Final Project

D. Choo-Kang

12/13/2022

Summary

This study explores student retention in STEM. They examine the effects of a psychological intervention on students in a gateway biology course. The differences between affirmed and unaffirmed students in number of friends, position in social networks, and persistence in the biology track were measured through questionnaires and calculated. The paper used multiple regression analyses for various end-of-semester (time 2) dependent variables, such as closeness, number of old friends, number of new friends, and tie strength, with the intervention condition as the critical predictor and course selection and the initial (time 1) measurements as covariates. The equation used for the model is y1 = beta0 + beta1 * Z + beta2 * S + beta3 * y0, where y1 is the post-treatment value of the dependent variables of interest; Z is the intervention condition, the covariate S is the students' course selection, and y0 is the pre-intervention (baseline) value of the dependent variable, where applicable. The model demonstrates the effects of and relationships between the affirmation condition, track persistence, and various friendship variables. On average, affirmed students had 29% more friends than controls and came to hold more central positions in social networks. They were also 11.7% more likely to continue with the next biology course, an effect mediated by time 2 friendships. Based on these findings, they concluded that the affirmation had an effect on social networks and these differences led to biology track persistence.

Introduction

Based on the available data, my paper will focus on number of friendship variables (number of old friends, number of new friends, and proportion of old to new) as the dependent variables of the aforementioned regression models. Therefore, since these do not have separate t1 and t2 values, beta3 and y0 will be excluded from the model. To evaluate the model, I removed participants who did not complete both the time 1 and time 2 questionnaires, which yielded similar results to the original models, and used biology track persistence as the response variable in the models, including friendship variables instead as predictor variables, which resulted in coefficients with greater significance. Overall, my attempted reproductions of selected results suggest that the model is resistant to changes in the data but not very significant, though in general, the original paper had lower p-values than my reproduction.

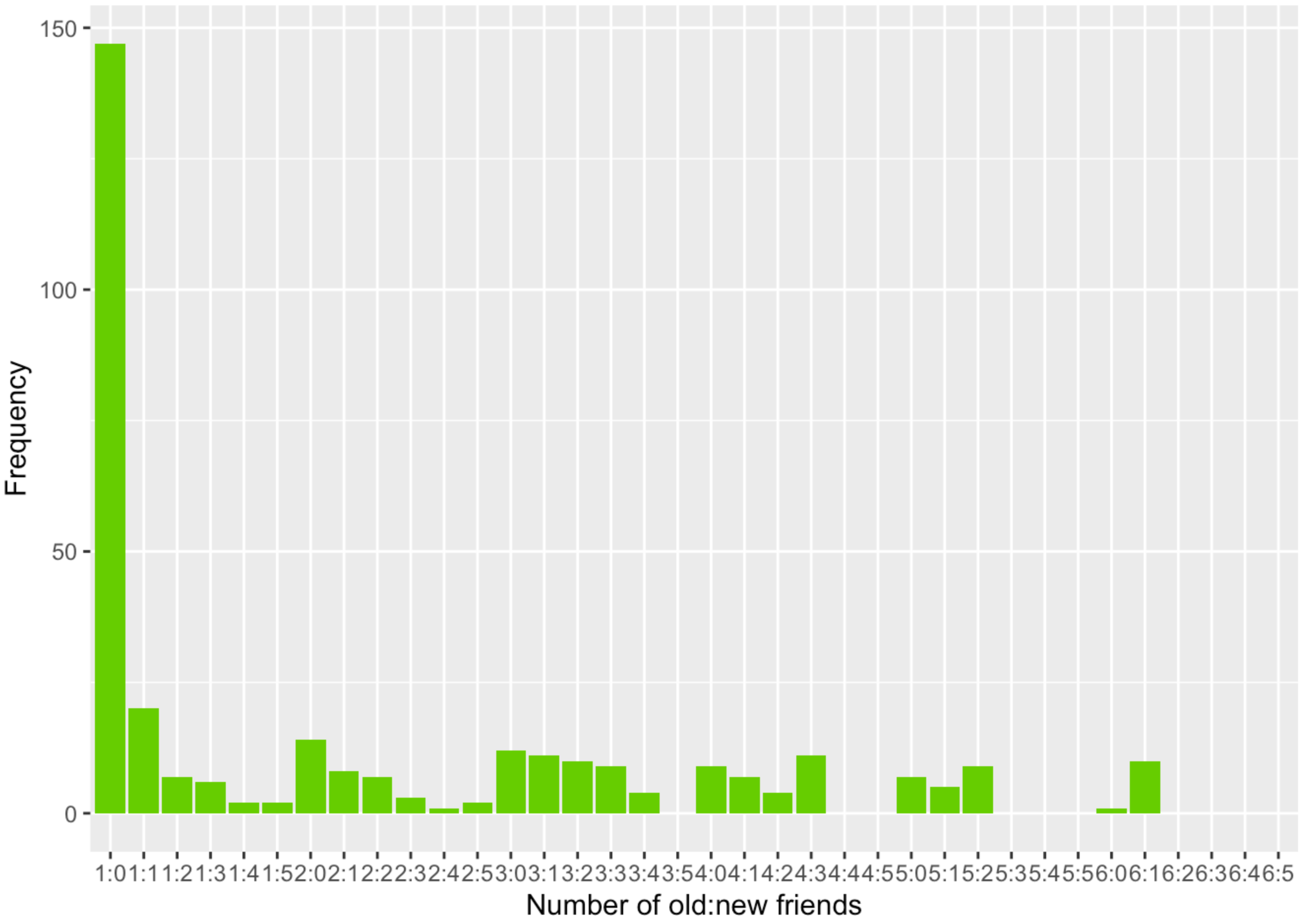
Description about the data

The relevant variables for these reproductions include: PPID, a unique identifier assigned to each participant; number of old friends, the number of friends listed in both the time 1 and time 2 questionnaires, with a maximum of 6; number of new friends, the number of friends listed in the time 2 questionnaire but not the time 1 questionnaire, with a maximum of 6; ratio of old : new friends; intervention condition, whether the students completed the affirmation or the control activity; CompletedPreQ, whether or not the participant completed the time 1 questionnaire; CompletedPostQ, whether or not the participant completed the time 2 questionnaire; and TookSpringClass, whether or not the participant enrolled in the next course in the biology track.

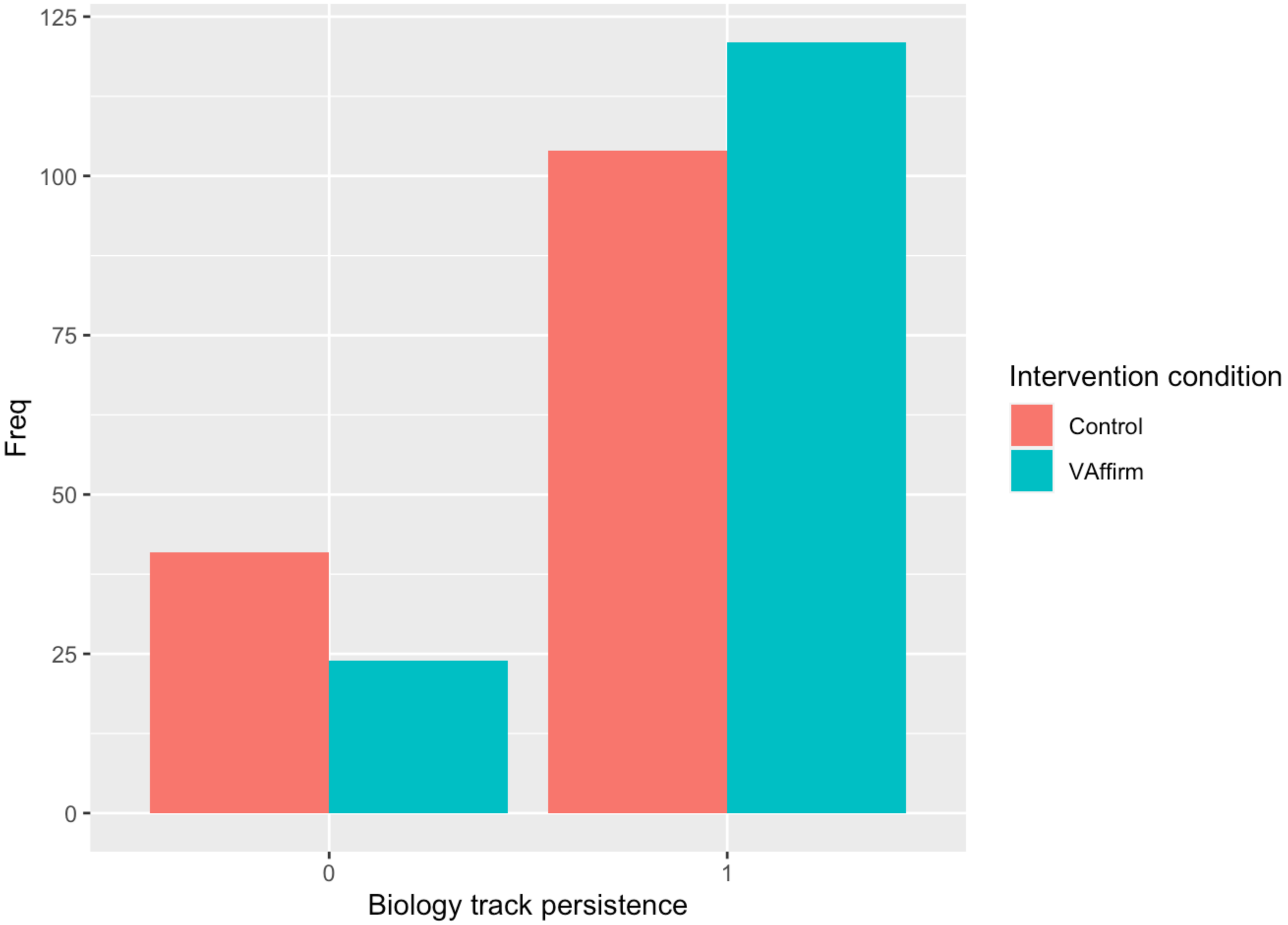
The ideal data set would have no participant attrition between time 1 and time 2, as well as no missing values. It would also have some standard for what counts as a friend or a metric for closeness that is taken into account, as well as no 6-individual limitation on how many can be listed. To be relevant to STEM attrition overall, using data that records intended STEM majors and their later outcomes (employment in related fields post-graduation) may be better suited to the research question. However, such a large-scale experiment would have many other factors involved (ex: friendships in multiple courses, rather than just one) and take at least four years to complete instead of just a semester. The data collected does demonstrate continuation on one STEM course sequence but is limited to one class of students, and though the dates the data was collected are not included, the paper was published at the end of 2020, meaning COVID and the transition to online learning may have had an effect on many of the relevant variables, especially between the beginning and end of the semester measurements. Additionally, the ways by which the affirmation affects social networks remain unclear.

```r
# reading the data
attributes <- read.csv("Attributes.csv")
prepost <- read.csv("PrePostMeasures.csv")
sna_t1 <- read.csv("SNA_T1.csv")
sna_t2 <- read.csv("SNA_T2.csv")

# organizing the data
t1_data <- sna_t1[1:7]

# renaming t2 columns with "2" in front so it can be merged with t1 data based only on PPID
names(sna_t2)[2:7] <- c("2Friend1Name", "2Friend2Name", "2Friend3Name", "2Friend4Name", "2Friend5Name", "2Friend6Name")

t2_data <- sna_t2[1:7]

big_data <- merge(t1_data, t2_data, all=TRUE)
ppl <- nrow(big_data)

# calculating the number of friends listed in t2 but not t1 as new friends
# calculating the number of friends listed in both t1 and t2 as old friends
big_data$old <- 0
big_data$new <- 0
`%!in%` <- Negate(`%in%`)

for(i in seq_len(ppl)){
  t1_friends <- c()
  t2_friends <- c()
  t1_friends <- big_data[i,2:7]
  t2_friends <- big_data[i,8:13]
  for(j in t2_friends){
    if(j %in% t1_friends){
      big_data$old[i] <- big_data$old[i] + 1
    }
    if(is.na(j) == FALSE & j %!in% t1_friends){
      big_data$new[i] <- big_data$new[i] + 1
    }
  }
}

# calculating the ratio of old vs. new friends
big_data$ratio <- big_data$new / big_data$old
big_data$ratio[big_data$ratio == "Inf" | big_data$ratio == "NaN"] <- NA

friend_data <- big_data[c(1, 14:16)]

# merging relevant friend data with attributes
full_data <- merge(friend_data, attributes, all=TRUE)

# full_data <- full_data[complete.cases(full_data),]
# creating a dummy variable for intervention condition (where 1 = affirmed, 0 = contr
ol)
affirm <- ifelse(full_data$Intervention == 'VAffirm', 1, 0)
full_data$int_cond <- affirm
```

```
library(ggplot2)

# visualizing key variables: number of old friends and number of new friends
freq <- table(full_data$old, full_data$new)
freq_df <- as.data.frame(freq)

col <- names(freq_df)[0:1]
freq_df[col] <- lapply(freq_df[col], as.numeric)

for(i in seq_len(nrow(freq_df))){
  freq_df$breakdown[i] <- paste(freq_df[i,"Var1"], freq_df[i,"Var2"], sep = ":")
}

rat_plot <- ggplot(freq_df, aes(x=breakdown, y=Freq)) +
geom_bar(stat="identity", fill="chartreuse3") + labs(x="Number of old:new friends", y
="Frequency")
rat_plot
```



```
# intervention condition and biology course persistence
var_freq <- table(full_data$Intervention, full_data$TookSpringClass)
var_freq_df <- as.data.frame(var_freq)

var_plot <- ggplot(data=var_freq_df, aes(x=Var2, y=Freq, fill=Var1)) + geom_bar(stat
="identity", position=position_dodge()) + labs(x="Biology track persistence", fill="I
ntervention condition")
var_plot
```



Reproducing the results

I am attempting to reproduce the regression results for the following dependent variables: Number of old friends, number of new friends, proportion of old (versus new) friends.

https://www.science.org/doi/10.1126/sciadv.aba9221#T2 (https://www.science.org/doi/10.1126/sciadv.aba9221#T2)

```
# models + results
old_friends <- lm(old ~ affirm + TookSpringClass, data=full_data)
new_friends <- lm(new ~ affirm + TookSpringClass, data=full_data)
ratio_friends <- lm(ratio ~ affirm + TookSpringClass, data=full_data)

summary(old_friends)$coefficients[2,]
```

```
##   Estimate Std. Error    t value   Pr(>|t|)
## 0.32083686 0.17777245 1.80476139 0.07216012
```

```
old_p <- summary(old_friends)$coefficients[2,4]
old_coef <- summary(old_friends)$coefficients[2,1]
# the results reported in the paper are:
# b1: 0.48, SE: 0.21, t(209): 2.26, P: 0.02
summary(new_friends)$coefficients[2,]
```

```
##   Estimate Std. Error    t value   Pr(>|t|)
## 0.24675762 0.14226130 1.73453787 0.08389635
```

```
new_p <- summary(new_friends)$coefficients[2,4]
new_coef <- summary(new_friends)$coefficients[2,1]
# b1: 0.33, SE: 0.18, t(209): 1.83, P: 0.07
summary(ratio_friends)$coefficients[2,]
```

```
##   Estimate Std. Error    t value   Pr(>|t|)
##  0.1227412  0.1582250  0.7757384  0.4392911
```

```
rat_p <- summary(ratio_friends)$coefficients[2,4]
rat_coef <- summary(ratio_friends)$coefficients[2,1]
# b1: -0.03, SE: 0.06, t(150):-0.57, P: 0.57

# control <- subset(full_data, Intervention == "Control")
# affirmed <- subset(full_data, Intervention == "VAffirm")
```

The results I produced are not exactly the same as those reported in the paper. Though the variables are ambiguously named (i.e. does “number of old friends” include only lasting friendships or also lost friends who were listed in t1 but not in t2?), they are defined by the paper, so the data used in my reproduction should be the same. For the ratio of old : new friends variable, I changed all Inf or NaN values to NAs and excluded them from my model. This might explain why my results were the most different from those of the original paper for the ratio model. Another potential issue is they did not specify whether the number of friends data had any weight based on closeness (though they describe using weights in their representations of network centrality). If taken into account, that would explain a discrepancy between their report and my results, as well as the differences in mean and standard error values. One significant inconsistency was that my model included more data than theirs. This might be because they ran regressions for more dependent variables than I did, so based on how they organized their data, they would have more NA values (fewer complete cases), which results in more data being excluded from the model.

Evaluating the result

Testing the result’s sensitivity to model choice

Logically, I thought participants who do not plan on taking the next course may be less likely to complete the second questionnaire because they may be less invested in the course overall. Some participants might have even completed the time 1 questionnaire and then dropped the course. However, pre and post-questionnaire completion was not accounted for in the analyses, so participants who did not complete the post-questionnaire were considered to just have no (old or new) friends since the numbers were calculated based on both time 1 and time 2 data. I decided to test the sensitivity of my results (above) to data by removing participants who did not complete both questionnaires from the model.

```
# removing participants who did not complete both the pre and post-questionnaires
eval_data <- full_data
eval_data$CompletedPreQ[eval_data$CompletedPreQ == "0"] <- NA
eval_data$incomplete <- eval_data$CompletedPreQ > eval_data$CompletedPostQ

eval_df <- subset(eval_data, incomplete != TRUE)

eval_df$incomplete[eval_df$incomplete == ""] <- NA

affirm <- ifelse(eval_df$Intervention == 'VAffirm', 1, 0)
eval_df$int_cond <- affirm

old_eval <- lm(old ~ affirm + TookSpringClass, data=eval_df)
new_eval <- lm(new ~ affirm + TookSpringClass, data=eval_df)
ratio_eval <- lm(ratio ~ affirm + TookSpringClass, data=eval_df)

summary(old_eval)$coefficients[2,]
```

```
##   Estimate Std. Error    t value   Pr(>|t|)
##  0.3135396  0.2076100  1.5102340  0.1323994
```

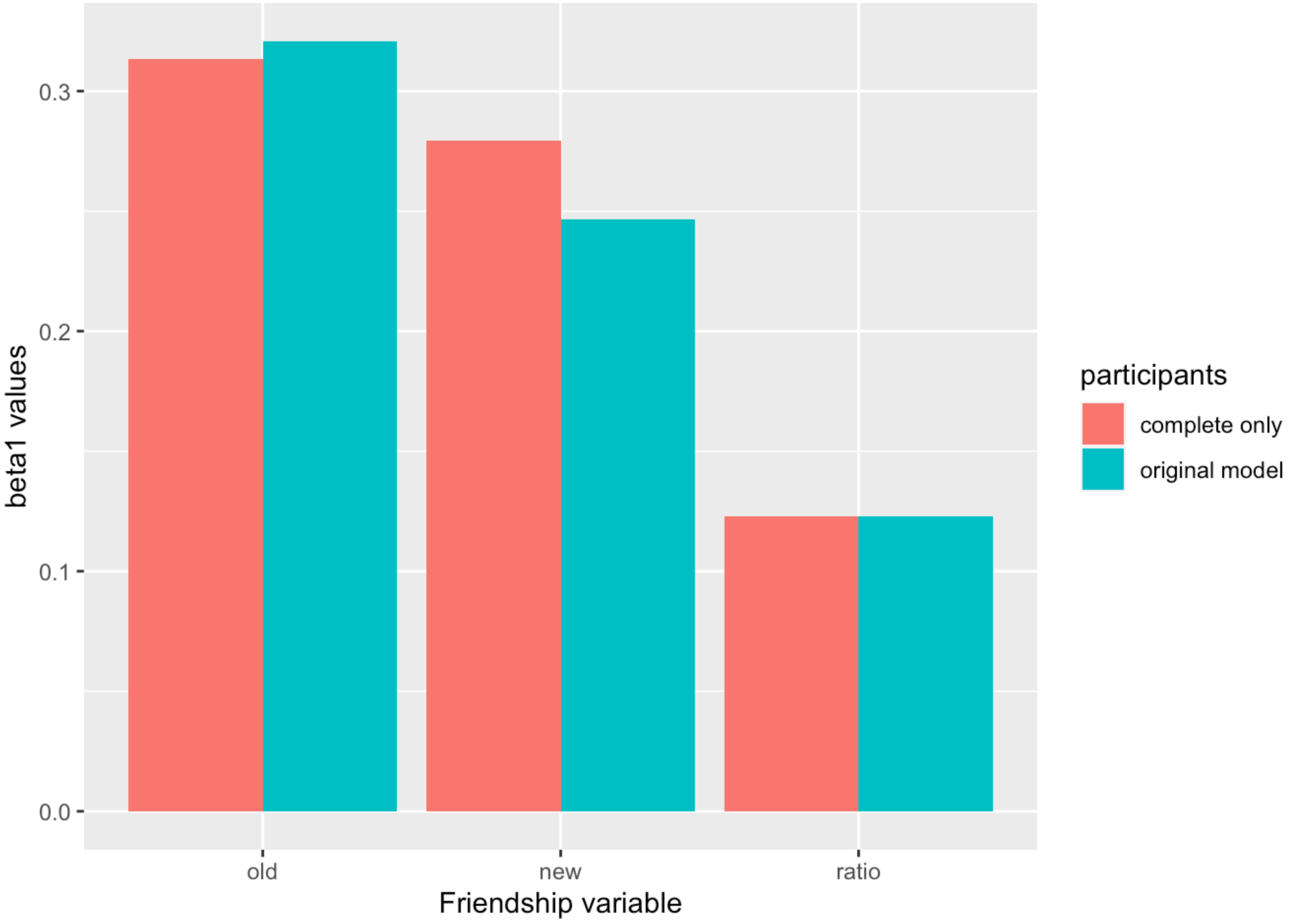
```
old_eval_coef <- summary(old_eval)$coefficients[2,1]
summary(new_eval)$coefficients[2,]
```

```
##   Estimate Std. Error    t value   Pr(>|t|)
## 0.27936946 0.16582259 1.68474913 0.09343593
```

```
new_eval_coef <- summary(new_eval)$coefficients[2,1]
summary(ratio_eval)$coefficients[2,]
```

```
##   Estimate Std. Error    t value   Pr(>|t|)
##  0.1227412  0.1582250  0.7757384  0.4392911
```

```r
rat_eval_coef <- summary(ratio_eval)$coefficients[2,1]

eval_plot_df <- data.frame(
  participants = factor(c("original model", "original model", "original model", "complete only", "complete only", "complete only")),
  type = factor(c("old", "new", "ratio", "old", "new", "ratio"), levels=c("old", "new", "ratio")),
  values = c(old_coef, new_coef, rat_coef, old_eval_coef, new_eval_coef, rat_eval_coef)
)

eval_plot <- ggplot(data=eval_plot_df, aes(x=type, y=values, fill=participants)) + geom_bar(stat="identity", position=position_dodge()) + labs(y="beta1 values", x="Friendship variable")
eval_plot
```



The models that included and excluded incomplete questionnaires yielded similar results, in terms of the affirm coefficient estimates, standard errors, t-values, and p-values. The beta1 values (intervention condition coefficients) are visualized in the above graph. Note that because participants who did not complete the post-questionnaire were considered to have 0 friends, the models with ratio of old : new friends as the dependent variable are automatically excluded, as they equal 0/0 and have NaN values. Therefore, identical data was used for both the original and complete only models with ratio of old : new friends as the response variable. This evaluation demonstrates that the result is not very sensitive to the removal of some of the data based on questionnaire/study completion.

Testing the sensitivity to model choice

Instead of using the friendship variables as the response variables in the model, I will try using biology track persistence. Because biology track persistence only has two outcomes (taking the spring class or not), I also use changed my models to be logistic.

```
flip_old <- glm(TookSpringClass ~ old + int_cond, data=full_data, family=binomial(lin
k="logit"))
flip_new <- glm(TookSpringClass ~ new + int_cond, data=full_data, family=binomial(lin
k="logit"))
flip_rat <- glm(TookSpringClass ~ ratio + int_cond, data=full_data, family=binomial(l
ink="logit"))

summary(flip_old)
```

```
## 
## Call:
## glm(formula = TookSpringClass ~ old + int_cond, family = binomial(link = "logit"), 
##     data = full_data)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -2.3171   0.4259   0.6407   0.7193   0.8972  
## 
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)    
## (Intercept)   0.7022     0.2052   3.423  0.00062 ***
## old           0.2590     0.1113   2.327  0.01995 *  
## int_cond      0.6168     0.2932   2.103  0.03544 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 308.62  on 289  degrees of freedom
## Residual deviance: 296.76  on 287  degrees of freedom
##   (38 observations deleted due to missingness)
## AIC: 302.76
## 
## Number of Fisher Scoring iterations: 4
```

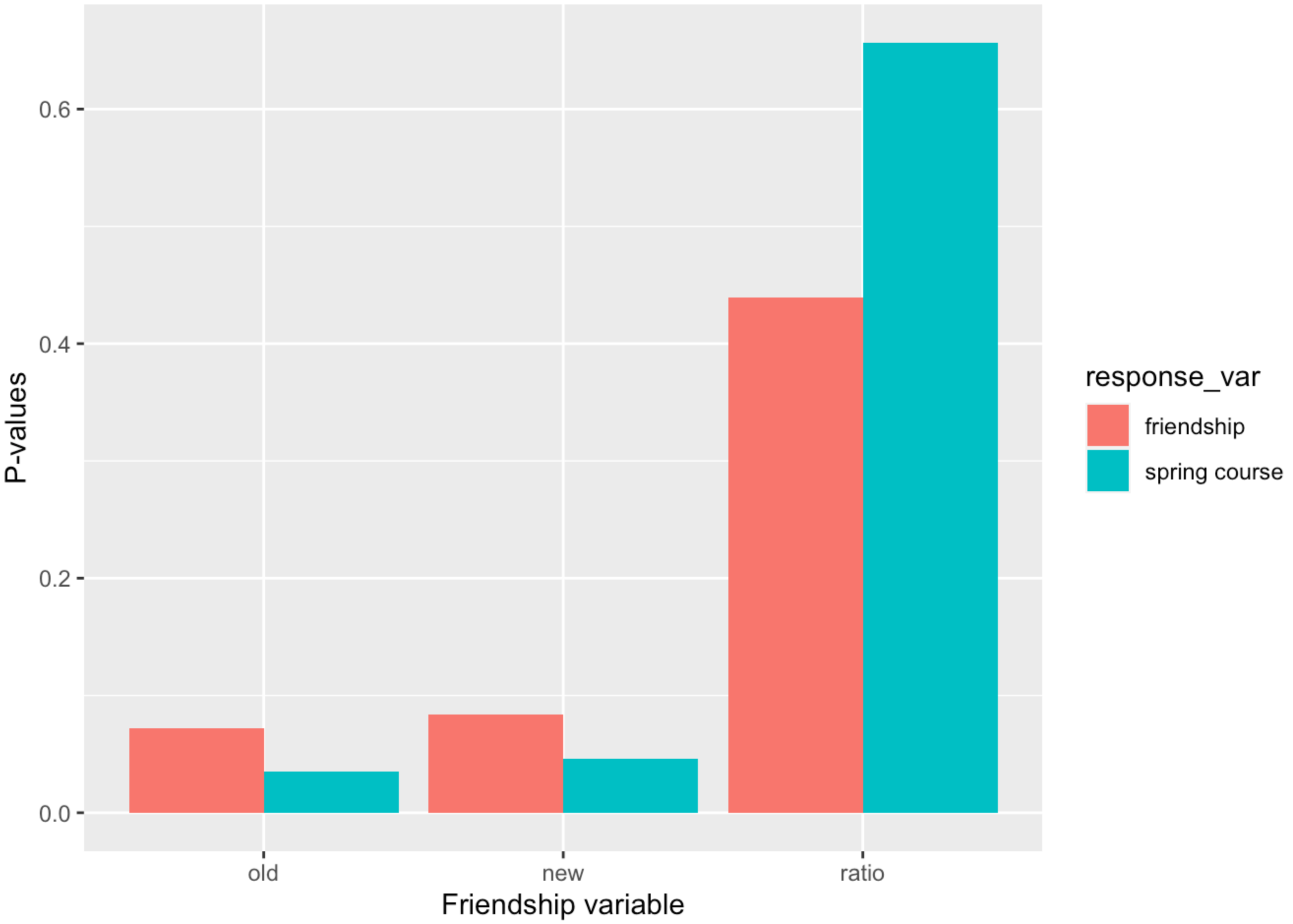
```
flip_old_p <- summary(flip_old)$coefficients[3,4]
summary(flip_new)
```

```
## 
## Call:
## glm(formula = TookSpringClass ~ new + int_cond, family = binomial(link = "logit"), 
##     data = full_data)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -2.2864   0.3561   0.5809   0.7563   0.9041  
## 
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)    
## (Intercept)   0.6835     0.2004   3.410 0.000649 ***
## new           0.4219     0.1544   2.733 0.006275 ** 
## int_cond      0.5885     0.2951   1.994 0.046125 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 308.62  on 289  degrees of freedom
## Residual deviance: 293.64  on 287  degrees of freedom
##   (38 observations deleted due to missingness)
## AIC: 299.64
## 
## Number of Fisher Scoring iterations: 4
```

```
flip_new_p <- summary(flip_new)$coefficients[3,4]
summary(flip_rat)
```

```
## 
## Call:
## glm(formula = TookSpringClass ~ ratio + int_cond, family = binomial(link = "logi
t"), 
##     data = full_data)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -2.1963   0.3870   0.4977   0.5629   0.6272  
## 
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)    
## (Intercept)   1.7622     0.4435   3.973  7.1e-05 ***
## ratio         0.5278     0.4235   1.246    0.213    
## int_cond     -0.2359     0.5308  -0.444    0.657    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 102.21  on 134  degrees of freedom
## Residual deviance: 100.06  on 132  degrees of freedom
##   (193 observations deleted due to missingness)
## AIC: 106.06
## 
## Number of Fisher Scoring iterations: 5
```

```
flip_rat_p <- summary(flip_rat)$coefficients[3,4]

newmod_df <- data.frame(
  response_var = factor(c("friendship", "friendship", "friendship", "spring course", 
"spring course", "spring course")),
  type = factor(c("old", "new", "ratio", "old", "new", "ratio"), levels=c("old", "ne
w", "ratio")),
  values = c(old_p, new_p, rat_p, flip_old_p, flip_new_p, flip_rat_p)
)

newmod_graph <- ggplot(data=newmod_df, aes(x=type, y=values, fill=response_var)) + ge
om_bar(stat="identity", position=position_dodge()) + labs(y="P-values", x="Friendship
variable")
newmod_graph
```



The new logistic models I proposed with spring course outcome as the response variable and number of [old or new] friends as predictor variables had coefficients that were more significant than in the model used in the paper. The visual displays the p-values for each model and demonstrates how the coefficients for my updated logistic model with spring course outcome as the response variable are more significant for every friendship variable, though only number of old friends and number of new friends were significant at the 0.05 level, out of all six models. These results suggest that intervention condition may be a better predictor of spring course outcome than of the various friendship variables. The directions and ties between these variables might be complex and unclear. One thing to note, however, is that the study authors found the greater likelihood of affirmed students to take the spring course was statistically mediated by time 2 friendships.

Conclusion

Some of the data for this study was not made publicly available, but I think it would be interesting to use a model that takes some other variables into account, such as gender and race, since gender and race-based disparities in STEM are known to persist, or to remove some data points based on these categorizations. Being able to weigh friendship variables based on closeness would also be interesting, and I am curious as to whether including that would make my results more similar to those I attempted to reproduce or if something else caused the differences. Though the models and my attempted reproductions of them did not show highly significant and strong associations between the variables, the implications on STEM attrition, the power of psychological interventions, and social networks are fascinating to consider.