The effects of categorization on active memory. Final Report Writeup

Noah Levy, John V. Tabbone, Utthaman Thirunavukkarasu

April 23, 2018

Research Question

A body of neurological research developed over the past twenty years suggests that many mental activities commonly thought of as occurring in our free thinking and conscious minds are actually heavily or completely, supported by underlying physical structures of the brain. The theory is that much of what we interpret as free will occurs because our brains are hard wired to behave that way.

One important such activity is our mind's preference to work with groups or categories. A thought experiment will make this clear. Suppose you are driving home from work and your spouse calls with a list of twenty items to pick up from the grocery store on the way home. Your spouse gives you the list in random order, 'turkey', 'apples', 'milk'.... and so on. It is unlikely that you will bring home many correct items. However, if your spouse gives you a categorized list such as "Bring home these fruits: apples, bannanas, grapes, mangos." "Bring home these meats: turkey, chicken,lamb, beef...." It would seem more likely that you would remember more correct items. This is the phenomena that our team tested, and the issue that became our research question:

Can people can recall catgeorized information better than non-categorized information.

Research Design

Treatment

To assess differences in the ability to recall categorized and uncategorized lists, we created a list of 20 common food items that could be found at a grocery store in the U.S. Then, we created two different versions of that list for the different experimental groups in our study. The control version of the survey listed all of the items in random order, while the treatment version of the survey listed all of the items by category. We then created audio recordings reading the two lists. The audio file for the control survey just read the randomized list of items, while the audio file for the treatment survey announced the relevant category and then listed the items in that category (e.g. "Vegatable category: lettuce, beans,... "). We chose audio presentation of the data to deliver a fixed dosage, roughly 15 seconds of information. Had we provided a written list to the subject, some may have repeated the items more often than others, exercising memory attributed to repitition. Some may have a better visual memory than other. Providing audio was a way tio fix the subject's exposure time. The two lists are illustrated in Tables 1 and 2:

Table 1: Randomly-ordered list of items. Note: numbers not included in audio recording

1. chicken	2. toast	3. cucumbers	4. cabbage	5. cheese
6. coffee	7. lettuce	8. pineapple	9. mango	10. eggs
11. lamb	12. bananas	13. turkey	14. steak	15. apples
16. beans	17. ice cream	18. milk	19. cereal	20. yogurt

Table 2: Categorized list of items. Each category was read in audio file before the corresponding items were listed

category	items
Vegetable	lettuce,beans,cabbage,cucumbers
Fruit	apples,bananas,pineapple,mango
Meat	steak,turkey,chicken,lamb
Breakfast	cereal,eggs,toast,coffee
Dairy	milk,cheese,yogurt,ice cream

Survey Overview

We hosted our survey on Qualtrics. For each subject who took our survey, the site would randomly provide either the categorized recording or the uncategorized recording. Each iteration of the survey asked participants to do the following:

1. Provide email address or Mechanical Turk Worker ID.
2. Listen to the audio recording.
3. Answer a few brainteaser questions.
4. List as many items from the audio recording as you can remember.
5. Provide feedback on the survey.

The dependent variable of the study was the proportion of the 20 listed items that each participant was able to recall. The ROXO design for this survey is as follows:

Table 3: ROXO Grammar

group	grammar
Categorized	R X O
Uncategorized	R - O

Survey Administration

Pilot Study

Overview

We did a pilot study with 11 participants before we administered the main survey. This preliminary study had two goals:

1. Identify potential issues with our survey design.
2. Obtain estimates for the population variances of the proportion of correct items recalled in the randomized and categorized treatment options. These estimates could then be used for power calculations.

We asked the following brainteaser question to distract the subjects after they heard the audio file:

If it takes 1 machine 1 minute to produce 1 good, how many minutes does it take 500 machines to produce 500 goods?

a. 500 Minutes
b. 60 Minutes
c. 1 Minute

Additionally, after the participants listed the items that they could remember, we asked them two questions about their experience taking the survey:

Did you understand the voice in the audio recording?

Do you think that the survey asked you to recall a reasonable number of items? Did you think that the list of items in the audio recordings was too long?

Evaluating Correct Responses

We manually counted the proportion of items that each subject recalled correctly. Errors in spelling or singular vs. plural were counted as correct as long as we were able to identify the item that the subject was referring to.

Results and Analysis

The ATE of hearing the categorized list rather than the random list of the percentage of items recalled, as well as the sample variances of each group, are listed in Table 4:

Table 4: Pilot Survey results

ATE	0.160
Cat. Sample Variance	0.095
Uncat. Sample Variance	0.048
Cat. Sample Size	6.000
Uncat. Sample Size	5.000

The ATE is fairly large (16% more items recalled correctly) , but it's not statistically significant because the sample is so small.

Based on this data, we were able to perform a power calculation to determine the sample sizes needed to have an 80% chance of detecting an effect size of at least 0.1. We used the methodology outlined in List et al., 2009:

```r
#get standard deviation of each group
sd_cat<-sqrt(var_cat)
sd_uncat<-sqrt(var_uncat)

#calculate minimum necessary sample size
pi_0<-sd_uncat/(sd_cat+sd_uncat)
pi_1<-sd_cat/(sd_cat+sd_uncat)

N<-(((1.96+0.84)/0.1)^2)*(var_uncat/pi_0 + var_cat/pi_1)
num_uncat_subjects<-round(N*pi_0,0)
paste('minimum number of subjects in random order group: ',num_uncat_subjects)
```

```
## [1] "minimum number of subjects in random order group:  91"
```

```r
num_cat_subjects<-round(N-num_uncat_subjects,0)
paste('minimum number of subjects in categorized group: ',num_cat_subjects)
```

```
## [1] "minimum number of subjects in categorized group:  127"
```

Based on this analysis, we determined that we would need at least 127 + 91 = 218 subjects in the main experiment to have a decent chance of detecting an effect size of at least 10%.

Other Takeaways from Power Study

None of the subjects indicated in the survey that they couldn't understand the audio recording, but a few of Noah's co-workers told him that they had trouble hearing the files while they were connected to their employer's corporate network. This alerted us to the possibility of firewall issues, so we thought it would be prudent to also include a question about understanding the audio recording in the final survey too.

A few participants also indicated that certain parts of the audio recording sounded blurry. Upon further review, we decided that participants might have an easier time understanding a human voice than a robotic voice, and we modified the audio files for the final survey accordingly.

Before the pilot study, we had considered asking subjects to recall items from two different lists in the final survey. We initially planned to randomly assign participants to receive either the randomized or the categorized grocery list, and then assign them to the opposite condition for a separate list of unrelated items like tv shows. This design would have enabled us to make within-subjects comparisons in the final analysis. However, the power calculation from the pilot study indicated that we would need more than 200 subjects just to have a decent chance of rejecting the null hypothesis for one comparison. Thus, we didn't add a second list for the final survey.

Finally, several subjects indicated that they thought that the length of the list of items that they were asked to remember was too long. However,several participants correctly remember 70-90% of the items correctly, and no participant managed to recall 100% of the items. Based on these findings, we decided that 20 items would be reasonable for the final survey.

Main Study

Gathering Participants

Given the sample size requirements for adequate power, we used Amazon's Mechanical Turk service to recruit 199 subjects for the main survey, and gathered the remaining participants from Noah and John's social and professional networks. The Mechanical Turk participants were restricted to the U.S., because we were concerned that some of the items in the list wouldn't be familiar to participants in other countries.

To incentivize people in our social networks to take our survey, we promised to give $50.00 Amazon gift cards to four randomly-selected participants. Ultimately, 34 individuals from our combined networks participated in our study.

Covariate Questions

In addition to asking for participants' email addresses or Mechanical Turk worker IDs, we asked them to provide basic demographic information that we thought might be correlated with the percentage of items that they recall. We asked each subject to specify their education level, gender, and age range. The full list of possible options is as follows:

Table 5: Education Level Options

No high school
some high school, no diploma
high school diploma or equivalent certificate
trade/vocational training
some college
some graduate school
graduate degree
Prefer not to specify

Table 6: Age Range Options

Under 12 years
12 - 17 years
18 - 24 years
25 - 34 years
35 - 44 years
45 - 54 years
55 - 64 years
65 -74 years
Over 75 Years
Prefer not to specify

Table 7: Gender Options

Male
Female
Prefer not to specify

Brainteaser

We included the brainteaser from the pilot survey in the final study, and we also added one additional brainteaser:

If it were two hours later, it would be half as long until midnight as it would be if it were an hour later. What time is it now?

a. 11 PM
b. 9 PM
c. 10 PM

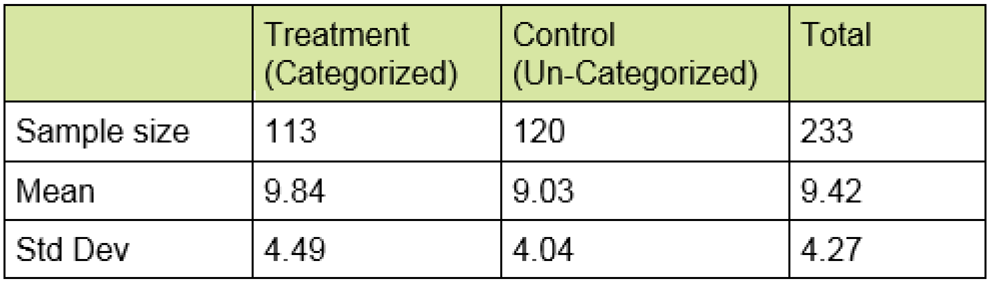
Validation

In order to identify participants who couldn't understand the audio recording because of either a lack of clarity or a corporate firewall issue, we included a question about whether or not the participant was able to understand the voice in the audio recording. The message that we used to recruit participants specifically instructed people to answer "No" in that question if their corporate firewall had blocked the audio files.

Analysis and Results

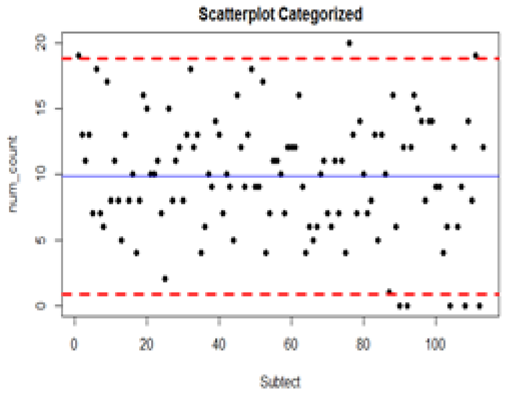
Descriptive Statistics

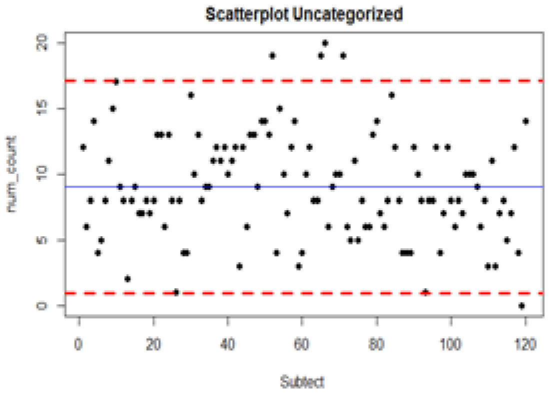
As mentioned earlier in the report we had a total of 233 subjects, 199 from Mechanical Turk and 34 from our social and professional networks. The randomization produced fairly balanced treatment groups: 120 subjects received the randomized survey and 113 participants received the categorized survey:

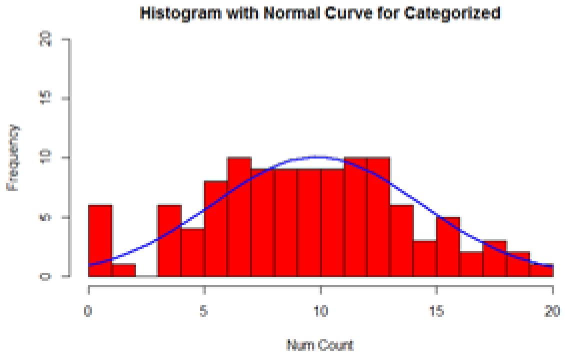


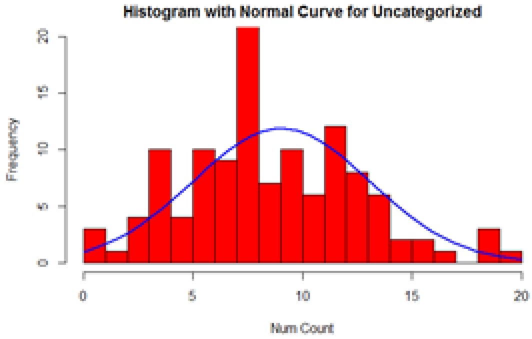
	Treatment (Categorized)	Control (Un-Categorized)	Total
Sample size	113	120	233
Mean	9.84	9.03	9.42
Std Dev	4.49	4.04	4.27

The scatterplots and histograms of the randomized and the categorized groups show that the outcomes are rougly normally distributed. We don't see any sytematic patterns in the outcome:



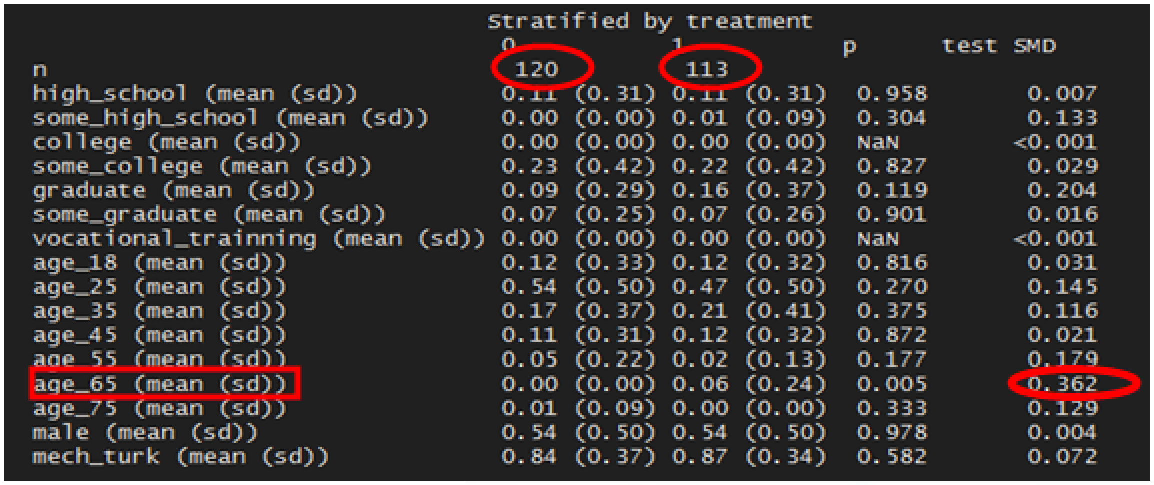






Randomization and Covariate Balance Checks

For our experiment we didn't collect data on baseline characteristics of subjects before treatment assignment ,and since the survey in our experiment was a self-selected one we didn't know the details about the subjects until we saw the survey results. The results of balance check were encouraging: almost all the covariates were well-balanced except for the fact that everyone in the 65 to 74 age group received the categorized list. We think that this imbalance occured by chance though, because the Qualtrics question assignment is completely random. The table below illustrates the covariate balance levels:



```
                                 Stratified by treatment
                                  0             1             p      test SMD
  n                                120           113
  high_school (mean (sd))         0.11 (0.31) 0.11 (0.31)  0.958        0.007
  some_high_school (mean (sd))    0.00 (0.00) 0.01 (0.09)  0.304        0.133
  college (mean (sd))             0.00 (0.00) 0.00 (0.00)  NaN         <0.001
  some_college (mean (sd))        0.23 (0.42) 0.22 (0.42)  0.827        0.029
  graduate (mean (sd))            0.09 (0.29) 0.16 (0.37)  0.119        0.204
  some_graduate (mean (sd))       0.07 (0.25) 0.07 (0.26)  0.901        0.016
  vocational_trainning (mean (sd)) 0.00 (0.00) 0.00 (0.00)  NaN         <0.001
  age_18 (mean (sd))              0.12 (0.33) 0.12 (0.32)  0.816        0.031
  age_25 (mean (sd))              0.54 (0.50) 0.47 (0.50)  0.270        0.145
  age_35 (mean (sd))              0.17 (0.37) 0.21 (0.41)  0.375        0.116
  age_45 (mean (sd))              0.11 (0.31) 0.12 (0.32)  0.872        0.021
  age_55 (mean (sd))              0.05 (0.22) 0.02 (0.13)  0.177        0.179
  age_65 (mean (sd))              0.00 (0.00) 0.06 (0.24)  0.005        0.362
  age_75 (mean (sd))              0.01 (0.09) 0.00 (0.00)  0.333        0.129
  male (mean (sd))                0.54 (0.50) 0.54 (0.50)  0.978        0.004
  mech_turk (mean (sd))           0.84 (0.37) 0.87 (0.34)  0.582        0.072
```

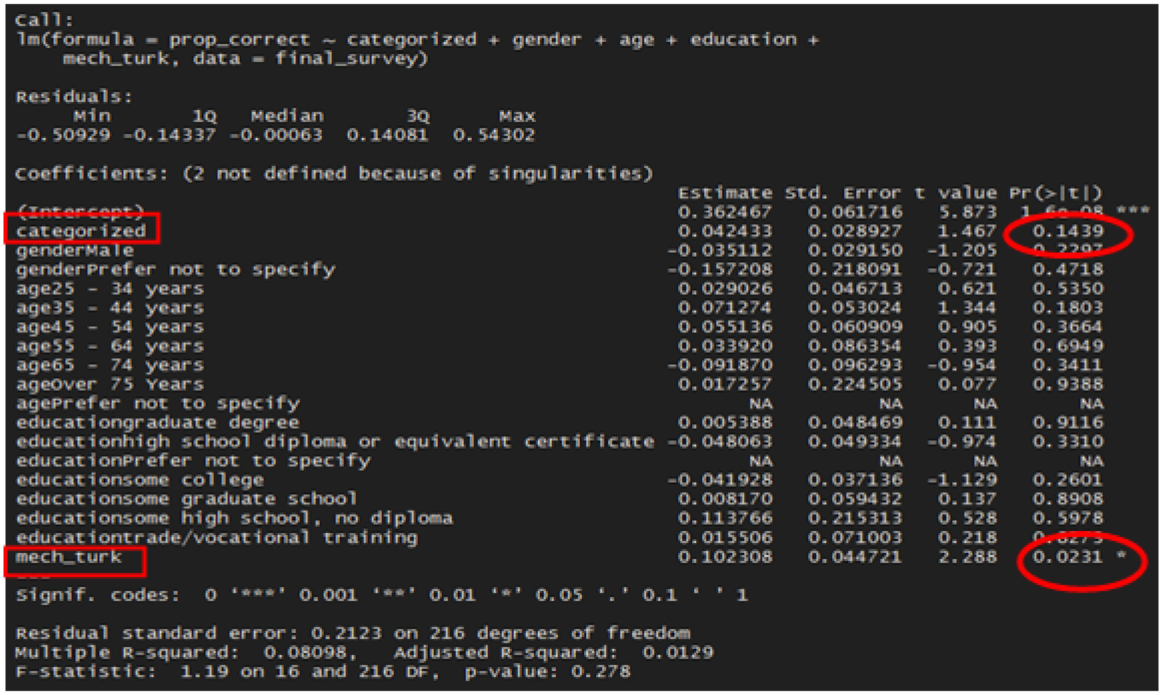
The calculation methodology for the standardized mean difference (SMD) is as follows:

$$smd = \frac{\bar{X}_{treatment} - \bar{X}_{control}}{\sqrt{\frac{s^2_{treatment} + s^2_{control}}{2}}}$$

An SMD with an absolute value greater than 0.2 indicates a serious covariate imbalance. The only group that clearly meets this criteria is the 65-74 age cohort.

Raw Data Analysis

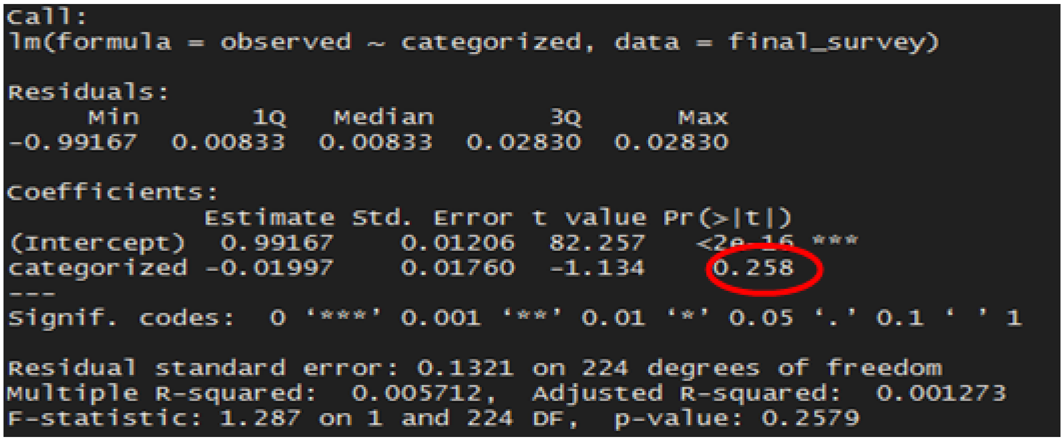
The dependent variable in our analysis is the proportion of the items in the list recalled correctly, i.e. the number of correct responses divided by 20. Our regression on the raw data without any data manipulation shows that the Average Treatment Effect (ATE) is 0.042 with a Std. Error of 0.0289 and a p-value of 0.14 . We do see a strong correlation between "mech_turk" and the outcome which is due to majority data (199) of the total (233) coming from Mechanical Turk. We ran more analysis on "mech_turk" and concluded that there is no treatment heterogeneity or significant interaction between 'mech_turk' and other covariates. We can see from the regression results that none of the other covariates have any significant effect on the outcome of the experiment.



```
Call:
lm(formula = prop_correct ~ categorized + gender + age + education +
    mech_turk, data = final_survey)

Residuals:
     Min       1Q   Median       3Q      Max
-0.50929 -0.14337 -0.00063  0.14081  0.54302

Coefficients: (2 not defined because of singularities)
                                                     Estimate Std. Error t value Pr(>|t|)
(Intercept)                                          0.362467   0.061716   5.873  1.6e-08 ***
categorized                                          0.042433   0.028927   1.467   0.1439
genderMale                                          -0.035112   0.029150  -1.205   0.2297
genderPrefer not to specify                         -0.157208   0.218091  -0.721   0.4718
age25 - 34 years                                     0.029026   0.046713   0.621   0.5350
age35 - 44 years                                     0.071274   0.053024   1.344   0.1803
age45 - 54 years                                     0.055136   0.060909   0.905   0.3664
age55 - 64 years                                     0.033920   0.086354   0.393   0.6949
age65 - 74 years                                    -0.091870   0.096293  -0.954   0.3411
ageOver 75 Years                                     0.017257   0.224505   0.077   0.9388
agePrefer not to specify                                   NA         NA      NA       NA
educationgraduate degree                             0.005388   0.048469   0.111   0.9116
educationhigh school diploma or equivalent certificate -0.048063   0.049334  -0.974   0.3310
educationPrefer not to specify                             NA         NA      NA       NA
educationsome college                               -0.041928   0.037136  -1.129   0.2601
educationsome graduate school                        0.008170   0.059432   0.137   0.8908
educationsome high school, no diploma                0.113766   0.215313   0.528   0.5978
educationtrade/vocational training                   0.015506   0.071003   0.218   0.8275
mech_turk                                            0.102308   0.044721   2.288   0.0231 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2123 on 216 degrees of freedom
Multiple R-squared:  0.08098,   Adjusted R-squared:  0.0129
F-statistic:  1.19 on 16 and 216 DF,  p-value: 0.278
```

Attrition

Based on our validation check, we determined that 6 subjects couldn't take the survey due to technical issues with their web browsers or corporate firewall issues. We analyzed whether the attrition was dependant on the treatment assignment or potential outcome. We ran a regression with a dummy “observed”" variable (1 if the outcome was observed and 0 if the outcome was missing) as the dependant variable and the group assignment as the independent variable .Results show that attrition was random and treatment assignment didn't have any significant effect on attrition (coefficient: -0.019997, Std. Error : 0,01760, p-value: 0.258).The randomness of attrition allows us to use the observed outcomes as good substitutes for the missing values:

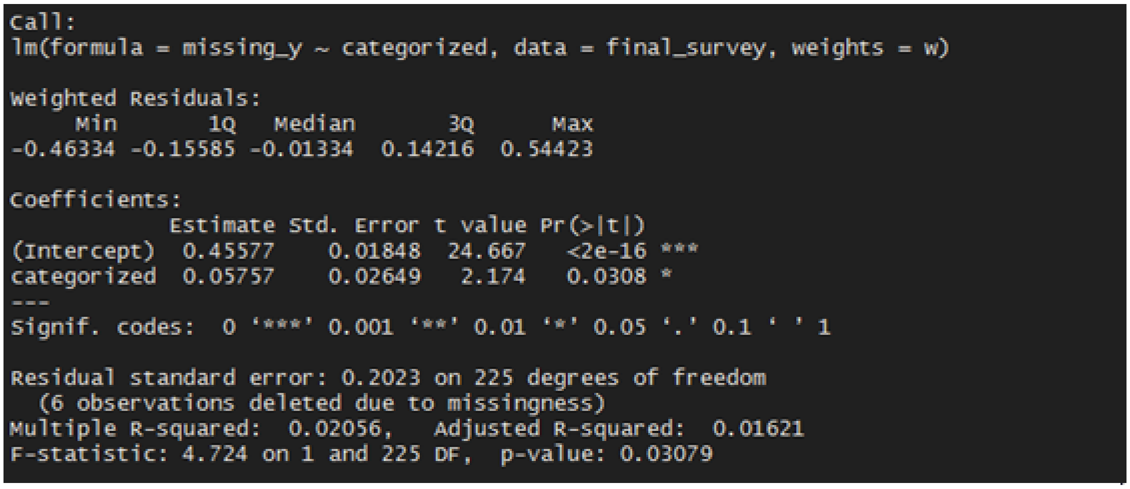


```
Call:
lm(formula = observed ~ categorized, data = final_survey)

Residuals:
     Min       1Q   Median       3Q      Max
-0.99167  0.00833  0.00833  0.02830  0.02830

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.99167    0.01206  82.257   <2e-16 ***
categorized -0.01997    0.01760  -1.134    0.258
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1321 on 224 degrees of freedom
Multiple R-squared:  0.005712,  Adjusted R-squared:  0.001273
F-statistic: 1.287 on 1 and 224 DF,  p-value: 0.2579
```

.

To account for this apparent attrition, we applied an inverse probability weighting scheme to recalculate our observed data. In effect, the weighting operation tries to recapture the missing values by replacing them with weighted values of the observed data. This method produces accurate estimates of average treatment effects when the non-missing observations are in fact good substitutes of the observations that are missing. Once attrition has been accounted for, we see that a simple linear regression on treatment gives us a significant

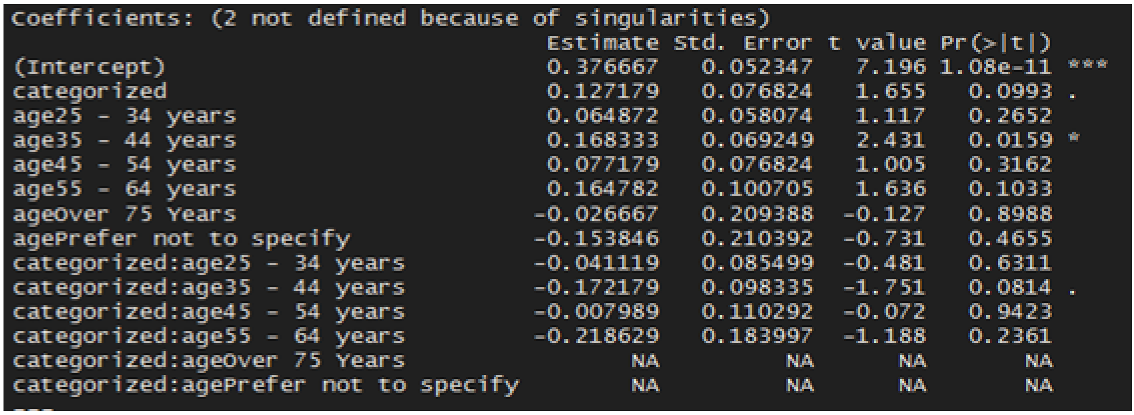
ATE which of 0.057 (Std. Error : 0,02649, p-value: 0.0308).



```
Call:
lm(formula = missing_y ~ categorized, data = final_survey, weights = w)

Weighted Residuals:
     Min       1Q   Median       3Q      Max 
-0.46334 -0.15585 -0.01334  0.14216  0.54423 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.45577    0.01848  24.667   <2e-16 ***
categorized  0.05757    0.02649   2.174   0.0308 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2023 on 225 degrees of freedom
  (6 observations deleted due to missingness)
Multiple R-squared:  0.02056,	Adjusted R-squared:  0.01621 
F-statistic: 4.724 on 1 and 225 DF,  p-value: 0.03079
```

Heterogeneous Treatment Effects

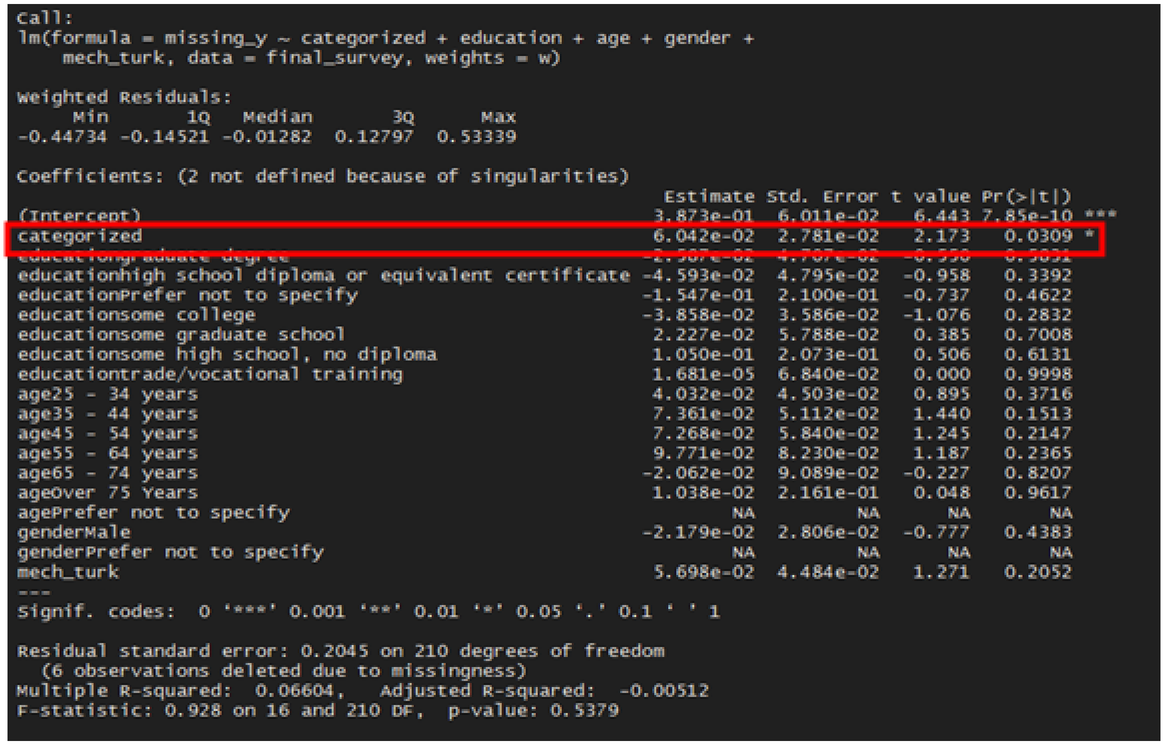
Having cleaned the data and identified a significant ATE, we then checked whether our treatment effect varied across the covariates. We ran a regression on all the covariates separately and in a combined multiple regression to see whether there were any significant interaction terms in our regression results. Of all the covariates tested, we found that none of the interaction terms were significant at the 5% level:



```
Coefficients: (2 not defined because of singularities)
                                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)                          0.376667   0.052347   7.196 1.08e-11 ***
categorized                          0.127179   0.076824   1.655   0.0993 .  
age25 - 34 years                     0.064872   0.058074   1.117   0.2652    
age35 - 44 years                     0.168333   0.069249   2.431   0.0159 *  
age45 - 54 years                     0.077179   0.076824   1.005   0.3162    
age55 - 64 years                     0.164782   0.100705   1.636   0.1033    
ageOver 75 Years                    -0.026667   0.209388  -0.127   0.8988    
agePrefer not to specify            -0.153846   0.210392  -0.731   0.4655    
categorized:age25 - 34 years        -0.041119   0.085499  -0.481   0.6311    
categorized:age35 - 44 years        -0.172179   0.098335  -1.751   0.0814 .  
categorized:age45 - 54 years        -0.007989   0.110292  -0.072   0.9423    
categorized:age55 - 64 years        -0.218629   0.183997  -1.188   0.2361    
categorized:ageOver 75 Years               NA         NA      NA       NA    
categorized:agePrefer not to specify       NA         NA      NA       NA    
```

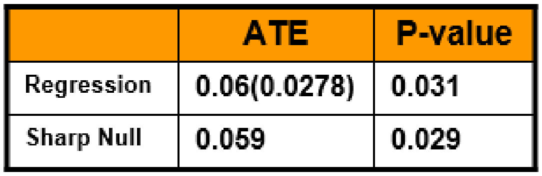
Final Analysis

After confirming the absence of heterogeneous treatment effects in our observed covariates, we moved on to our final regression analysis with all the co-variates to get our final ATE. The result was very positive .The final ATE was 0.064 with a Std. Error of 0.027 and a p-value of 0.031. Thus, our experiment indicates that individuals who receive the categorized list will remember 6.4% more items on average than people who receive the randomized list.



```
Call:
lm(formula = missing_y ~ categorized + education + age + gender +
    mech_turk, data = final_survey, weights = w)

Weighted Residuals:
     Min       1Q   Median       3Q      Max
-0.44734 -0.14521 -0.01282  0.12797  0.53339

Coefficients: (2 not defined because of singularities)
                                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                                     3.873e-01  6.011e-02   6.443 7.85e-10 ***
categorized                                     6.042e-02  2.781e-02   2.173   0.0309 *
educationgraduate degree                       -2.587e-02  4.707e-02  -0.550   0.5831
educationhigh school diploma or equivalent certificate -4.593e-02  4.795e-02  -0.958   0.3392
educationPrefer not to specify                 -1.547e-01  2.100e-01  -0.737   0.4622
educationsome college                          -3.858e-02  3.586e-02  -1.076   0.2832
educationsome graduate school                   2.227e-02  5.788e-02   0.385   0.7008
educationsome high school, no diploma           1.050e-01  2.073e-01   0.506   0.6131
educationtrade/vocational training              1.681e-05  6.840e-02   0.000   0.9998
age25 - 34 years                                4.032e-02  4.503e-02   0.895   0.3716
age35 - 44 years                                7.361e-02  5.112e-02   1.440   0.1513
age45 - 54 years                                7.268e-02  5.840e-02   1.245   0.2147
age55 - 64 years                                9.771e-02  8.230e-02   1.187   0.2365
age65 - 74 years                               -2.062e-02  9.089e-02  -0.227   0.8207
ageOver 75 Years                                1.038e-02  2.161e-01   0.048   0.9617
agePrefer not to specify                               NA         NA      NA       NA
genderMale                                     -2.179e-02  2.806e-02  -0.777   0.4383
genderPrefer not to specify                            NA         NA      NA       NA
mech_turk                                       5.698e-02  4.484e-02   1.271   0.2052
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2045 on 210 degrees of freedom
  (6 observations deleted due to missingness)
Multiple R-squared:  0.06604,	Adjusted R-squared:  -0.00512
F-statistic: 0.928 on 16 and 210 DF,  p-value: 0.5379
```

We also calculated the ATE using randomization inference. As the table below shows, the results from the difference in mean analysis is very close to the value in the regression and the p-value of 0.029 indicates that we can reject the null hypothesis that (H0 = H1).



	ATE	P-value
Regression	0.06(0.0278)	0.031
Sharp Null	0.059	0.029

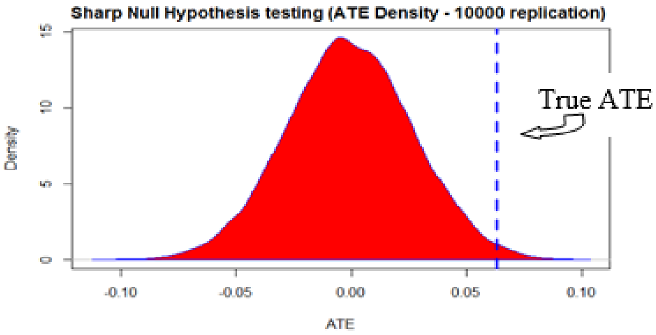


Figure 8 :ATE and distribution under sharp null hypothesis testing

Remarkable Observations

Arguably, several subjects demonstrated something similar to a treatment effect when providing an incorrect answer.

8 subjects omitted ‘toast’ and responded with ‘bread’ 8 subjects omitted ‘steak’ and responded with ‘beef’

All of the other items in the ‘meat’ category were broad, categorical items themselves except for ‘steak’ (i.e. turkey, chicken, lamb). ‘Beef’ seems to be more appropriate than ‘steak’. Similarly, all of the other ‘breakfast’ items were also categories (‘eggs’, ‘coffee’,‘cereal’).

However instead of accepting these as correct answers, we acknowledge that the subjects are pointing out that the items in the survey were not named consistently. If this experiment is repeated due attention should be payed to the naming of items.

Final Conclusion

We have demonstrated that individuals will recall 6.4% more data when it is categorized than those who receive the same randomized information. The results suggest that the organization of data plays a role in our recollection of it. We respond to meta-information, not just content information and it goes unnoticed by our conscious minds. We don't have to think harder to recall categorized data, it happens effortlessly.

Our experiment is a small first step in showing that there are specialized neuro-structures that assist in the recollection of categorized data. A path of continued research might fork into two directions. One could be localizing the actual physical structures that assist in categorical recollection using an MRI. Another path might be to explore more consequences of categorization. For example, the idea of category is very similar to the idea of stereo-type. Future research might explore whether the manifestation of such biases is as innate as the effects of categorization on active memory.

Citations

List, John A. et al. "So you want to run an experiment, now what? Some Simple Rules of Thumb for Optimal Experimental Design". *Collegio Carlo Alberto.* 2009.