36-309 / 36-749 Homework 2: Your Experiences are Valid, but are Your Experiments?

Code

Due Wednesday, September 21, 11:59pm on Gradescope

Sanaz Saadatifar

Question 1: Considering Experimental Design Principles in a Real Study (50 points)

For this part, you must read the paper, “The Tennessee Study of Class Size in the Early School Grades” by Frederick Mosteller (1995) published in the academic journal *The Future of Children*. (The paper is available on Canvas.) This paper discusses the results from a famous experiment, the Tennessee STAR experiment (where STAR stands for Student-Teacher Achievement Ratio). In this experiment, children from kindergarten to 3rd grade across Tennessee were randomly assigned to one of three types of classrooms:

1. Small classes of 13-17 students
2. Regular-sized classes of 22-25 students
3. Regular-sized classes of 22-25 students with an additional teacher’s aide in the class.

The goal of the experiment was to assess if reducing class sizes or adding more class staff affected students’ achievement. The author of this paper (Frederick Mosteller) was one of the most influential statisticians of the 20th century – he attended what was the Carnegie Institute of Technology and was the founding chairman of Harvard’s Statistics Department. He also cared deeply about education. So, when reading this paper, you will get to see how an expert statistician interpreted the design and analysis of an important experiment. Despite Mosteller’s technical expertise, this paper was written for education researchers without a statistical background (in the paper, Mosteller even gives tutorials on what we mean by “standard deviation” and “effect size,” which you might find helpful).

What follows are questions about the design and analysis of the Tennessee STAR experiment. You should be able to answer some of the questions just by reading the paper; however, some questions will take a bit more thinking about the course material. **YOUR ANSWER TO EACH QUESTION CANNOT BE MORE THAN 3 SENTENCES LONG. MANY QUESTIONS DO NOT REQUIRE MORE THAN 1-2 SENTENCES.** This is meant to assure you that only short answers - not essays - are required for each question.

a. (5pts) In the “Study Design and Execution” section (starting on Page 115), it is discussed that students and teachers were randomly assigned to one of the three types of classes. What

complications would have arose if only students or only teachers had been randomized to the three types of classes? Discuss how your answer relates to the concept of **internal validity**.

[In that case, it would not be possible to ensure that classes came from equivalent populations and that teachers did not choose their classes. In a study of this kind, randomization protects against all variables that might matter, whether they have been identified or not. Hence, randomization can bring internal validity but it does not mean that randomization will always bring internal validity.]

b. (10pts) Discuss how the paragraph on Page 115 (starting with "To be eligible to participate in the experiment…") relates to one or more (but not necessarily all) of the following concepts: **internal validity**, **external validity**, and **construct validity**. (In other words: For each concept you think is relevant for this paragraph, mention it and explain why. However, if you don't think a concept is relevant for this paragraph, do not mention it.)

[This paragraph is related to "interval Validity", because they are setting constraints in terms of selecting their participants. These constraints are to assure the internal validity or in other words the ability to make comparisons among the three kinds of classes within a single school. Therefore, there is nothing to do with the generalizability, external validity, or construct validity.]

c. (5pts) On Page 116, it is stated that, "During the first year, the study involved about 6,400 pupils in 108 small classes, 101 regular-sized classes, and 99 regular-sized classes with teachers' aides." The goal of the experiment was to compare these three groups, so it may be sensible to use ANOVA to compare the mean outcomes of these three groups. Remember that ANOVA has three assumptions: Normality, equal variance, and independent measurements. For this question, let's give special attention to the independent measurements assumption. Consider two scenarios:

- *Scenario 1*: We treat "classroom" as the subject of the experiment, such that there are treatment group sizes of 108, 101, and 99 subjects.
- *Scenario 2*: We treat "students" as the subject of the experiment – meaning that there are about 6,400 total subjects, with some number of subjects in each of the three treatment groups.

Do you think the independent measurements assumption is (1) More Plausible in Scenario 1 than Scenario 2, (2) More Plausible in Scenario 2 than Scenario 1, or (3) Equally Plausible in Scenario 1 and Scenario 2? State one of the three answers, and then give a 1-3 sentence explanation.

[Independent measurement is more plausible in Scenario 2 than Scenario 1. Because an ANOVA assumes that the observations in each group are independent of each other. Therefore, here the students are subjects that ate grouped into three classes, where students are (subjects) are independent.]

d. (10pts) Discuss how the following paragraph (Page 117) relates to one or more (but not necessarily all) of the following concepts: **internal validity**, **external validity**, and **construct validity**.

"In assessing student performance, two types of tests were used: (1) standardized tests, which have the advantage of being used nationally but the disadvantage of not being directly related to any particular curriculum or course of study; and (2) curriculum-based tests, which reverse the advantages and disadvantages of standardized tests. Curriculum-based tests measure more directly the student's increased knowledge of what was actually taught, but they give little indication of where local results stand in the national picture."

(In other words: For each concept you think is relevant for this paragraph, mention it and explain why. However, if you don't think a concept is relevant for this paragraph, do not mention it.)

[Using the "standardized test" is related to "external validity" as it is used as an explanatory variable that is used nation-wide. Hence to generalize the local results in the national picture it is used. However, the curriculum-based test, as it is directly related to the students of specific schools it can be more related to having a same explanatory variable across data sets that we have, not nationwide. Therefore, it can be related to "internal validity".]

e. (5pts) Please look at Table 2 of the paper (Page 121). This table suggests that there is a positive effect size on student performance from (1) smaller class sizes and (2) including a teacher's aide in a regular-sized class. As seen in Project Challenge, the state of Tennessee took this as enough evidence to install smaller class sizes in many school districts. However, is this also enough evidence to suggest that it would be helpful to install smaller classes and teacher's aides in those smaller classes? If so, why? If not, why not, and how would you have changed the STAR experiment to address that issue?

[Effect size is not enough to generalize the results. Because we do not know what is the probability of seeing such an effect size. It would be better to add a step in experiment to run probability simulations and hypothesis testing to see what is the p-Value and based on that to decide wether to add small classes with teacher aids to other schools.]

f. (5pts) As discussed in Box 1 and Page 122 of the paper, after the STAR experiment, Tennessee implemented Project Challenge, where small classes were installed in kindergarten and 1st, 2nd, and 3rd grades for the 17 lowest-income school districts. In this case, these 17 school districts can be considered the "teatment group." As discussed on Page 122, studies claimed that installing small classes for these districts were effective by comparing to some kind of control group. What is the implicit control group for such a comparison? Why does Mosteller say that these comparisons "must be regarded as weaker [than the STAR experiment]"? (**Hint**: Just saying "because this new investigation is less well controlled" isn't an adequate answer.)

[The control group consists of the schools that no treatment applied or in other words the schools that no small classes were installed in kindergarten and 1st, 2nd, and 3rd grades. Probably around 17 of them have been selected to be considered as control group. The reason that Mosteller say that these comparisons "must be regarded as weaker is that this study is not an experiment, it is more of a observational study where the internal validity conditions are not checked to see whether is true or not.]

g. (5pts) It is stated on Page 115 that, "No new textbooks or curricula were to be introduced [during the Tennessee STAR experiment]." Why did the researchers of this experiment make sure that no new textbooks or curricula were introduced during the experiment?

[They wanted to have a controlled experiment, and not allow for any addition of any new factor or variable that might affect the experiment results. Adding new books could affect the results, hence they wanted to prevent it and highlight that they had a controlled experiment. it can also help in terms of preserving the validity of the experiment internally and externally.]

h. (5pts) Overall, the Tennessee STAR experiment suggests that smaller class sizes improve students' achievement. Does this suggest that "smaller is always better"? For example, does the experiment suggest that, ideally, classes would consist of one teacher and one student? Why or why not?

[No. the experiment results can be interpreted in the range of the explanatory variables. The explanatory variables were categorial not nominal, and we do not have any data regarding the one student and one teacher. Hence we cannot make any inferences about the one student and one teacher case with the current data set.]

Question 2: Assessing the Validity of Two Experiments (50 points)

In this question, we will consider two simulated experiments. Let's say that a school board is considering redeveloping their literacy program for first-grade students. In particular, they are considering three different programs: The "standard" program that has been implemented for years, a new "writing" program that tries to improve students' literacy with novel writing exercises, and a new "reading" program that tries to improve students' literacy with novel reading exercises. The school board would like to know if the different programs lead to significantly different outcomes in terms of students' literacy ability.

To help them answer this question, the school board enlists Professor E. Greenwich (https://en.wikipedia.org/wiki/Ellie_Greenwich) and Professor K. Deal (https://en.wikipedia.org/wiki/Kim_Deal) to each conduct a randomized experiment in a given school district. For each experiment, students take a literacy pre-test to measure their literacy at the beginning of the experiment. Then, students are randomly assigned to one of three literacy programs ("standard", "writing", or "reading"). After students go through their respective programs, they take another test (a "post-test") to measure their literacy at the end of the experiment.

Professor Deal conducted her experiment in one school district, and Professor Greenwich conducted her experiment in another school district. You are given the data that resulted from each of their experiments:

Hide

```
#Loading the Deal experiment
deal = read.csv("https://raw.githubusercontent.com/zjbranson/stat309fall2022/master/deal.csv")
#Loading the Greenwich experiment
greenwich = read.csv("https://raw.githubusercontent.com/zjbranson/stat309fall2022/master/greenwich.csv")
```

Both datasets contain three variables: `pretest`, `posttest`, and `program`. Professor Deal and Professor Greenwich randomized `program` within their respective school districts, and thus both of these datasets come from a randomized experiment. We'll assess how well each of these experiments were conducted, and how to interpret them. So, *you don't want to mix up these two datasets*. Throughout this problem, I'll call the `deal` dataset the "Deal experiment" and the `greenwich` dataset the "Greenwich experiment."

a. (6pts) First, perform one-way ANOVA on the Deal experiment and on the Greenwich experiment **using only the post-test and literacy program as variables** (i.e., ignore pre-test for now). After running your one-way ANOVA analyses, state the scientific conclusion of both ANOVAs.

Hide

```
summary(aov(posttest~program, data = deal))
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## program       2    0.8   0.390   0.047  0.955
## Residuals   147 1233.2   8.389
```

Hide

```
summary(aov(posttest~program, data = greenwich))
```

```
##              Df Sum Sq Mean Sq F value  Pr(>F)
## program       2  127.4   63.71   7.122 0.00112 **
## Residuals   147 1315.1    8.95
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

[The one-way ANOVA results shows that there is no significant difference between the mean post test results of following three groups, “writing”, “reading”, and “standard”, in professor Deal’s data set as the P-value is above 0.05 (0.955). However, The one-way ANOVA results for professor Greenwich shows that there is significant difference between the mean post test results of following three groups, “writing”, “reading”, and “standard”, as the P-value is less than 0.05 (0.00112).]

b. (14pts) Now, **again using only the post-test and program as variables** (i.e., ignoring pre-test) perform three t-tests, comparing standard vs writing, writing vs reading, and standard vs reading, for the Deal experiment and Greenwich experiment. Thus, in this part, you should perform six t-tests. (As we’ve talked about in class, there are technically multiple testing issues when we do this, but let’s ignore that for this homework.) For each t-test, when you use `t.test()`, you can set `var.equal = TRUE` or `var.equal = FALSE` (either is fine, because the results will be the same regardless). Please be sure to display the output from all of your t-tests. After running your six t-tests, answer the following: What is the p-value and scientific conclusion of each of the six resulting t-tests?

Hint: To help you with this part, here is the code to create six different datasets (a dataset comparing just the standard and writing programs, just the writing and reading programs, and just the standard and reading programs, for both experiments):

Hide

```
#standard and writing (for Deal)
deal.SW = subset(deal, program != "reading")
#writing and reading (for Deal)
deal.WR = subset(deal, program != "standard")
#standard and reading (for Deal)
deal.SR = subset(deal, program != "writing")

#standard and writing (for Greenwich)
greenwich.SW = subset(greenwich, program != "reading")
#writing and reading (for Greenwich)
greenwich.WR = subset(greenwich, program != "standard")
#standard and reading (for Greenwich)
greenwich.SR = subset(greenwich, program != "writing")
```

Hide

```
t.test(posttest~program, data = deal.SW)
```

```
##
##  Welch Two Sample t-test
##
## data:  posttest by program
## t = 0.21305, df = 91.232, p-value = 0.8318
## alternative hypothesis: true difference in means between group standard and group writi
ng is not equal to 0
## 95 percent confidence interval:
##  -1.047525  1.299239
## sample estimates:
## mean in group standard  mean in group writing
##               7.179600               7.053743
```

Hide

```
t.test(posttest~program, data = deal.WR)
```

```
##
##  Welch Two Sample t-test
##
## data:  posttest by program
## t = -0.072568, df = 94.937, p-value = 0.9423
## alternative hypothesis: true difference in means between group reading and group writin
g is not equal to 0
## 95 percent confidence interval:
##  -1.262510  1.173468
## sample estimates:
## mean in group reading mean in group writing
##              7.009221              7.053743
```

Hide

```
t.test(posttest~program, data = deal.SR)
```

```
##
##  Welch Two Sample t-test
##
## data:  posttest by program
## t = -0.32121, df = 97.077, p-value = 0.7487
## alternative hypothesis: true difference in means between group reading and group standa
rd is not equal to 0
## 95 percent confidence interval:
##  -1.2231215  0.8823647
## sample estimates:
##  mean in group reading mean in group standard
##               7.009221               7.179600
```

Hide

```
t.test(posttest~program, data = greenwich.SW)
```

```
##
##  Welch Two Sample t-test
##
## data:  posttest by program
## t = 2.112, df = 97.766, p-value = 0.03724
## alternative hypothesis: true difference in means between group standard and group writi
ng is not equal to 0
## 95 percent confidence interval:
##  0.07276064 2.33841063
## sample estimates:
## mean in group standard  mean in group writing
##               8.234682               7.029096
```

Hide

```
t.test(posttest~program, data = greenwich.WR)
```

```
##
##  Welch Two Sample t-test
##
## data:  posttest by program
## t = -1.7364, df = 95.757, p-value = 0.08571
## alternative hypothesis: true difference in means between group reading and group writin
g is not equal to 0
## 95 percent confidence interval:
##  -2.2510425  0.1504221
## sample estimates:
## mean in group reading mean in group writing
##              5.978786              7.029096
```

Hide

```
t.test(posttest~program, data = greenwich.SR)
```

```
##
##  Welch Two Sample t-test
##
## data:  posttest by program
## t = -3.6507, df = 96.934, p-value = 0.0004239
## alternative hypothesis: true difference in means between group reading and group standa
rd is not equal to 0
## 95 percent confidence interval:
##  -3.482342 -1.029449
## sample estimates:
##  mean in group reading mean in group standard
##               5.978786               8.234682
```

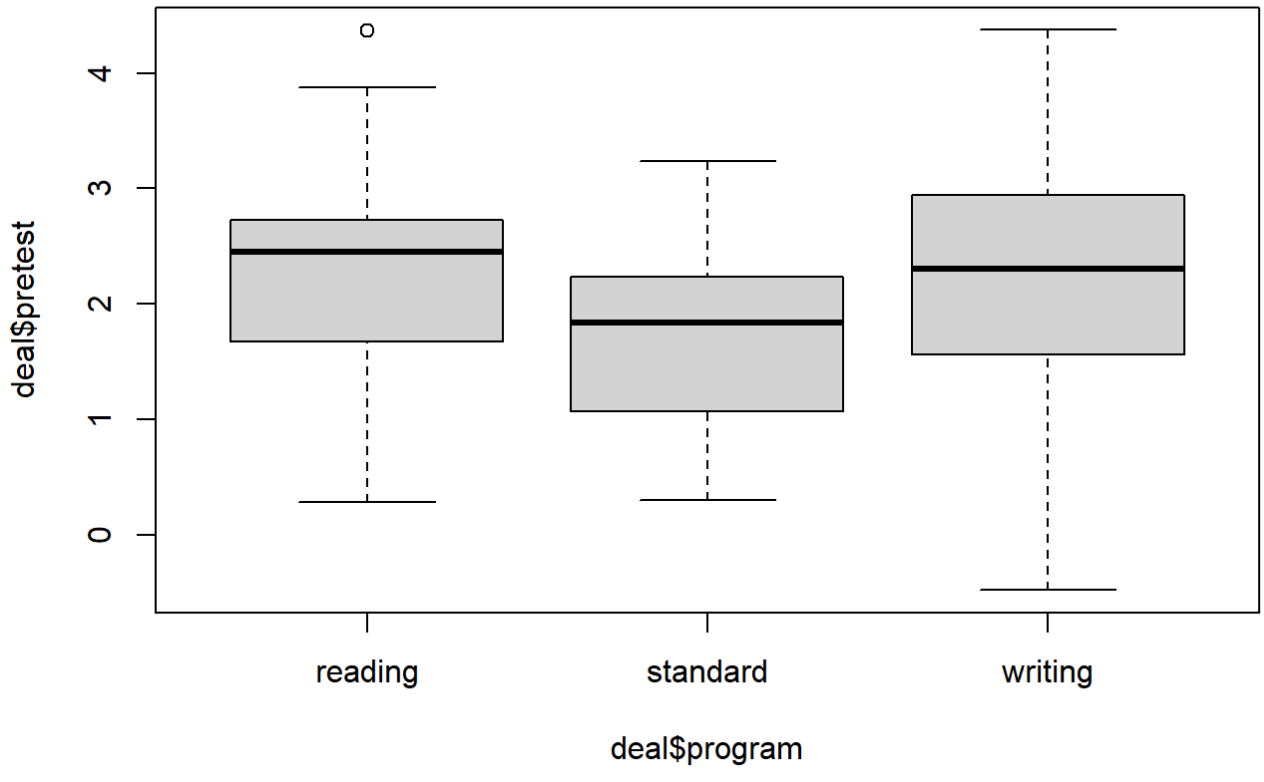
[the t-test results were in consistent with the one-way ANOVA results. Since for the Deal data set, the ANOVA did not show a difference between the posttest means of three groups, the t-test also proves that as none of the “standard vs writing”, “writing vs reading”, and “standard vs reading”

comparisons have a P-value less than 0.05. their P-values are 0.8318, 0.9423, and 0.7487, respectively. Also the CIs includes 0. For the Greenwich dataset, also the ANOVA showed that there is significant difference between the means of three groups, however it was not obvious which groups' means are actually different. The T-test results showed that there is significant difference between the results of the "standard vs writing", and "standard vs reading", as their P-values are less than 0.05, respectively 0.03724, and 0.0004239. The writing group's mean posttest results is less than standard group's mean posttest results in (0.07276064 , 2.33841063) range with 95% confidence interval. The reading group's mean posttest results is less than standard group's mean posttest results with 95% confidence interval and the confidence interval range is (-3.482342, -1.029449). finally there is no significant difference between the reading and writing groups in Greenwich dataset as the P-value was 0.08571 and the CI includes 0.]

c. (10pts) Now let's consider the pre-test variable. In both datasets, the students were randomized to the three programs. Because of this, on average, the students in each of the three programs should have similar pre-test scores. To assess if the randomization within each experiment "worked," use EDA **AND** formal statistical tests to assess if the average pre-test scores are similar in each of the three programs for the Deal experiment and the Greenwich experiment. After conducting your EDA **AND** tests for **EACH** experiment, explain in 1-3 sentences why these EDA and tests are appropriate for assessing if the pre-test scores are, on average, similar across programs. Then, using your EDA and tests, answer the following: Did one experiment work better than another in terms of balancing the average pre-test scores among the three groups? Explain in 1-2 sentences.

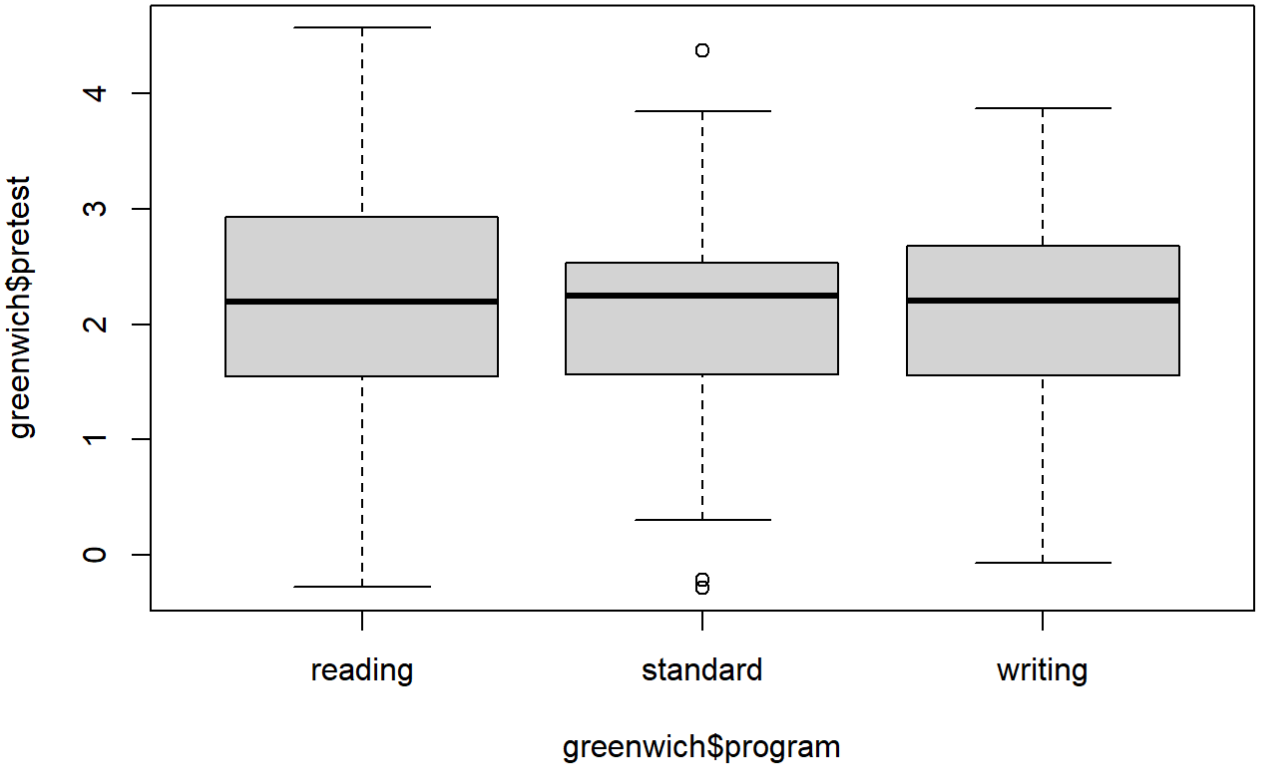
Hide

```
boxplot(deal$pretest~deal$program)
```



Hide

```
boxplot(greenwich$pretest~greenwich$program)
```



Hide

```
summary(aov(pretest~program, data = deal))
```

```
##              Df Sum Sq Mean Sq F value  Pr(>F)
## program       2  10.29   5.145   5.845 0.00361 **
## Residuals   147 129.39   0.880
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hide

```
summary(aov(pretest~program, data = greenwich))
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## program       2   1.36  0.6818   0.718   0.49
## Residuals   147 139.68  0.9502
```

Hide

```
t.test(pretest~program, data = deal.SW)
```

```
##
##  Welch Two Sample t-test
##
## data:  pretest by program
## t = -2.0728, df = 87.597, p-value = 0.04113
## alternative hypothesis: true difference in means between group standard and group writi
ng is not equal to 0
## 95 percent confidence interval:
##  -0.79372834 -0.01668948
## sample estimates:
## mean in group standard  mean in group writing
##               1.716396               2.121605
```

Hide

```
t.test(pretest~program, data = deal.WR)
```

```
##
##  Welch Two Sample t-test
##
## data:  pretest by program
## t = 1.1365, df = 91.091, p-value = 0.2587
## alternative hypothesis: true difference in means between group reading and group writin
g is not equal to 0
## 95 percent confidence interval:
##  -0.1705853  0.6268321
## sample estimates:
## mean in group reading mean in group writing
##              2.349728              2.121605
```

Hide

```
t.test(pretest~program, data = deal.SR)
```

```
##
##  Welch Two Sample t-test
##
## data:  pretest by program
## t = 3.8458, df = 97.43, p-value = 0.0002145
## alternative hypothesis: true difference in means between group reading and group standa
rd is not equal to 0
## 95 percent confidence interval:
##  0.3065002 0.9601645
## sample estimates:
##  mean in group reading mean in group standard
##               2.349728               1.716396
```

[Both EDA and ANOVA and T-test are appropriate to test whether the randomization actually resulted in similar explanatory variables in two data sets for comparing between three groups. the EDAs and ANOVA and T-tests can help in showing whether the pretest results are same between the three groups in two datasets. The Boxplots is used to compare the mean pre-test of each group, and It shows that the means of "reading", "Standard" and "writing" groups are not that similar in deal data set compared to the Greenwich dataset. However, to be sure, the ANOVA and t-Test are used. The anova shows that prof.Greenwich worked better than prof.Deal in terms of balancing the average pre-test scores among the three groups as there was no significant difference between the Greenwich data set groups (P-value = 0.49 > 0.05). however there is a big difference between the Deal data set's groups (P-value = 0.00361 <0.05). the difference is especially between the "reading vs standard", and "Standard vs writing". There is no significant difference between the pre-test mean results of the reading and writing groups in Greenwich dataset.]

d. (10pts) For this part, answer the following questions:

- How does your answer to Part C relate to the concept of **internal validity**? Explain in 1-2 sentences.

[Internal validity says that In order to trust the a hypothesis testing result of an outcome variable in a dataset, the explanatory variables should be similar as they affect the results of outcome variables. The answer to the part C showed that in the Deal data set, since the mean pretest is not similar for the three groups, the internal validity is violated.]

- Which of the two experiments has higher internal validity? Explain in 1-2 sentences.

[The Geenwich has higher internal validity as there in no significant difference between the mean of the pre-test results (as the explanatory variable) between its three groups. this is not the case for the Deal data set.]

- Which experimental results would you "trust" more – the Deal experiment or the Greenwich experiment? Explain in 1-2 sentences.

[I would trust the Greenwich experiment results because it had higher internal validity. And based on internal validity, In order to trust the a hypothesis testing result of an outcome variable in a dataset, the explanatory variables should be similar as they affect the results of outcome variables.]

e. (10pts) Now look at the p-values and scientific conclusions you reported in Part B. For which comparisons (standard vs writing, writing vs reading, standard vs reading) do the Deal experiment and Greenwich experiment agree (in terms of scientific conclusion) and for which do they disagree? Explain why there may be these agreements and disagreements. (**Hint**: To come up with an explanation, it may be helpful to look back at your answer for Part C.)

[They agree that there is no significant difference between the posttest results of reading vs writing groups, as the P-value for both data set were above 0.05. however, they disagreed regarding the standard vs writing, and standard vs reading. The Greenwich showed a difference between these two, but the Deal data set did not. Results of part C showed that the reason this disagreement happened was that the Deal data set, although used random assignment approach, its groups were not similar in terms of their pre-test results (or in other words the Deal data set had low interval validity).]