Methods of Transcriptions and Their Effects on Memory

Yigit Demiralp, Jeffrey Leung, John McCoy, Chiebuka Onwuzurike,
Ruchika Venkateswaran, Phoenix Wang

Introduction

Background

Many experiments have tested the effects of typing versus handwriting on memory and comprehension. Smoker, Murphy, and Rockwell (2009) showed that participants had better recognition and recall of words when asked to write them versus typing them. Many of these studies postulate that the improved performance in memory is due to increased semantic processing from handwriting. Mueller and Oppenheimer (2014) demonstrated that students who took notes by hand performed better on conceptual questions than students who took notes on a laptop. The authors noted that laptop note takers were more likely to copy notes word-for-word, indicating that they were not processing and truly understanding the information as well as longhand note takers.

Research Question

Our objective of this experiment is to study the effect of memory on survey participants who type in information, viz-a-viz those who copy and paste information. In this experiment, we study the effect of copying and pasting (treatment group) versus typing (control group) on memory.

Given the background research, we expect that copy-pasting would result in even less semantic processing than typing, and that our participants asked to copy-paste information would have worse memory than the typing participants.

Hypothesis

Our main hypothesis is that users who copy and paste information are less likely than users who type information to remember what they have learned. In our experiment, we set up a survey to test whether an individual when presented with a block of text is able to retain information. There are two different conditions that the survey takers are exposed to:

- The first is the "Copy & Paste" (C&P) condition, where the survey participants can copy and paste the answer to the questions based on the block of text that they need to read.

- The second is the "Typing" (T) condition, where the text is presented as a picture so the only viable option that the participants have is to type in the information that is asked of them. The "Experimental Design" section provides more details about the survey design and execution.

We expect that if a participant copy-pastes a long block of text, they are less likely to remember the information than if they typed it. On the other hand, if a participant is forced to type, we expect that they would spend more time processing the information which would lead to better memory.

Experimental Design

Qualtrics Workflow

When participants started the survey, they were directed to a default question block that collected basic information: age, gender, country of origin, native English speaker, student or not. Afterwards, Qualtrics randomized who would be in the C&P (treatment) group and who would be in the T (control) group. Both groups were shown two excerpts: a paragraph regarding emoji usage and a conversation including emojis. Having read each excerpt, the survey taker answered four questions each: one was for transcribing a piece of the text they just read, and the other three were distraction questions, which were also on emojis. The difference between the C&P and T group was that the C&P group's excerpts were highlightable for copy and pasting, whereas the T group's excerpts were screenshots.

Having completed the two reading and transcribing activities, all survey takers were shown the same multiple choice comprehension where they were asked questions specific to the information they had transcribed earlier. Two questions were based on numbers from the excerpts while the other two were based on text. Besides one correct option and four wrong options, we also included a "Not sure" option to prevent participants from answering correctly by taking a wild guess. At the end of the entire survey, two questions were asked regarding difficulty and whether or not they copy/pasted or typed. The last question specifically helped in measuring how many complied to C&P, since we expect that not all participants who had the option to C&P would actually do so. (The excerpts and survey questions can be found in Appendix A.)

Initial Findings from Pilot

The purpose of a pilot survey is to test the correctness of the instructions, examine if the results are varied, and identify potential issues. The pilot survey was sent to 10 of our batchmates and we received 10 valid responses within 48 hours. 6 of the 10 pilot survey takers were assigned to the T (control) group and 4 of them were assigned to the C&P (treatment) group. In this pilot stage, the T group answered the four questions with an average accuracy of 91.7%, while the C&P group had an accuracy rate of 68.8%. As we hypothesized, the C&P group appeared to have a lower accuracy rate than the T group. Since we observed varied results from the T and C&P groups and our pilot survey takers indicated that the instructions were clear, we proceeded to distribute the same survey on a larger scale without making any changes.

## Experimental Results

Survey Participants

For experiment participants, we gathered survey responses from 3 main sources, our MSBA classmates, Reddit forums, and Facebook groups. Our criteria for choosing these three channels were based on size, convenience, experiment stage, and diversity. We aimed for a control (T) vs treatment (C&P) split of 50%; however, we achieved a split of 57.9% control to 42.1% treatment. In total, 140 survey respondents fully completed the survey. 81 were in the control group and 59 were in the treatment group. The survey demographic breakdown is as follows:

- English as first language - Yes (56.6%), No (44.4%)
- Gender - Men (35.2%), Women (59.2%), Non-Binary (3.5%), Perfer not to say (1.4%), Other (0.7%)
- Student - Yes (65.5%), No (34.5%)

Randomization Check

Since we did a simple randomization, we needed to conduct randomization checks. We performed *t*-tests on four controlled variables: age, gender, English as first language, and if the participant was a student or not. The *t*-tests compared these variables between both the treatment and control groups and found no statistical significance, indicating that our groups were well-balanced. (Detailed results can be found in Appendix C.)

Results

Performance by Question
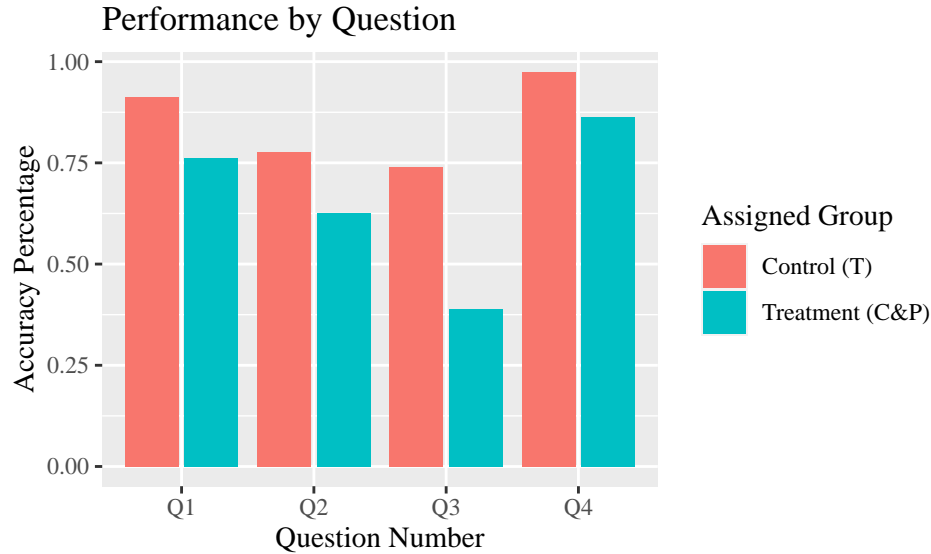
Average Treatment Effect

```
##                      All_Questions    Text_Questions  Numeric_Questions
## Dependent Var.:      Total_Accuracy     Text_Accuracy       Num_Accuracy
##
## (Intercept)       0.8519*** (0.0236)  0.8272*** (0.0331) 0.8765*** (0.0241)
## Treatment        -0.1908*** (0.0464) -0.2509*** (0.0583)  -0.1308* (0.0503)
## _____  _____  _____ _____
```

To calculate the average treatment effect (ATE), we ran regressions of accuracy scores (range from 0% to 100%) on the treatment. We saw that on average, the treatment (C&P) group scored 19.08% lower than the participants placed in the control (T) group. When we examined the ATE by question type, we saw that our treatment (C&P) group scored 25.09% lower than the control (T) group on textual questions whereas on numerical questions, the treatment (C&P) group only scored 13.08% lower than the control (T) group. It is worth noting that the ATE for numerical questions was not statistically significant.

Complier Average Causal Effect

```
##                   CACE_AllQuestions  CACE_TextQuestions  CACE_NumQuestions
## Dependent Var.:      Total_Accuracy       Text_Accuracy       Num_Accuracy
##
## (Intercept)       0.8519*** (0.0236)  0.8272*** (0.0331) 0.8765*** (0.0241)
## Compliance       -0.2047*** (0.0484) -0.2691*** (0.0608) -0.1403** (0.0533)
## _____  _____  _____ _____
```

The compliance rate in our experiment is 93.2%. For the Complier Average Causal Effects (CACE), we concluded that participants who copy-pasted the information scored 20.5% less than the participants who were forced to type. The accuracy of the treated (C&P) participants for textual questions were 26.91% lower than the people who were in the control (T) group, while for the numerical questions the treatment group scored 14.03% lower. The CACE for numerical questions was not significant at a significance level of 0.01.

Heterogeneous Treatment Effect

We also checked for heterogeneous treatment effects, but found no statistical significance. The treatment effects did not differ by whether the participant is a Native speaker, whether the participant is a student, or

how old the participant is. (Detailed results can be found in Appendix C.)

Limitations and Potential Errors

1. **Randomization** - We elected to perform simple randomization since we did not know ahead of time who would take and finish our survey. One issue we faced was that Qualtrics was vague about how it randomly assigned participants. There was an option to "evenly present elements", which we chose not to use since it was unclear if that would violate rules for experimental randomization (assigning every other participant to treatment). As a result we ended up with an uneven split of participants in our treatment (59) and control groups (81). However, since we had such a large treatment effect, our experiment still had very strong statistical power (98.24%). If we had an even split (70/70) and the same treatment effect, statistical power would only slightly increase to 98.45%.

2. **Possible Spillovers** - Since we randomized at the individual level, there is a chance that participants in the treatment and control groups interacted with one another before the survey was taken. This was somewhat mitigated by the distraction emoji questions that obscured the true motivation of the survey. We could have further mitigated this risk by performing block randomization, but this would have come at the expense of potentially having dissimilar participants in the treatment and control groups.

3. **Missing Covariates** - We collected various demographic information from participants to use as covariates in our analysis, but one variable we neglected to collect was whether the survey was completed on a phone/laptop/tablet. The result might have been different depending on the device used, so this was a missed opportunity to more precisely measure the factors that affected memory.

Conclusion

As paper becomes a thing of the past for note-taking and transcribing the switch to digital mediums becomes more and more prominent. After setting up a randomized control trial using surveys to test whether a participant was less likely to remember information that they had copy pasted rather than typed in, we found that, among the compliers who copy-pasted had an accuracy score of 20.5% less. We used demographic variables such as age, gender, and whether the participants native language was English or not in order to ensure randomization. We found that participants who have typed in information are more likely to remember said info compared with those who copy pasted it. In terms of the question types, the treatment effect for numerical questions was not as significant as it was for textual questions. How well people memorize texts versus numbers can be further investigated in the future.

References

Smoker TJ, Murphy CE, Rockwell AK. Comparing Memory for Handwriting versus Typing. Proceedings of the Human Factors and Ergonomics Society Annual Meeting. 2009;53(22):1744-1747. doi:10.1177/154193 120905302218

Mueller PA, Oppenheimer DM. The Pen Is Mightier Than the Keyboard: Advantages of Longhand Over Laptop Note Taking. Psychological Science. 2014;25(6):1159-1168. doi:10.1177/0956797614524581

Appendix A - Survey

Excerpt 1

Does Emoji Use Differ Based on User Location and Culture?

A group of researchers analyzed 427 million messages worldwide and ranked countries in order of the percentage of messages including emojis. The top 5 countries that love using emojis the most are France (20%), Russia (11%), U.S. (9%), Mexico (8%), and Turkey (6%). They also found that countries with high individualism use more happy emojis, and countries where ties between individuals are tight use more sad/angry emojis.

Participants were asked to copy-paste or type the top 5 countries and their percentages.

Question 1 (Textual): Which country uses emojis the most?

Options: Japan, Turkey, France, Austria, United States, Not sure

Question 2 (Numerical): What was the highest percentage of emoji use exhibited in the study?

Options: 5%, 20%, 8%, 14%, 25%, Not sure

Excerpt 2

Please read this conversation between Leon Stefan, Bo Seong-Min, and Justino Tadeo that occurred on 4/14/21:

Leon Stefan: Hello guys, how's everybody doing? 😊
Bo Seong-Min: I'm ok. How about you?
Justino Tadeo: Yea I'm fine, but what's the plan today? 🤔
Leon Stefan: I don't know the plan but I am hungry right now
Bo Seong-Min: What's wrong with you, you're literally always hungry
Justino Tadeo: You are 😂
Leon Stefan: 😅 Whatever lol. Where do you guys want to eat?
Justino Tadeo: I don't care
Bo Seong-Min: Just not seafood 🤢 you know what it does to my stomach
Leon Stefan: How about Chipotle??? 👀
Bo Seong-Min: Yes! Yes! Yes!
Justino Tadeo: I could go for some Chipotle 🤤
Bo Seong-Min: Sounds like a plan then!
Leon Stefan: 👍

Participants were asked to copy-paste or type the date and the names of the people in the conversation.

Question 3 (Textual): Who was in the lunch conversation?

Options: Justin Todeo, Bo Seong-Min, Leon Tadeo, Leo Stefano, Bo Eun-Jung, Not sure

Question 4 (Numerical): When did the food conversation happen?

Options: 4/14/21, 2/15/21, 3/14/20, 7/11/20, Not sure

Appendix B - Data Dictionary

Variable	Type	Description
ID	integer	Unique identifier of participant
Latitude	numeric	Lattitude that the participant took the survey
Longitude	numeric	Longtitude that the participant took the survey
Age	numeric	Age of participant
Gender	character	Gender of participant
Nationality	character	Nationality of participant
English_First_Language	character	"Yes" if English is the first language of participant, "No" if not
Student	numeric	"1" if participant is a student, "0" if not
Treatment	integer	"1" if participant was in the treatment group, "0" if not
Q1_Score	integer	"1" if participant answers correctly, "0" if not
Q2_Score	integer	"1" if participant answers correctly, "0" if not
Q3_Score	integer	"1" if participant answers correctly, "0" if not
Q4_Score	integer	"1" if participant answers correctly, "0" if not
Total_Score	integer	Total score of participant (ranges from 0 to 4)
Group	character	"CP" if the participant copy-pasted, "T" if typed
Native	numeric	"1" if the participant is a native English speaker, "0" if not
Total_Accuracy	numeric	Total accuracy score of participant (ranges from 0 to 1)
Text_Accuracy	numeric	Accuracy score for textual questions (ranges from 0 to 1)
Num_Accuracy	numeric	Accuracy score for numerical questions (ranges from 0 to 1)

Appendix C - Code

```
df_clean <- fread("results_clean.csv")
```

```
# Randomization Check - Age
t.test(df_clean[Treatment == 0, Age],
       df_clean[Treatment == 1, Age])
```

```
##
##  Welch Two Sample t-test
##
## data:  df_clean[Treatment == 0, Age] and df_clean[Treatment == 1, Age]
## t = -0.90048, df = 98.417, p-value = 0.3701
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.846255  1.820816
## sample estimates:
## mean of x mean of y
##   25.57500  27.08772
```

```
# Randomization Check - Gender (Woman)
dummy <- dummyVars(" ~ Gender", data = df_clean)
dummy_df <- data.frame(predict(dummy, newdata = df_clean))
gender_df <- as.data.table(cbind(dummy_df, Treatment = df_clean[, Treatment]))

t.test(gender_df[Treatment == 0, GenderWoman],
       gender_df[Treatment == 1, GenderWoman])
```

```
##
##  Welch Two Sample t-test
##
## data:  gender_df[Treatment == 0, GenderWoman] and gender_df[Treatment == 1, GenderWoman]
## t = -1.2651, df = 128.32, p-value = 0.2081
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.27040981  0.05948703
## sample estimates:
## mean of x mean of y
## 0.5555556 0.6610169
```

```
# Randomization Check - Gender (Man)
t.test(gender_df[Treatment == 0, GenderMan],
       gender_df[Treatment == 1, GenderMan])
```

```
##
##  Welch Two Sample t-test
##
## data:  gender_df[Treatment == 0, GenderMan] and gender_df[Treatment == 1, GenderMan]
## t = 1.6994, df = 131.47, p-value = 0.09161
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.02234777  0.29478971
## sample estimates:
```

```
## mean of x mean of y
## 0.4074074 0.2711864
```

```
# Randomization Check - Gender (Non binary)
t.test(gender_df[Treatment == 0, GenderNon.binary],
       gender_df[Treatment == 1, GenderNon.binary])
```

```
##
##  Welch Two Sample t-test
##
## data:  gender_df[Treatment == 0, GenderNon.binary] and gender_df[Treatment == 1, GenderNon.binary]
## t = -0.77702, df = 98.234, p-value = 0.439
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.09295509  0.04064289
## sample estimates:
##  mean of x  mean of y
## 0.02469136 0.05084746
```

```
# Randomization Check - Gender (Agender)
t.test(gender_df[Treatment == 0, GenderAgender],
       gender_df[Treatment == 1, GenderAgender])
```

```
##
##  Welch Two Sample t-test
##
## data:  gender_df[Treatment == 0, GenderAgender] and gender_df[Treatment == 1, GenderAgender]
## t = -1, df = 58, p-value = 0.3215
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.05087657  0.01697826
## sample estimates:
##  mean of x  mean of y
## 0.00000000 0.01694915
```

```
# Randomization Check - Gender - (Prefer not to say)
t.test(gender_df[Treatment == 0, GenderPrefer.not.to.say],
       gender_df[Treatment == 1, GenderPrefer.not.to.say])
```

```
##
##  Welch Two Sample t-test
##
## data:  gender_df[Treatment == 0, GenderPrefer.not.to.say] and gender_df[Treatment == 1, GenderPrefer
## t = 1, df = 80, p-value = 0.3203
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.01222301  0.03691436
## sample estimates:
##  mean of x  mean of y
## 0.01234568 0.00000000
```

```
# Randomization Check - English as first language
df_clean[, Native := ifelse(df_clean[, English_First_Language] == "Yes", 1, 0)]
t.test(df_clean[Treatment == 0, Native],
       df_clean[Treatment == 1, Native])
```

```
##
##  Welch Two Sample t-test
##
## data:  df_clean[Treatment == 0, Native] and df_clean[Treatment == 1, Native]
## t = -0.73115, df = 126.09, p-value = 0.466
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.2311324  0.1064201
## sample estimates:
## mean of x mean of y
## 0.5308642 0.5932203
```

```
# Randomization Check - Student status
df_clean[, Student := ifelse(df_clean[, Student] == "Yes", 1, 0)]
t.test(df_clean[Treatment == 0, Student],
       df_clean[Treatment == 1, Student])
```

```
##
##  Welch Two Sample t-test
##
## data:  df_clean[Treatment == 0, Student] and df_clean[Treatment == 1, Student]
## t = -0.29124, df = 126.22, p-value = 0.7713
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1843122  0.1370220
## sample estimates:
## mean of x mean of y
## 0.6543210 0.6779661
```

```
# Average Treatment Effect
df_clean[, Total_Accuracy := Total_Score/4]
df_clean[, Text_Accuracy := (Q1_Score + Q3_Score)/2]
df_clean[, Num_Accuracy := (Q2_Score + Q4_Score)/2]

All_Questions = feols(Total_Accuracy ~ Treatment, data = df_clean, se = 'white')
Text_Questions = feols(Text_Accuracy ~ Treatment, data = df_clean, se = 'white')
Numeric_Questions = feols(Num_Accuracy ~ Treatment, data = df_clean, se = 'white')
etable(All_Questions, Text_Questions, Numeric_Questions)
```

```
##                         All_Questions       Text_Questions   Numeric_Questions
## Dependent Var.:         Total_Accuracy      Text_Accuracy         Num_Accuracy
##
## (Intercept)        0.8519*** (0.0236)  0.8272*** (0.0331) 0.8765*** (0.0241)
## Treatment         -0.1908*** (0.0464) -0.2509*** (0.0583)  -0.1308* (0.0503)
## _____   _____ _____  _____
## S.E. type         Heteroskedast.-rob. Heteroskedast.-rob. Heteroskedas.-rob.
## Observations                     140                 140                 140
```

```
## R2                               0.12057              0.12539              0.05291
## Adj. R2                          0.11420              0.11905              0.04604
```

```r
# Power Analysis
ate = All_Questions$coefficients[2]
cohenD = ate/df_clean[, sd(Total_Accuracy)]

pwr.t2n.test(n1 = df_clean[Treatment==1, .N], n2 = df_clean[Treatment==0, .N],
d = cohenD, sig.level = .05, power = NULL)
```

```
##
##      t test power calculation
##
##              n1 = 59
##              n2 = 81
##               d = 0.7006818
##       sig.level = 0.05
##           power = 0.9823619
##     alternative = two.sided
```

```r
# Heterogeneity - English as first language
All_Questions = feols(Total_Accuracy ~ Treatment*Native, data = df_clean, se = 'white')
Text_Questions = feols(Text_Accuracy ~ Treatment*Native, data = df_clean, se = 'white')
Numeric_Questions = feols(Num_Accuracy ~ Treatment*Native, data = df_clean, se = 'white')
etable(All_Questions, Text_Questions, Numeric_Questions)
```

```
##                         All_Questions     Text_Questions  Numeric_Questions
## Dependent Var.:         Total_Accuracy      Text_Accuracy       Num_Accuracy
##
## (Intercept)      0.8750*** (0.0350)  0.8421*** (0.0503) 0.9079*** (0.0319)
## Treatment        -0.2917*** (0.0750) -0.3838*** (0.0933)  -0.1996* (0.0794)
## Native             -0.0436 (0.0474)    -0.0282 (0.0670)   -0.0591 (0.0478)
## Treatment x Native  0.1746. (0.0947)    0.2270. (0.1180)    0.1222 (0.1032)
## ------------------ -------------------- -------------------- ------------------
## S.E. type        Heteroskedast.-rob. Heteroskedast.-rob. Heteroskedas.-rob.
## Observations                    140                 140                140
## R2                          0.14797             0.15917            0.06442
## Adj. R2                     0.12917             0.14062            0.04378
```

```r
# Heterogeneity - Age
All_Questions = feols(Total_Accuracy ~ Treatment*Age, data = df_clean, se = 'white')
Text_Questions = feols(Text_Accuracy ~ Treatment*Age, data = df_clean, se = 'white')
Numeric_Questions = feols(Num_Accuracy ~ Treatment*Age, data = df_clean, se = 'white')
etable(All_Questions, Text_Questions, Numeric_Questions)
```

```
##                         All_Questions     Text_Questions  Numeric_Questions
## Dependent Var.:         Total_Accuracy      Text_Accuracy       Num_Accuracy
##
## (Intercept)      0.8669*** (0.0826) 0.8575*** (0.1194) 0.8763*** (0.0765)
## Treatment        -0.3228** (0.1227)  -0.4193* (0.1634)  -0.2264. (0.1259)
## Age                -0.0007 (0.0029)    -0.0013 (0.0043)   -4.92e-5 (0.0028)
## Treatment x Age     0.0052 (0.0039)     0.0068 (0.0055)    0.0036 (0.0040)
```

```
## _____  _____  _____  _____
## S.E. type           Heteroskedas.-rob. Heteroskedas.-rob. Heteroskedas.-rob.
## Observations                       137                137                137
## R2                             0.12548            0.12608            0.05867
## Adj. R2                        0.10576            0.10637            0.03744
```

```
# Heterogeneity - Student
All_Questions = feols(Total_Accuracy ~ Treatment*Student, data = df_clean, se = 'white')
Text_Questions = feols(Text_Accuracy ~ Treatment*Student, data = df_clean, se = 'white')
Numeric_Questions = feols(Num_Accuracy ~ Treatment*Student, data = df_clean, se = 'white')
etable(All_Questions, Text_Questions, Numeric_Questions)
```

```
##                             All_Questions     Text_Questions   Numeric_Questions
## Dependent Var.:             Total_Accuracy      Text_Accuracy         Num_Accuracy
##
## (Intercept)           0.8571*** (0.0372) 0.8393*** (0.0516) 0.8750*** (0.0415)
## Treatment               -0.1203. (0.0715)  -0.1814. (0.0993)    -0.0592 (0.0794)
## Student                 -0.0081 (0.0481)   -0.0185 (0.0672)     0.0024 (0.0512)
## Treatment x Student     -0.1038 (0.0929)   -0.1019 (0.1227)    -0.1056 (0.1019)
## _____  _____  _____  _____
## S.E. type           Heteroskedas.-rob. Heteroskedas.-rob. Heteroskedas.-rob.
## Observations                       140                140                140
## R2                             0.13631            0.13666            0.06537
## Adj. R2                        0.11726            0.11761            0.04475
```

```
# CACE
df_clean[, Compliance := ifelse((df_clean[, Treatment]==1 & Group=='CP'), 1, 0)]
CACE_AllQuestions = feols(Total_Accuracy ~ 1 | 0 | Compliance ~ Treatment, data=df_clean, se='white')
CACE_TextQuestions = feols(Text_Accuracy ~ 1 | 0 | Compliance ~ Treatment, data=df_clean, se='white')
CACE_NumQuestions = feols(Num_Accuracy ~ 1 | 0 | Compliance ~ Treatment, data=df_clean, se='white')
etable(CACE_AllQuestions, CACE_TextQuestions, CACE_NumQuestions)
```

```
##                   CACE_AllQuestions   CACE_TextQuestions   CACE_NumQuestions
## Dependent Var.:       Total_Accuracy        Text_Accuracy         Num_Accuracy
##
## (Intercept)       0.8519*** (0.0236)  0.8272*** (0.0331) 0.8765*** (0.0241)
## Compliance        -0.2047*** (0.0484) -0.2691*** (0.0608) -0.1403** (0.0533)
## _____  _____  _____  _____
## S.E. type         Heteroskedast.-rob. Heteroskedast.-rob. Heteroskedas.-rob.
## Observations                     140                140                140
## R2                           0.15926            0.16287            0.07211
## Adj. R2                      0.15316            0.15681            0.06539
```