
Shade		−0.083*	
		(0.050)	
Sunlight		0.154***	
		(0.047)	
Ambient White			0.112
			(0.073)
Shade Black			−0.169**
			(0.073)
Shade White			0.107
			(0.081)
Sunlight Black			0.134*
			(0.069)
Sunlight White			0.282***
			(0.076)
Attention Pass	−0.232***	−0.203**	−0.203**
	(0.083)	(0.084)	(0.083)
Little Experience	0.321***	0.329***	0.336***
	(0.090)	(0.091)	(0.091)
Moderate Experience	0.171**	0.203**	0.207***
	(0.081)	(0.081)	(0.080)
Viewed on Phone	0.090*	0.108**	0.115**
	(0.050)	(0.048)	(0.048)
Viewed on Tablet	0.064	0.071	0.082
	(0.114)	(0.118)	(0.120)
Constant	1.103***	1.050***	1.084***
	(0.095)	(0.105)	(0.111)
Observations	352	352	352
R^2	0.163	0.222	0.230
Residual Std. Error	0.372 (df = 343)	0.360 (df = 341)	0.359 (df = 339)
Note:			*p<0.1; **p<0.05; ***p<0.01
			Includes survey source fixed effects

Among the covariates, Passing the attention test showed significantly decreased RMSE values (a very statistically significant as well as large effect), Viewing the survey on a phone showed increased error per user. Users with professional experience reading test strips performed better than users with less experience, though the result was interestingly not statistically significant.

Intercept models without Covariates

Table 6: Subject Regression Line Intercept based on Treatment

	Outcome: Intercept		
		intercept	
	BG only	2-level	1-level
	(1)	(2)	(3)
Background white (not black)	−0.368***	−0.368***	
	(0.088)	(0.083)	
Shade		−0.023	
		(0.106)	
Sunlight		−0.630***	
		(0.108)	
Ambient White			−0.752***
			(0.178)
Shade Black			−0.196
			(0.181)
Shade White			−0.633***
			(0.188)
Sunlight Black			−1.059***
			(0.195)
Sunlight White			−1.007***
			(0.180)
Constant	−1.204***	−0.984***	−0.776***
	(0.074)	(0.112)	(0.163)
Observations	352	352	352
R^2	0.049	0.174	0.212
Residual Std. Error	0.808 (df = 350)	0.756 (df = 348)	0.740 (df = 346)
Note:			*p<0.1; **p<0.05; ***p<0.01

In terms of the intercept outcome variable, the different models behaved somewhat differently than when looking at the RMSE outcome variable. Similar to RMSE, both Background-only and Background + Lighting conditions showed significance (though only two of the three lighting conditions showed significant effects). The largest difference is observable with the flat treatment group model, in which almost every condition showed statistically significant and different treatment effects for the intercept.

Intercept models with Covariates

Table 7: Subject Regression Line Intercept based on Treatment with covariates

	Outcome: Intercept		
		intercept	
	BG only	2-level	1-level, All Groups
	(1)	(2)	(3)
Background white (not black)	−0.377***	−0.376***	
	(0.090)	(0.084)	
Shade		−0.003	
		(0.101)	
Sunlight		−0.646***	
		(0.107)	
Ambient White			−0.734***
			(0.176)
Shade Black			−0.158
			(0.170)
Shade White			−0.613***
			(0.179)
Sunlight Black			−1.049***
			(0.189)
Sunlight White			−1.018***
			(0.180)
Attention Pass	−0.182	−0.237	−0.218
	(0.279)	(0.266)	(0.266)
Little Experience	−0.316*	−0.320*	−0.337**
	(0.184)	(0.171)	(0.167)
Moderate Experience	−0.320*	−0.398**	−0.375**
	(0.169)	(0.158)	(0.149)
Viewed on Phone	0.136	0.070	0.062
	(0.119)	(0.112)	(0.115)
Viewed on Tablet	−0.005	−0.007	−0.071
	(0.145)	(0.149)	(0.151)
Constant	−0.651**	−0.349	−0.182
	(0.273)	(0.283)	(0.321)
Observations	352	352	352
R^2	0.089	0.220	0.254
Residual Std. Error	0.800 (df = 343)	0.742 (df = 341)	0.727 (df = 339)

Note: *p<0.1; **p<0.05; ***p<0.01
Includes survey source fixed effects

Unlike RMSE, the addition of covariates to the the intercept models did not show sensitivity to the type of viewing device, though user experience was found to be significant. Passing the attention test signifcantly changed the performance of the people in the group.

Conclusions & Discussion

Further changes to analysis

While linear models on the RMSE values did not show a statistically significant difference between lighting conditions, it was a simple observation that the the best-performers groups showed far fewer subjects from the Sunlight condition. Another approach to analyzing the differences could be to section the user ratings into three groups:

1. **Low**: Cassettes with true scores between 0 to 3
2. **Medium**: Cassettes with true scores bewteen 4 to 7
3. **High**: Cassettes with true scores between 8 to 10

Separate linear models could be made for each cassette grouping, treatment, etc. This is beyond the scope of this paper, but the relevant analysis may possibly be performed by following this line of reasoning.

The two background conditions, Black and White, showed consistent differences in average scores. It is unknown whether this was potentially due to differences in automatic exposure settings from the camera, or whether the source is in psychological measurement estimation. Further experimentation could pursue and answer this question by forcing the exposure settings to be identical for both backgrounds within each lighting condition.

Image quantification as additional analysis

One note on the true values assigned to the various cassettes: the actual cassettes are made by striping gold nanoparticles. At diameters near 25nm, solutions of gold nanoparticles have an overall reddish color. These gold nanoparticles are very common in the industry, but there are other kinds of nanoparticles used for generating signals in lateral flow immunoassays (latex, fluroescent molecules, quantum dots, other metallic suspensions). Of note is the fact that the strips were value-assigned a single visual grade (e.g. 5), but this value assignment is itself a subject act. One further evolution of this project would be to assign the true values by image processing. This could result in shifts relative to the numbers given as true in this experiment.

Much larger studies with more conditions, and perhaps with better image processing techniques applied would be necessary to make more pointed and statistically significant recommendations.

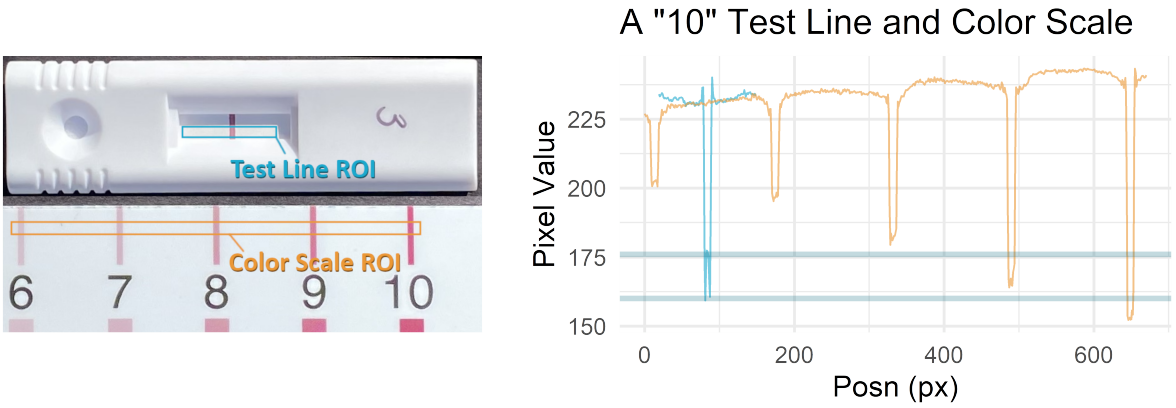

Figure 15: A single cassette from the Sunlight Black condition, shown near the relevent end of the color scale. From the image on the left, different Regions of Interest (ROIs) can be obtained, and the pixel values of the test line versus the test strip background compared to the actual color scale. The ROIs can then be processed by averaging each vertical column in grayscale format. In the graph on the right, the orange line corresponds to the oranage ROI, and the teal line to the teal ROI. The test line itself can be represented by two possible intensities, shown as horizontal lines. These lines roughly correspond to 9 or 10. Further analysis could be done to quantify exactly where between 9 and 10 the test line truly falls.

The above image and graph show how image quantification may take place. Methods related to this were not used in this experiment due to time constraints, but further analysis along this line would help adjust the X-position for each cassette when graphed as the "True Score."

Despite the imprecision and subjectivity in the initial value assignment, the various patterns between treatment groups should not be affected by this. The Y-values would not change, but there would be shifts in the X-values.

Additionally, since the test line of the cassette and the printed ink on the color grading card may react somewhat differently to different lighting conditions, it may be that there are actual inconsistencies between the same cassette imaged under different conditions. This would need to be explored quantitatively as well.

Recommendations

Overall, if a recommendation were to be made as to ideal lighting conditions to give the most consistent and least variable results at grading cassettes, it would be the following:

Take the image of the cassettes with a black background, but not under direct sunlight.

Appendix

This section contains extra information that is related but not required to understand the experiment, including the exact wording of the covariate questions, code for processing the data in this experiment, and (quite importantly), code used to select a winner for our $20 raffle meant to encourage participation of fellow MIDS students and coworkers.

Question encodings

These are the encodings for the raw data.

First, a mapping of the original column names of covariates collected for each subject to the column names in the data table used in the analysis.

Raw data column name	Cleaned data column name
Q0_Browser	browser
Q0_Operating System	os
Q0_Resolution	screen_resolution
Q1.2	age_group
Q1.3	gender
Q1.4	colorblind
Q1.5	eyewear
Q1.6	device
Q1.7	cassette_experience
Various	attention_check_passed

Second, a list of the question content and answers.

Q1.2

How old are you?

Answers:

- Under 18
- 18 - 24
- 25 - 34
- 35 - 44
- 45 - 54
- 55 - 64
- 65+

Q1.3

How would you describe yourself

Answers:

Survey response	Encoded response
Male	Male
Female	Female
Prefer to self describe	Other
Prefer not to say	Other

Q1.4

Do you have color blindness?

Answers:

- Yes
- No
- Not sure

Q1.5

Do you use a visual aid (glasses, contacts, etc.) for seeing up close?

Answers:

- Yes, and I'm wearing them
- Yes, but I'm not wearing them
- No

Q1.6

How are you viewing this survey?

Answers:

- Laptop / Desktop
- Phone
- Tablet
- Other

Q1.7

Describe your level of experience with reading lateral flow test strips such as the one below

Answers:

Survey response	Encoded response
I read them professionally or very frequently	high
I read them sometimes	moderate
I have never read one	little
Not sure	little

Attention Check

This raw data column name varied as it depended on which treatment block of questions the user was assigned to answer.

What grade is the test strip above? This question is an attention check.

For this question, you must select the grade that was given in the instructions above the image.

Survey response	Encoded response
9	Pass
Other values	Fail

Data processing functions

Copied from earlier in this file for reference.

```
source(here("utils.R"))

# File names
file_prolific <- here('data', '20221202_prolific.csv')
file_mids <- here('data', '20221205_mids.csv')
file_personal <- here('data', '20221205_personal.csv')

# Load all raw data
raw_data_prolific <- load_raw_qualtrics_data(file_prolific, "prolific")
raw_data_mids     <- load_raw_qualtrics_data(file_mids, "mids")
raw_data_personal <- load_raw_qualtrics_data(file_personal, "personal")

# Combine all raw data
raw_data_all <- rbindlist(list(raw_data_prolific,
                               raw_data_mids,
                               raw_data_personal), fill=TRUE)

# Clean and reshape raw data
d <- clean_qualtrics_raw_data(raw_data_all)

# Create a new data table aggregating by `subject_id`
# Select desired covariate and treatment columns
# Create a new column, `rmse` to aggregate the cassette scores
d_agg <- d[, .(rmse=rmse(user_score, true_score)),
            by=.(subject_id, source,
                 treatment_group, completion_time,
                 browser, os, screen_resolution,
                 age_group, gender, colorblind,
                 eyewear, device, cassette_experience,
                 attention_check_passed, width, height,
                 MP, background, lighting)]

for (sid in unique(d$subject_id)) {
  s_lm <- d[subject_id == sid, lm(user_score ~ true_score)]
  d_agg[subject_id == sid, rsq:= summary(s_lm)$r.squared]
  d_agg[subject_id == sid, intercept:= summary(s_lm)$coefficients[1,1]]
  d_agg[subject_id == sid, intercept_se:= summary(s_lm)$coefficients[1,2]]
  d_agg[subject_id == sid, slope:= summary(s_lm)$coefficients[2,1]]
  d_agg[subject_id == sid, slope_se:= summary(s_lm)$coefficients[2,2]]
}

d_agg <- d_agg[order(rmse),]
```

Copied from the `utils.R` file for reference

```
# new column names for questions
data_cols <- c('0', '1', '2', '3', '4', '5', '6', '7', '8', '9', '10')

# answers to the the questions
true_values <- c(9, 2, 5, 8, 9, 7, 10, 1, 4, 6, 3)
cass_nums <- c(1, 13, 15, 21, 28, 3, 30, 5, 6, 9)

# extra information columns
```

```
info_cols <- c("source", "treatment_group", "Duration (in seconds)",
               "Q0_Browser", "Q0_Operating System", "Q0_Resolution",
               "Q1.2", "Q1.3", "Q1.4", "Q1.5", "Q1.6", "Q1.7")

# sequence of treatments
treatments <- c('Ambient Black', 'Ambient White',
                'Shade Black', 'Shade White',
                'Sunlight Black', 'Sunlight White')
```

```
# load Qualtrics data from a file name into raw form & aggregate the answers in a more useful format
load_raw_qualtrics_data <- function(filename, data_source) {
  raw_data <- fread(filename)
  raw_data[, treatment_group := "NA"]
  raw_data[, source := data_source]
  raw_data[, data_cols] <- 0

  start_col <- 29
  if (data_source == "prolific") {
    # Account for extra question in the Prolific survey
    start_col <- start_col + 1
  }

  for (index in 3:nrow(raw_data)) {
  # Iterate through each block (1-6) of data
    for (i in 1:6) {
      # Each block has 11 columns of data
      # 1 attention check, 10 cassette grades
      row <- as.numeric(unlist(
        raw_data[index,
              (start_col + 11*(i-1)):(start_col + 11*i - 1)]))

      # If a non-null block is detected, extract the data
      # Append to the end of the data table
      if (!is.na(row[1])) {
        raw_data[index, 'treatment_group'] <- treatments[i]
        for (col_n in 1:length(data_cols)) {
          raw_data[index, data_cols[col_n]] <- row[col_n]
        }
      }
    }
  }

  # Remove the first two header rows
  return (raw_data[3:.N,])
}

# Assign "other" for non- "male" and "female" responses
gender_labeler <- function(x) {
  gender <- ifelse(x == "Male", "Male",
                   ifelse(x == "Female", "Female",
                          "Other"))
  return(gender)
}
```

```
# Assign values for experience level
exp_labeler <- function(x) {
  exp <- ifelse(
    x == "I read them professionally or very frequently", "High",
    ifelse(x == "I read them sometimes", "Moderate",
           "Little"))
  return(exp)
}

clean_qualtrics_raw_data <- function(raw_data) {
  # Concatenate the demographic and survey columns
  cols <- c(info_cols, data_cols)

  # Remove first 2 (non-data) rows
  d_clean <- raw_data[-(1:2), ..cols]

  # Rename columns
  setnames(
    x = d_clean,
    old = c("Duration (in seconds)", "Q0_Browser", "Q0_Operating System",
            "Q0_Resolution", "Q1.2", "Q1.3", "Q1.4",
            "Q1.5", "Q1.6", "Q1.7"),
    new = c("completion_time", "browser", "os", "screen_resolution",
            "age_group", "gender", "colorblind",
            "eyewear", "device", "cassette_experience")
  )

  # Assign an internal subject ID to each person
  d_clean[, subject_id := 1:.N]

  # Evaluate whether each subject passed our attention test (Question #0)
  d_clean[, 'attention_check_passed'] <- ifelse(d_clean$'0' == 9, "Pass", "Fail")

  # Get numerical values for screen size (height, width, megapixels)
  nums <- type.convert(str_split_fixed(d_clean$screen_resolution, 'x', 2), as.is=TRUE)
  d_clean[, ':=' (width = nums[,1], height=nums[,2])]
  d_clean[, ':=' (MP = width*height/1000000)]

  # Convert completion_time column values to numeric
  d_clean[, 'completion_time'] <- as.numeric(unlist(d_clean[, 'completion_time']))

  # correctly label gender & experience
  d_clean[, gender := gender_labeler(gender)]
  d_clean[ , cassette_experience := exp_labeler(cassette_experience)]

  d <- reshape(data=d_clean,
               idvar="subject_id",
               varying=data_cols,
               v.name=c("value"),
               times=data_cols,
               new.row.names = 1:(nrow(d_clean)*11),
               direction="long")
```

```
  # Rename the user question # and scores columns, then sort by user / question #
  colnames(d)[colnames(d) == "time"] <- "q_num"
  colnames(d)[colnames(d) == "value"] <- "user_score"
  d <- d[order(d$'subject_id', as.numeric(d$'q_num'),)]

  # Append the true grades to the user data in order
  for (question_number in 1:11) {
    d[q_num == (question_number-1), 'true_score'] <- true_values[question_number]
  }

  # Add delta columns = how off the "true value" the user guess was
  d[, ':=' (delta = user_score - true_score)]

  d <- d[treatment_group != "NA",]
  d <- d[user_score != "NA",]

  # Drop any users with less than the full number of questions answered
  counts <- d[ , .(count = length(q_num)), by = subject_id]
  incomplete_sub_ids <- counts[count < length(true_values), subject_id]
  d <- d[subject_id %!in% incomplete_sub_ids]

  # Assign groups based on our two-factor design
  d[treatment_group %in% list('Ambient Black', 'Shade Black', 'Sunlight Black'), background:='Black']
  d[treatment_group %in% list('Ambient White', 'Shade White', 'Sunlight White'), background:='White']
  d[treatment_group %in% list('Ambient Black', 'Ambient White'), lighting:='Ambient']
  d[treatment_group %in% list('Shade Black', 'Shade White'), lighting:='Shade']
  d[treatment_group %in% list('Sunlight Black', 'Sunlight White'), lighting:='Sunlight']

  return (d)
}
```

Figure 16: Marketing material used to catch the eye and the green-loving souls of MIDS students and coworkers. This campaign was quite effective, close to outpacing the ROI of Prolific.

```
emails_string <-
"email1,
email2,
emailn"

emails <- str_split(emails_string, '\n')[[1]]
num_users <- length(emails)

set.seed(num_users)

choice <- sample(emails, 1)
choice
```