Assessing the Effects of Color and Context on Analytical Problem Solving Ability

Maria Auslander, Praveen Joseph, Kevin Jun, Julie Nguyen

Research Question

This paper seeks to examine the research question, "Can the use of color and/or the use of questions with context help students improve analytical problem solving performance?". In order to address this question, a 2 x 3 factorial design with a within-subjects component is utilized. Each subject is asked six unique analytical problem solving questions each with 2 potential color conditions– 'no color' or 'with color', and 3 potential context conditions– 'question without context', 'question with context', and 'question with context and diagram'. We took steps to mitigate the carryover and learning effect of within subject studies, by using counterbalancing. The treatment conditions are assigned randomly per question and the order of each question is randomized. We discovered that the Color conditions were not found to be significant in improving the outcome of analytical problem solving questions. The third context condition, question context with diagram, had a significant positive effect on the question score outcome in some cases, however, there is a possibility that a treatment effect does not exist for this treatment condition either.

Introduction / Motivation

The topic for this experimental study bears resemblance to an influential study conducted in 1956 by George A. Miller - a cognitive psychologist from Harvard University's Department of Psychology. This study on human cognition led to Miller's law in psychology. The study titled "The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information" is one of the most highly cited papers in psychology published in Psychological Review (1956).

Our study follows the preamble of Miller's study and seeks to answer questions of human cognition when participants in the study are subject to treatment with colour. We would then devise a method to measure and test the cognitive capacity for processing (analyzing) and remembering through the process of learning-information.

The reason why this is an interesting experiment to design and execute is because the results could have a meaningful impact for delivering learning materials. As online learning becomes more prevalent, we seek to answer a pertinent question through this study; Can the use of color and/or the use of questions with context help students improve analytical problem solving performance?

Experimental Design

Our experiment aims to see if there is any significant improvement in analytical problem solving by (a) adding color, (b) adding context to the question, and/or (c) implementing both onto a straightforward black and white baseline question format.

The first arm of treatment is to add 'color', in which text-only problems are now shown on a cyan background with white letters and problems with diagrams are shown in their full colored version opposed to grey-scale.

Of note, the color cyan was selected randomly out of many purely to test the hypothesis and because it is distinguishable to color-blind respondents, if any.

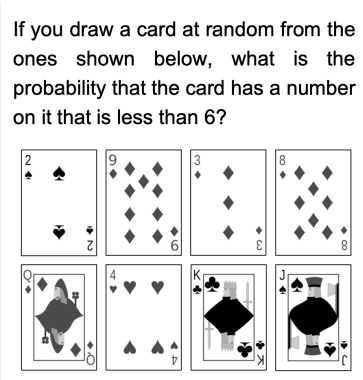


Figure 1: Sample Question Without Color, and Context plus Diagram

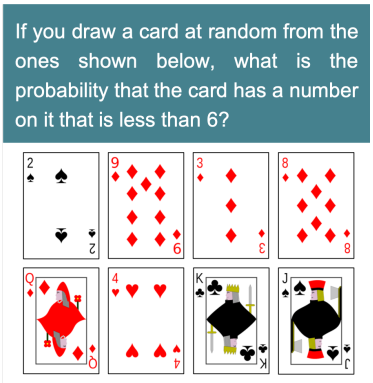


Figure 2: Sample Question With Color, and Context plus Diagram

Here, the expectation is that the colored version of the question will drive more engagement from the viewer and act as a better stimulus to understand the problem and hence solve it better.

The next arm of the treatment conditions is adding context to questions. The context 'treatment' consists of two steps: the first is by adding situational context to the question.

If you choose a number at random from the set (2, 9, 3, 8, 12, 4, 13, 11), what is the probability that the number is less than 6?

Figure 3: Sample Question Without Color and Without Context

If you draw a card at random from the following set (2♠, 9◇, 3◇, 8◇, Q◇, 4♡, K♣, J♠), what is the probability that the card has a number on it that is less than 6?

Figure 4: Sample Question Without Color and With Context

This part of the experiments benchmarks from the infamous 'selection task' devised by Peter Cathcart Wason in 1966. In this task, subjects are shown a set of 4 cards on the table, each of which has a figure on one side and a colored patch on the other side. People have to answer the logical question following: "Which cards must be turned over to test the idea that if a card shows an even number on one face, then the opposite face is red?" In this test, less than 10% of the subjects found the correct solution. However, subjects find the task much easier to solve if it is placed in the context of a social rule that they are asked to enforce. One of these experiments (Cosmides & Tooby, 1992) is using a set of 4 other cards on the table: two card have an age on one side and beverage on the other, e.g., "16", "25", "coke", "beer". When being asked "Which cards must be turned over to test the idea that if you are drinking alcohol then you must be over 18?". Most people correctly pick the correct cards ("16" and "beer"). Cosmides and Tooby's experiment supports Watson's assumption that the Watson selection task is highly content-dependant.

The key idea is similar in our case too, in that layering a relatable context to the analytical problem is intended to aid the participant. Like in the example illustrated above, instead of listing down a set of numbers, we put them into context as 'six-sided dice', 'prices of items', 'cards', 'people', or 'pets'. By giving tangible examples, the expectation is that the respondent will be able to understand the question better.

The second step of the context treatment is, adding to the context-treated problem a 'diagram' that visually illustrates the problem.

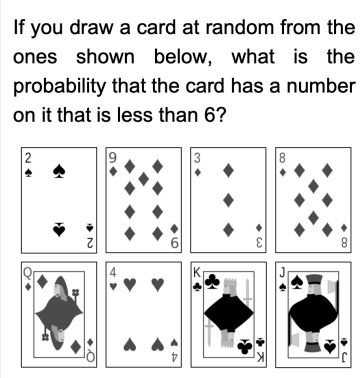


Figure 5: Sample Question Without Color and with Context plus Diagram

This stimuli is expected to provide visual cues to understand the question even more easily.

Since we are looking at two vectors of treatment, our NULL hypothesis can be divided into two:

1. *Color* has no effect on a subjects' ability to answer an analytical question
2. *Additional context* to the question has no effect on subjects' ability to answer an analytical question

In order to test this hypothesis, a 2 x 3 factorial design with a within-subjects component is utilized. Each subject is asked six unique analytical problem solving questions each with 2 potential color conditions– 'no color' or 'with color', and 3 potential context conditions– 'question without context', 'question with context', and 'question with context plus diagram'.

	No.Context	With.Context	Context.and.Diagram
No Color	Control (A-0)	Treatment (A-1)	Treatment (A-2)
With Color	Treatment (B-0)	Treatment (B-1)	Treatment (B-2)

For each participant, we ask a set of different analytical problem solving questions, each randomly assigned one of the six conditions from our 2x3 design, with replacement. We also randomize the order of the questions to even out any 'training effect' where respondents could increasingly become better at solving math problems throughout the survey itself. Lastly, participants were offered a reward for scoring above a certain threshold (*getting an overall score of 5 or higher*), as an incentive to do well on the problems presented in the study and avoid non-compliance (where a subject could simply click random answers to the questions without actually making an effort to solve them).

Pilot Review

A pilot study was conducted in order to determine the feasibility and data collection practices for the final study. The pilot consisted of two surveys sent to differing groups–one to friends and family, and another to subjects through prolific and google surveys. The surveys were created through Qualtrics - an online survey platform. Each survey had three questions and each possible color and context combination of conditions was tested through the surveys. Every subject completed three analytical problem solving questions and at the end of the survey each subject was asked demographic questions as well as questions on the difficulty of the problem-solving tasks. The aim of asking demographic questions was to test the effect of covariates on outcomes. Because our experiment consists of a within-subjects design, we wanted the questions to be of

similar difficulty coming out of the pilot study and into the full study to make a more clear comparison of outcomes amongst differing questions.

Pilot Study Results

As part of our analysis on the pilot study, we reviewed the response rate for questions. The response rate by question number per survey is listed below:

Study Version	Question Number	Response Rate
1	1	0.9615385
1	2	0.9615385
1	3	0.9615385
2	1	0.7142857
2	2	0.8571429
2	3	0.8571429

Study version 1 was sent to respondents on qualtrics and google surveys whereas study version 2 was sent only to friends and family. Because the analytical problem solving questions were not a requirement to submit the survey, users could submit the survey without answering questions. Looking at the table above, it is clear that question 1 from study version 2 had the lowest response rate. Additionally we received a higher frequency of comments on the difficulty of this question than any other question listed. Here is an example of one such statement when a subject was asked to comment on the difficulty of questions, "First question was impossible to comprehend". Upon further investigation we came to the conclusion that question 1 was worded poorly, hence we decided to refrain from using this question in the final study. Question 1 of the pilot study is listed in the appendix under A7.

Further analysis is conducted below on the inclusion of questions. Models are listed below per question number which regress dummy variables indicating whether the question number was the one of interest or not against the `correct` variable, indicating whether a question was correct. A sample formula is below (with a dummy variable for question 1): `correct~question_number==1`

The models are created using clustered standard errors, clustering on `ResponseId`, a variable which indicates a specific respondent. Clustered standard errors are used to control for differences among subjects. The results for study version 1 models are below, where the models are listed in order of question number.

```
##
## ============================================================
##                                   Dependent variable:
##                               -----------------------------
##                                          correct
##                                  (1)       (2)       (3)
## ------------------------------------------------------------
## question_number_factor == 1   -0.250**
##                                (0.102)
##
## question_number_factor == 2               0.154
##                                          (0.119)
##
## question_number_factor == 3                         0.096
##                                                    (0.120)
##
## Constant                      0.519***  0.385***  0.404***
##                                (0.074)   (0.069)   (0.069)
```

```
##
## ------------------------------------------------------------
## Observations                      78        78        78
## R2                               0.056     0.021     0.008
## Adjusted R2                      0.044     0.009    -0.005
## Residual Std. Error (df = 76)    0.488     0.497     0.500
## F Statistic (df = 1; 76)        4.550**    1.661     0.640
## ============================================================
## Note:                          *p<0.1; **p<0.05; ***p<0.01
```

From the results above, we concluded that question 1 has a significant, negative effect on whether a question is answered correctly. Questions 2 and 3 do not have significant effects on whether a question is answered correctly. Question 1 in study version 1 is not used in the final study. The results for study version 2 models are in A10 of the appendix. In the results for study version 2, it is evident that question 1 has a significant, negative effect on whether a question is answered correctly. Questions 2 and 3 do not have significant effects on whether a question is answered correctly. It is also notable that the low response rate for question 1 in study version 2 may make the comparison of outcomes for question 1 with the other question outcomes a less valid comparison, this is under the implication that subjects may have been less likely to answer questions they did not understand.

In determining the questions to include in the final study, in addition to comparing question results, we also want to avoid ceiling or floor effects. According to Garin, "The ceiling effect is said to occur when participants' scores cluster toward the high end (or best possible score) of the measure/instrument. The opposite is the floor effect." A ceiling effect can create an issue where a normal distribution predicts outcomes above the maximum level where a floor effect can create an issue where a normal distribution predicts outcomes below the minimum level. To avoid ceiling and floor effects, we examined questions to determine if too high a proportion of subjects were answering a question correctly or incorrectly. The same models listed above regressing question number dummy variables on `correct` (example: correct ~ question number == 1) can be used to assess the likelihood of ceiling and floor effects.

Below the proportions of users answering questions correctly and incorrectly by study number and question number are listed. Ultimately, to avoid ceiling effects it would be best if the proportion of subjects who answered a question correctly at around 50% and the proportion of subjects who answered a question incorrectly at around 50% as well. Looking at the table below, we do not see any overt cases for a question having a ceiling or floor effect (<20% correct or <20% correct). However, question 1 in study 1 has a fairly low percentage of subjects who've answered the question correctly (26.92 %) and question 3 in study 2 has a fairly high percentage of subjects who've answered the question correctly (71.43 %) .

Study Version	Question Number	Proportion Correct	Proportion Incorrect
1	1	0.2692	0.7308
1	2	0.5385	0.4615
1	3	0.5	0.5
2	1	0.4286	0.5714
2	2	0.619	0.381
2	3	0.7143	0.2857

After assessing results on the question level to determine which questions are appropriate for the final study, we look to determine whether the randomization method implemented by Qualtrics is viable. Because we did not develop the software/code that Qualtrics uses to randomize treatment, we created models for each color treatment condition and each treatment condition which regress covariates against treatment conditions to ensure covariate values did not affect treatment. The simple formulas for the models described are as follows (without coefficients):

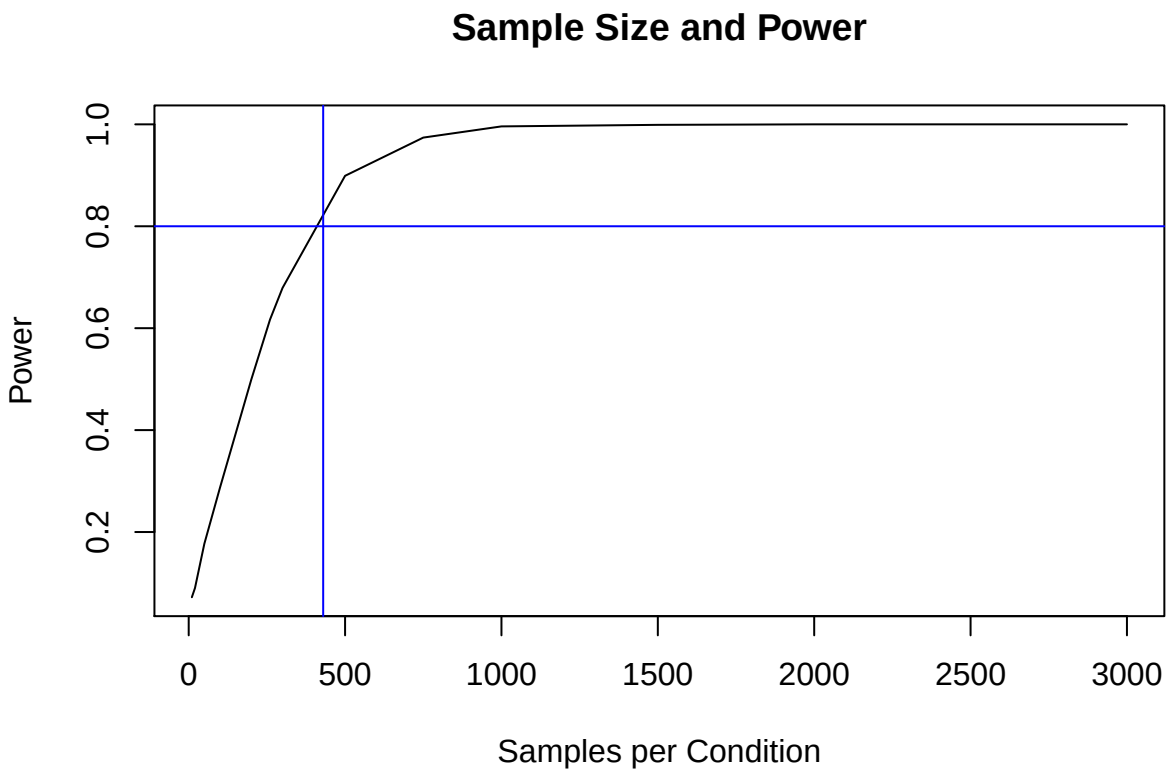
- color ~ gender + education +age
- context condition ~ gender
- color ~ age
- context condition ~ age
- color ~ education
- context condition ~ education

There is no reason to believe randomization of treatment conditions would be affected by Qualtrics software, Prolific or Google Surveys, however, because we did not design the software ourselves, we've decided to check that covariates have no effect on treatments assigned. The models described for gender are listed below, showing no significant effects of covariates on treatment conditions. The models for age and education are in A11 of the appendix. As with the gender models, no covariates had a significant effect on the treatment conditions assigned.

```
##
## gender models:
## ===============================================================
##                                    Dependent variable:
##                                ----------------------------
##                                    color         context
##                                     (1)            (2)
## ---------------------------------------------------------------
## Female                             -0.111         0.000
##                                   (0.157)        (0.257)
##
## Male                               -0.110         0.000
##                                   (0.160)        (0.262)
##
## Constant                          0.583***      2.000***
##                                   (0.146)        (0.238)
##
## ---------------------------------------------------------------
## Observations                        141            141
## R2                                 0.004          0.000
## Adjusted R2                        -0.011        -0.014
## Residual Std. Error (df = 138)     0.504          0.825
## F Statistic (df = 2; 138)          0.264          0.000
## ===============================================================
## Note:                           *p<0.1; **p<0.05; ***p<0.01
```

Looking at the outcomes above, no covariates have a significant effect on treatment condition selection. We assume the randomization of treatment leads to expected balance on covariates.

A short analysis of power based on pilot results is displayed below. However, the results in our pilot are not expected to perfectly match the results of the final study, particularly taking into account differences in outcomes (question score) by question number where the questions in the final study will not perfectly match those of the pilot. According to the results of the pilot study, in order to have 80% power, there will need to be about 430 samples per condition. In the final study, as many subjects as possible are recruited according to budgetary constraints. The lines represent the point at which 80% power will be achieved.



Final Research Methodology

Experimental Design Review

As previously mentioned, the final version of the study consists of a 2x3 factorial design where treatment can have one of two color conditions and one of three question context conditions. The two color conditions are without color (control) and with color. The three question context conditions are: without context (control), with context, and with context plus diagram. Six questions were selected to be used in assessing the effects of treatment conditions on subjects. These questions, each with all available conditions listed, are available in the appendix (A1-A6).

Subject Recruitment and Treatment Randomization

The research design has a within-subjects component–each subject is asked all six potential questions with differing, randomized treatment conditions in a randomized order. An illustration of our treatment randomization is below.

Based on the results of the pilot study, the final study was implemented using Qualtrics to create the survey and Prolific - an online participant recruitment platform - to recruit users to the survey. The experiment was limited in budget, so the sample size was determined based on what was afforded within this experimental constraint. We analyzed results from 260 subjects recruited through prolific. There was an initial pool of 132,393 subjects in Prolific. The total number of potential subjects was reduced to 101,740 when a condition was placed where subjects needed to be fluent in English in order to understand the questions at hand, and total number of potential subjects was reduced again to 93,903 when only subjects with an Prolific approval

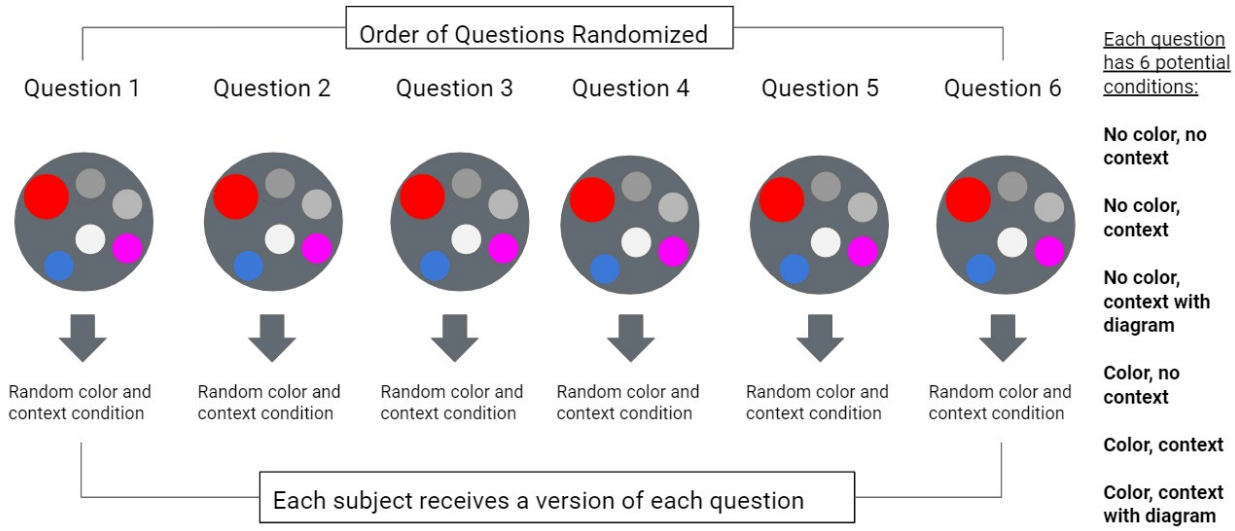


Figure 6: Treatment Randomization Illustration

rate greater than or equal to 50% were selected. 260 subjects were analyzed from the total pool of 93,903 subjects; these subjects were selected randomly. 266 subjects were initially selected for treatment but 6 subjects revoked consent. A consort diagram detailing subject selection is below.

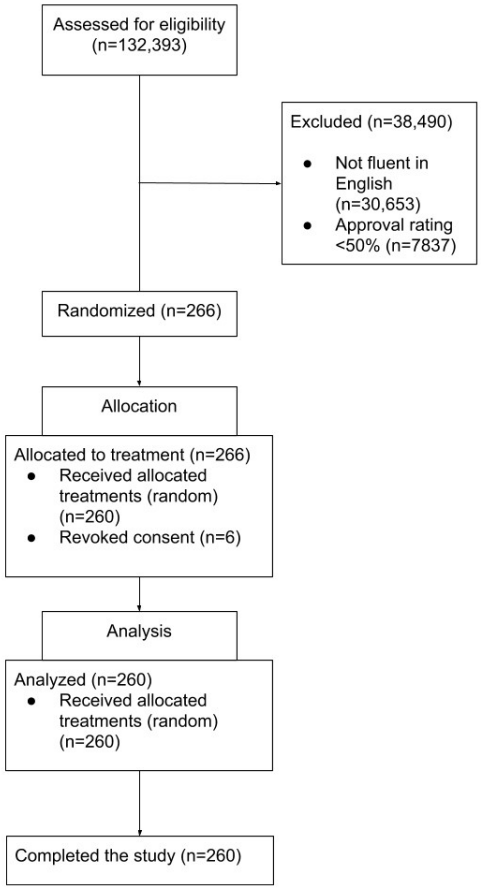


Figure 7: Consort Diagram

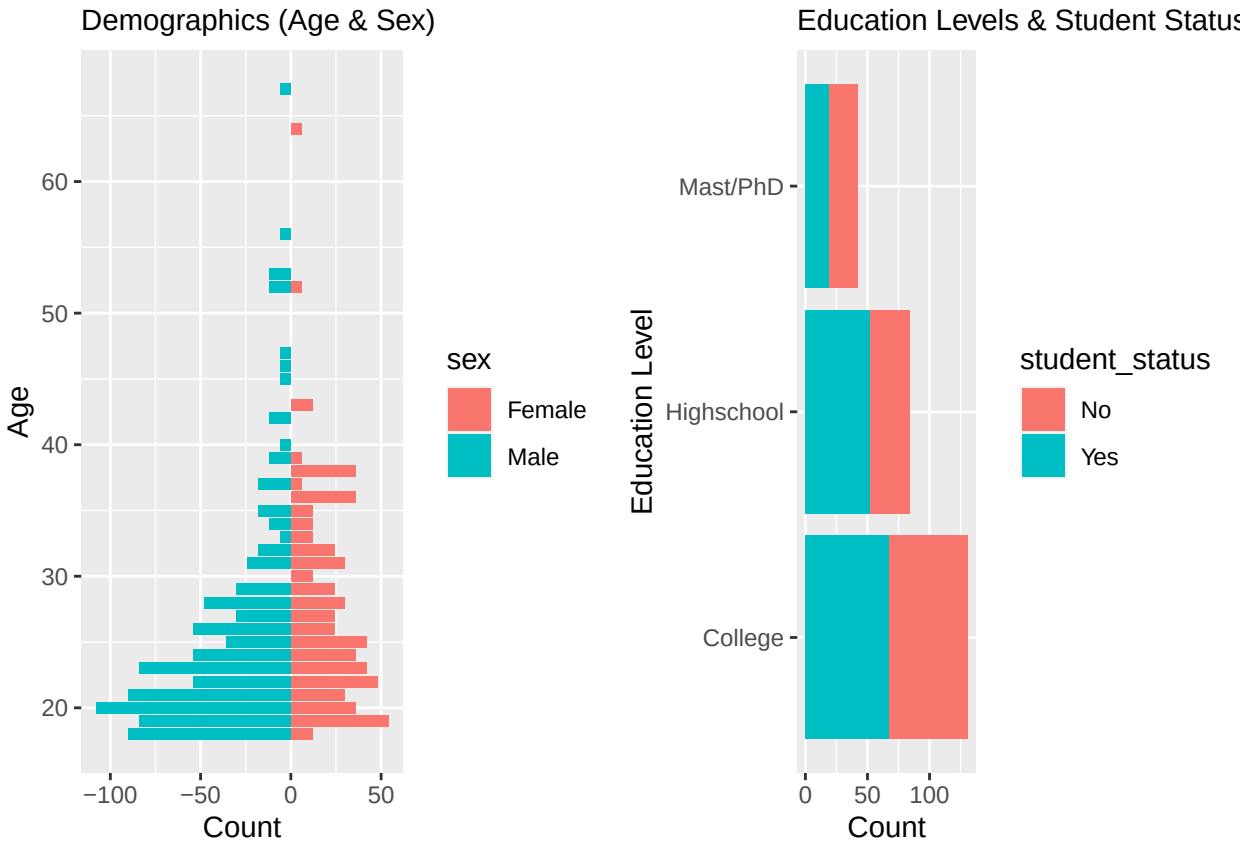
Outcome Measurement

In this study, each question answered by a subject consists of an observation and whether the question is answered correctly or not is the measured outcome. In the analysis the outcome variable (a boolean indicating whether a question was answered correctly) is listed as `score`. Potential treatment conditions whose outcomes are assessed are listed below:

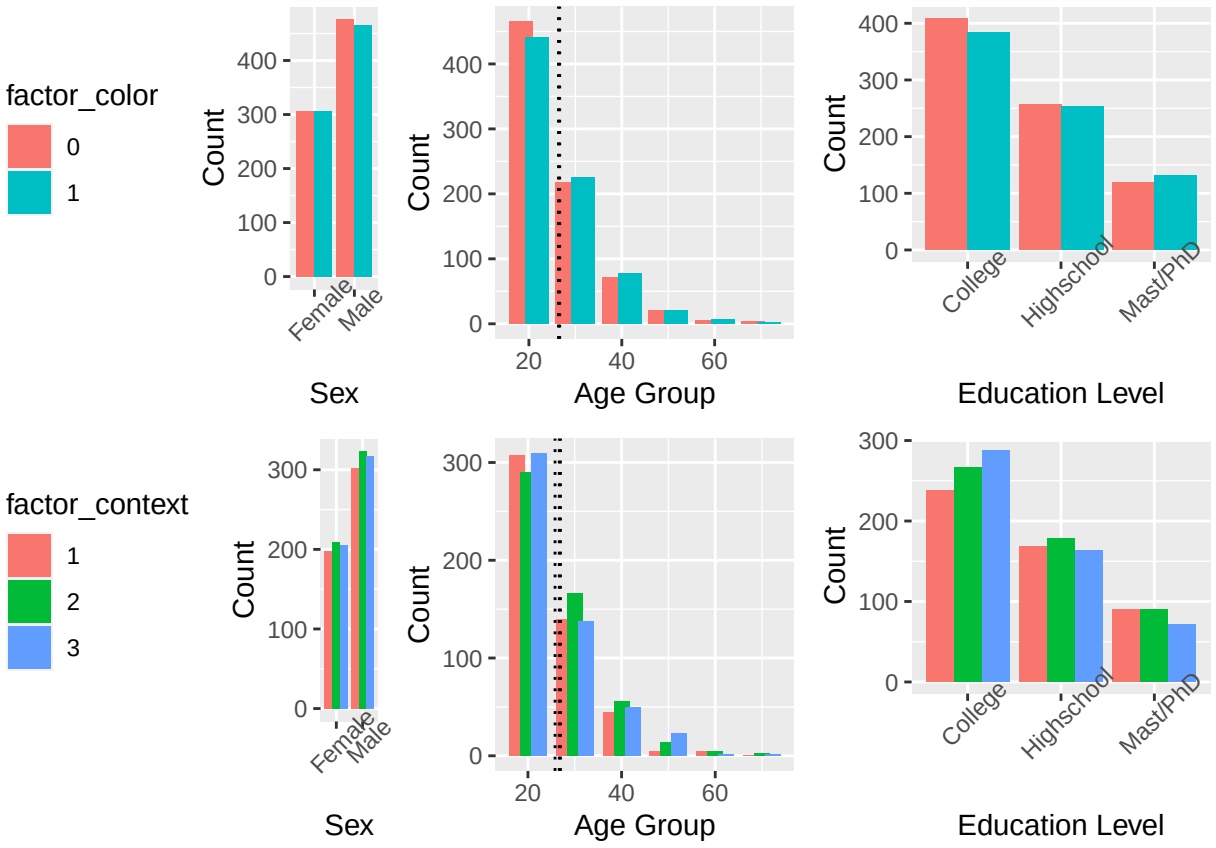
	No.Context	With.Context	Context.and.Diagram
No Color	Control (A-0)	Treatment (A-1)	Treatment (A-2)
With Color	Treatment (B-0)	Treatment (B-1)	Treatment (B-2)

Exploratory Data Analysis (EDA)

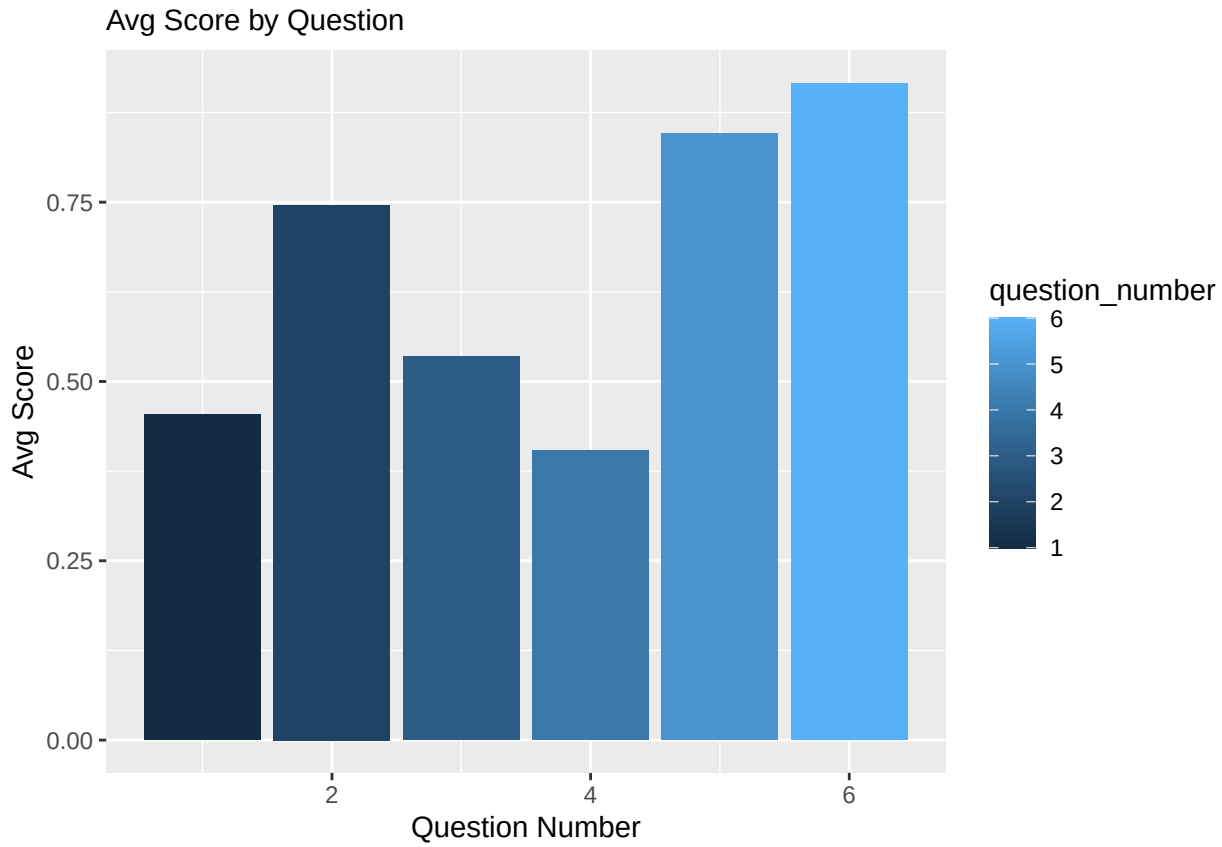
A total of 260 unique respondents completed our survey on prolific, the demographic was skewed towards younger respondents spread across different education levels (and nationalities mainly European, led by UK, Poland, and Portuguese). (We also saw that roughly half of all respondents are currently students, more prominently in Poland and Portugal.)



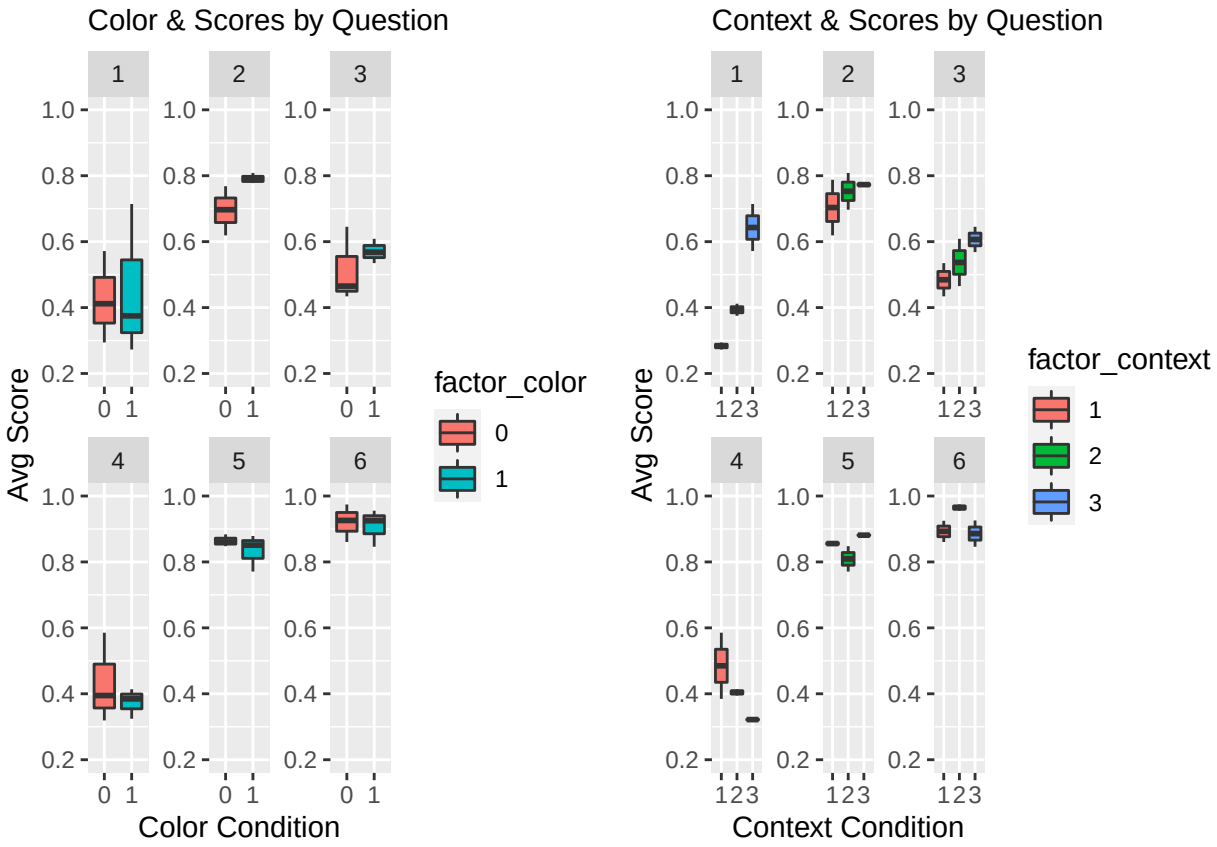
Through the Qualtrics survey platform, we were able to programmatically randomize the distribution of each condition of our treatment. For reassurance, we next visually check to see if either the color treatment arm or context treatment arm was skewed toward a certain covariate. The plots below show no major imbalances between treatment conditions across gender, age, education level, as the heights of each bar are similar within each breakout.



Each question as an aggregate yielded varying average scores. The average scores for questions 5 and 6 are very high, so there may be a potential ceiling effect. Within-subjects regression results control for differences in question number to eliminate this issue.



When we split the scores between those with 'color' treatment vs. those without, we do not see a consistent uplift when applying the treatment; only questions 2 and 3 seem to show some improvement in scores. Looking at scores among different 'context' treatments, we see that scores from 'context 3' questions yielded higher than the baseline ('context 1') for question 1, 2, 3, and 5.



Regression Results

Before using regression models to analyze the results, we review two primary assumptions of a within-subjects design, the no-anticipation and no-persistence assumptions.

Assessment of No-Anticipation and No-Persistence Assumptions

A within-subjects design relies on the no-anticipation assumption as well as the no-persistence assumption. According to Gerber and Green (2012), "the no-anticipation assumption states that potential outcomes are unaffected by treatments that are administered in the future" and "no-persistence assumption requires that potential outcomes in one period are unaffected by treatments administered in prior periods". The treatments in this case are the two color conditions (no color, and color) as well as the question context conditions (1: no context, 2: context, and 3: context with diagram). To assess whether the no-persistence and no-anticipation assumptions associated with within-subject designs hold, the existence of future and previous conditions were regressed against question `score`. The model formula is below:

$$\begin{aligned} score = & future\_question\_color + previous\_question\_color + previous\_question\_context1 \\ & + previous\_question\_context2 + previous\_question\_context3 + future\_question\_context1 \\ & + future\_question\_context2 + future\_question\_context3 \end{aligned}$$

The summary of the model in the table below shows no previous or future treatment conditions had a significant effect on question `score`. This indicates that the no-persistence and no-anticipation assumptions

of the within-subjects design hold. The model was created using `participant_id` fixed effects and clustered standard errors to control for differences amongst subjects.

```
##
## Model to Assess Within-Subjects Assumptions
## ========================================================
##                                Dependent variable:
##                            ----------------------------
##                                        score
## --------------------------------------------------------
## future.question.color                   0.003
##                                        (0.045)
##
## previous.question.color                 0.051
##                                        (0.041)
##
## previous.question.context1             -0.006
##                                        (0.041)
##
## previous.question.context2              0.002
##                                        (0.040)
##
## previous.question.context3             -0.040
##                                        (0.039)
##
## future.question.context1                0.035
##                                        (0.042)
##
## future.question.context2                0.042
##                                        (0.043)
##
## future.question.context3               -0.046
##                                        (0.043)
##
## Constant                              0.840***
##                                        (0.069)
##
## --------------------------------------------------------
## Observations                            1,554
## R2                                      0.313
## Adjusted R2                             0.170
## Residual Std. Error                0.434 (df = 1287)
## F Statistic                    2.200*** (df = 266; 1287)
## ========================================================
## Note:                         *p<0.1; **p<0.05; ***p<0.01
```

Stage I: Analysing the treatment effect at the individual question level.

In the within-subject design, since each question can be treated as a separate experiment, we analyze the treatment effect on the outcome variable (score) for each experiment. In stage I, by considering each question as a 'between-subjects' experiment we look for the average treatment effect (ATE) of color or context condition on score , which is a measure of an individuals' ability to answer analytical problems.

```
##
```

```
## ================================================================================================
##
##                          question 1 question 2 question 3 question 4 question 5 question 6
## ------------------------------------------------------------------------------------------------
## Constant                  0.294***   0.619***   0.434***   0.585***   0.860***   0.860***
##                           (0.081)    (0.077)    (0.069)    (0.079)    (0.054)    (0.054)
##
## color                      -0.021     0.169      0.101     -0.201*     -0.009     0.065
##                           (0.106)    (0.106)    (0.104)    (0.112)    (0.076)    (0.069)
##
## context2                   0.118      0.078      0.031     -0.191*     -0.013     0.114*
##                           (0.107)    (0.113)    (0.104)    (0.113)    (0.077)    (0.060)
##
## context3                  0.277**     0.149      0.211*    -0.266**     0.023      0.065
##                           (0.118)    (0.092)    (0.113)    (0.105)    (0.074)    (0.065)
##
## color:context2             -0.015     -0.057     0.043      0.220      -0.068     -0.083
##                           (0.149)    (0.147)    (0.150)    (0.154)    (0.112)    (0.080)
##
## color:context3             0.164      -0.159     -0.178     0.206      0.004      -0.144
##                           (0.150)    (0.138)    (0.157)    (0.154)    (0.108)    (0.098)
##
## N                           260        260        260        260        260        260
## Adjusted R2                0.090      0.002      0.003      0.012      -0.008     0.007
## ================================================================================================
## Stat. significance (p)              ***Significant at the 1 percent level.
##                                      **Significant at the 5 percent level.
##                                      *Significant at the 10 percent level.
```

- We find that the treatment of color has no significant treatment effect for any of the question
- However, the treatment effect for condition *3: Problem description with context and diagram* is statistically significant for questions 1, 3 and 4.

This is an encouraging result of the treatment effect on the outcome 'score' and we choose the explore the study results at the aggregate level by pooling the data together.

Stage II: Pool the data and test the treatment effect by considering 3 classes of models:

Framework for the analysis of results

According to Gerber and Green (2012), "the allure of within-subject design is their capacity to generate precise treatment estimates with a single subject" (p 273). Within-subjects regression requires including individual subject level fixed effects (using id which is unique to each participant). In the EDA we discovered that the average scores for questions 5 and 6 are very high, so there may be a potential ceiling effects. To control for the differences in question number we decided to also incorporate question number as a fixed effect in the model. We also test for the presence of HTE by including an interaction term. We do this using the following hierarchy of linear models:

1. **Simple model**: outcome ~ Treatment conditions
2. **Fixed effects model**: Outcome ~ Treat + Fixed Effects
3. **Fully saturated model**: Outcome ~ Treat + Fixed Effects + Interaction effects

Statistical significance and reporting standard errors

In within-subjects studies, "the key assumption is that the errors are uncorrelated across clusters while errors for individuals belonging to the same cluster may be correlated" (Cameron and Miller [2015], p. 320). Clustering is an experimental design feature because the outcome is correlated at the subject level since we observe the same individual's reponse to multiple questions, therefore we must cluster the results using subject id to ensure that the statistical significance of treatment effects are not overstated. The stargazer model results using clustered standard errors (clustered by id).

```
##
## OLS regression results with std errors clustered by participant ID
## ================================================================================
##
##                        [SIMPLE MODEL] [FIXED EFFECT MODEL] [FULLY SATURATED MODEL]
## --------------------------------------------------------------------------------
## Constant                   0.615***          0.596***              0.619***
##                            (0.028)           (0.039)               (0.047)
##
## color                       0.005             0.006                -0.030
##                            (0.024)           (0.023)               (0.044)
##
## context2                    0.021             0.007                -0.021
##                            (0.033)           (0.029)               (0.042)
##
## context3                   0.075**           0.067**                0.044
##                            (0.030)           (0.029)               (0.040)
##
## color:context2                                                     0.057
##                                                                   (0.059)
##
## color:context3                                                     0.048
##                                                                   (0.058)
##
## Fixed Effects                NO                YES                  YES
## N                          1,560             1,560                1,560
## Adjusted R2                0.002             0.380                0.380
## ================================================================================
## Stat. significance (p)           ***Significant at the 1 percent level.
##                                   **Significant at the 5 percent level.
##                                   *Significant at the 10 percent level.
```

Typically, the motivation given for the clustering adjustments is that unobserved components in outcomes for units within clusters are correlated. However, because correlation may occur across more than one dimension (for e.g. subject id and question number). Clustering standard errors by both dimensions of 'id' and 'question number' will produce results under a more conservative reporting standard, ensuring that the significance of the treatment effects are not overstated. We take the view that this second perspective best fits the experimental setting our of within-subjects study where clustering adjustments are used. Below, we present a stargazer output of the same hierarchy of linear models using clustered standard errors (clustered by both id and question number).

```
##
## OLS regression results with std errors clustered by both participant ID and question number
## ================================================================================
```

```
##
##                         [SIMPLE MODEL] [FIXED EFFECT MODEL] [FULLY SATURATED MODEL]
## ------------------------------------------------------------------------------------
## Constant                   0.615***           0.596                  0.619
##                            (0.102)
##
## color                       0.005             0.006                 -0.030
##                            (0.023)           (0.020)                (0.039)
##
## context2                    0.021             0.007                 -0.021
##                            (0.035)           (0.036)                (0.039)
##
## context3                    0.075             0.067                  0.044
##                            (0.071)           (0.064)                (0.051)
##
## color:context2                                                       0.057
##                                                                     (0.051)
##
## color:context3                                                       0.048
##                                                                     (0.073)
##
## Fixed Effects                NO                YES                   YES
## N                          1,560             1,560                 1,560
## Adjusted R2                0.002             0.380                 0.380
## ====================================================================================
## Stat. significance (p)           ***Significant at the 1 percent level.
##                                   **Significant at the 5 percent level.
##                                   *Significant at the 10 percent level.
```

The findings of the analysis and the interpretation of the model results:

1. The 'simple model' shows that the treatment condition *3: Problem description with context and diagram* is statistically significant but the test is able to explain very little changes in outcome due to treatment low adj. R-squared (~0%)
2. The more complex 'Fixed effects model' correctly captures the within-subject variation at the participant levels and produces a better fitting model with a higher adj. R-squared (37%) and a statistically significant treatment effect for condition 3.
3. The 'fully saturated model' which nests the 'fixed effects model' aims to test for additional interaction effects between the color and context treatment conditions. The model output shows that there is no heterogeneous treatment effect (HTE) between color and context and surprisingly also no statistically significant treatment effect when the interaction term is included in the model. This makes us skeptical of the results from the simpler nested model and conclude that the potential treatments effects might not actually exist.

4. We decided to re-run all 3 models with more stringent standard errors (clustered by both id and question number). To our surprise we found that none of the model results were statistically significant and the model inference is that none of the treatment effects are statistically significant.

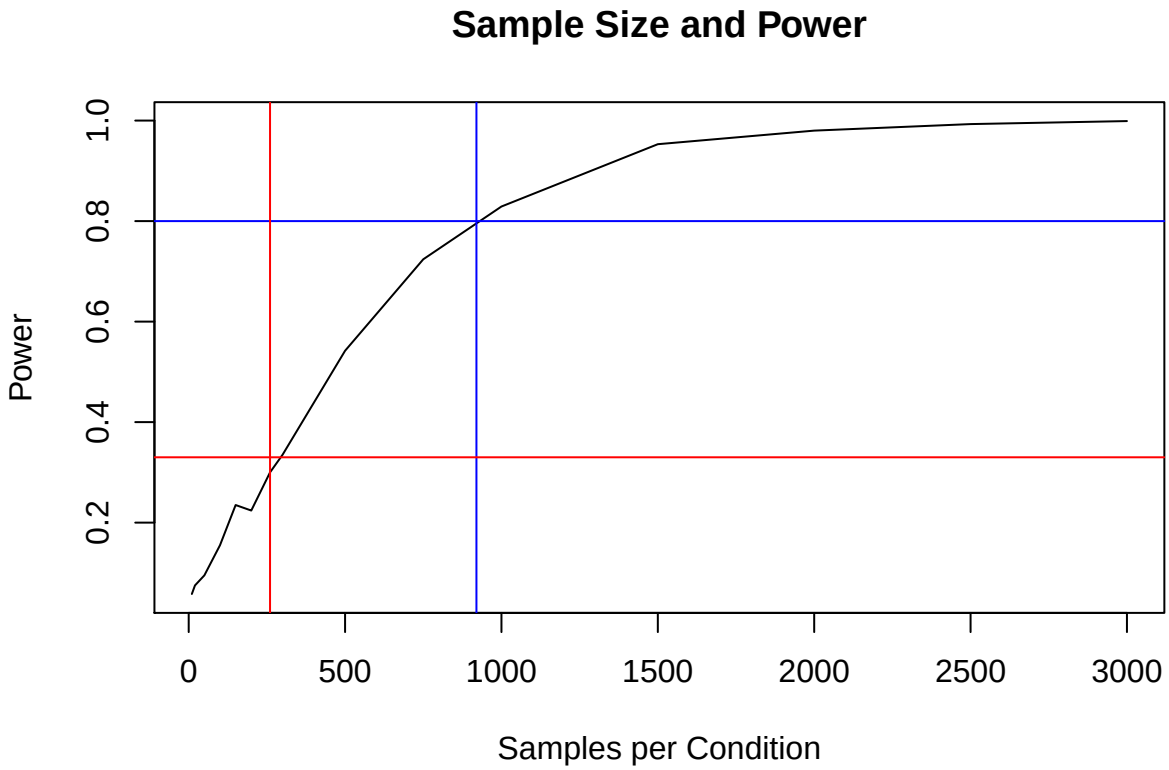
Estimating Statistical Power

Power analysis is an important aspect of experimental design. Statistical power or the power of a hypothesis test is the probability that the test correctly rejects the null hypothesis. It can be used as a tool to estimate

the number of observations or sample size required in order to detect an effect in an experiment.

In general, within-subject design results in greater statistical power than between-subject designs as in within-subject designs, each participant receives multiple conditions or treatments. In our example, by having 2x3 or 6 types of questions or conditions, we get 1560 observations upon 260 participants.

Though it is true that we need fewer participants in a within-subject design in order to find statistically significant effects, the power of the study depends on the average ATE as well as the standard deviation of the ATE. The smaller the ATE and the larger the standard deviations, the lower the power of the test. Our test is under-powered due to the budgetary constraints. We are uncertain that we have enough power to detect any real treatment effects in this study. It would be ideal to collect more samples or add more questions in order to get a higher power for this experiment. In the below graph, blue lines represent the number of samples needed to have 80% power for our overall study, and the red lines delineate the current level of power in our study given the number of samples per treatment condition.



Limitations of within-subjects Design

A potential limitation of the within-subjects research design is the problem of "carryover effect" and "learning effect".

- **Carryover effect**: In this experiment where each subject is administered 6 analytical questions, with multiple conditions, the participants may fatigue as they approach the end of the study and start to lose focus. This could potentially decrease their performance on the last questions in the study as compared to the earlier questions in the study.
- **Learning effect**: As the subject progresses through the study, the practice effect might mean that they are more confident and accomplished after the first treatment condition, and the experience of

having solved some analytical problems makes them more confident and better prepared to answer later questions in the study.

We account for "carryover effect" and "learning effect" by using counterbalance as part of the design. We implement counterbalancing where the order of thequestiona and the order of the treatments is varied. i.e. we randomized the order of the 6 questions to ensure that the order in which the questions are administered is fully randomized. At the same time we also randomize the treatment condition so each condition is different and the subject doesn't get the opportunity to get familiar or become practiced at answering analytical questions.

Advantages of within-subjects design

- A major advantage of within-subjects design is it eliminates all problems concerning individual differences, using a person as a control for himself reduces variance and produces the ultimate paired-subjects design as the same subject receives both control and treatment conditions. Another important advantage that the within subject design has over the between subject design is that it requires fewer participants, to have adequately powered-tests making the process much more streamlined and less resource heavy.
- For example, in this study we chose to test 6 conditions, using 6 types of questions for 260 participants. By treating each round of questions as a separate independent experiment we were able to get 1560 observations for this study, boosting the power of the results. Ease was not the only advantage, through carefully planned within subject design we were able to implement counterbalancing to lower the possibility of individual differences skewing the results.

Generalizability concerns

One factor that can affect the usefulness of our study, regardless of the strength of the design, is its generalizability. We believe the results of our study can be generalized across a large population but it's restricted to a very specific condition of analytical questions. We would be careful about extending the finding of the study outside the scope of the conditions in the context of online learning applications. For e.g., the study results would be difficult to extend to reading comprehension or other non-analytical problems since the diagram and context conditions are not easily applicable beyond the specific scope of the analytical problems which we tested.

Conclusion

At the close of this study, it was found that there was no significant treatment effect for the color conditions. The third context condition, question with context plus diagram, appeared to be statistically significant, but true treatment effects may not actually exist when more conservative reporting standards are used for measuring treatment effects. Therefore, we fail to reject both null hypothesis:

1. *Color* has no effect on a subjects' ability to answer an analytical question
2. *Additional context* to the question has no effect on subjects' ability to answer an analytical question

We conclude that the "use of color and/or the use of questions with context cannot be proven to help students improve analytical problem solving performance". In order to have more confidence in the results of a similar study, it would be beneficial to increase the number of subjects involved in the study or the number

of the questions in the study to increase the power of the tests. When adding more questions to increase the power of effects, it is necessary to further assess the possibility of no-persistence or no-anticipation violations. According to Gerber and Green (2012) "when both assumption hold, the within subjects desgin provides unbiased estimates of the ATE" (pg 275). Without a violation of the no-anticipation and no-persistence assumptions, the within-subjects design also leads to a higher power associated with treatment effects while being able to utilize the results of fewer subjects than the alternative between-subjects design.

References:

Barkow, J. H., Cosmides, L., & Tooby, J. (1992). The Adapted Mind: Evolutionary Psychology and the Generation of Culture. Oxford University Press, USA.

Garin O. (2014) Ceiling Effect. In: Michalos A.C. (eds) Encyclopedia of Quality of Life and Well-Being Research. Springer, Dordrecht. https://doi.org/10.1007/978-94-007-0753-5_296

Gerber, Alan S., and Donald P. Green. *Field Experiments: Design, Analysis, and Interpretation.* Norton & c., 2012.

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychological Review, 63(2), 81–97. https://doi.org/10.1037/h0043158

Wason, P. C. (1966). In B. Foss (ed.), New Horizons in Psychology. Harmondsworth: Penguin Books. pp. 135-151.

Appendix:

A1 through A6 in the appendix have the different questions asked in the full study with all potential conditions listed. The question conditions are listed in the following order per question: 1. No color, no context 2. Color, no context 3. No color, context 4. Color, context 5. No color, context with diagram 6. Color, context with diagram

Each question was asked with five multiple choice answers available. These options are listed after the questions with their conditions are listed.

A1, Question 1

X can be any one of the following numbers with equal probability (1, 2, 3, 4, 5, 6), Y can be any one of the following numbers with equal probability (1, 2, 3, 4, 5, 6). What is the probability X+Y=9?

Figure 8: Question 1, No Color, No Context:

X can be any one of the following numbers with equal probability (1, 2, 3, 4, 5, 6), Y can be any one of the following numbers with equal probability (1, 2, 3, 4, 5, 6). What is the probability X+Y=9?

Figure 9: Question 1, With Color, No Context:

If given two six-sided dice, what is the probability that the sum of the two numbers rolled will equal 9?

Figure 10: Question 1, No Color, With Context:

If given two six-sided dice, what is the probability that the sum of the two numbers rolled will equal 9?

Figure 11: Question 1, With Color, With Context:

Answer Options:

- 1/36
- 1/24
- 1/6
- 1/9 (Correct)
- 1/18

If given two six-sided dice, what is the probability that the sum of the two numbers rolled will equal 9?

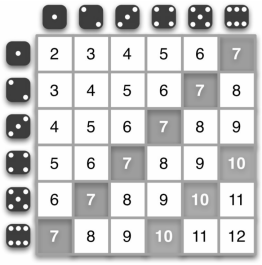


Figure 12: Question 1, No Color, With Context Plus Diagram:

If given two six-sided dice, what is the probability that the sum of the two numbers rolled will equal 9?

Figure 13: Question 1, With Color, With Context Plus Diagram:

A2, Question 2

There are 9 items of type A, 15 items of type B, 2 items of type C, and 6 items of type D. What is the probability that an item picked at random will NOT be of type B or of type C?

Figure 14: Question 2, No Color, No Context:

There are 9 items of type A, 15 items of type B, 2 items of type C, and 6 items of type D. What is the probability that an item picked at random will NOT be of type B or of type C?

Figure 15: Question 2, With Color, No Context:

Answer Options:

Sally wants to adopt a pet, but she's not sure which animal she wants. Her local animal shelter has 9 dogs, 15 cats, 2 hamsters, and 6 rabbits. What is the probability that Sally will NOT randomly pick a cat or hamster at the shelter?

Figure 16: Question 2, No Color, With Context:

Sally wants to adopt a pet, but she's not sure which animal she wants. Her local animal shelter has 9 dogs, 15 cats, 2 hamsters, and 6 rabbits. What is the probability that Sally will NOT randomly pick a cat or hamster at the shelter?

Figure 17: Question 2, With Color, With Context:

Sally wants to adopt a pet, but she's not sure which animal she wants. Her local animal shelter has 9 dogs, 15 cats, 2 hamsters, and 6 rabbits. What is the probability that Sally will NOT randomly pick a cat or hamster at the shelter?

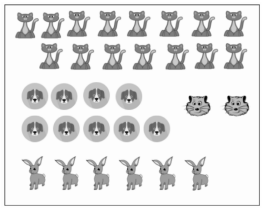


Figure 18: Question 2, No Color, With Context Plus Diagram:

Sally wants to adopt a pet, but she's not sure which animal she wants. Her local animal shelter has 9 dogs, 15 cats, 2 hamsters, and 6 rabbits. What is the probability that Sally will NOT randomly pick a cat or hamster at the shelter?

Figure 19: Question 2, With Color, With Context Plus Diagram:

- 1/3
- 15/32 (Correct)
- 3/16

- 5/16
- 1/2

A3, Question 3

There is a process consisting of steps X, Y, and Z. All steps together take 50 seconds. Step X takes 20 seconds, step Y takes 25 seconds, and step Z takes the remaining amount of time. Over a period of 100 seconds, what is the probability that the process will be at step Z at a randomly chosen time?

Figure 20: Question 3, No Color, No Context:

There is a process consisting of steps X, Y, and Z. All steps together take 50 seconds. Step X takes 20 seconds, step Y takes 25 seconds, and step Z takes the remaining amount of time. Over a period of 100 seconds, what is the probability that the process will be at step Z at a randomly chosen time?

Figure 21: Question 3, With Color, No Context:

A street light in Anytown, USA, completes a cycle from red to green to yellow, and back to red, in 50 seconds. During this time, the light will be red for 20 seconds, green for 25 seconds, and yellow for the remaining time. Over a period of 100 seconds, what is the probability that the light will be yellow at a randomly chosen time?

Figure 22: Question 3, No Color, With Context:

Answer Options:

- 1/50
- 1/20
- 1/10 (correct)
- 1/5
- 1/25

A street light in Anytown, USA, completes a cycle from red to green to yellow, and back to red, in 50 seconds. During this time, the light will be red for 20 seconds, green for 25 seconds, and yellow for the remaining time. Over a period of 100 seconds, what is the probability that the light will be yellow at a randomly chosen time?

Figure 23: Question 3, With Color, With Context:

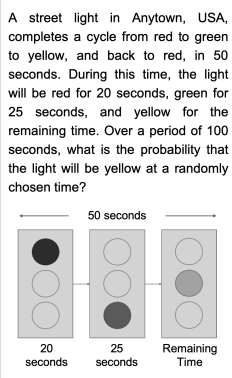


Figure 24: Question 3, No Color, With Context Plus Diagram:

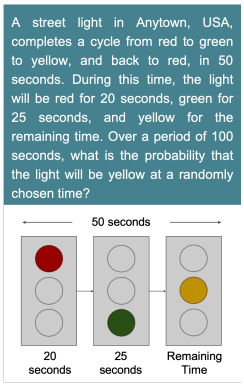


Figure 25: Question 3, With Color, With Context Plus Diagram:

A4, Question 4

Answer Options

- A
- B
- C
- D
- E (correct)

Which is the third highest number?

A: 25 * 0.8
B: 21
C: 20 * 0.6
D: 17
E: 16 * 1.1

Figure 26: Question 4, No Color, No Context:

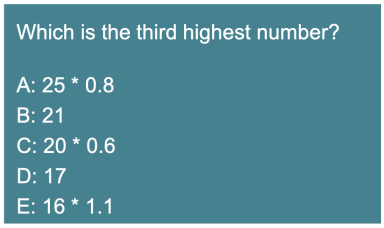


Figure 27: Question 4, With Color, No Context:

Which is the third highest priced item?

A: $25 shoes with 20% discount
B: $21 wine glass
C: $20 gloves with 40% discount
D: $17 noodles
E: $16 earphones with 10% markup

Figure 28: Question 4, No Color, With Context:

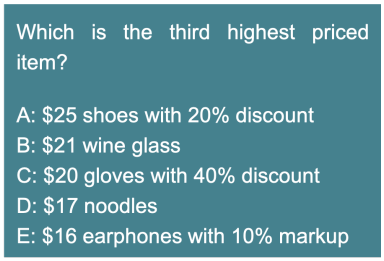


Figure 29: Question 4, With Color, With Context:

A5, Question 5

Answer Options:

- 1/4
- 3/10
- 3/8 (correct)
- 2/5
- 1/2

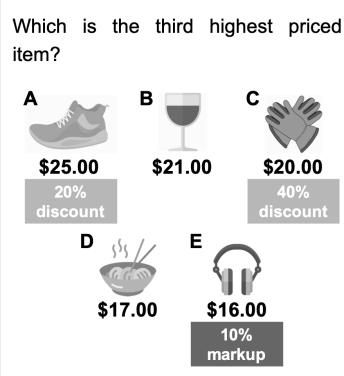


Figure 30: Question 4, No Color, With Context Plus Diagram:

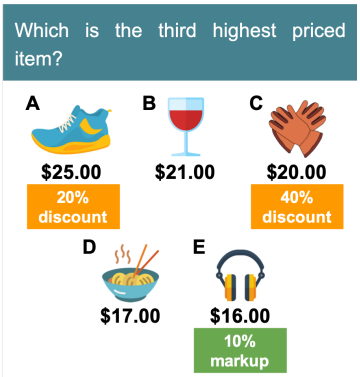


Figure 31: Question 4, With Color, With Context Plus Diagram:

If you choose a number at random from the set (2, 9, 3, 8, 12, 4, 13, 11), what is the probability that the number is less than 6?

Figure 32: Question 5, No Color, No Context:

If you choose a number at random from the set (2, 9, 3, 8, 12, 4, 13, 11), what is the probability that the number is less than 6?

Figure 33: Question 5, With Color, No Context:

If you draw a card at random from the following set (2♠, 9◇, 3◇, 8◇, Q◇, 4♡, K♣, J♠), what is the probability that the card has a number on it that is less than 6?

Figure 34: Question 5, No Color, With Context:

If you draw a card at random from the following set (2♠, 9◇, 3◇, 8◇, Q◇, 4♡, K♣, J♠), what is the probability that the card has a number on it that is less than 6?

Figure 35: Question 5, With Color, With Context:

If you draw a card at random from the ones shown below, what is the probability that the card has a number on it that is less than 6?

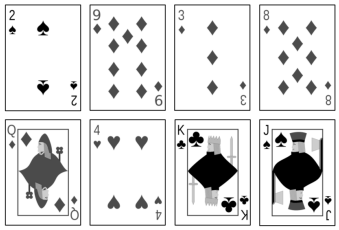


Figure 36: Question 5, No Color, With Context Plus Diagram:

If you draw a card at random from the ones shown below, what is the probability that the card has a number on it that is less than 6?

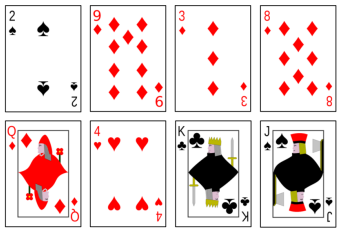


Figure 37: Question 5, With Color, With Context Plus Diagram:

A6, Question 6

A is bigger than S. M is bigger than A but smaller than K. K is bigger than S. S is smaller than M. G is the biggest. Who is the smallest?

Figure 38: Question 6, No Color, No Context:

Answer Options:

- Karish / K
- Asha / A

A is bigger than S. M is bigger than A but smaller than K. K is bigger than S. S is smaller than M. G is the biggest. Who is the smallest?

Figure 39: Question 6, With Color, No Context:

Asha is older than Swati. Mohtarma is older than Asha but younger than Karish. Karish is older than Swati. Swati is younger than Mohtarma. Gauri is the oldest. Who is the youngest?

Figure 40: Question 6, No Color, With Context:

Asha is older than Swati. Mohtarma is older than Asha but younger than Karish. Karish is older than Swati. Swati is younger than Mohtarma. Gauri is the oldest. Who is the youngest?

Figure 41: Question 6, With Color, With Context:

Asha is older than Swati. Mohtarma is older than Asha but younger than Karish. Karish is older than Swati. Swati is younger than Mohtarma. Gauri is the oldest. Who is the youngest?

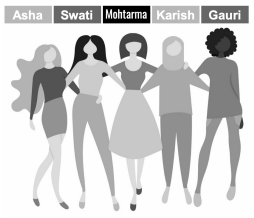


Figure 42: Question 6, No Color, With Context Plus Diagram:

- Mohtarma / M
- Swati / S
- Gauri / G

Appendices A7 through A9 show the three questions that were not included in the full study. Images are listed with the three context conditions in the following order: no context, context, and context with diagram. Answer options are available in question images.

A7

A7 in the appendix lists question 1 associated with both version 1 and version 2 of the pilot study.

Asha is older than Swati. Mohtarma is older than Asha but younger than Karish. Karish is older than Swati. Swati is younger than Mohtarma. Gauri is the oldest. Who is the youngest?

Figure 43: Question 6, With Color, With Context Plus Diagram:

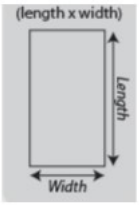
If x is chosen at random from the set (4, 6, 7, 9, 11) and y is chosen at random from the set (12, 13, 15, 17) then what is the probability that x*y is odd?

a. 6/10
b. 11/20
c. 3/10
d. 9/20
e. 9/10

You are in the process of buying a picture frame, but you're unsure of which size picture frame you'd like. The possible options for the length of the picture frame are (4, 6, 7, 9, 11), all in inches. The possible options for the width of the picture frame are (12, 13, 15, 17), all in inches. You'd like the area of the picture frame to be odd, how many picture frame options do you have? (Picture frames are available for every combination of length and width.)

a. 6/10
b. 11/20
c. 3/10
d. 9/20
e. 9/10

You are in the process of buying a picture frame, but you're u
picture frame you'd like. The possible options for the length o
are (4, 6, 7, 9, 11), all in inches. The possible options for the
frame are (12, 13, 15, 17), all in inches. You'd like the area of
be odd, how many picture frame options do you have? (Pictu
available for every combination of length and width.)



a. 6/10
b. 11/20
c. 3/10
d. 9/20
e. 9/10

A8

What is 55 x 2 x 12?

Answer options:

A. 1120
B. 1320
C. 1597
D. 1980
E. 2100

Answer: B

What is the *annu
two bags of pet foo
each bag costs $5

Answer options:

A. $1120
B. $1320
C. $1597
D. $1980
E. $2100

Answer: B

A9

Six cards are each written A, B, C, D, E, and F on the front. If cards “A”, “C”, and “E” each showed “b”, “d”, and “f” on their backs, and each letter is only written once, how many cards are written with the same letter on both sides?

Answer options:

A. 0
B. 1
C. 2
D. 3
E. 4

Answer: A 0

Six veg
vegetab
bell pe
vegetab
bin labe
“bell pe
has squ
How ma

Answe
A.
B.
C.
D.
E.

Answe

A10

```
##
## Question Models
## ============================================================
##                                    Dependent variable:
##                                ----------------------------
##                                           correct
##                                  (1)        (2)      (3)
## ------------------------------------------------------------
## question_number_factor == 1   -0.238***
##                                (0.089)
##
## question_number_factor == 2                0.048
##                                           (0.134)
##
## question_number_factor == 3                          0.190
##                                                     (0.131)
##
## Constant                      0.667***  0.571***  0.524***
##                               (0.101)    (0.077)  (0.076)
```

```
##
## ------------------------------------------------------------
## Observations                        63        63        63
## R2                                0.052     0.002     0.033
## Adjusted R2                       0.036    -0.014     0.017
## Residual Std. Error (df = 61)     0.487     0.500     0.492
## F Statistic (df = 1; 61)          3.344*    0.127     2.099
## ============================================================
## Note:                            *p<0.1; **p<0.05; ***p<0.01
```

A11

```
##
## age models:
## ================================================================
##                                      Dependent variable:
##                                  ---------------------------
##                                      color        condition
##                                       (1)            (2)
## ----------------------------------------------------------------
## shared_df[[cov]]18 to 24 years       -0.250        -0.000
##                                     (0.172)        (0.286)
##
## shared_df[[cov]]25 to 34 years       -0.028        -0.000
##                                     (0.161)        (0.267)
##
## shared_df[[cov]]35 to 44 years       -0.155        -0.000
##                                     (0.182)        (0.303)
##
## shared_df[[cov]]45 to 54 years       -0.028        -0.000
##                                     (0.222)        (0.369)
##
## shared_df[[cov]]55 to 64 years       -0.083        -0.000
##                                     (0.205)        (0.342)
##
## shared_df[[cov]]Age 65 or older      -0.250        -0.000
##                                     (0.325)        (0.541)
##
## Constant                            0.583***      2.000***
##                                     (0.145)        (0.242)
##
## ----------------------------------------------------------------
## Observations                          141            141
## R2                                   0.036          0.000
## Adjusted R2                          -0.007        -0.045
## Residual Std. Error (df = 134)       0.503          0.838
## F Statistic (df = 6; 134)            0.827          0.000
## ================================================================
## Note:                              *p<0.1; **p<0.05; ***p<0.01
##
## education models:
## ============================================================================
##                                                    Dependent variable:
```

```
##                                                 ----------------------------
##                                                     color        condition
##                                                      (1)            (2)
## ------------------------------------------------------------------------------
## shared_df[[cov]]Associate degree                    -0.250         0.000
##                                                    (0.253)        (0.422)
##
## shared_df[[cov]]Bachelor's degree                   -0.083         0.000
##                                                    (0.173)        (0.288)
##
## shared_df[[cov]]Completed some college              -0.250        -0.000
##                                                    (0.188)        (0.314)
##
## shared_df[[cov]]Completed some high school          -0.183         0.000
##                                                    (0.196)        (0.327)
##
## shared_df[[cov]]Completed some postgraduate         -0.028         0.000
##                                                    (0.223)        (0.372)
##
## shared_df[[cov]]High school graduate                -0.167         0.000
##                                                    (0.179)        (0.298)
##
## shared_df[[cov]]Master's degree                      0.042        -0.000
##                                                    (0.179)        (0.298)
##
## shared_df[[cov]]Ph.D., law or medical degree        0.083         0.000
##                                                    (0.326)        (0.545)
##
## Constant                                           0.583***      2.000***
##                                                    (0.146)        (0.244)
##
## ------------------------------------------------------------------------------
## Observations                                         141            141
## R2                                                  0.043          0.000
## Adjusted R2                                         -0.015        -0.061
## Residual Std. Error (df = 132)                      0.505          0.844
## F Statistic (df = 8; 132)                           0.739          0.000
## ==============================================================================
## Note:                                      *p<0.1; **p<0.05; ***p<0.01
```