Self-Efficacy and Performance of Research Skills Among First-Semester Bioscience Doctoral Students

This is an R Markdown Notebook detailing all of the data analysis completed for the paper entitled "Self-Efficacy and Performance of Research Skills Among First-Semester Bioscience Doctoral Students".

All analyses were performed by K. Lachance. This notebook was compiled on May 26, 2020.

Library Import

Import libraries and packages required for analysis and visualization and set seed for reproducibility.

```
# Install packages, if not already installed
list.of.packages <- c("reshape2", "ggplot2", "gridExtra", "grid", "orddom", "svglite", "ggdendro", "plo
new.packages <- list.of.packages[!(list.of.packages %in% installed.packages()[,"Package"])]
if(length(new.packages)) install.packages(new.packages)

# Load packages
library(reshape2) # Data structure manipulation
library(ggplot2) # Plotting
library(gridExtra) # Plotting
library(grid) # Plotting
library(orddom) # Statistics
```

```
## Loading required package: psych

##
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
library(svglite) # Plotting
library(ggdendro) # Plotting dendrograms
library(plotly) # Plotting dendrograms
```

```
##
## Attaching package: 'plotly'

## The following object is masked from 'package:ggplot2':
##
##     last_plot
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

```
## The following object is masked from 'package:graphics':
##
##     layout
```

```
# Make jitter plots and other analyses / visualizations reproducible
set.seed(123)
```

Data Import

Import data, contained in ExperimentalDesignData.csv file. Each row corresponds to a student (identified with an anonymous code) and each column contains information about that student's response about research skills self-efficacy, experimental design aptitude, experience, and demographic information. In total, 103 students were surveyed (only students that completed all pre- and post-test assessments were included in this study).

```
# Main data
data = read.csv(file = "./StudentData.csv", header = TRUE, row.names = 1, stringsAsFactors = FALSE) # I
numStudents = nrow(data) # Calculate number of students in cohort to be analyzed

# Print output
cat(paste("There are ", numStudents, " student responses contained in the following analyses.\n", sep =
```

```
## There are 103 student responses contained in the following analyses.
```

```
# Factors in experience and comfort
data2 = read.csv(file = "./ExperienceComfortFactors.csv", header = TRUE, stringsAsFactors = FALSE) # Im

# Demographics by year for Table S1
data3 = read.csv(file = "./DemographicsByYear.csv", header = TRUE, row.names = 1, stringsAsFactors = FA

# Demographics by sex for Table S4
data4 = read.csv(file = "./DemographicsBySex.csv", header = TRUE, row.names = 1, stringsAsFactors = FAL
```

Figure 1

Length of time spent doing pre-doctoral research is not predictive of student research skills self-efficacy or self-reported experience and comfort with experimental design.

Figure 1A

```
# Make a data frame holding frequency of students' previous research experience
labExp <- melt(table(data$LabExp))
colnames(labExp) = c("Years", "Freq")
```

```r
labExp$Years = as.character(labExp$Years)
labExp$Percent = round((labExp$Freq / numStudents) * 100, 0)

# Make pie chart
labExp$Years <- factor(labExp$Years, levels = c("7", "6", "5", "4", "3", "2", "1")) # Re-order for plot
fig1a <- ggplot(labExp, aes(x="", y=Percent, fill=Years))+
  geom_bar(width = 1, stat = "identity") +
  coord_polar("y", start=0) +
  scale_fill_manual("Previous Lab Experience", values=c("#333333", "#4C4C4C", "#666666", "#808080", "#9
                    labels=c(paste("> 7 years (", labExp$Percent[7], "%)", sep=""),
                             paste("6 - 7 years (", labExp$Percent[6], "%)", sep=""),
                             paste("5 - 6 years (", labExp$Percent[5], "%)", sep=""),
                             paste("4 - 5 years (", labExp$Percent[4], "%)", sep=""),
                             paste("3 - 4 years (", labExp$Percent[3], "%)", sep=""),
                             paste("2 - 3 years (", labExp$Percent[2], "%)", sep=""),
                             paste("< 2 years (", labExp$Percent[1], "%)", sep="")))+
  theme_minimal() +
  theme(axis.title = element_blank(), panel.border = element_blank(), panel.grid=element_blank(), axis.
  ggtitle("Figure 1A") + theme(plot.title = element_text(hjust = 0.5))

ggsave(filename = "./Figures/Fig1A.svg", plot = fig1a, width = 6, height = 4)
plot(fig1a)
```

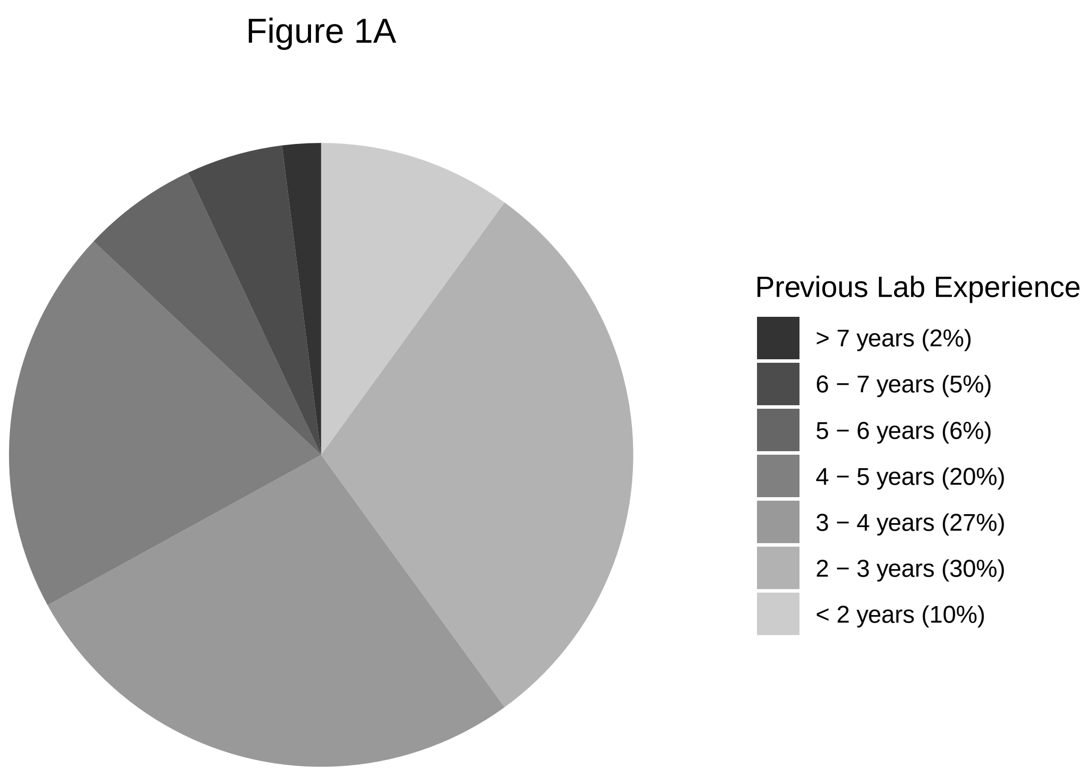


Figure 1B

```
# Split data by responses to self-efficacy questionaire before and after the first semester of graduate
SE_pre = data[,14:27]
SE_post = data[,28:41]
SE_pre$LabExp = as.character(data$LabExp)
SE_post$LabExp = as.character(data$LabExp)
# Melt responses
mSE_pre = melt(SE_pre)
```

```
## Using LabExp as id variables
```

```
mSE_post = melt(SE_post)
```

```
## Using LabExp as id variables
```

```
# Add question number label to each pre / post array
mSE_pre$Question = rep(c("Q1", "Q2", "Q3", "Q4", "Q5", "Q6", "Q7", "Q8", "Q9", "Q10", "Q11", "Q12", "Q13
mSE_pre$Test = "Pre"
mSE_post$Question = rep(c("Q1", "Q2", "Q3", "Q4", "Q5", "Q6", "Q7", "Q8", "Q9", "Q10", "Q11", "Q12", "Q1
mSE_post$Test = "Post"

# Change labels to strings for categorical plotting
mSE_pre[mSE_pre$value==1,"value"] = "a1" # Not at all, pre-test
mSE_pre[mSE_pre$value==2,"value"] = "b1" # A little, pre-test
mSE_pre[mSE_pre$value==3,"value"] = "c1" # A moderate amount, pre-test
mSE_pre[mSE_pre$value==4,"value"] = "d1" # A lot, pre-test
mSE_pre[mSE_pre$value==5,"value"] = "e1" # A great deal, pre-test

mSE_post[mSE_post$value==1,"value"] = "a2" # Not at all, post-test
mSE_post[mSE_post$value==2,"value"] = "b2" # A little, post-test
mSE_post[mSE_post$value==3,"value"] = "c2" # A moderate amount, post-test
mSE_post[mSE_post$value==4,"value"] = "d2" # A lot, post-test
mSE_post[mSE_post$value==5,"value"] = "e2" # A great deal, post-test

# Stacked bar chart
labexp.labs = c("< 2 years", "2 - 3 years", "3 - 4 years", "4 - 5 years", "5 - 6 years", "6 - 7 years",
names(labexp.labs) = as.character(1:7)

fig1b <- ggplot(mSE_pre, aes(Test)) +
  geom_bar(aes(fill=value), position = "fill") + facet_grid(~ LabExp, labeller = labeller(LabExp = labe
  theme_bw() +
  scale_fill_manual("Research Skills Self-Efficacy", values = c("a1" = "#DAE3F3", "b1" = "#B4C7E7", "c1
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank()) +
  theme(axis.title=element_blank(), axis.text=element_blank(), axis.ticks=element_blank(), strip.backgr
  ggtitle("Figure 1B") + theme(plot.title = element_text(hjust = 0.5))

ggsave(fig1b, filename="./Figures/Fig1B.svg", width = 4, height = 3)
plot(fig1b)
```

Figure 1B

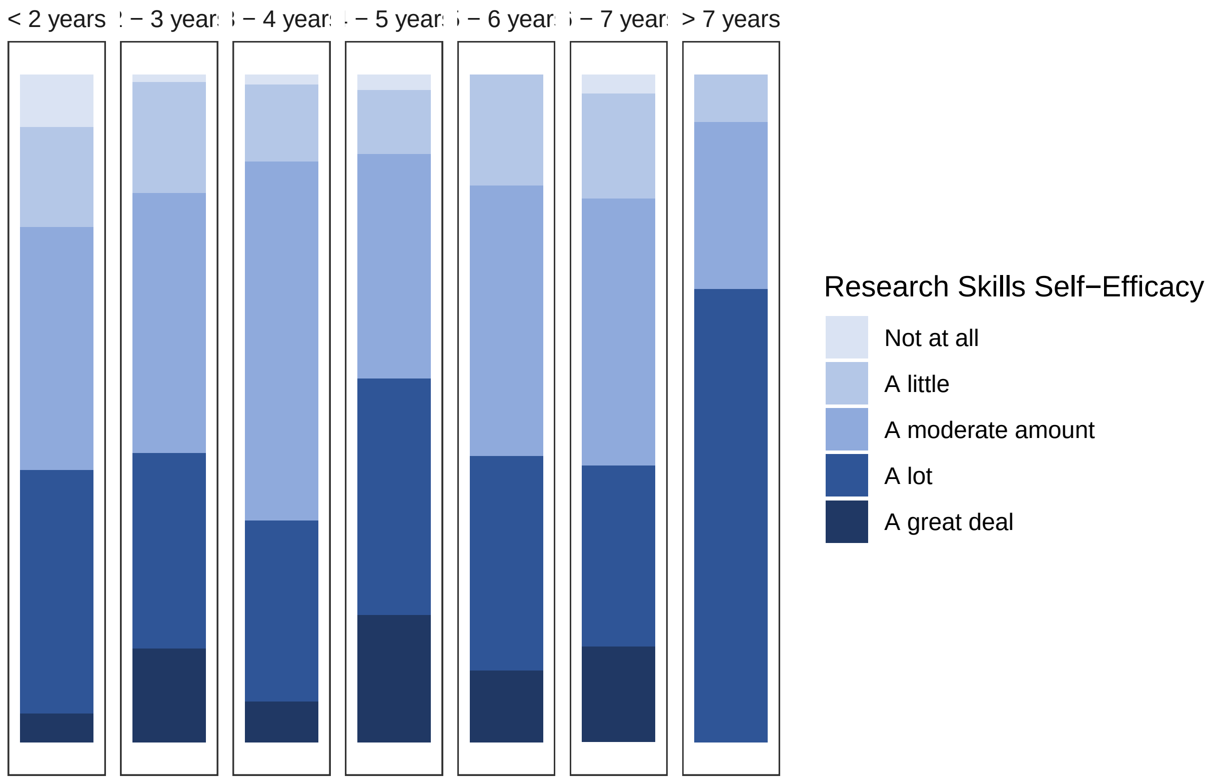


Figure 1Ci

```
# Calculate linear regression
lm_exp = lm(formula = ExperimetalDeignExperience ~ LabExp, data = data)
summary(lm_exp) # Adjusted R-squared:  0.06328
```

```
##
## Call:
## lm(formula = ExperimetalDeignExperience ~ LabExp, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.29102 -0.79102 -0.08613  0.70898  1.91387
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.67634    0.24372  10.981  < 2e-16 ***
## LabExp       0.20490    0.07294   2.809  0.00597 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.013 on 101 degrees of freedom
## Multiple R-squared:  0.07247,    Adjusted R-squared:  0.06328
## F-statistic: 7.891 on 1 and 101 DF,  p-value: 0.005966
```

```
# Scatter plot
fig1ci <- ggplot(data, aes(x = LabExp, y = ExperimetalDeignExperience)) +
  geom_jitter(width = 0.25, height = 0.25, size = 2, color = "#808080") +
  scale_x_continuous("Previous years working in lab", breaks=c(1:7), limits=c(0.5,7.5), labels = c("< 2
  scale_y_continuous("Self-reported experience\nwith experimental design", breaks=c(1:5), limits=c(0.5,5
  theme_bw() +
  theme(panel.grid.minor = element_blank()) +
  ggtitle(expression(paste("Figure 1Ci (", R^2," = ", 0.06, ")", sep=""))) + theme(plot.title = element_

ggsave(fig1ci, filename="./Figures/Fig1Ci.svg", width = 4, height = 3)
plot(fig1ci)
```

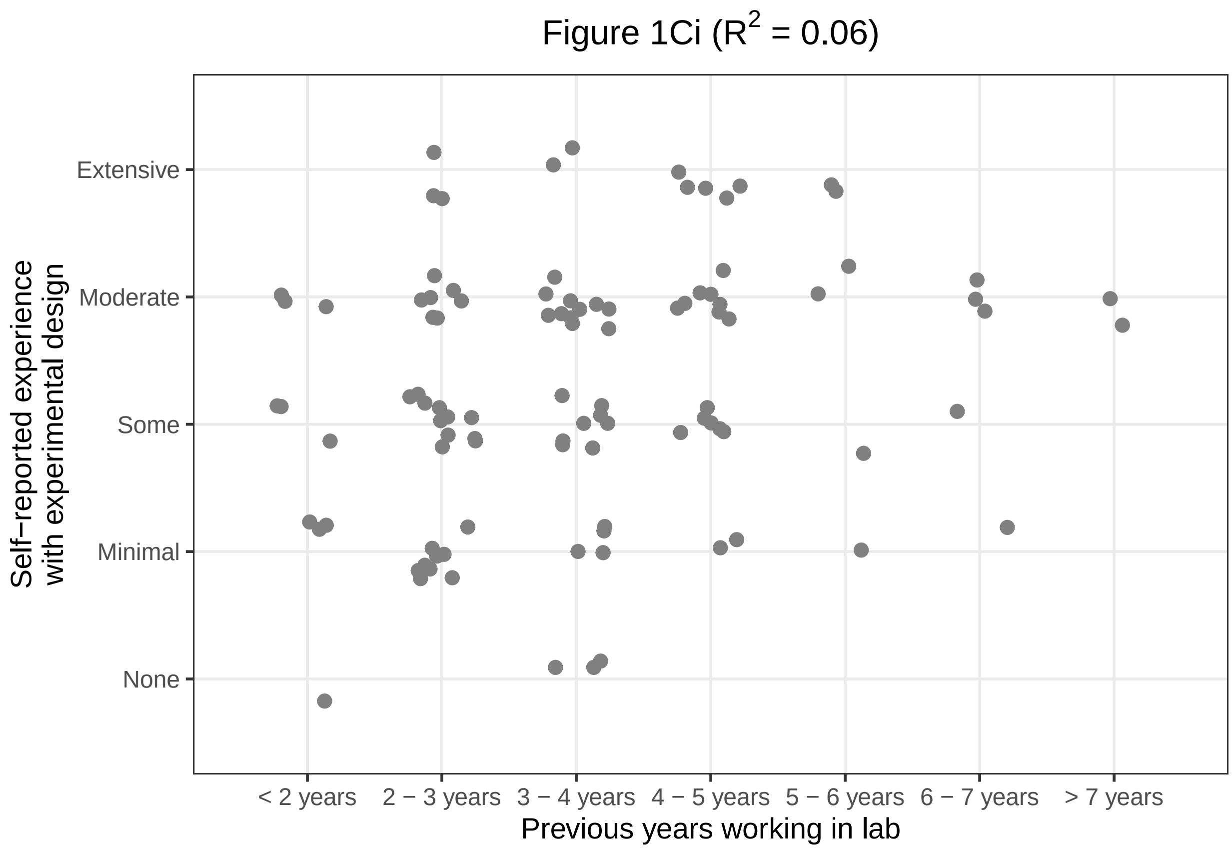
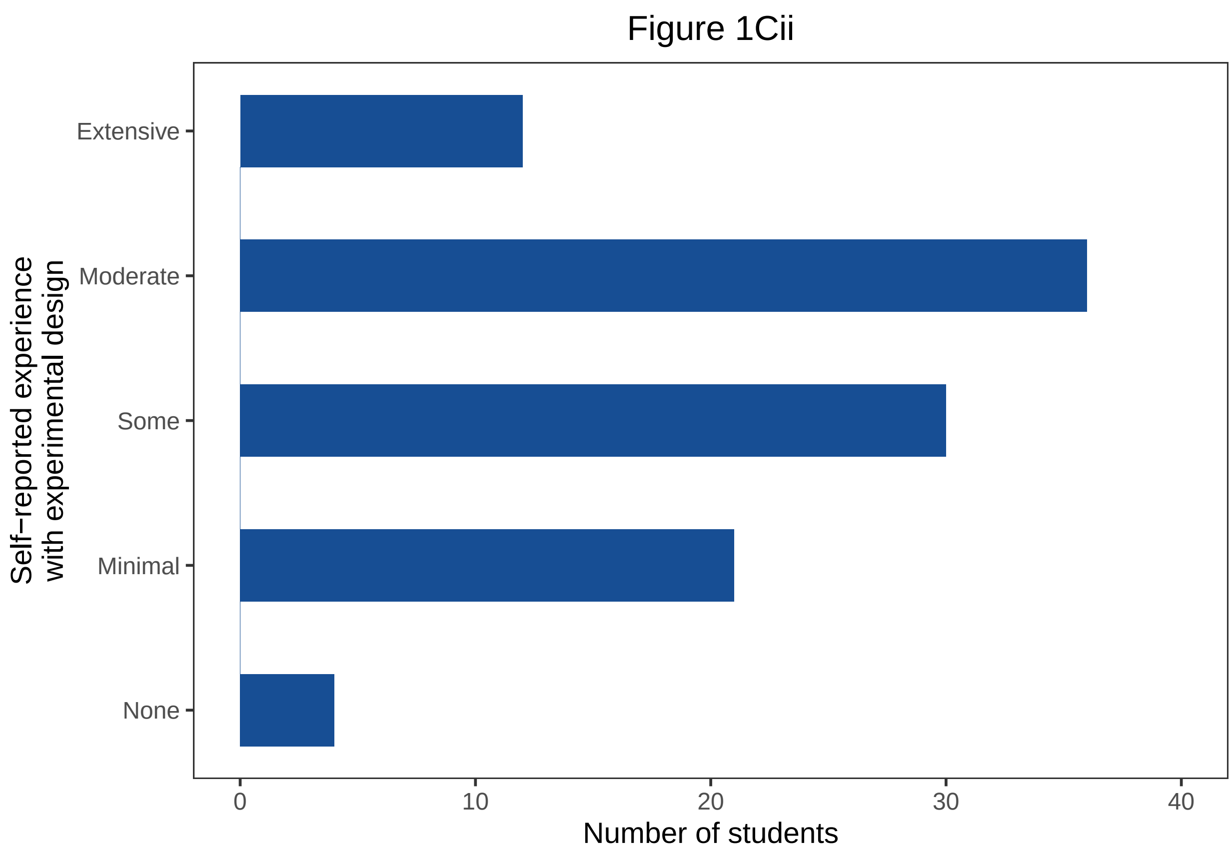


Figure 1Cii

```
# Histogram
fig1cii <- ggplot(data, aes(x = ExperimetalDeignExperience)) +
  geom_histogram(binwidth = 0.5, fill = "#174E94") +
  theme_bw() +
  scale_x_continuous("Self-reported experience\nwith experimental design", breaks=c(1:5), labels = c("N
  ylim(0, 40) + ylab("Number of students") +
  coord_flip() +
  theme(panel.grid.minor = element_blank(), panel.grid.major = element_blank()) +
  ggtitle("Figure 1Cii") + theme(plot.title = element_text(hjust = 0.5))

plot(fig1cii)
```




```
ggsave(fig1cii, filename="./Figures/Fig1Cii.svg", width = 4, height = 3)
```

Figure 1Di

```
# Calculate linear regression
lm_comf = lm(formula = ExperimetalDeignComfort ~ LabExp, data = data)
summary(lm_comf) # Adjusted R-squared:  0.05171
```

```
##
## Call:
## lm(formula = ExperimetalDeignComfort ~ LabExp, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.736 -1.148  0.381  1.264  2.852
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.5593     0.3838   9.274 3.57e-15 ***
## LabExp        0.2942     0.1149   2.562   0.0119 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.595 on 101 degrees of freedom
## Multiple R-squared:  0.06101,    Adjusted R-squared:  0.05171
## F-statistic: 6.562 on 1 and 101 DF,  p-value: 0.01189
```

```r
# Scatter plot
fig1di <- ggplot(data, aes(x = LabExp, y = ExperimetalDeignComfort)) +
  geom_jitter(width = 0.25, height = 0.25, size = 2, color = "#808080") +
  scale_x_continuous("Previous years working in lab", breaks=c(1:7), limits=c(0.5,7.5), labels = c("< 2
  scale_y_continuous("Self-reported comfort\nwith experimental design", breaks=c(1:7), limits=c(0.5,7.5
  theme_bw() +
  theme(panel.grid.minor = element_blank()) +
  ggtitle(expression(paste("Figure 1Di (", R^2," = ", 0.05, ")", sep=""))) + theme(plot.title = element_

ggsave(fig1di, filename="./Figures/Fig1Di.svg", width = 4, height = 3)
plot(fig1di)
```

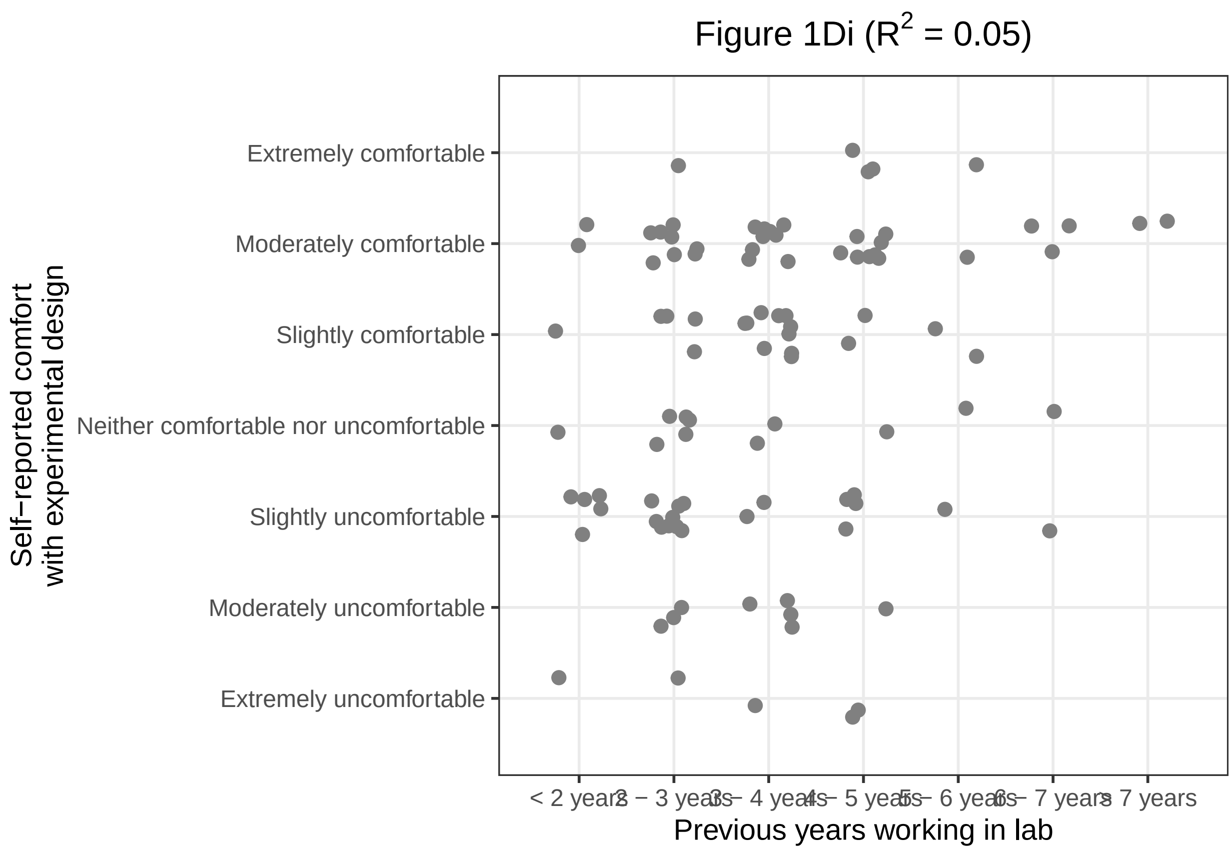


Figure 1Dii

```r
# Histogram
fig1dii <- ggplot(data, aes(x = ExperimetalDeignComfort)) +
  geom_histogram(binwidth = 0.5, fill = "#174E94") +
  theme_bw() +
  scale_x_continuous("Self-reported Comfort\nwith experimental design", breaks=c(1:7), labels = c("Extre
  ylim(0, 40) + ylab("Number of students") +
  coord_flip() +
  theme(panel.grid.minor = element_blank(), panel.grid.major = element_blank()) +
  ggtitle("Figure 1Dii") + theme(plot.title = element_text(hjust = 0.5))

ggsave(fig1dii, filename="./Figures/Fig1Dii.svg", width = 4, height = 3)
plot(fig1dii)
```

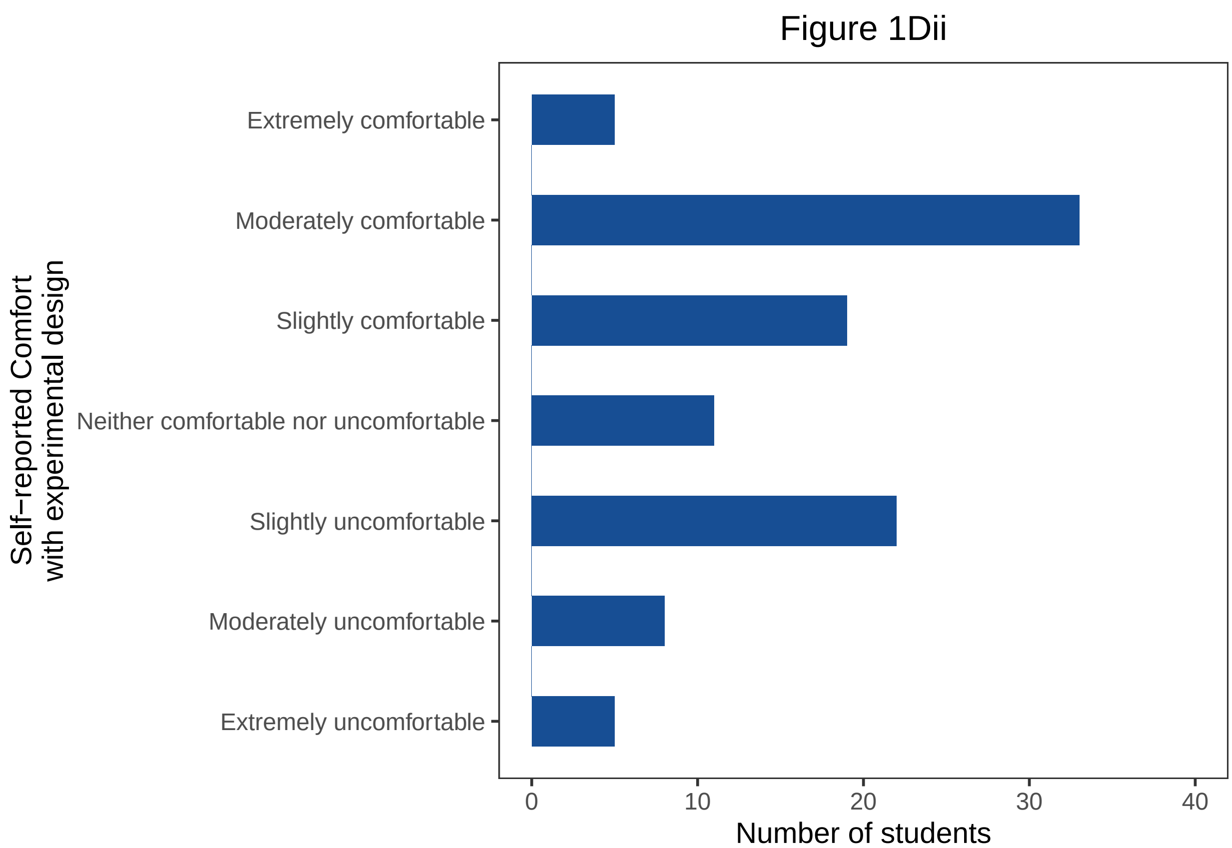


Figure 2

Students improved their performance of experimental design over the first semester of doctoral training.

Figure 2A

```
# Note that only 2017 data can be used in BEDCI calculations
data_2017 = data[data$Year == "2017",]
numStudents_2017 = nrow(data_2017) # Calculate number of students in cohort to be analyzed (2017)
# Print output
cat(paste("There are ", numStudents_2017, " student responses contained in the following BEDCI analyses
```

```
## There are 45 student responses contained in the following BEDCI analyses.
```

```
# Calculate linear regression
lm_bedci = lm(formula = Pre_CITotal ~ LabExp, data = data_2017)
summary(lm_bedci) # Adjusted R-squared:  0.018
```

```
##
## Call:
## lm(formula = Pre_CITotal ~ LabExp, data = data_2017)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -3.9738 -1.5019  0.2621  1.2621  3.0262
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.2660     0.6137  13.470   <2e-16 ***
## LabExp        0.2359     0.1755   1.344    0.186
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.878 on 43 degrees of freedom
## Multiple R-squared:  0.04032,    Adjusted R-squared:  0.018
## F-statistic: 1.806 on 1 and 43 DF,  p-value: 0.186
```

```
# Plot histograms
fig2a <- ggplot(data_2017, aes(x = (Pre_CITotal / 14) * 100, y = LabExp)) +
  geom_jitter(width = 0.25, height = 0.25, size = 2, color = "#174E94") +
  coord_flip() +
  scale_x_continuous("Total BEDCI score (%)", breaks=seq(40, 100, 20), limits=c(30, 90)) +
  scale_y_continuous("Previous years working in lab", breaks=c(1:7), limits=c(0.5,7.5), labels = c("< 2
  theme_bw()+
  theme(panel.grid.minor = element_blank()) +
  ggtitle(expression(paste("Figure 2A (", R^2," = ", 0.02, ")", sep=""))) + theme(plot.title = element_

ggsave(fig2a, filename="./Figures/Fig2A.svg", width = 4, height = 3)
plot(fig2a)
```

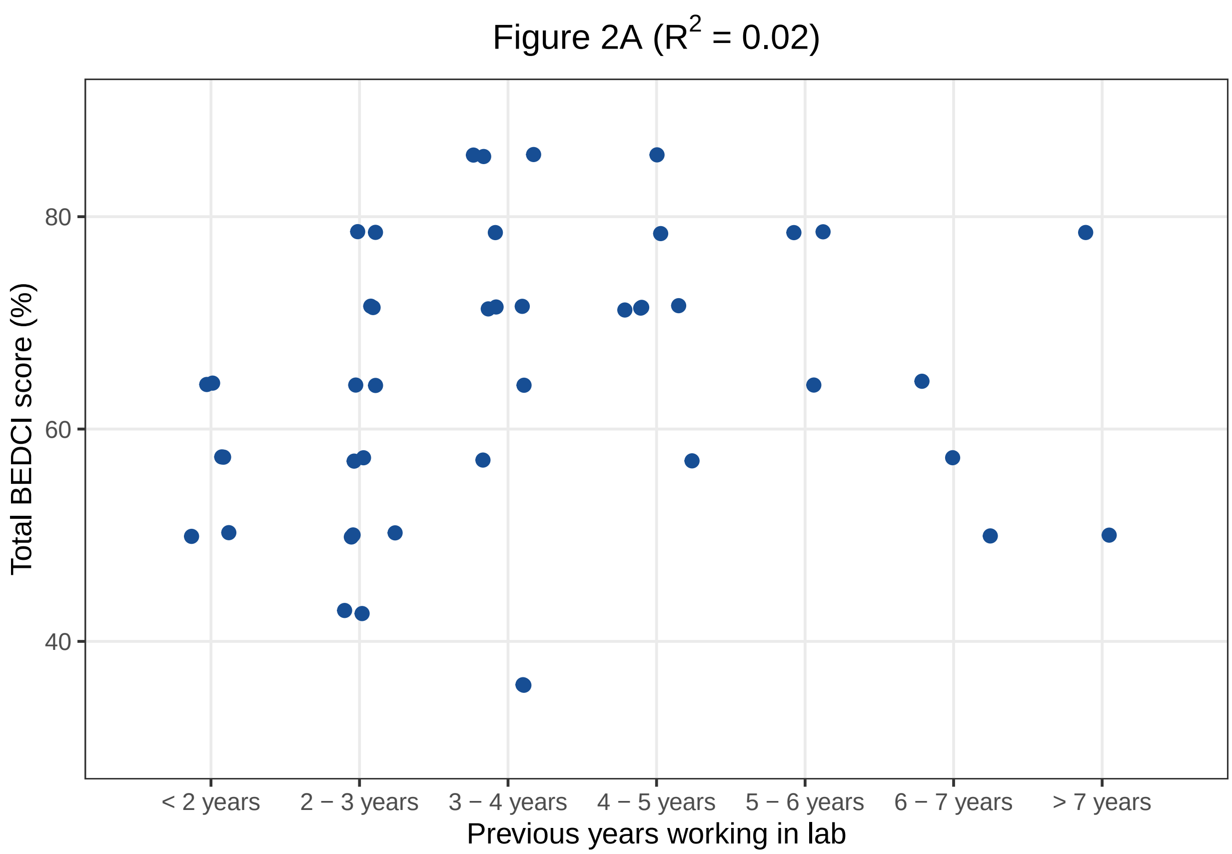


Figure 2B

```
# Calculate scores (in percent)
pre_score = data.frame(Score = (data_2017$Pre_CITotal / 14) * 100, Test = "Pre", stringsAsFactors = FALS
post_score = data.frame(Score = (data_2017$Post_CITotal / 14) * 100, Test = "Post", stringsAsFactors = F
BEDCI_score = rbind(pre_score, post_score)

# Perform wilcoxon signed-rank test
wTest = wilcox.test(data_2017$Pre_CITotal, data_2017$Post_CITotal, paired=TRUE, alternative = "two.sidec

# Determine significance level
sig1 = "is NOT a statistically significant difference"
sig2 = ""
if (wTest$p.value < 0.001) {
  sig1 = "IS a statistically significant difference"
  sig2 = "***"
} else if (wTest$p.value < 0.01) {
  sig1 = "IS a statistically significant difference"
  sig2 = "**"
} else if (wTest$p.value < 0.05) {
  sig1 = "IS a statistically significant difference"
  sig2 = "*"
}

# Print output
cat(paste("By a Wilcoxon signed-rank test, there ", sig1, " between student performance on the BEDCI con
```

```
## By a Wilcoxon signed-rank test, there IS a statistically significant difference between student perfo
```

```
# Order for plotting
BEDCI_score$Test <- factor(BEDCI_score$Test, levels = c('Pre','Post'), ordered = TRUE)

# Create violin plot
fig2b <- ggplot(BEDCI_score, aes(x = Test, y = Score)) +
  geom_violin(draw_quantiles = c(0.5), scale = "width", aes(fill = Test)) +
  geom_jitter(height = 3, width = 0.1) +
  scale_fill_manual(values = c("#174E94", "#EE7D31")) +
  scale_y_continuous("Total BEDCI score (%)", breaks = seq(40, 100, 20), labels = seq(40, 100, 20)) +
  scale_x_discrete(labels = c("Pre-test", "Post-test")) +
  theme_bw() +
  theme(panel.grid.minor = element_blank(), panel.grid.major = element_blank(), axis.title.x = element_
  ggtitle("Figure 2B") + theme(plot.title = element_text(hjust = 0.5))

ggsave(fig2b, filename="./Figures/Fig2B.svg", width = 2, height = 3)
plot(fig2b)
```

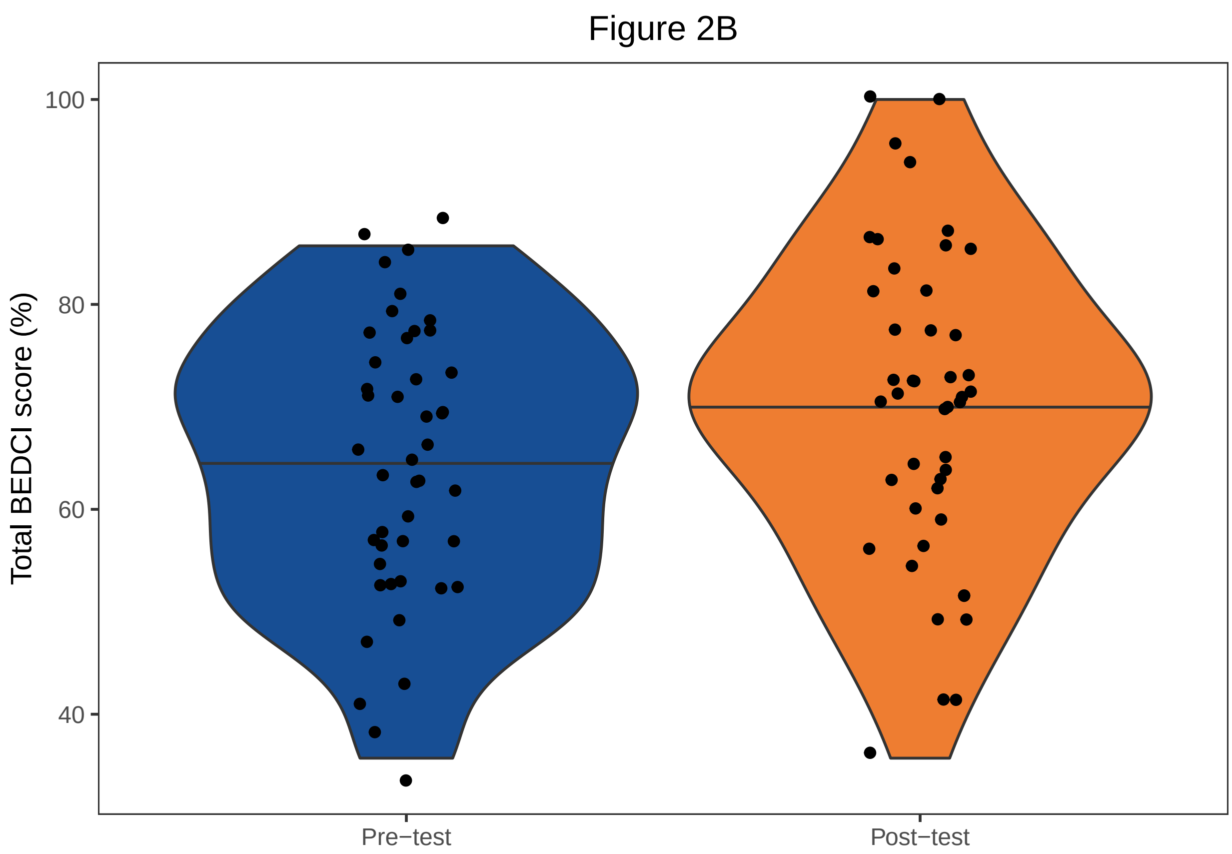


Figure 2C

```
# Function for calculating standard error of the mean
stderr_perc <- function(x, na.rm=FALSE) {
  if (na.rm) x <- na.omit(x)
  y = (x / numStudents_2017) * 100
  sqrt(var(y))
}

# Make data frame suitable for plotting
Q = c("Q1", "Q2", "Q3", "Q4", "Q5", "Q6", "Q7", "Q8", "Q9", "Q10", "Q11", "Q12", "Q13", "Q14") # Order
mCI_pre = data.frame(Question = Q, Test = "Pre", Score = (colSums(data_2017[,42:55]) / numStudents_2017)
mCI_post = data.frame(Question = Q, Test = "Post", Score = (colSums(data_2017[,57:70]) / numStudents_20

# Combine
mCI = rbind(mCI_pre, mCI_post)

# Function to perform McNemar's chi-squared test for each question on the BEDCI pre- and post-tests
mcNemarChiSquaredTest <- function(dat, Q) {

  # Get relevant data for each question
  pre = dat[,(Q+41)]
  post = dat[,(Q+56)]

  # Calculate values to populate matrix
  c_c = sum(pre + post == 2) # Correct, correct
```

```r
  i_c = sum(pre - post == -1) # Incorrect, correct
  c_i = sum(pre - post == 1) # Correct, incorrect
  i_i = sum(pre + post == 0) # Incorrect, incorrect

  # Create confusion matrix
  mat = matrix(c(c_c, i_c, c_i, i_i), nrow = 2, dimnames = list("Pre" = c("Correct", "Incorrect"), "Pos

  # Calculate significance
  mcnemar.test(mat)
}

# Calculate the p-value for each question and save to an output table
coreConcepts = c("Controls", "", "Hypotheses", "", "Biological variation", "", "Accuracy", "Extraneous :
questions = c(1, 5, 2, 9, 3, 10, 4, 6, 14, 7, 12, 8, 13, 11) # Questions ordered by subject
sigOut = data.frame(CoreConcept = coreConcepts, Question = questions, Chi = rep(0, 14), pValue = rep(0,
i = 1 # Iterate over every row of output
for (q in questions) {
  mnTest = mcNemarChiSquaredTest(data_2017, q)
  sigOut$Chi[i] = round(mnTest$statistic, 2)
  sigOut$pValue[i] = round(mnTest$p.value, 4)
  i = i + 1
}

pValues = p.adjust(sigOut$pValue, method = "fdr")
for (i in 1:14) {
  sigOut$pAdj[i] = round(pValues[i], 4)
  if (sigOut$pAdj[i] < 0.001) {
    sigOut$Significance[i] = "***"
  } else if (sigOut$pAdj[i] < 0.01) {
    sigOut$Significance[i] = "**"
  } else if (sigOut$pAdj[i] < 0.05) {
    sigOut$Significance[i] = "*"
  }
}

# Print signficance table
print(sigOut) # Also Table S3
```

```
##               CoreConcept Question  Chi pValue Significance   pAdj
## 1                Controls        1 4.00 0.0455              0.1592
## 2                                5 0.10 0.7518              1.0000
## 3              Hypotheses        2 5.88 0.0153              0.1592
## 4                                9 0.00 1.0000              1.0000
## 5    Biological variation        3 0.00 1.0000              1.0000
## 6                               10 4.00 0.0455              0.1592
## 7                Accuracy        4 4.92 0.0265              0.1592
## 8      Extraneous factors        6 0.12 0.7237              1.0000
## 9                               14 0.12 0.7237              1.0000
## 10   Independent sampling        7 0.00 1.0000              1.0000
## 11                              12 0.75 0.3865              1.0000
## 12        Random sampling        8 0.00 1.0000              1.0000
## 13                              13 0.00 1.0000              1.0000
## 14 Purpose of experiments       11 0.08 0.7728              1.0000
```

```
# Reorder for plotting
mCI$Question <- factor(mCI$Question, levels = c('Q1', 'Q5', 'Q2', 'Q9', 'Q3', 'Q10', 'Q4', 'Q6', 'Q14',
mCI$Test <- factor(mCI$Test, levels = c('Pre','Post'), ordered = TRUE)

# Create bar chart
fig2c <- ggplot(mCI, aes(x = Question, y = Score, fill = Test, group = Test)) +
  geom_bar(position = position_dodge(), stat = "identity") +
  geom_errorbar(aes(ymin=Score-SEM, ymax=Score+SEM), colour="black", width = 0.3, position = position_d
  scale_fill_manual("", values = c("#174E94", "#EE7D31"), labels = c("Pre-test", "Post-test")) +
  ylab("% of students providing expert-like answers") +
  theme_bw() +
  theme(panel.grid.minor = element_blank(), panel.grid.major.x = element_blank(), axis.title.x = elemen
  ggtitle("Figure 2C") + theme(plot.title = element_text(hjust = 0.5))

ggsave(fig2c, filename="./Figures/Figure2C.svg", width = 5, height = 3)
plot(fig2c)
```

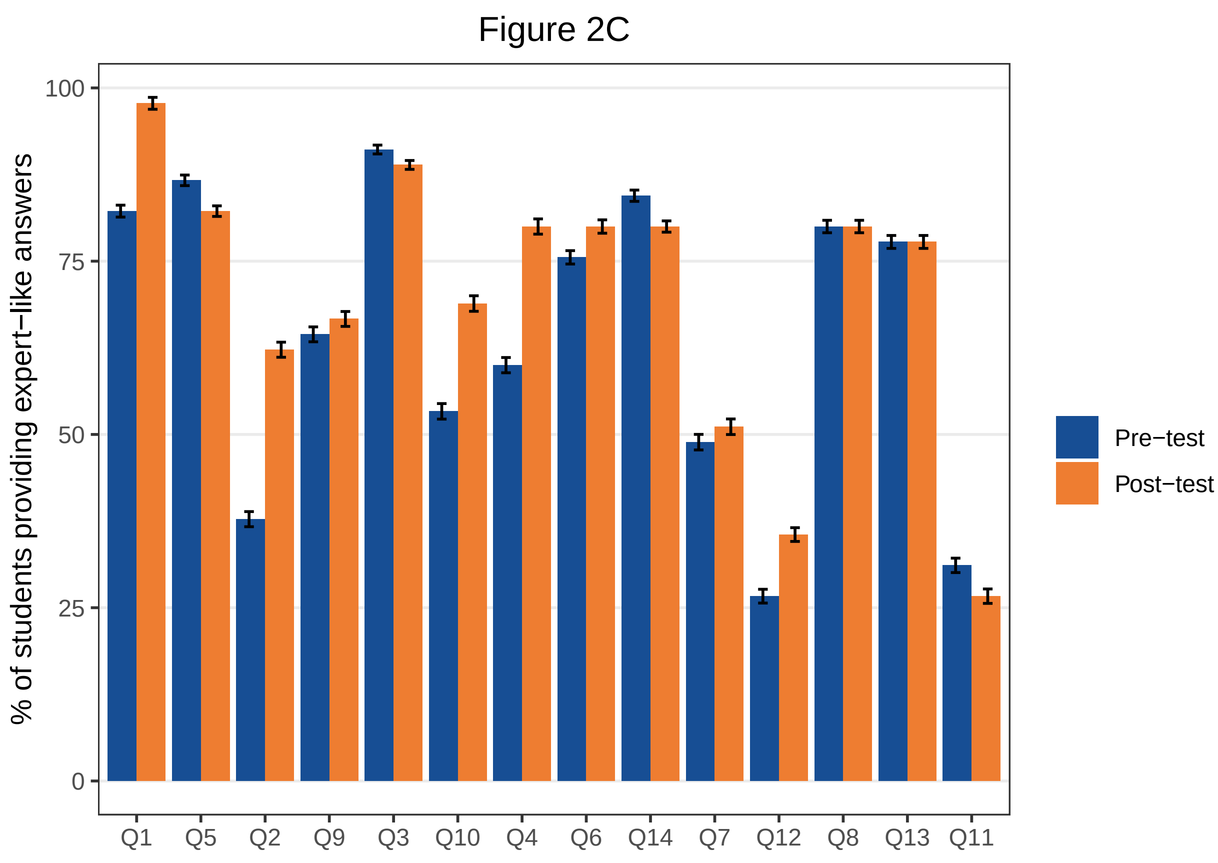


Figure 3

Students significantly improved in many aspects of research skills self-efficacy during their first semester of doctoral training.

Figure 3B

```r
# Test for significance
wilcoxonSignedRankTest <- function(dat, Q) {

# Get the data of interest for each question
  pre = dat[,Q+13]
  post = dat[,(Q+27)]

  # Perform wilcoxon signed-rank test for significance
  wilcox.test(pre, post, paired=TRUE, alternative = "two.sided", exact=FALSE)
}

# Calculate the p-value for each question and save to an output table
Items = c("Understand contemporary concepts in your field", "Make use of the primary scientific researc
sigOut = data.frame(Item = Items, Question = 1:14, V = rep(0, 14), pValue = rep(0, 14), Significance =
for (i in 1:14) {
  wTest = wilcoxonSignedRankTest(data, i)
  sigOut$V[i] = wTest$statistic
  sigOut$pValue[i] = round(wTest$p.value, 4)
}

wTest_Total = wilcox.test(melt(SE_pre)$value, melt(SE_post)$value, paired=TRUE, alternative = "two.side
```

```
## Using LabExp as id variables
## Using LabExp as id variables
```

```r
tmp = data.frame(Item = "Total", Question = 15, V = wTest_Total$statistic, pValue = round(wTest_Total$p
sigOut[15,] = tmp

pValues = p.adjust(sigOut$pValue, method = "fdr")
for (i in 1:15) {
  sigOut$pAdj[i] = round(pValues[i], 4)
  if (sigOut$pAdj[i] < 0.001) {
    sigOut$Significance[i] = "***"
  } else if (sigOut$pAdj[i] < 0.01) {
    sigOut$Significance[i] = "**"
  } else if (sigOut$pAdj[i] < 0.05) {
    sigOut$Significance[i] = "*"
  }
}

# Print signficance table
print(sigOut) # This is also Table S2
```

```
##                                                                                               Item
## 1                                                   Understand contemporary concepts in your field
## 2  Make use of the primary scientific research literature in your field (e.g., journal articles)
## 3              Identify a specific question for investigation based on the research in your field
## 4                                             Formulate a research hypothesis based on a specific question
## 5                                               Design an experiment or theoretical test of the hypothesis
## 6                                                   Understand the importance of 'controls' in research
## 7                                                                             Observe and collect data
## 8                                                                           Statistically analyze data
```

```
## 9                                  Interpret data by relating results to the original hypothesis
## 10                                 Reformulate your original research hypothesis (as appropriate)
## 11                                      Relate your results to the 'bigger picture' in your field
## 12                                            Orally communicate the results of research projects
## 13                                                         Write a research paper for publication
## 14                                                                          Think independently
## V                                                                                         Total
##    Question       V pValue Significance   pAdj
## 1         1   524.5 0.0071            * 0.0118
## 2         2   488.0 0.0784              0.0905
## 3         3   677.0 0.0164            * 0.0224
## 4         4   429.5 0.0000          *** 0.0000
## 5         5   339.5 0.0000          *** 0.0000
## 6         6   389.0 0.0033           ** 0.0062
## 7         7   607.5 0.3072              0.3072
## 8         8   498.0 0.0004           ** 0.0012
## 9         9   385.5 0.0028           ** 0.0060
## 10       10   286.0 0.0000          *** 0.0000
## 11       11   506.0 0.1198              0.1284
## 12       12   413.0 0.0024           ** 0.0060
## 13       13   411.0 0.0208            * 0.0260
## 14       14   619.0 0.0114            * 0.0171
## V        15 90521.0 0.0000          *** 0.0000
```

```r
# Combine melted self-efficacy data frames
mSE = rbind(mSE_pre, mSE_post)

# Re-order for plotting
mSE$Test = factor(mSE$Test, levels = c("Pre", "Post"))
mSE$Question = factor(mSE$Question, levels = c("Q1", "Q2", "Q3", "Q4", "Q5", "Q14", "Q6", "Q7", "Q8", "(
mSE$value = factor(mSE$value, levels = c("a1", "b1", "c1", "d1", "e1", "a2", "b2", "c2", "d2", "e2"))

# Stacked bar chart
fig3bi <- ggplot(mSE, aes(Test)) +
  geom_bar(aes(fill=value), position = "fill") + facet_grid(~ Question) +
  theme_bw() +
  scale_fill_manual("Research Skills\nSelf-Efficacy",
                    values = c("a1" = "#DAE3F3", "a2" = "#FBE5D6", "b1" = "#B4C7E7", "b2" = "#F8CBAD",
  guides(fill=guide_legend(ncol = 2)) +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank()) +
  theme(axis.title=element_blank(), axis.text=element_blank(), axis.ticks=element_blank(), strip.backgro
  ggtitle("Figure 3Bi") +  theme(plot.title = element_text(hjust = 0.5))

ggsave(fig3bi, filename="./Figures/Figure3Bi.svg", width = 5, height = 3)
plot(fig3bi)
```

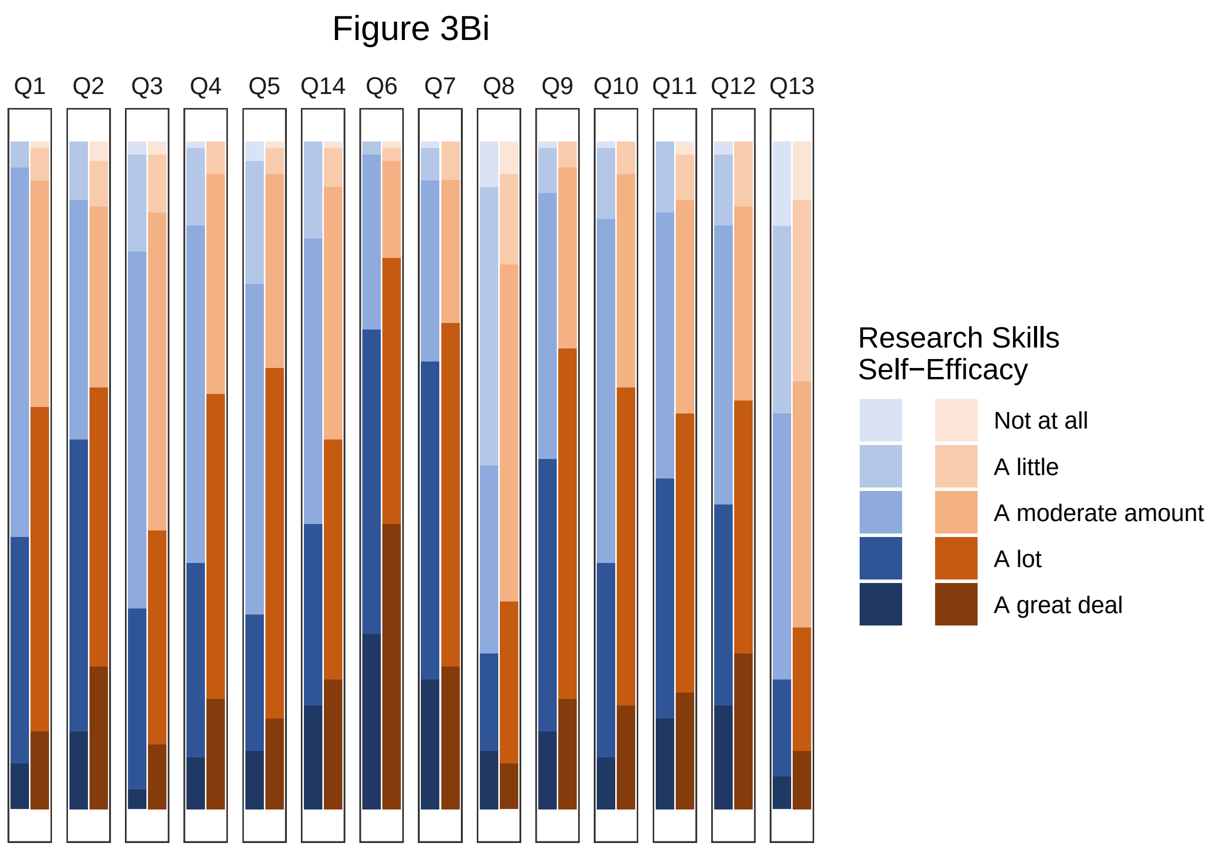


Figure 3Bii

```r
# Get total self-efficacy calculation
mSE_total = data.frame(Test = rep(c("Pre", "Post"), each = 5), Lab = row.names(table(mSE$value)), Freq =
mSE_total$Test <- factor(mSE_total$Test, levels = c("Pre", "Post"), ordered = TRUE)
mSE_total$Lab = factor(mSE_total$Lab, levels = c("a1", "b1", "c1", "d1", "e1", "a2", "b2", "c2", "d2",

# Stacked bar chart
fig3bii <- ggplot(mSE_total, aes(x=Test, y=Freq)) +
  geom_bar(aes(fill=Lab), stat = "identity", position = "fill") +
  theme_bw() +
scale_fill_manual("Research Skills\nSelf-Efficacy",
                  values = c("a1" = "#DAE3F3", "a2" = "#FBE5D6", "b1" = "#B4C7E7", "b2" = "#F8CBAD",
  guides(fill=guide_legend(ncol = 2)) +
  xlab("Total") +
  theme(panel.grid.minor = element_blank(), panel.grid.major = element_blank(), axis.text=element_blank
  ggtitle("Figure 3Bii") +  theme(plot.title = element_text(hjust = 0.5))

ggsave(fig3bii, filename="./Figures/Figure3Bii.svg", width = 1, height = 3)
plot(fig3bii)
```

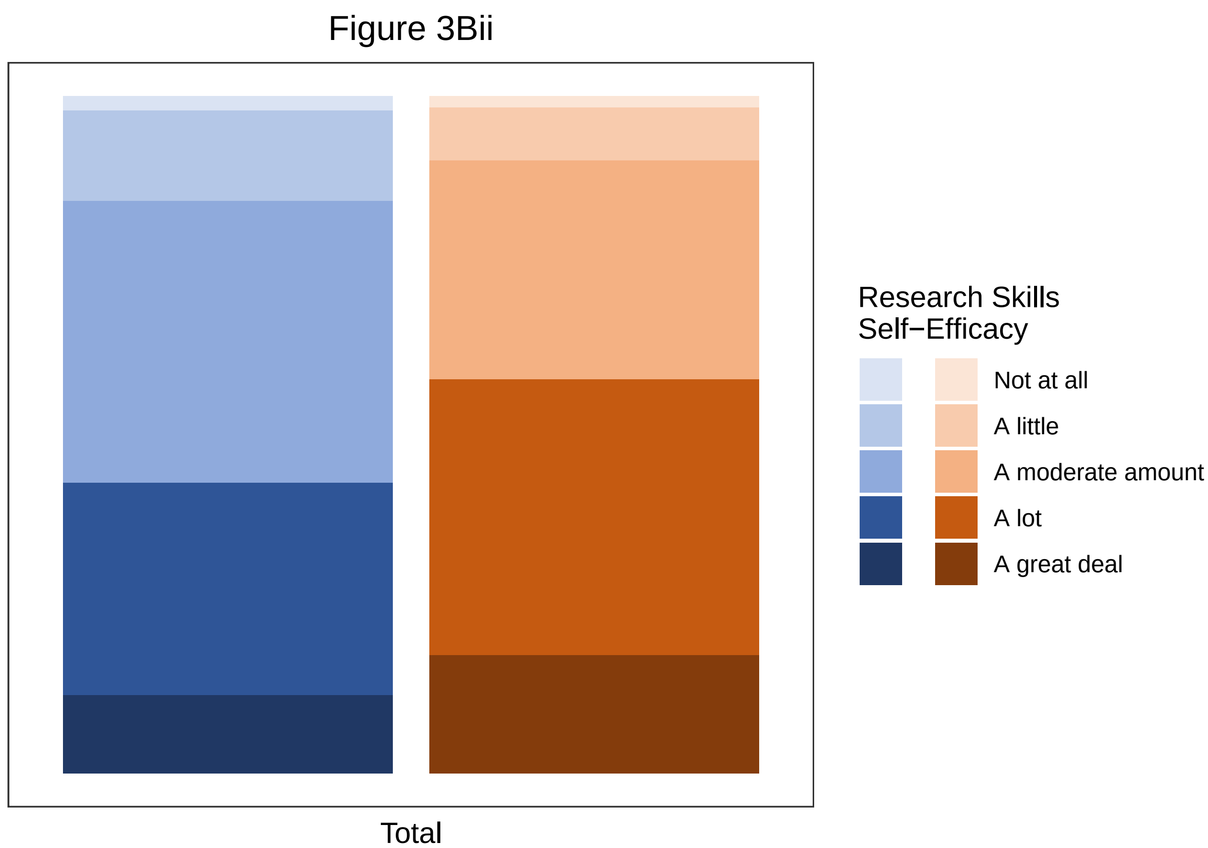


Figure 3C

```
SE_pre$Sex = as.character(data$Sex)
SE_post$Sex = as.character(data$Sex)

# Remove NA Sex
SE_pre = SE_pre[!is.na(SE_pre$Sex),]
SE_post = SE_post[!is.na(SE_post$Sex),]

# Remove lab experience columns
SE_pre = SE_pre[,-15]
SE_post = SE_post[,-15]

male_SE_pre = SE_pre[SE_pre$Sex == "Male",]
female_SE_pre = SE_pre[SE_pre$Sex == "Female",]
male_SE_post = SE_post[SE_post$Sex == "Male",]
female_SE_post = SE_post[SE_post$Sex == "Female",]

# Melt responses
male_mSE_pre = melt(male_SE_pre)
```

```
## Using Sex as id variables
```

```
male_mSE_post = melt(male_SE_post)
```

```
## Using Sex as id variables
```

```
female_mSE_pre = melt(female_SE_pre)
```

```
## Using Sex as id variables
```

```
female_mSE_post = melt(female_SE_post)
```

```
## Using Sex as id variables
```

```
nMale = nrow(male_SE_pre)
nFemale = nrow(female_SE_pre)

# Add question number label to each pre / post array
male_mSE_pre$Question = rep(c("Q1", "Q2", "Q3", "Q4", "Q5", "Q6", "Q7", "Q8", "Q9", "Q10", "Q11", "Q12"
male_mSE_pre$Test = "Pre"
male_mSE_post$Question = rep(c("Q1", "Q2", "Q3", "Q4", "Q5", "Q6", "Q7", "Q8", "Q9", "Q10", "Q11", "Q12
male_mSE_post$Test = "Post"

female_mSE_pre$Question = rep(c("Q1", "Q2", "Q3", "Q4", "Q5", "Q6", "Q7", "Q8", "Q9", "Q10", "Q11", "Q1
female_mSE_pre$Test = "Pre"
female_mSE_post$Question = rep(c("Q1", "Q2", "Q3", "Q4", "Q5", "Q6", "Q7", "Q8", "Q9", "Q10", "Q11", "Q
female_mSE_post$Test = "Post"

# Change labels to strings for categorical plotting
male_mSE_pre[male_mSE_pre$value==1,"value"] = "m_a1" # Not at all, pre-test
male_mSE_pre[male_mSE_pre$value==2,"value"] = "m_b1" # A little, pre-test
male_mSE_pre[male_mSE_pre$value==3,"value"] = "m_c1" # A moderate amount, pre-test
male_mSE_pre[male_mSE_pre$value==4,"value"] = "m_d1" # A lot, pre-test
male_mSE_pre[male_mSE_pre$value==5,"value"] = "m_e1" # A great deal, pre-test

male_mSE_post[male_mSE_post$value==1,"value"] = "m_a2" # Not at all, post-test
male_mSE_post[male_mSE_post$value==2,"value"] = "m_b2" # A little, post-test
male_mSE_post[male_mSE_post$value==3,"value"] = "m_c2" # A moderate amount, post-test
male_mSE_post[male_mSE_post$value==4,"value"] = "m_d2" # A lot, post-test
male_mSE_post[male_mSE_post$value==5,"value"] = "m_e2" # A great deal, post-test

# Change labels to strings for categorical plotting
female_mSE_pre[female_mSE_pre$value==1,"value"] = "f_a1" # Not at all, pre-test
female_mSE_pre[female_mSE_pre$value==2,"value"] = "f_b1" # A little, pre-test
female_mSE_pre[female_mSE_pre$value==3,"value"] = "f_c1" # A moderate amount, pre-test
female_mSE_pre[female_mSE_pre$value==4,"value"] = "f_d1" # A lot, pre-test
female_mSE_pre[female_mSE_pre$value==5,"value"] = "f_e1" # A great deal, pre-test

female_mSE_post[female_mSE_post$value==1,"value"] = "f_a2" # Not at all, post-test
female_mSE_post[female_mSE_post$value==2,"value"] = "f_b2" # A little, post-test
female_mSE_post[female_mSE_post$value==3,"value"] = "f_c2" # A moderate amount, post-test
female_mSE_post[female_mSE_post$value==4,"value"] = "f_d2" # A lot, post-test
female_mSE_post[female_mSE_post$value==5,"value"] = "f_e2" # A great deal, post-test

# Combine pre- and post-test dataframes for plotting
mSE <- rbind(male_mSE_pre, male_mSE_post, female_mSE_pre, female_mSE_post)
head(mSE)
```

```
##    Sex variable value Question Test
## 1 Male   Pre_Q1  m_d1       Q1  Pre
## 2 Male   Pre_Q1  m_c1       Q1  Pre
## 3 Male   Pre_Q1  m_c1       Q1  Pre
## 4 Male   Pre_Q1  m_c1       Q1  Pre
## 5 Male   Pre_Q1  m_d1       Q1  Pre
## 6 Male   Pre_Q1  m_c1       Q1  Pre
```

```
mSE_total = data.frame(Sex = rep(c("Female", "Male"), each = 10), Test = rep(c("Pre", "Post"), 5), Lab =
mSE_total$Test <- factor(mSE_total$Test, levels = c("Pre", "Post"), ordered = TRUE)

# Stacked bar chart
tests.labs = c("Pre-test", "Post-test")
names(tests.labs) = c("Pre", "Post")

# Stacked bar chart
fig3c <- ggplot(mSE_total, aes(x=Sex, y=Freq)) +
  geom_bar(aes(fill=Lab), stat = "identity", position = "fill") +
  theme_bw() +
  facet_grid(~ Test, labeller = labeller(Test = tests.labs)) +
  scale_fill_manual("legend", values = c("f_a1" = "#DAE3F3", "f_b1" = "#B4C7E7", "f_c1" = "#8FAADC", "f_
  scale_x_discrete(labels = c("Women", "Men", "Women", "Men")) +
  theme(panel.grid.minor = element_blank(), panel.grid.major = element_blank(), axis.title=element_blank
  theme(strip.background = element_blank(), panel.border = element_rect(colour = "white"), legend.positi
  ggtitle("Figure 3C") + theme(plot.title = element_text(hjust = 0.5))

ggsave(fig3c, filename="./Figures/Figure3C.svg", width = 4, height = 3)
plot(fig3c)
```

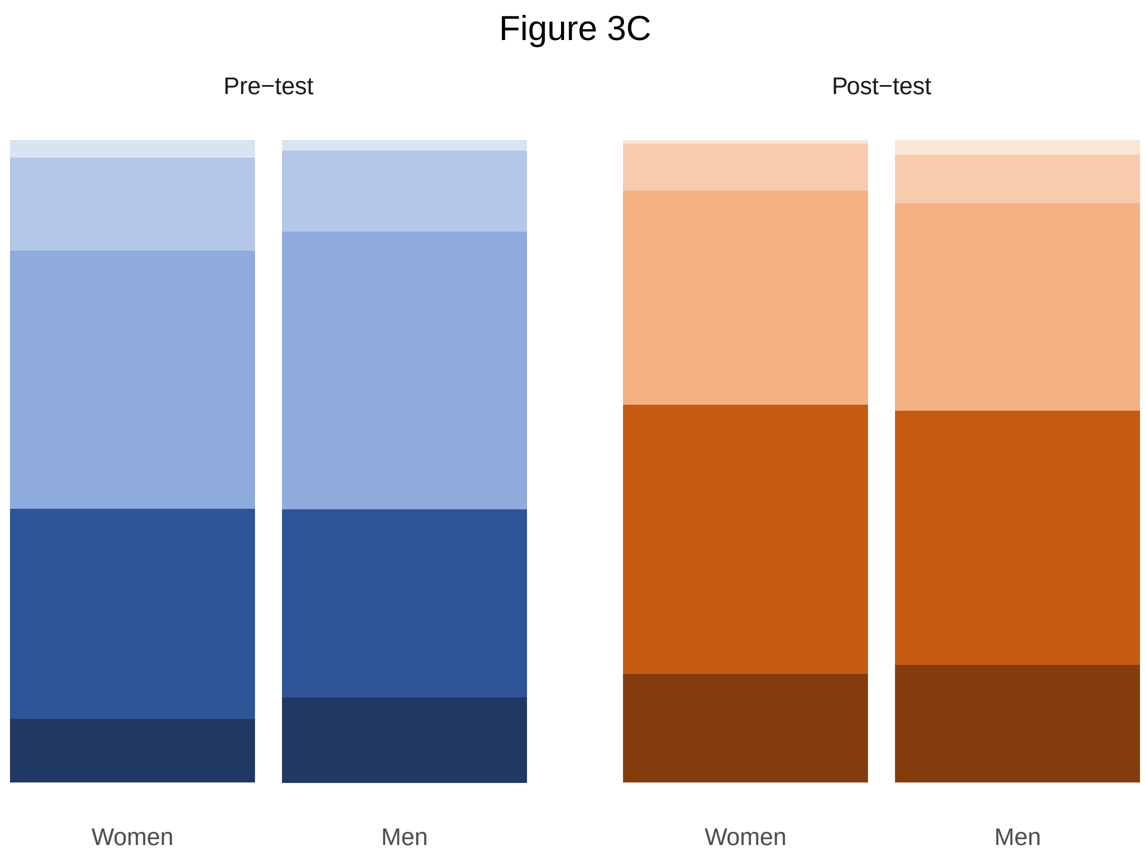


Figure 4

Students with a net increase in research skills self-efficacy select different factors as the most important for contributing to their experience or comfort with experimental design.

Figure 4A

```
data2$ExpPercent = round((data2$ExperienceFreq / sum(data2$ExperienceFreq)) * 100, 2)
data2$Category <- factor(data2$Category, levels = c("Sem", "CC", "Discussing", "Present", "Lab", "Read"

fig4a <- ggplot(data2, aes(x="", y=ExpPercent, fill=Category))+
  geom_bar(width = 1, stat = "identity") +
  coord_polar("y", start=0) +
  scale_fill_manual("Most important factor", values = c("Sem" = "#BEBEBE", "CC" = "#D64051", "Discussin
                                            paste("Completing coursework (", round(data2$ExpPercent[2
                                            paste("Discussing scientific topics with colleagues (", r
                                            paste("Giving scientific presentations (", round(data2$Ex
                                            paste("Participating in laboratory research (", round(dat
                                            paste("Reading scientific literature (", round(data2$ExpP
                                            paste("Recieving advice from mentors (", round(data2$ExpP
                                            paste("Writing a project proposal (", round(data2$ExpPerc
  theme_minimal() +
  theme(axis.title = element_blank(), panel.border = element_blank(), panel.grid=element_blank(), axis.
  ggtitle("Figure 4A") + theme(plot.title = element_text(hjust = 0.5))
```

```r
ggsave(fig4a, filename="./Figures/Figure4A.svg", width = 3, height = 3)
```

```
## Warning: Removed 10 rows containing missing values (position_stack).
```

```r
plot(fig4a)
```

```
## Warning: Removed 10 rows containing missing values (position_stack).
```

Figure 4A

Figure 4B

```r
data2$ComfPercent = round((data2$ComfortFreq / sum(data2$ComfortFreq)) * 100, 2)

fig4b <- ggplot(data2, aes(x="", y=ComfPercent, fill=Category))+
  geom_bar(width = 1, stat = "identity") +
  coord_polar("y", start=0) +
  scale_fill_manual("Most important factor", values = c("Sem" = "#BEBEBE", "CC" = "#D64051", "Discussin
                                                      paste("Completing coursework (", round(data2$ComfPercent[
                                                      paste("Discussing scientific topics with colleagues (", r
                                                      paste("Giving scientific presentations (", round(data2$Co
                                                      paste("Participating in laboratory research (", round(dat
                                                      paste("Reading scientific literature (", round(data2$Comf
                                                      paste("Recieving advice from mentors (", round(data2$Comf
                                                      paste("Writing a project proposal (", round(data2$ComfPer
  theme_minimal() +
```

```
  theme(axis.title = element_blank(), panel.border = element_blank(), panel.grid=element_blank(), axis.t
  ggtitle("Figure 4B") + theme(plot.title = element_text(hjust = 0.5))

ggsave(fig4b, filename="./Figures/Figure4B.svg", width = 3, height = 3)
```

```
## Warning: Removed 10 rows containing missing values (position_stack).
```

```
plot(fig4b)
```

```
## Warning: Removed 10 rows containing missing values (position_stack).
```

Figure 4B

Figure S1

65% of students had a net increase in performance of experimental design over the semester.

Figure S1A

```
# Calculate total change in BEDCI score (post - pre)
CI_delta = data.frame(Score = (data_2017$Post_CITotal - data_2017$Pre_CITotal), Gender = data_2017$Sex,

# Add color labels
CI_delta$Col = "orange"
```

```
CI_delta[CI_delta$Score < 0, 'Col'] ="red"
CI_delta[CI_delta$Score > 0, 'Col'] = "yellow"

cat(paste(round(sum((CI_delta$Score > 0) / numStudents_2017) * 100, 0), "% of students had a net increa
```

```
## 64% of students had a net increase in performance of experimental design over the semester.
```

```
# Create histogram
figS1a = ggplot(CI_delta, aes(x = Score, fill = Col)) +
  geom_histogram(binwidth = 0.5) +
  scale_fill_manual("Net change in\nBEDCI score", values = c("yellow" = "#FDCB41", "orange" = "#F37627"
                    labels = c("Net decrease in concept inventory score", "No change in concept invento
  scale_x_continuous(breaks=seq(-4, 4, 1)) +
  theme_bw() +
  xlab("Change in BEDCI score") + ylab("Number of students") +
  theme(panel.background=element_blank(), panel.grid.minor=element_blank(), plot.background=element_bla
  ggtitle("Figure S1A") + theme(plot.title = element_text(hjust = 0.5))

# Plot figure
ggsave(figS1a, filename="./Figures/FigureS1A.svg", width = 6, height = 4)
plot(figS1a)
```

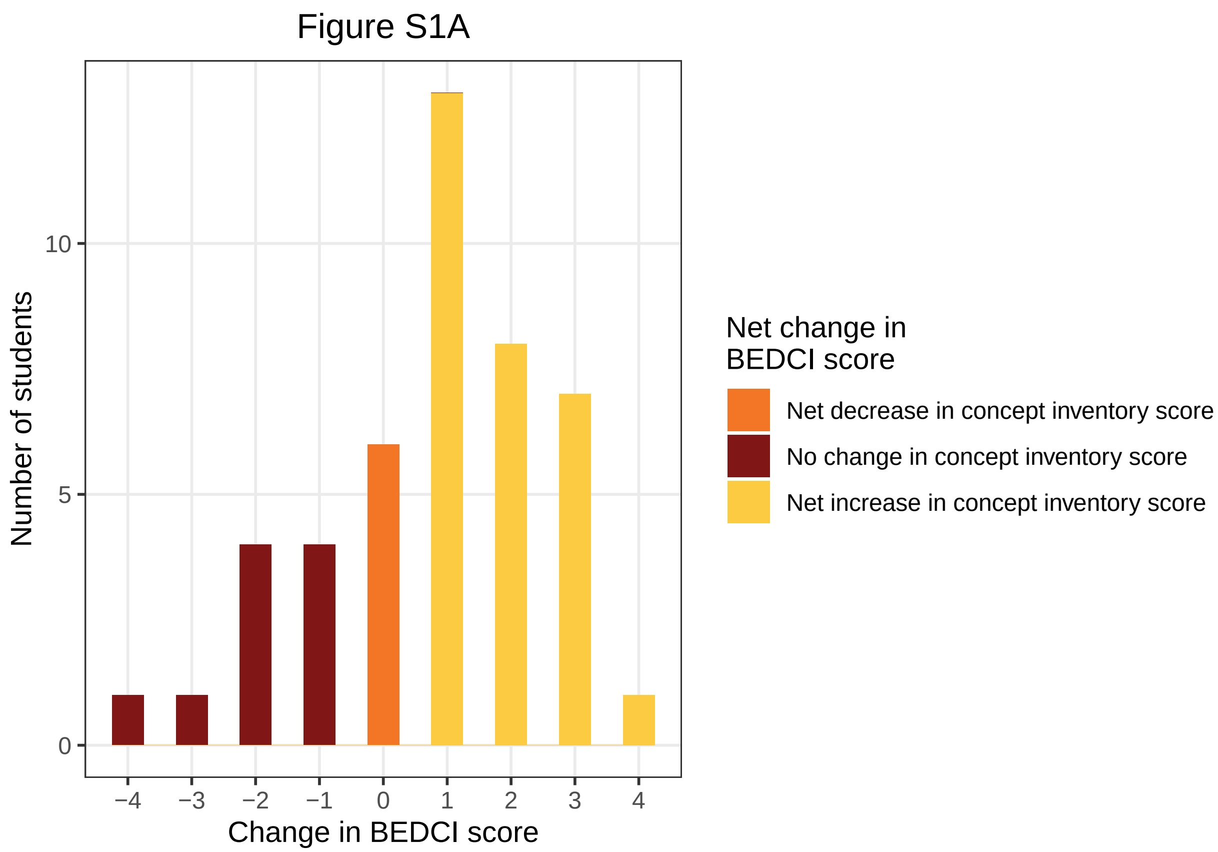


Figure S1Bi

```
negBEDCI_gender = data.frame(Var1 = c("Female", "Male"), Freq = c(nrow(CI_delta[CI_delta$Score < 0 & CI_

figS1bi <- ggplot(negBEDCI_gender, aes(x="", y=Freq, fill = Var1))+
  geom_bar(width = 1, stat = "identity") +
  coord_polar("y", start=0) +
  scale_fill_manual("Gender of students with\nnet negative BEDCI scores", values = c("magenta4", "green
                    labels = c(paste("Women (", round((negBEDCI_gender$Freq[1] / sum(negBEDCI_gender$Fr
                               paste("Men (", round((negBEDCI_gender$Freq[2] / sum(negBEDCI_gender$Freq
  theme_minimal() +
  theme(axis.title = element_blank(), panel.border = element_blank(), panel.grid=element_blank(), axis.
  ggtitle("Figure S1Bi") + theme(plot.title = element_text(hjust = 0.5))

ggsave(figS1bi, filename="./Figures/FigureS1Bi.svg", width = 3, height = 3)
plot(figS1bi)
```

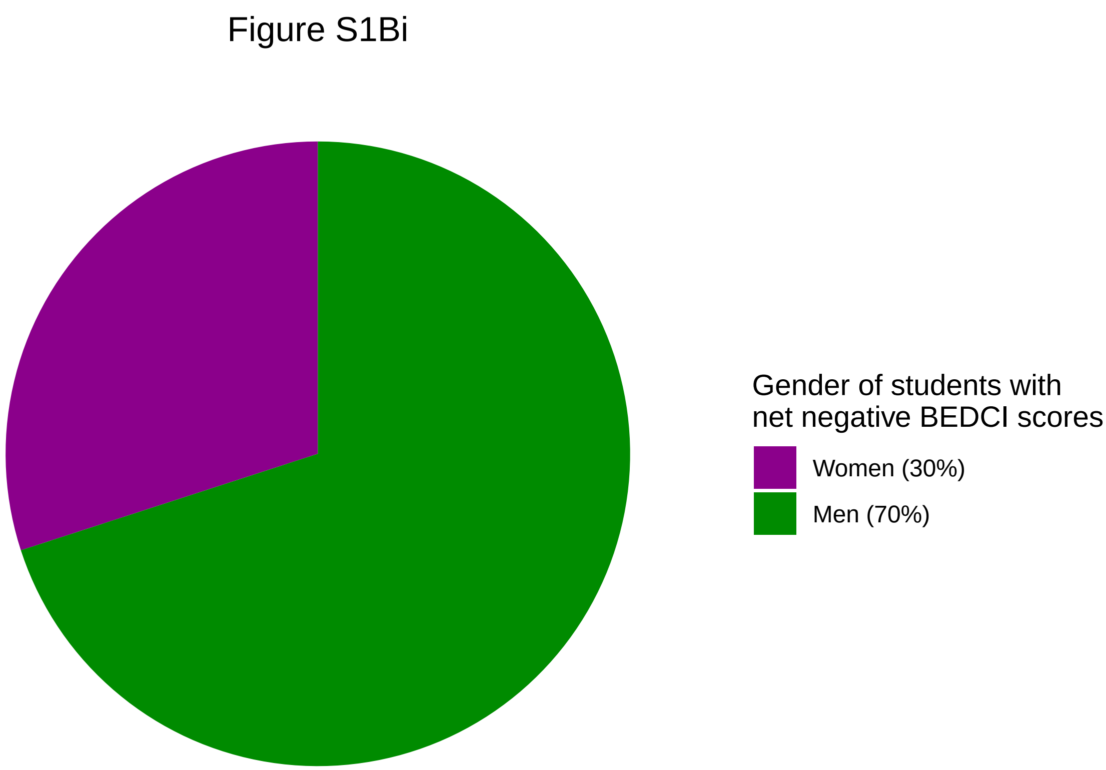


Figure S1Bii

```
negBEDCI_labExp = data.frame(Var1 = as.character(1:7), Freq = c(nrow(CI_delta[CI_delta$Score < 0 & CI_d
negBEDCI_labExp$Var1 <- factor(negBEDCI_labExp$Var1, levels = as.character(1:7))

figS1bii <- ggplot(negBEDCI_labExp, aes(x="", y=Freq, fill = Var1))+
  geom_bar(width = 1, stat = "identity") +
  coord_polar("y", start=0, direction = -1) +
  scale_fill_manual("Lab experience of students with\nnet negative BEDCI scores", values=c("7" = "#3333
                    labels = c(paste("< 2 years (", round((negBEDCI_labExp$Freq[1] / sum(negBEDCI_labEx
                               paste("2 - 3 years (", round((negBEDCI_labExp$Freq[2] / sum(negBEDCI_lab
```

```
                              paste("3 - 4 years (", round((negBEDCI_labExp$Freq[3] / sum(negBEDCI_lab
                              paste("4 - 5 years (", round((negBEDCI_labExp$Freq[4] / sum(negBEDCI_lab
                              paste("5 - 6 years (", round((negBEDCI_labExp$Freq[5] / sum(negBEDCI_lab
                              paste("6 - 7 years (", round((negBEDCI_labExp$Freq[6] / sum(negBEDCI_lab
                              paste("> 7 years (", round((negBEDCI_labExp$Freq[7] / sum(negBEDCI_labEx
  theme_minimal() +
  theme(axis.title = element_blank(), panel.border = element_blank(), panel.grid=element_blank(), axis.
  ggtitle("Figure S1Bii") + theme(plot.title = element_text(hjust = 0.5))

ggsave(figS1bii, filename="./Figures/FigureS1Bii.svg", width = 3, height = 3)
plot(figS1bii)
```

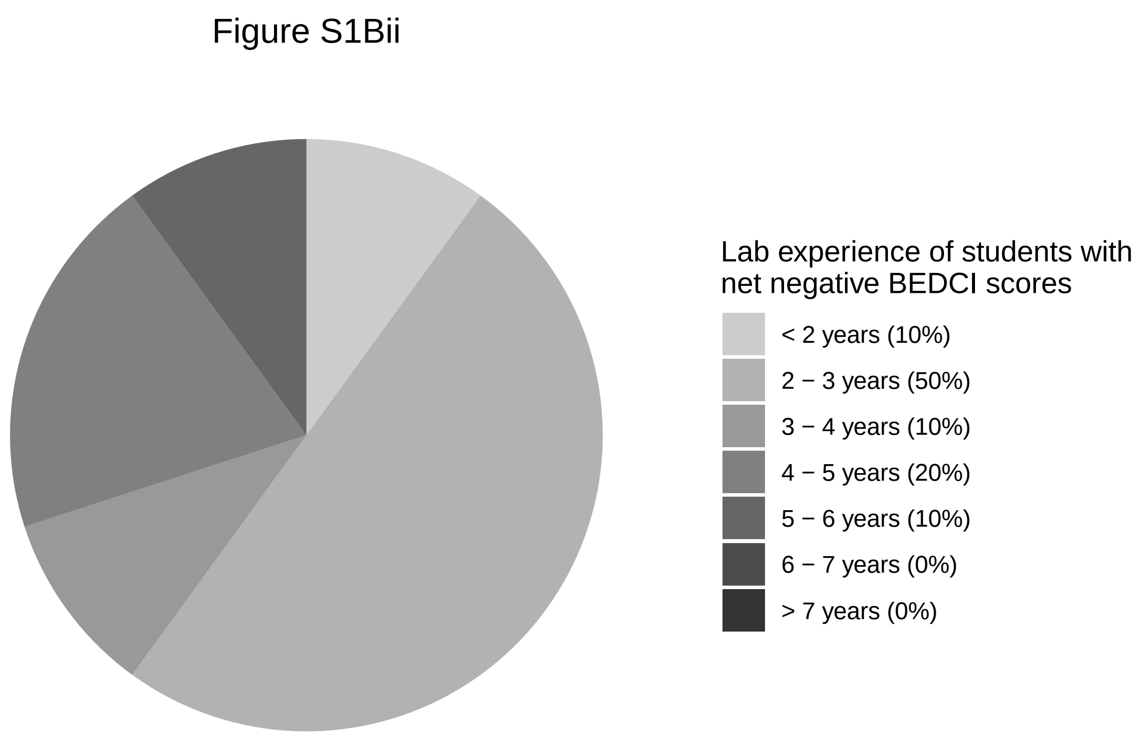


Figure S2

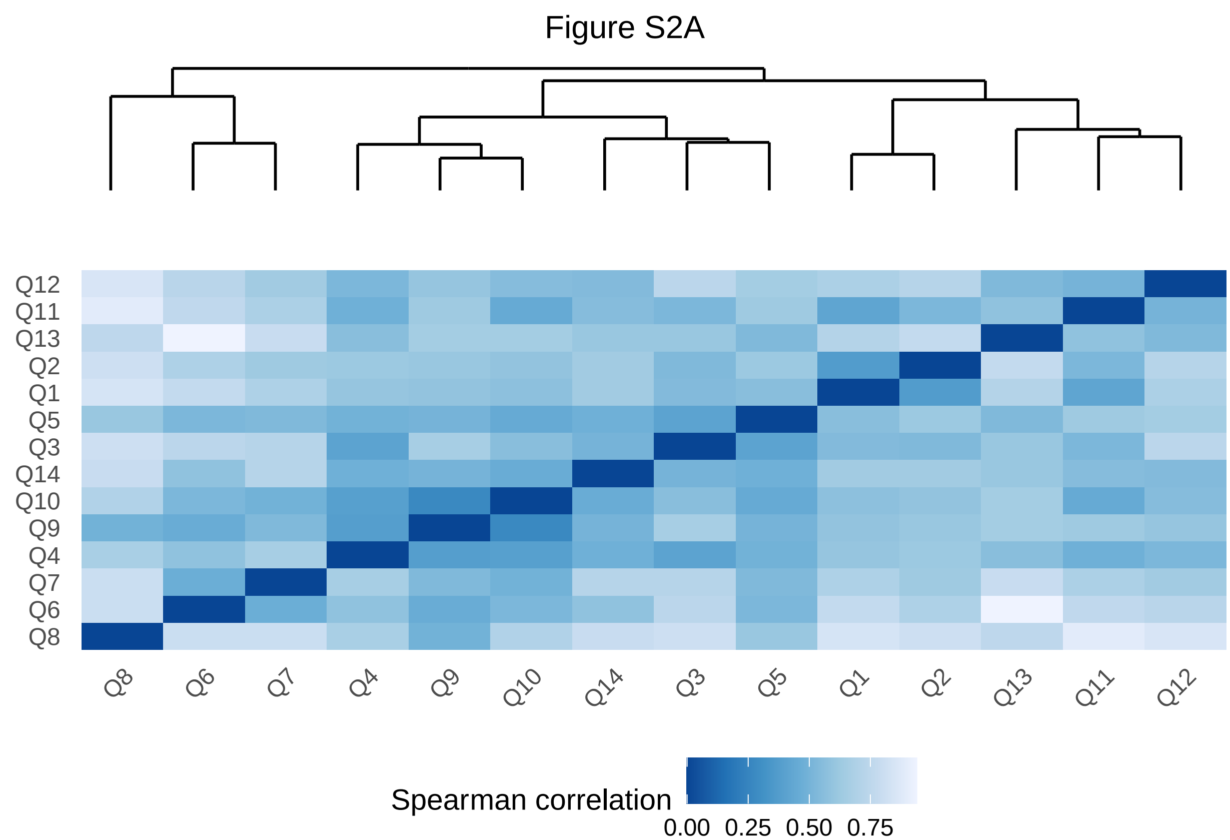
Correlation generated sever topical groups of research skills self-efficacy questions.

Figure S2A

```
# Split data by responses to self-efficacy questionaire before and after the first semester of graduate
SE_pre = data[,14:27]

# Melt responses
mSE_pre = melt(SE_pre)
```

```
## No id variables; using all as measure variables
```

```
# Add question number label to each pre / post array
Questions = c("Q1", "Q2", "Q3", "Q4", "Q5", "Q6", "Q7", "Q8", "Q9", "Q10", "Q11", "Q12", "Q13", "Q14")
mSE_pre$Question = rep(Questions, each = numStudents)

# Make data frame for clustering
out = data.frame(matrix(0, nrow = length(Questions), ncol = length(Questions)), row.names = Questions, s
colnames(out) = Questions

i = 1
for (Qi in Questions) {
  j = 1
  for (Qj in Questions) {
    out[i, j] = round(cor(mSE_pre[mSE_pre$Question == Qi, "value"], mSE_pre[mSE_pre$Question == Qj, "val
    j = j + 1
  }
  i = i + 1
}

# Cluster
row.order <- hclust(dist(out, method = "euclidean"), method="ward.D")$order
col.order <- hclust(dist(t(out), method = "euclidean"), method="ward.D")$order
out_new <- out[row.order, col.order]
m_out <- melt(as.matrix(out_new))
names(m_out)[c(1:2)] <- c("Qx", "Qy")

# Dendrogram
dd.row <- as.dendrogram(hclust(dist(t(out), method = "euclidean"), method="ward.D"))
dx <- dendro_data(dd.row)

# Helper function for creating dendograms
ggdend <- function(df) {
  ggplot() +
    geom_segment(data = df, aes(x=x, y=y, xend=xend, yend=yend)) +
    theme_bw() +
    labs(x = "", y = "") + theme_minimal() +
    theme(axis.text = element_blank(), axis.ticks = element_blank(), panel.grid = element_blank())
}

# x/y dendograms
figS2a_dendrogramX = ggdend(dx$segments)

figS2a_heatmap = ggplot(data = m_out, aes(x = Qx, y = Qy, fill = 1 - value)) +
  geom_raster() +
  theme_bw() +
  scale_fill_distiller("Spearman correlation", palette = "Blues",) + # pre = blues, post = oranges, del
  theme(axis.text.x = element_text(angle = 45, hjust = 1), axis.title = element_blank(), axis.ticks =ele
  theme(panel.border = element_blank(), panel.grid.major = element_blank(), panel.grid.minor = element_

figS2a = grid.arrange(figS2a_dendrogramX, figS2a_heatmap, ncol = 1, heights = c(1, 3), top ="Figure S2A
```




```
ggsave(figS2a, filename="./Figures/FigureS2A.svg", width = 6, height = 4)
```

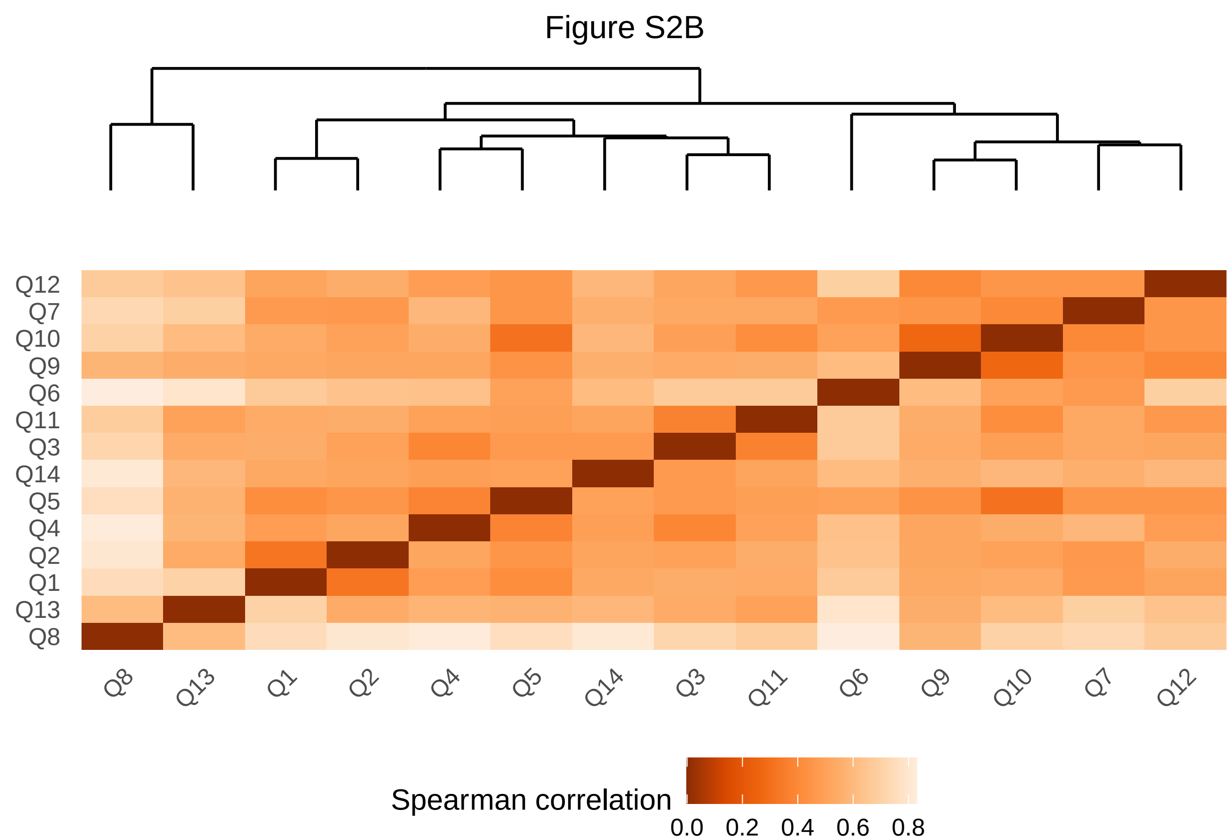
Figure S2B

```
# Split data by responses to self-efficacy questionaire before and after the first semester of graduate
SE_post = data[,28:41]

# Melt responses
mSE_post = melt(SE_post)
```

```
## No id variables; using all as measure variables
```

```
# Add question number label to each pre / post array
Questions = c("Q1", "Q2", "Q3", "Q4", "Q5", "Q6", "Q7", "Q8", "Q9", "Q10", "Q11", "Q12", "Q13", "Q14")
mSE_post$Question = rep(Questions, each = numStudents)

# Make data frame for clustering
out = data.frame(matrix(0, nrow = length(Questions), ncol = length(Questions)), row.names = Questions, s
colnames(out) = Questions

i = 1
for (Qi in Questions) {
  j = 1
  for (Qj in Questions) {
```

```
    out[i, j] = round(cor(mSE_post[mSE_post$Question == Qi, "value"], mSE_post[mSE_post$Question == Qj,
    j = j + 1
  }
  i = i + 1
}

# Cluster
row.order <- hclust(dist(out, method = "euclidean"), method="ward.D")$order
col.order <- hclust(dist(t(out), method = "euclidean"), method="ward.D")$order
out_new <- out[row.order, col.order]
m_out <- melt(as.matrix(out_new))
names(m_out)[c(1:2)] <- c("Qx", "Qy")

# Dendrogram
dd.row <- as.dendrogram(hclust(dist(t(out), method = "euclidean"), method="ward.D"))
dx <- dendro_data(dd.row)

# x/y dendograms
figS2b_dendrogramX = ggdend(dx$segments)

figS2b_heatmap = ggplot(data = m_out, aes(x = Qx, y = Qy, fill = 1 - value)) +
  geom_raster() +
  theme_bw() +
  scale_fill_distiller("Spearman correlation", palette = "Oranges",) + # pre = blues, post = oranges, d
  theme(axis.text.x = element_text(angle = 45, hjust = 1), axis.title = element_blank(), axis.ticks =ele
  theme(panel.border = element_blank(), panel.grid.major = element_blank(), panel.grid.minor = element_b

figS2b = grid.arrange(figS2b_dendrogramX, figS2b_heatmap, ncol = 1, heights = c(1, 3), top ="Figure S2B
```




```
ggsave(figS2b, filename="./Figures/FigureS2B.svg", width = 6, height = 4)
```

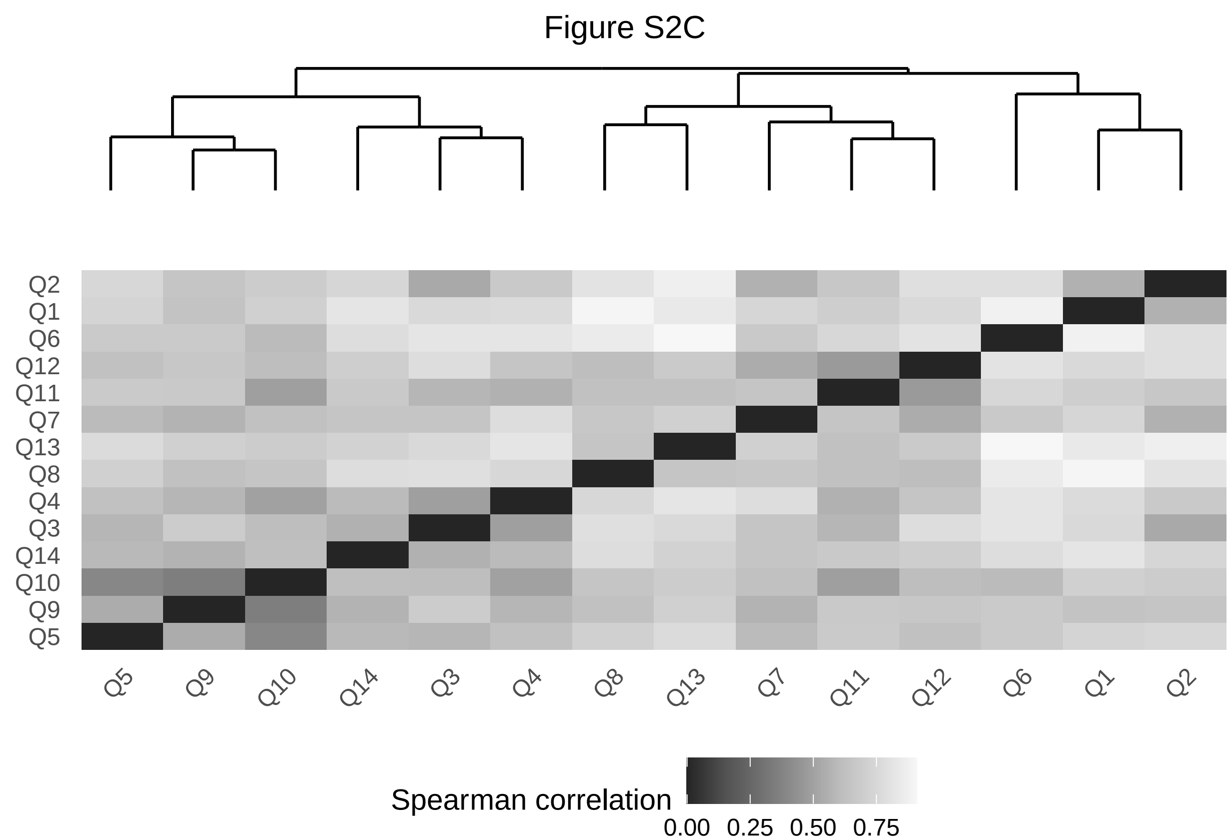
Figure S2C

```
# Split data by responses to self-efficacy questionaire before and after the first semester of graduate
SE_delta = data[,28:41] - data[,14:27]

# Melt responses
mSE_delta = melt(SE_delta)
```

```
## No id variables; using all as measure variables
```

```
# Add question number label to each pre / post array
Questions = c("Q1", "Q2", "Q3", "Q4", "Q5", "Q6", "Q7", "Q8", "Q9", "Q10", "Q11", "Q12", "Q13", "Q14")
mSE_delta$Question = rep(Questions, each = numStudents)

# Make data frame for clustering
out = data.frame(matrix(0, nrow = length(Questions), ncol = length(Questions)), row.names = Questions, s
colnames(out) = Questions

i = 1
for (Qi in Questions) {
  j = 1
  for (Qj in Questions) {
    out[i, j] = round(cor(mSE_delta[mSE_delta$Question == Qi, "value"], mSE_delta[mSE_delta$Question ==
```

```
    j = j + 1
  }
  i = i + 1
}

# Cluster
row.order <- hclust(dist(out, method = "euclidean"), method="ward.D")$order
col.order <- hclust(dist(t(out), method = "euclidean"), method="ward.D")$order
out_new <- out[row.order, col.order]
m_out <- melt(as.matrix(out_new))
names(m_out)[c(1:2)] <- c("Qx", "Qy")

# Dendrogram
dd.row <- as.dendrogram(hclust(dist(t(out), method = "euclidean"), method="ward.D"))
dx <- dendro_data(dd.row)

# x/y dendograms
figS2c_dendrogramX = ggdend(dx$segments)

figS2c_heatmap = ggplot(data = m_out, aes(x = Qx, y = Qy, fill = 1 - value)) +
  geom_raster() +
  theme_bw() +
  scale_fill_distiller("Spearman correlation", palette = "Greys",) + # pre = blues, post = oranges, del
  theme(axis.text.x = element_text(angle = 45, hjust = 1), axis.title = element_blank(), axis.ticks =ele
  theme(panel.border = element_blank(), panel.grid.major = element_blank(), panel.grid.minor = element_

figS2c = grid.arrange(figS2c_dendrogramX, figS2c_heatmap, ncol = 1, heights = c(1, 3), top ="Figure S2C
```




```
ggsave(figS2c, filename="./Figures/FigureS2C.svg", width = 6, height = 4)
```

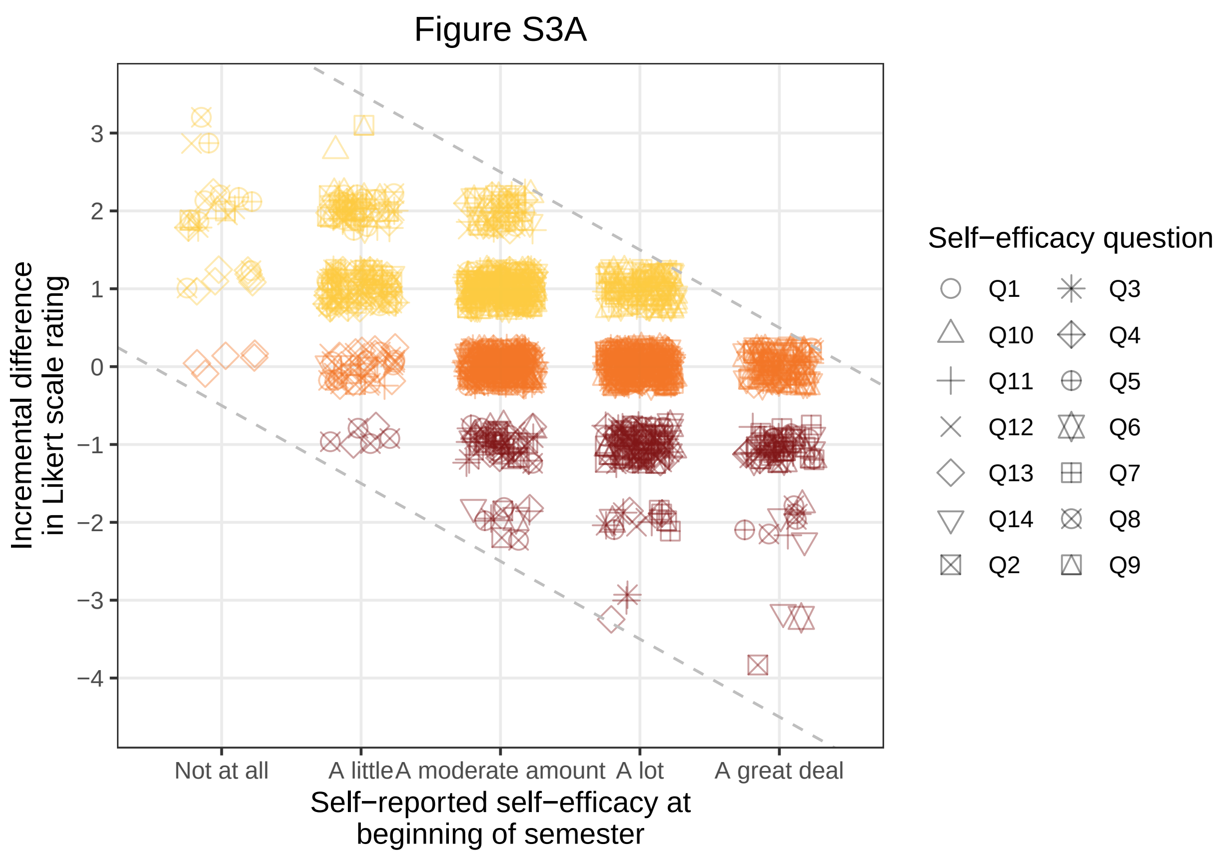
Figure S3

Over 20% of students decrease on at least one individual self-efficacy item from the beginning to the end of the semester.

Figure S3A

```
# Make combined data frame
mSE_pre_delta = cbind(mSE_pre, mSE_delta)[,c(3,2,5)]
colnames(mSE_pre_delta) <- c("Question", "Pre", "Delta")

# Create color scheme
mSE_pre_delta$Color = "orange" # If delta = 0
mSE_pre_delta[mSE_pre_delta$Delta > 0, "Color"] = "red" # if delta > 0
mSE_pre_delta[mSE_pre_delta$Delta < 0, "Color"] = "yellow" # if delta < 0

# Plot scatterplot
figS3a <- ggplot(mSE_pre_delta, aes(Pre, Delta, color = Color)) +
  geom_jitter(height = 0.25, width = 0.25, alpha = 0.4, size = 3, aes(shape = Question)) +
  scale_shape_manual(values=1:14) +
  scale_color_manual(values = c("yellow" = "#811617", "orange" = "#F37627", "red" = "#FDCB41"), guide =
  geom_abline(intercept = 0.5, slope = -1, linetype="dashed", color = "grey") +
```

```
  geom_abline(intercept = 5.5, slope = -1, linetype="dashed", color = "grey") +
  scale_x_continuous("Self-reported self-efficacy at\nbeginning of semester", limits = c(0.5, 5.5), brea
  scale_y_continuous("Incremental difference\nin Likert scale rating", limits = c(-4.5, 3.5), breaks = 
  theme_bw() +
  guides(shape=guide_legend(ncol=2)) + labs(shape = "Self-efficacy question") +
  theme(panel.grid.minor = element_blank()) +
  ggtitle("Figure S3A") + theme(plot.title = element_text(hjust = 0.5))

plot(figS3a)
```



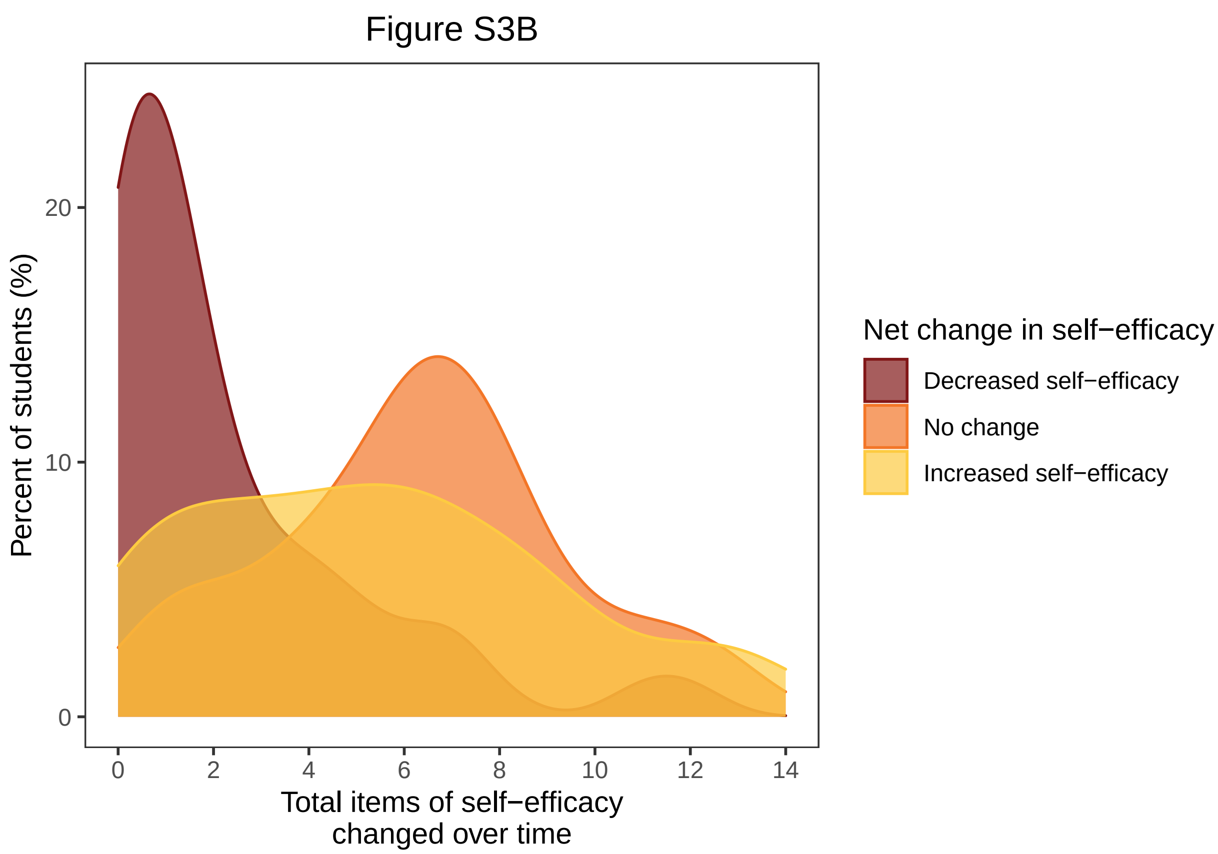

```
ggsave(figS3a, file="./Figures/FigS3A.svg", width = 6, height = 4)
```

Figure S3B

```
# Construct a dataframe for a density plot of changes in student responses
SE_change = data.frame(red = rowSums(SE_delta < 0), orange = rowSums(SE_delta  == 0), yelloq = rowSums(S

# Melt data frame for plotting
mSE_change = melt(SE_change)
```

```
## No id variables; using all as measure variables
```

```
figS3b <- ggplot(mSE_change, aes(x = value, fill = variable, color = variable)) +
  geom_density(alpha = 0.7) +
  theme_bw() +
  scale_fill_manual("Net change in self-efficacy", values = c("#811617", "#F37627", "#FDCB41"), labels =
  scale_color_manual("Net change in self-efficacy", values = c("#811617", "#F37627", "#FDCB41"), labels
  scale_x_continuous("Total items of self-efficacy\nchanged over time", breaks=seq(0, 14, 2), labels=se
  scale_y_continuous("Percent of students (%)", breaks=c(0, 0.1, 0.2), labels=c(0, 10, 20)) +
  theme(panel.grid.minor = element_blank(), panel.grid.major= element_blank()) +
  ggtitle("Figure S3B") + theme(plot.title = element_text(hjust = 0.5))

plot(figS3b)
```




```
ggsave(figS3b, file="./Figures/FigS3B.svg", width = 6, height = 4)
```

Figure S4

Over 70% of students had a net increase in research skills over the semester.

```
# Add columns to delta- dataframe to identify each student (currently row names)
SE_delta$Student = row.names(SE_delta)

# Change question names
colnames(SE_delta) = c("Q1", "Q2", "Q3", "Q4", "Q5", "Q6", "Q7", "Q8", "Q9", "Q10", "Q11", "Q12", "Q13"
```

```r
# Melt data
mSE_delta = melt(SE_delta)
```

```
## Using Student as id variables
```

```r
colnames(mSE_delta) <- c("Student", "Question", "Delta")

# Calcualte frequencies at which students improved are decreased in self-efficacy
delta_freq  = dcast(mSE_delta, Student~Delta)
```

```
## Using Delta as value column: use value.var to override.
```

```
## Aggregation function missing: defaulting to length
```

```r
# Calcualte net-change for each student, adding information about gender and years of expeirence workin
delta_net = rowSums(cbind(delta_freq[,2]*-4, delta_freq[,3]*-3, delta_freq[,4]*-2, delta_freq[,5]*-1, 0
delta_net = data.frame(Value = delta_net, Gender = data$Sex, Exp = data$LabExp, stringsAsFactors = F)

# Add color labels
delta_net$Color = "orange"
delta_net[delta_net$Value < 0, 'Color'] ="red"
delta_net[delta_net$Value > 0, 'Color'] = "yellow"

# Create histogram
figS4a = ggplot(delta_net, aes(x = Value, fill = Color)) +
  geom_histogram(binwidth = 0.5) +
  scale_fill_manual("Net change in\nself-efficacy", values = c("yellow" = "#FDCB41", "orange" = "#F3762
                    labels = c("Net decrease in self-efficacy score", "No change in self-efficacy score
  scale_x_continuous("Net change in self-efficacy", breaks=seq(-20, 20, 5)) +
  scale_y_continuous("Number of students", breaks=seq(0, 10, 2)) +
  theme_bw() +
  theme(panel.background=element_blank(), panel.grid.minor=element_blank(), plot.background=element_blan
  ggtitle("Figure S4A") + theme(plot.title = element_text(hjust = 0.5))

ggsave(figS4a, filename="./Figures/FigureS4A.svg", width = 6, height = 4)
plot(figS4a)
```

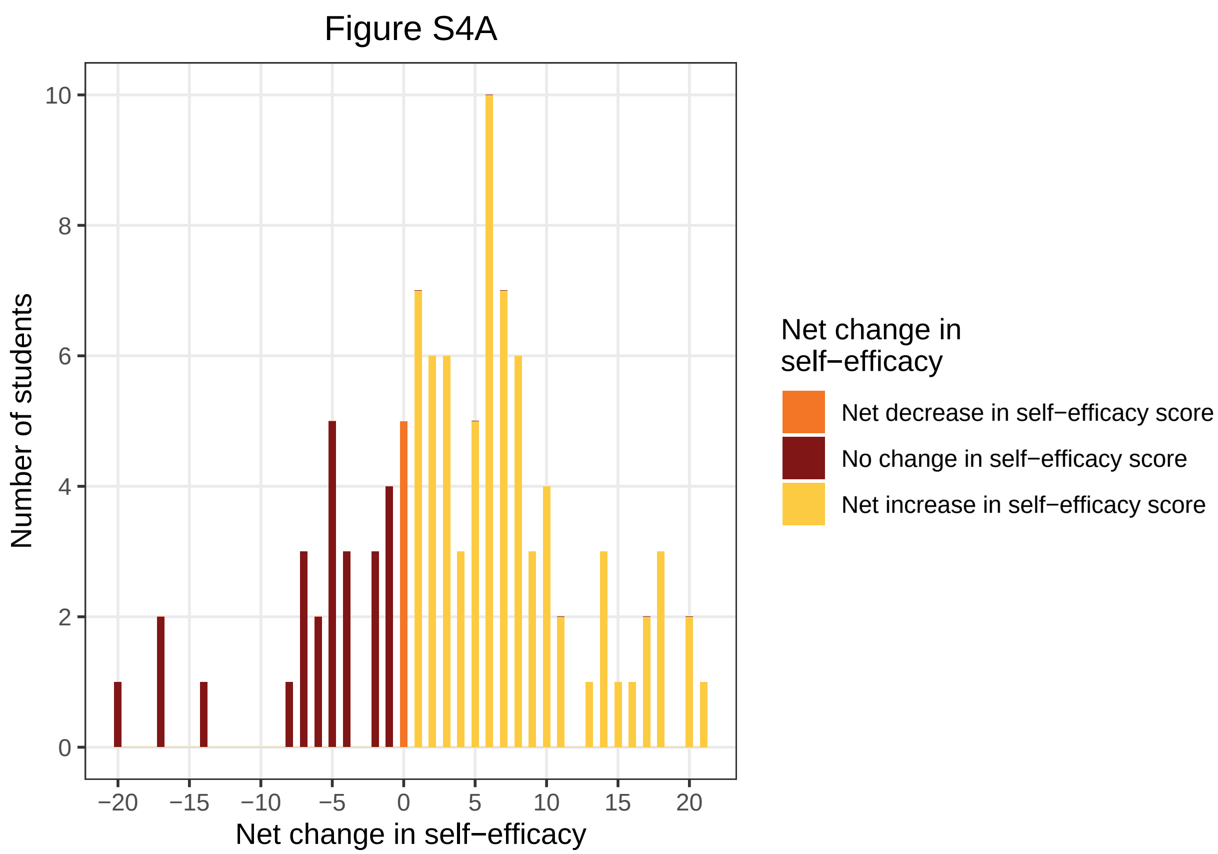


Figure S4Bi

```
negSE_gender = data.frame(Var1 = c("Female", "Male"), Freq = c(nrow(delta_net[delta_net$Value < 0 & del

figS4bi <- ggplot(negSE_gender, aes(x="", y=Freq, fill = Var1))+
  geom_bar(width = 1, stat = "identity") +
  coord_polar("y", start=0) +
  scale_fill_manual("Gender of students with\nnet negative self-efficacies", values = c("magenta4", "gr
                    labels = c(paste("Women (", round((negSE_gender$Freq[1] / sum(negSE_gender$Freq))*1
                               paste("Men (", round((negSE_gender$Freq[2] / sum(negSE_gender$Freq))*100
  theme_minimal() +
  theme(axis.title = element_blank(), panel.border = element_blank(), panel.grid=element_blank(), axis.
  ggtitle("Figure S4Bi") + theme(plot.title = element_text(hjust = 0.5))

ggsave(figS4bi, filename="./Figures/FigureS4Bi.svg", width = 3, height = 3)
plot(figS4bi)
```

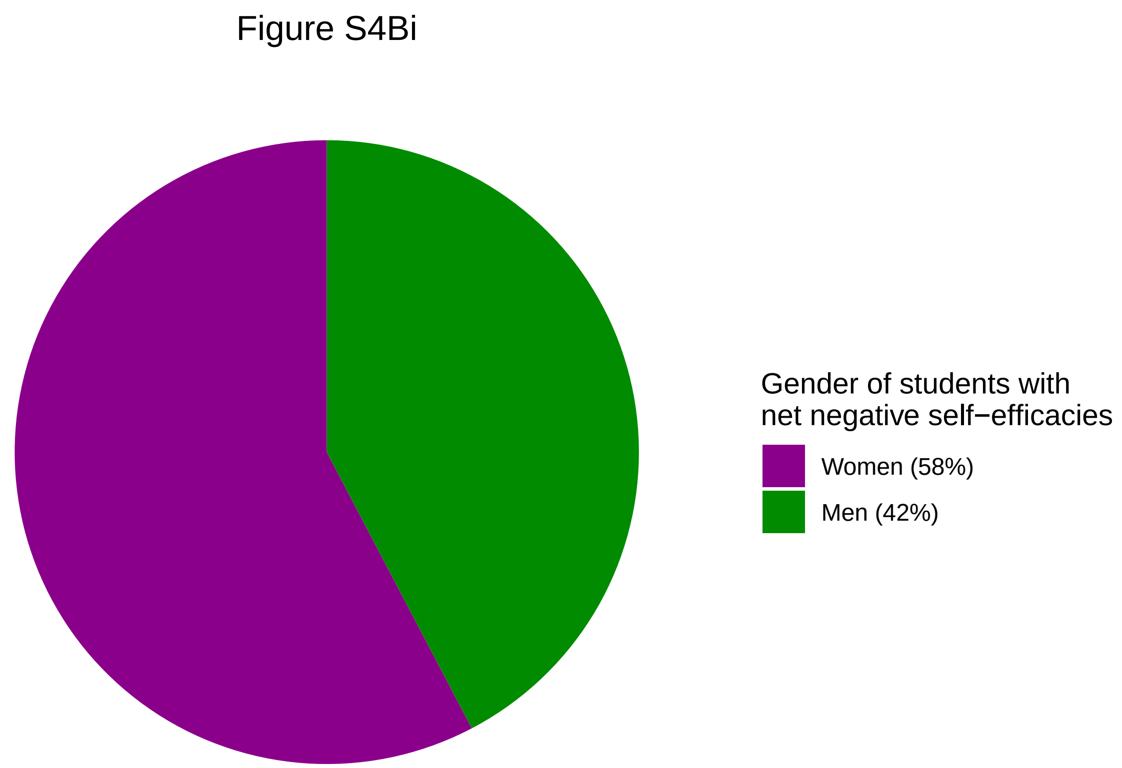


Figure S4Bii

```
negSE_labExp = data.frame(Var1 = as.character(1:7), Freq = c(nrow(delta_net[delta_net$Value < 0 & delta_
negSE_labExp$Var1 <- factor(negSE_labExp$Var1, levels = as.character(1:7))

figS4bii <- ggplot(negSE_labExp, aes(x="", y=Freq, fill = Var1))+
  geom_bar(width = 1, stat = "identity") +
  coord_polar("y", start=0, direction = -1) +
  scale_fill_manual("Lab experience of students with\nnet negative self-efficacies", values=c("7" = "#3
                    labels = c(paste("< 2 years (", round((negSE_labExp$Freq[1] / sum(negSE_labExp$Freq
                               paste("2 - 3 years (", round((negSE_labExp$Freq[2] / sum(negSE_labExp$Fr
                               paste("3 - 4 years (", round((negSE_labExp$Freq[3] / sum(negSE_labExp$Fr
                               paste("4 - 5 years (", round((negSE_labExp$Freq[4] / sum(negSE_labExp$Fr
                               paste("5 - 6 years (", round((negSE_labExp$Freq[5] / sum(negSE_labExp$Fr
                               paste("6 - 7 years (", round((negSE_labExp$Freq[6] / sum(negSE_labExp$Fr
                               paste("> 7 years (", round((negSE_labExp$Freq[7] / sum(negSE_labExp$Freq
  theme_minimal() +
  theme(axis.title = element_blank(), panel.border = element_blank(), panel.grid=element_blank(), axis.
  ggtitle("Figure S4Bii") + theme(plot.title = element_text(hjust = 0.5))

ggsave(figS4bii, filename="./Figures/FigureS4Bii.svg", width = 3, height = 3)
plot(figS4bii)
```

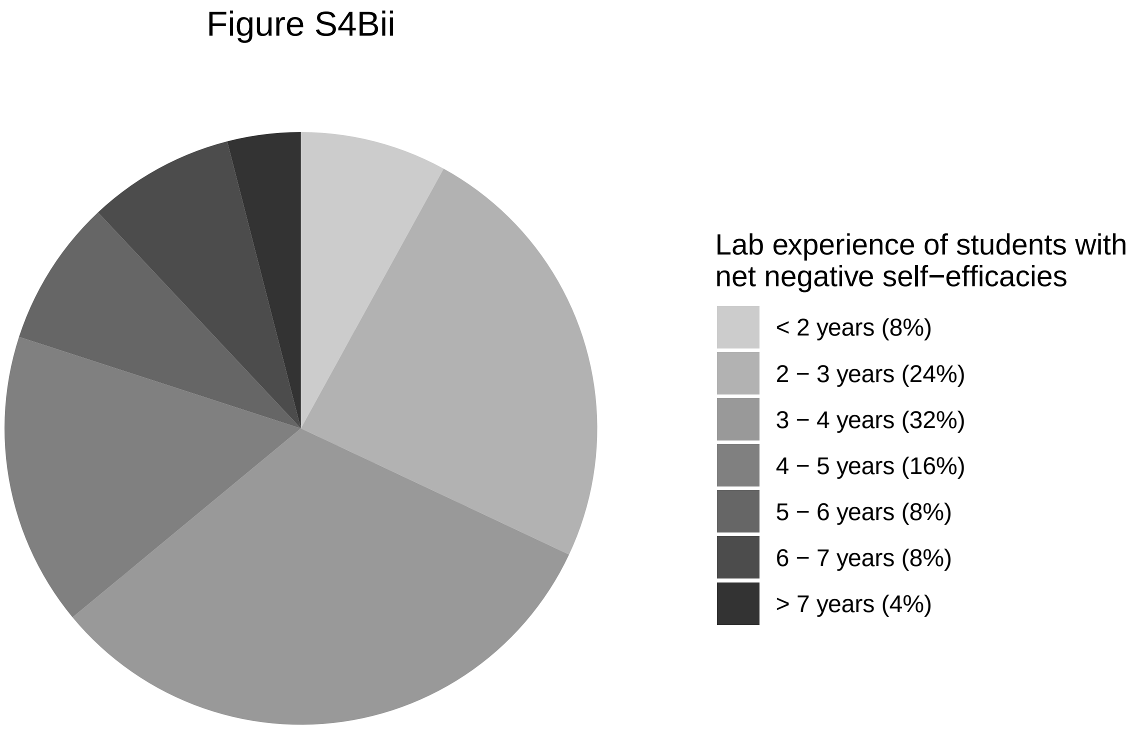


Figure S5

The only significant difference between men and women's self-efficacy is in experimental design at the beginning of the semester in 2017.

Figure S5A

```
# Split data by responses to self-efficacy questionaire before and after the first semester of graduate
SE_pre = data_2017[,c(2,14:27)] # Change 2017 / 2018

# Split by gender
femaleSE_pre = SE_pre[SE_pre$Sex == "Female",]
maleSE_pre = SE_pre[SE_pre$Sex == "Male",]

# Function to calculate the p-value of every factor on the post-test score
ordLogReg <- function(dat, Q, test) {

  # Get relevant data for each question
  pre = dat[,(Q+13)]
  post = dat[,(Q+27)]

  # Create data frame including pre-test score, post-test score, gender, and years of research experien
  mat = data.frame(Pre = pre, Post = post, Gender = dat$Sex, stringsAsFactors = FALSE)

  # Reorder
```

```r
  mat$Gender <- factor(mat$Gender, levels=c("Female", "Male"), ordered=TRUE)

  if (test == "pre") {
    summary(glm(Pre ~ Gender, data = mat))
  } else if (test == "post") {
    summary(glm(Post ~ Gender, data = mat))
  } else {
    cat("Please enter either 'pre' or 'post' as a selection of the test to correctly use this function.
  }
}

# Calculate the p-value for each question and save to an output table
Items = c("Understand contemporary concepts in your field", "Make use of the primary scientific research
sigOut = data.frame(Item = Items, Question = 1:14, tValue = rep(0, 14), pValue = rep(0, 14), Significanc
for (i in 1:14) {
  ordLogRegTest = as.vector(ordLogReg(data_2017, i, "pre")$coefficients)[c(6,8)]
  sigOut$tValue[i] = round(ordLogRegTest[1], 2)
  sigOut$pValue[i] = round(ordLogRegTest[2], 4)
}

for (i in 1:14) {
  if (sigOut$pValue[i] < 0.001) {
    sigOut$Significance[i] = "***"
  } else if (sigOut$pValue[i] < 0.01) {
    sigOut$Significance[i] = "**"
  } else if (sigOut$pValue[i] < 0.05) {
    sigOut$Significance[i] = "*"
  }
}

print(sigOut)
```

```
##                                                                                         Item
## 1                                                Understand contemporary concepts in your field
## 2  Make use of the primary scientific research literature in your field (e.g., journal articles)
## 3             Identify a specific question for investigation based on the research in your field
## 4                                    Formulate a research hypothesis based on a specific question
## 5                                      Design an experiment or theoretical test of the hypothesis
## 6                                             Understand the importance of 'controls' in research
## 7                                                                      Observe and collect data
## 8                                                                    Statistically analyze data
## 9                                   Interpret data by relating results to the original hypothesis
## 10                                 Reformulate your original research hypothesis (as appropriate)
## 11                                      Relate your results to the 'bigger picture' in your field
## 12                                            Orally communicate the results of research projects
## 13                                                        Write a research paper for publication
## 14                                                                         Think independently
##    Question tValue pValue Significance
## 1         1   1.26 0.2155
## 2         2   0.75 0.4595
## 3         3   1.40 0.1674
## 4         4   0.21 0.8330
## 5         5   2.94 0.0053           **
```

```
## 6          6   1.99 0.0531
## 7          7   0.88 0.3852
## 8          8   0.18 0.8602
## 9          9   1.52 0.1363
## 10        10   1.97 0.0549
## 11        11   1.01 0.3171
## 12        12  -0.99 0.3284
## 13        13   0.98 0.3334
## 14        14   1.25 0.2190
```

```
# Count number of students
numFemaleStudents = nrow(femaleSE_pre)
numMaleStudents = nrow(maleSE_pre)

# Melt responses
m_femaleSE_pre = melt(femaleSE_pre)
```

```
## Using Sex as id variables
```

```
m_maleSE_pre = melt(maleSE_pre)
```

```
## Using Sex as id variables
```

```
# Add question number label to each array
m_femaleSE_pre$variable = rep(c("Q1", "Q2", "Q3", "Q4", "Q5", "Q6", "Q7", "Q8", "Q9", "Q10", "Q11", "Q1
m_maleSE_pre$variable = rep(c("Q1", "Q2", "Q3", "Q4", "Q5", "Q6", "Q7", "Q8", "Q9", "Q10", "Q11", "Q12"

# Change labels to strings for categorical plotting
m_femaleSE_pre[m_femaleSE_pre$value==1,"value"] = "a1" # Not at all, female
m_femaleSE_pre[m_femaleSE_pre$value==2,"value"] = "b1" # A little, female
m_femaleSE_pre[m_femaleSE_pre$value==3,"value"] = "c1" # A moderate amount, female
m_femaleSE_pre[m_femaleSE_pre$value==4,"value"] = "d1" # A lot, female
m_femaleSE_pre[m_femaleSE_pre$value==5,"value"] = "e1" # A great deal, female

m_maleSE_pre[m_maleSE_pre$value==1,"value"] = "a2" # Not at all, male
m_maleSE_pre[m_maleSE_pre$value==2,"value"] = "b2" # A little, male
m_maleSE_pre[m_maleSE_pre$value==3,"value"] = "c2" # A moderate amount, male
m_maleSE_pre[m_maleSE_pre$value==4,"value"] = "d2" # A lot, male
m_maleSE_pre[m_maleSE_pre$value==5,"value"] = "e2" # A great deal, male

# Combine gender dataframes for plotting
m_gSE_pre <- rbind(m_femaleSE_pre, m_maleSE_pre)

# Order variables for plotting
m_gSE_pre$variable <- factor(m_gSE_pre$variable, levels = c("Q1", "Q2", "Q3", "Q4", "Q5", "Q6", "Q7", "
m_gSE_pre$Sex <- factor(m_gSE_pre$Sex , levels = c("Female", "Male"), ordered = TRUE)
m_gSE_pre$value = factor(m_gSE_pre$value, levels = c("a1", "b1", "c1", "d1", "e1", "a2", "b2", "c2", "d

# Create stacked bar chart for each question, for male and female students
figS5a <- ggplot(m_gSE_pre, aes(Sex)) +
  geom_bar(aes(fill=value), position = "fill") + facet_grid(~ variable) +
  theme_bw() +
```

```
  scale_fill_manual("Research skills\nself-efficacy", values = c("a1" = "#D5ACD9", "b1" = "#BE7DC4", "c
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank()) +
  theme(axis.title=element_blank(), axis.text=element_blank(), axis.ticks=element_blank(), strip.backgro
  guides(fill=guide_legend(ncol = 2)) +
  ggtitle("Figure S5A") + theme(plot.title = element_text(hjust = 0.5))

# Plot figure
ggsave(figS5a, filename="./Figures/FigureS5A.svg", width = 6, height = 4)
plot(figS5a)
```

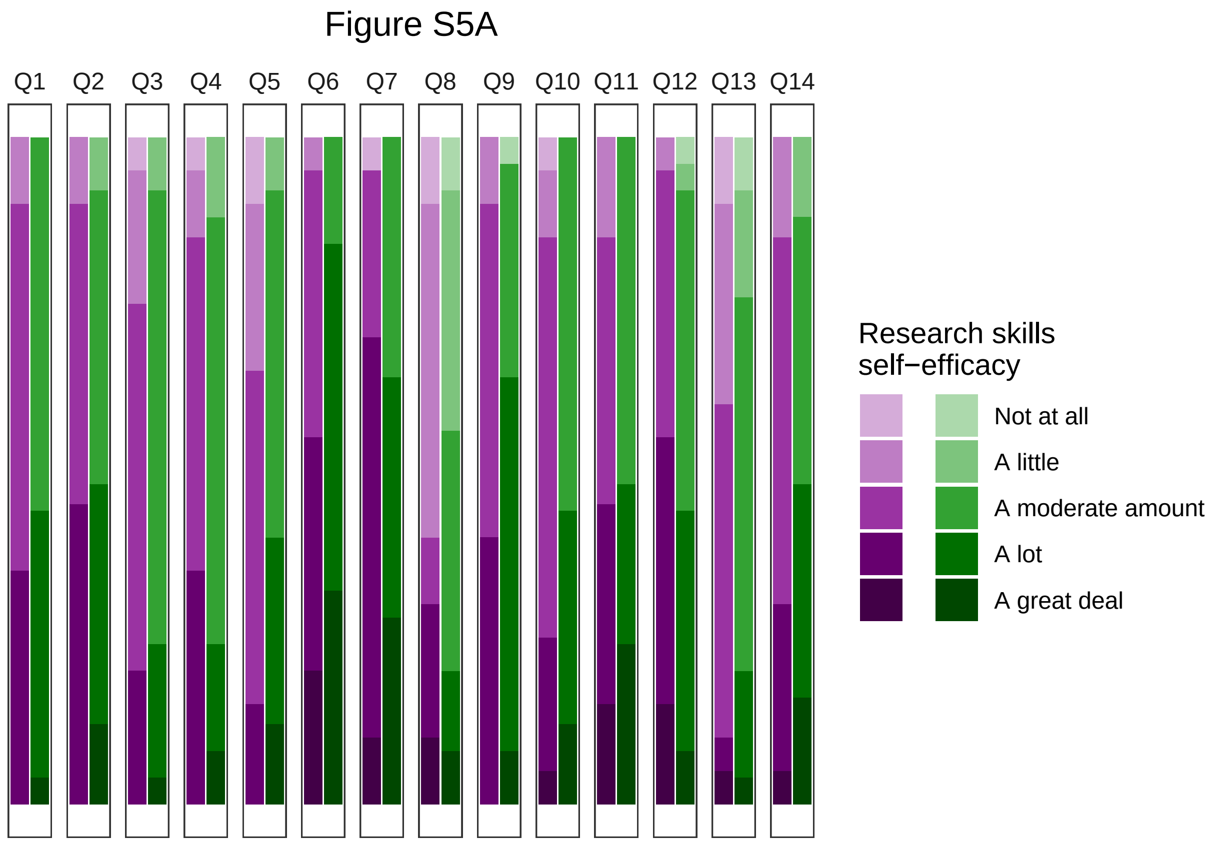


Figure S5B

```
# Split data by responses to self-efficacy questionaire before and after the first semester of graduate
SE_post = data_2017[,c(2,28:41)] # Change 2017 / 2018

# Split by gender
femaleSE_post = SE_post[SE_post$Sex == "Female",]
maleSE_post = SE_post[SE_post$Sex == "Male",]

# Calculate the p-value for each question and save to an output table
Items = c("Understand contemporary concepts in your field", "Make use of the primary scientific research
sigOut = data.frame(Item = Items, Question = 1:14, tValue = rep(0, 14), pValue = rep(0, 14), Significanc
for (i in 1:14) {
  ordLogRegTest = as.vector(ordLogReg(data_2017, i, "post")$coefficients)[c(6,8)]
  sigOut$tValue[i] = round(ordLogRegTest[1], 2)
  sigOut$pValue[i] = round(ordLogRegTest[2], 4)
```

```r
}

for (i in 1:14) {
  if (sigOut$pValue[i] < 0.001) {
    sigOut$Significance[i] = "***"
  } else if (sigOut$pValue[i] < 0.01) {
    sigOut$Significance[i] = "**"
  } else if (sigOut$pValue[i] < 0.05) {
    sigOut$Significance[i] = "*"
  }
}

print(sigOut)
```

```
##                                                                                        Item
## 1                                                Understand contemporary concepts in your field
## 2  Make use of the primary scientific research literature in your field (e.g., journal articles)
## 3             Identify a specific question for investigation based on the research in your field
## 4                                  Formulate a research hypothesis based on a specific question
## 5                                    Design an experiment or theoretical test of the hypothesis
## 6                                           Understand the importance of 'controls' in research
## 7                                                                       Observe and collect data
## 8                                                                     Statistically analyze data
## 9                                   Interpret data by relating results to the original hypothesis
## 10                                 Reformulate your original research hypothesis (as appropriate)
## 11                                      Relate your results to the 'bigger picture' in your field
## 12                                             Orally communicate the results of research projects
## 13                                                           Write a research paper for publication
## 14                                                                          Think independently
##    Question tValue pValue Significance
## 1         1   0.16 0.8708
## 2         2   0.62 0.5356
## 3         3   1.19 0.2400
## 4         4   1.84 0.0725
## 5         5   0.72 0.4732
## 6         6   0.64 0.5243
## 7         7   0.84 0.4065
## 8         8   0.18 0.8574
## 9         9  -0.72 0.4777
## 10       10   0.20 0.8421
## 11       11   0.03 0.9740
## 12       12  -0.15 0.8844
## 13       13   0.10 0.9187
## 14       14   1.21 0.2340
```

```r
# Count number of students
numFemaleStudents = nrow(femaleSE_post)
numMaleStudents = nrow(maleSE_post)

# Melt responses
m_femaleSE_post = melt(femaleSE_post)
```

```
## Using Sex as id variables
```

```
m_maleSE_post = melt(maleSE_post)
```

```
## Using Sex as id variables
```

```
# Add question number label to each array
m_femaleSE_post$variable = rep(c("Q1", "Q2", "Q3", "Q4", "Q5", "Q6", "Q7", "Q8", "Q9", "Q10", "Q11", "Q
m_maleSE_post$variable = rep(c("Q1", "Q2", "Q3", "Q4", "Q5", "Q6", "Q7", "Q8", "Q9", "Q10", "Q11", "Q12

# Change labels to strings for categorical plotting
m_femaleSE_post[m_femaleSE_post$value==1,"value"] = "a1" # Not at all, female
m_femaleSE_post[m_femaleSE_post$value==2,"value"] = "b1" # A little, female
m_femaleSE_post[m_femaleSE_post$value==3,"value"] = "c1" # A moderate amount, female
m_femaleSE_post[m_femaleSE_post$value==4,"value"] = "d1" # A lot, female
m_femaleSE_post[m_femaleSE_post$value==5,"value"] = "e1" # A great deal, female

m_maleSE_post[m_maleSE_post$value==1,"value"] = "a2" # Not at all, male
m_maleSE_post[m_maleSE_post$value==2,"value"] = "b2" # A little, male
m_maleSE_post[m_maleSE_post$value==3,"value"] = "c2" # A moderate amount, male
m_maleSE_post[m_maleSE_post$value==4,"value"] = "d2" # A lot, male
m_maleSE_post[m_maleSE_post$value==5,"value"] = "e2" # A great deal, male

# Combine gender dataframes for plotting
m_gSE_post <- rbind(m_femaleSE_post, m_maleSE_post)

# Order variables for plotting
m_gSE_post$variable <- factor(m_gSE_post$variable, levels = c("Q1", "Q2", "Q3", "Q4", "Q5", "Q6", "Q7",
m_gSE_post$Sex <- factor(m_gSE_post$Sex , levels = c("Female", "Male"), ordered = TRUE)
m_gSE_post$value = factor(m_gSE_post$value, levels = c("a1", "b1", "c1", "d1", "e1", "a2", "b2", "c2",

# Create stacked bar chart for each question, for male and female students
figS5b <- ggplot(m_gSE_post, aes(Sex)) +
  geom_bar(aes(fill=value), position = "fill") + facet_grid(~ variable) +
  theme_bw() +
  scale_fill_manual("Research skills\nself-efficacy", values = c("a1" = "#D5ACD9", "b1" = "#BE7DC4", "c
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank()) +
  theme(axis.title=element_blank(), axis.text=element_blank(), axis.ticks=element_blank(), strip.backgr
  guides(fill=guide_legend(ncol = 2)) +
  ggtitle("Figure S5B") + theme(plot.title = element_text(hjust = 0.5))

# Plot figure
ggsave(figS5b, filename="./Figures/FigureS5B.svg", width = 6, height = 4)
plot(figS5b)
```

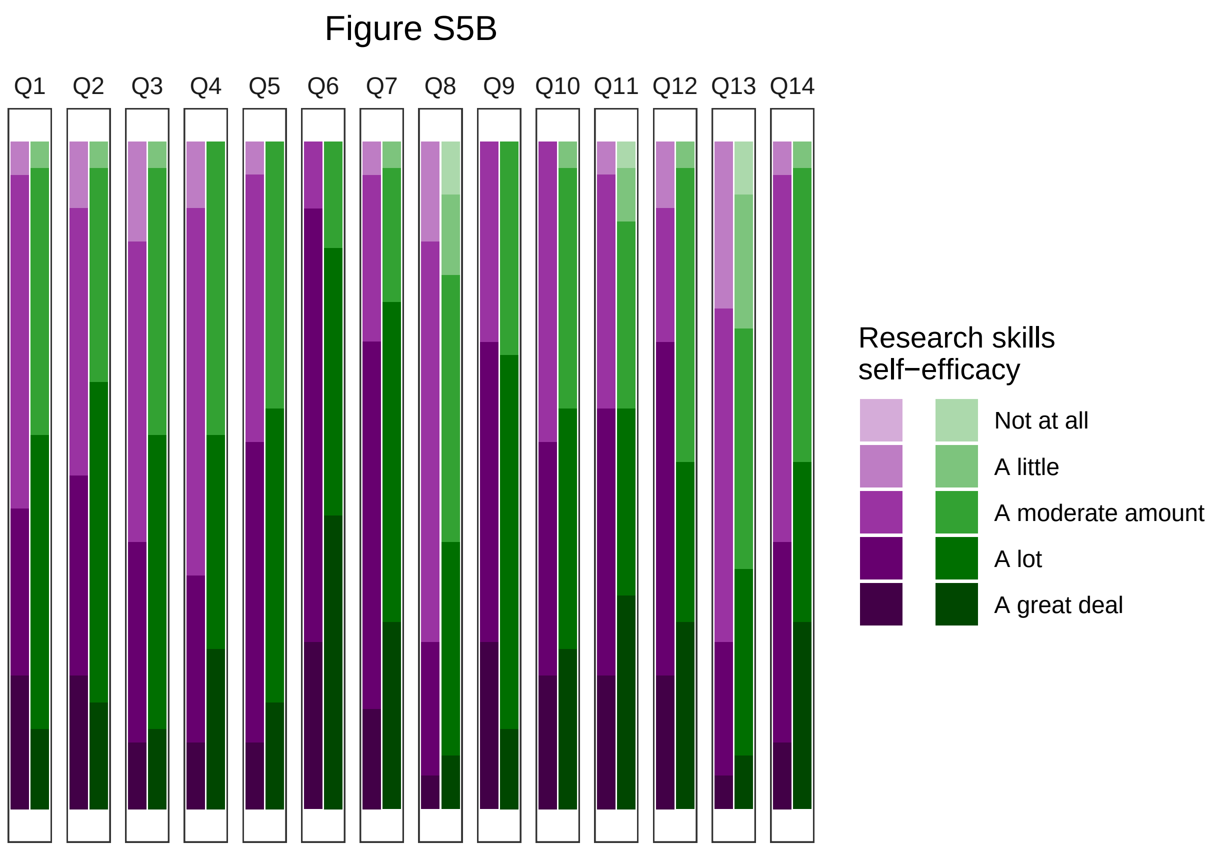


Figure S5C

```
# Split data by responses to self-efficacy questionaire before and after the first semester of graduate
data_2018 = data[data$Year == 2018,]
SE_pre = data_2018[,c(2,14:27)] # Change 2017 / 2018

# Remove if gender not specified
SE_pre = SE_pre[!is.na(SE_pre$Sex),]

# Split by gender
femaleSE_pre = SE_pre[SE_pre$Sex == "Female",]
maleSE_pre = SE_pre[SE_pre$Sex == "Male",]

# Calculate the p-value for each question and save to an output table
Items = c("Understand contemporary concepts in your field", "Make use of the primary scientific researc
sigOut = data.frame(Item = Items, Question = 1:14, tValue = rep(0, 14), pValue = rep(0, 14), Significan
for (i in 1:14) {
  ordLogRegTest = as.vector(ordLogReg(data_2017, i, "pre")$coefficients)[c(6,8)]
  sigOut$tValue[i] = round(ordLogRegTest[1], 2)
  sigOut$pValue[i] = round(ordLogRegTest[2], 4)
}

for (i in 1:14) {
  if (sigOut$pValue[i] < 0.001) {
    sigOut$Significance[i] = "***"
  } else if (sigOut$pValue[i] < 0.01) {
```

```r
    sigOut$Significance[i] = "**"
  } else if (sigOut$pValue[i] < 0.05) {
    sigOut$Significance[i] = "*"
  }
}

print(sigOut)
```

```
##                                                                                          Item
## 1                                                  Understand contemporary concepts in your field
## 2  Make use of the primary scientific research literature in your field (e.g., journal articles)
## 3            Identify a specific question for investigation based on the research in your field
## 4                                     Formulate a research hypothesis based on a specific question
## 5                                       Design an experiment or theoretical test of the hypothesis
## 6                                              Understand the importance of 'controls' in research
## 7                                                                           Observe and collect data
## 8                                                                         Statistically analyze data
## 9                                    Interpret data by relating results to the original hypothesis
## 10                                  Reformulate your original research hypothesis (as appropriate)
## 11                                       Relate your results to the 'bigger picture' in your field
## 12                                             Orally communicate the results of research projects
## 13                                                            Write a research paper for publication
## 14                                                                             Think independently
##    Question tValue pValue Significance
## 1         1   1.26 0.2155
## 2         2   0.75 0.4595
## 3         3   1.40 0.1674
## 4         4   0.21 0.8330
## 5         5   2.94 0.0053           **
## 6         6   1.99 0.0531
## 7         7   0.88 0.3852
## 8         8   0.18 0.8602
## 9         9   1.52 0.1363
## 10       10   1.97 0.0549
## 11       11   1.01 0.3171
## 12       12  -0.99 0.3284
## 13       13   0.98 0.3334
## 14       14   1.25 0.2190
```

```r
# Count number of students
numFemaleStudents = nrow(femaleSE_pre)
numMaleStudents = nrow(maleSE_pre)

# Melt responses
m_femaleSE_pre = melt(femaleSE_pre)
```

```
## Using Sex as id variables
```

```r
m_maleSE_pre = melt(maleSE_pre)
```

```
## Using Sex as id variables
```

```
# Add question number label to each array
m_femaleSE_pre$variable = rep(c("Q1", "Q2", "Q3", "Q4", "Q5", "Q6", "Q7", "Q8", "Q9", "Q10", "Q11", "Q12
m_maleSE_pre$variable = rep(c("Q1", "Q2", "Q3", "Q4", "Q5", "Q6", "Q7", "Q8", "Q9", "Q10", "Q11", "Q12"

# Change labels to strings for categorical plotting
m_femaleSE_pre[m_femaleSE_pre$value==1,"value"] = "a1" # Not at all, female
m_femaleSE_pre[m_femaleSE_pre$value==2,"value"] = "b1" # A little, female
m_femaleSE_pre[m_femaleSE_pre$value==3,"value"] = "c1" # A moderate amount, female
m_femaleSE_pre[m_femaleSE_pre$value==4,"value"] = "d1" # A lot, female
m_femaleSE_pre[m_femaleSE_pre$value==5,"value"] = "e1" # A great deal, female

m_maleSE_pre[m_maleSE_pre$value==1,"value"] = "a2" # Not at all, male
m_maleSE_pre[m_maleSE_pre$value==2,"value"] = "b2" # A little, male
m_maleSE_pre[m_maleSE_pre$value==3,"value"] = "c2" # A moderate amount, male
m_maleSE_pre[m_maleSE_pre$value==4,"value"] = "d2" # A lot, male
m_maleSE_pre[m_maleSE_pre$value==5,"value"] = "e2" # A great deal, male

# Combine gender dataframes for plotting
m_gSE_pre <- rbind(m_femaleSE_pre, m_maleSE_pre)

# Order variables for plotting
m_gSE_pre$variable <- factor(m_gSE_pre$variable, levels = c("Q1", "Q2", "Q3", "Q4", "Q5", "Q6", "Q7", "
m_gSE_pre$Sex <- factor(m_gSE_pre$Sex , levels = c("Female", "Male"), ordered = TRUE)
m_gSE_pre$value = factor(m_gSE_pre$value, levels = c("a1", "b1", "c1", "d1", "e1", "a2", "b2", "c2", "d2

# Create stacked bar chart for each question, for male and female students
figS5c <- ggplot(m_gSE_pre, aes(Sex)) +
  geom_bar(aes(fill=value), position = "fill") + facet_grid(~ variable) +
  theme_bw() +
  scale_fill_manual("Research skills\nself-efficacy", values = c("a1" = "#D5ACD9", "b1" = "#BE7DC4", "c
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank()) +
  theme(axis.title=element_blank(), axis.text=element_blank(), axis.ticks=element_blank(), strip.backgro
  guides(fill=guide_legend(ncol = 2)) +
  ggtitle("Figure S5C") + theme(plot.title = element_text(hjust = 0.5))

# Plot figure
ggsave(figS5c, filename="./Figures/FigureS5C.svg", width = 6, height = 4)
plot(figS5c)
```

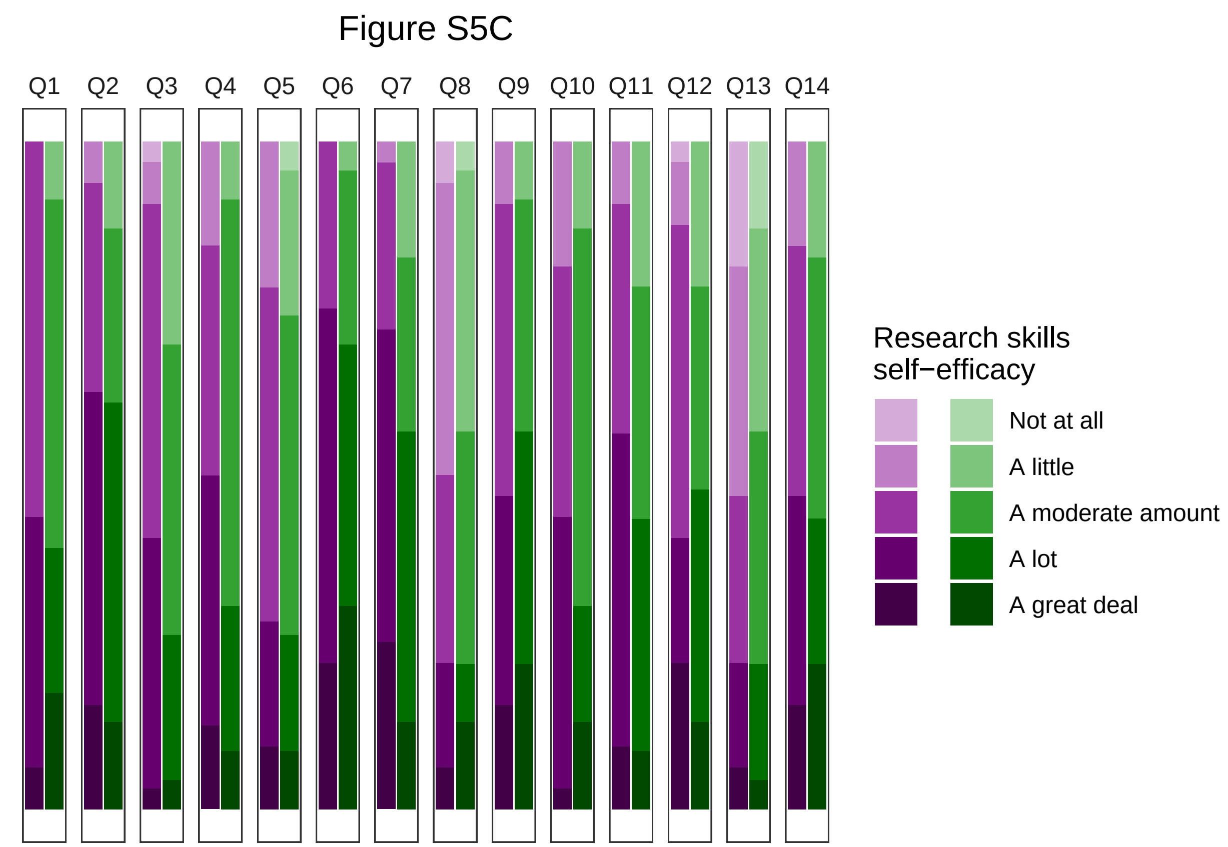


Figure S5D

```
# Split data by responses to self-efficacy questionaire before and after the first semester of graduate
SE_post = data_2018[,c(2,28:41)] # Change 2017 / 2018

# Remove if gender not specified
SE_post = SE_post[!is.na(SE_post$Sex),]

# Split by gender
femaleSE_post = SE_post[SE_post$Sex == "Female",]
maleSE_post = SE_post[SE_post$Sex == "Male",]

# Calculate the p-value for each question and save to an output table
Items = c("Understand contemporary concepts in your field", "Make use of the primary scientific researc
sigOut = data.frame(Item = Items, Question = 1:14, tValue = rep(0, 14), pValue = rep(0, 14), Significan
for (i in 1:14) {
  ordLogRegTest = as.vector(ordLogReg(data_2017, i, "post")$coefficients)[c(6,8)]
  sigOut$tValue[i] = round(ordLogRegTest[1], 2)
  sigOut$pValue[i] = round(ordLogRegTest[2], 4)
}

for (i in 1:14) {
  if (sigOut$pValue[i] < 0.001) {
    sigOut$Significance[i] = "***"
  } else if (sigOut$pValue[i] < 0.01) {
    sigOut$Significance[i] = "**"
```

```
  } else if (sigOut$pValue[i] < 0.05) {
    sigOut$Significance[i] = "*"
  }
}

print(sigOut)
```

```
##                                                                                          Item
## 1                                                    Understand contemporary concepts in your field
## 2  Make use of the primary scientific research literature in your field (e.g., journal articles)
## 3             Identify a specific question for investigation based on the research in your field
## 4                                    Formulate a research hypothesis based on a specific question
## 5                                      Design an experiment or theoretical test of the hypothesis
## 6                                             Understand the importance of 'controls' in research
## 7                                                                       Observe and collect data
## 8                                                                     Statistically analyze data
## 9                                  Interpret data by relating results to the original hypothesis
## 10                                Reformulate your original research hypothesis (as appropriate)
## 11                                     Relate your results to the 'bigger picture' in your field
## 12                                            Orally communicate the results of research projects
## 13                                                         Write a research paper for publication
## 14                                                                          Think independently
##    Question tValue pValue Significance
## 1         1   0.16 0.8708
## 2         2   0.62 0.5356
## 3         3   1.19 0.2400
## 4         4   1.84 0.0725
## 5         5   0.72 0.4732
## 6         6   0.64 0.5243
## 7         7   0.84 0.4065
## 8         8   0.18 0.8574
## 9         9  -0.72 0.4777
## 10       10   0.20 0.8421
## 11       11   0.03 0.9740
## 12       12  -0.15 0.8844
## 13       13   0.10 0.9187
## 14       14   1.21 0.2340
```

```
# Count number of students
numFemaleStudents = nrow(femaleSE_post)
numMaleStudents = nrow(maleSE_post)

# Melt responses
m_femaleSE_post = melt(femaleSE_post)
```

```
## Using Sex as id variables
```

```
m_maleSE_post = melt(maleSE_post)
```

```
## Using Sex as id variables
```

```r
# Add question number label to each array
m_femaleSE_post$variable = rep(c("Q1", "Q2", "Q3", "Q4", "Q5", "Q6", "Q7", "Q8", "Q9", "Q10", "Q11", "Q
m_maleSE_post$variable = rep(c("Q1", "Q2", "Q3", "Q4", "Q5", "Q6", "Q7", "Q8", "Q9", "Q10", "Q11", "Q12

# Change labels to strings for categorical plotting
m_femaleSE_post[m_femaleSE_post$value==1,"value"] = "a1" # Not at all, female
m_femaleSE_post[m_femaleSE_post$value==2,"value"] = "b1" # A little, female
m_femaleSE_post[m_femaleSE_post$value==3,"value"] = "c1" # A moderate amount, female
m_femaleSE_post[m_femaleSE_post$value==4,"value"] = "d1" # A lot, female
m_femaleSE_post[m_femaleSE_post$value==5,"value"] = "e1" # A great deal, female

m_maleSE_post[m_maleSE_post$value==1,"value"] = "a2" # Not at all, male
m_maleSE_post[m_maleSE_post$value==2,"value"] = "b2" # A little, male
m_maleSE_post[m_maleSE_post$value==3,"value"] = "c2" # A moderate amount, male
m_maleSE_post[m_maleSE_post$value==4,"value"] = "d2" # A lot, male
m_maleSE_post[m_maleSE_post$value==5,"value"] = "e2" # A great deal, male

# Combine gender dataframes for plotting
m_gSE_post <- rbind(m_femaleSE_post, m_maleSE_post)

# Order variables for plotting
m_gSE_post$variable <- factor(m_gSE_post$variable, levels = c("Q1", "Q2", "Q3", "Q4", "Q5", "Q6", "Q7",
m_gSE_post$Sex <- factor(m_gSE_post$Sex , levels = c("Female", "Male"), ordered = TRUE)
m_gSE_post$value = factor(m_gSE_post$value, levels = c("a1", "b1", "c1", "d1", "e1", "a2", "b2", "c2",

# Create stacked bar chart for each question, for male and female students
figS5d <- ggplot(m_gSE_post, aes(Sex)) +
  geom_bar(aes(fill=value), position = "fill") + facet_grid(~ variable) +
  theme_bw() +
  scale_fill_manual("Research skills\nself-efficacy", values = c("a1" = "#D5ACD9", "b1" = "#BE7DC4", "c
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank()) +
  theme(axis.title=element_blank(), axis.text=element_blank(), axis.ticks=element_blank(), strip.backgr
  guides(fill=guide_legend(ncol = 2)) +
  ggtitle("Figure S5D") + theme(plot.title = element_text(hjust = 0.5))

# Plot figure
ggsave(figS5d, filename="./Figures/FigureS5D.svg", width = 6, height = 4)
plot(figS5d)
```

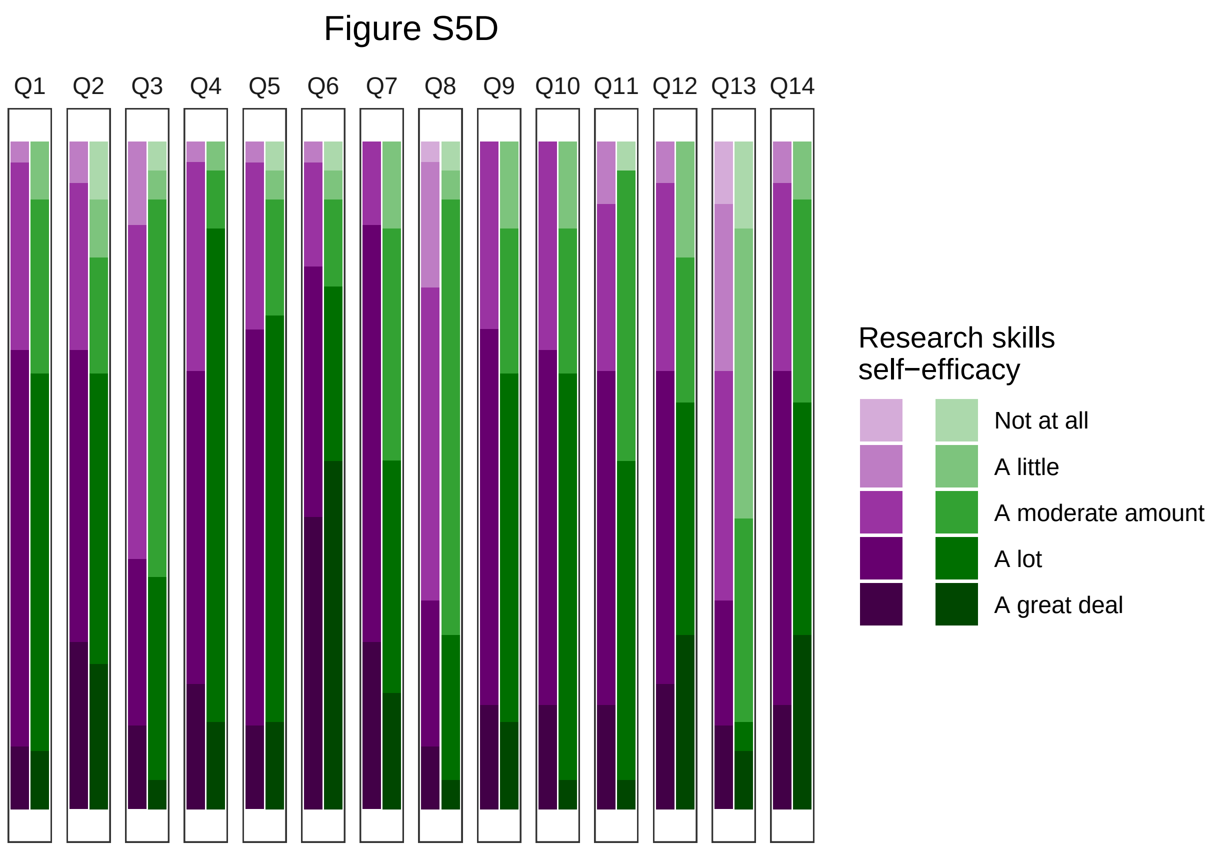


Figure S6

Women in the 2017 cohort were less confident in designing experiments at the beginning of the semester, despite performing comparably on an assessment of experimental design aptitude. The geneder gap in self-efficacy disappears by the end of the course.

Figure S6A

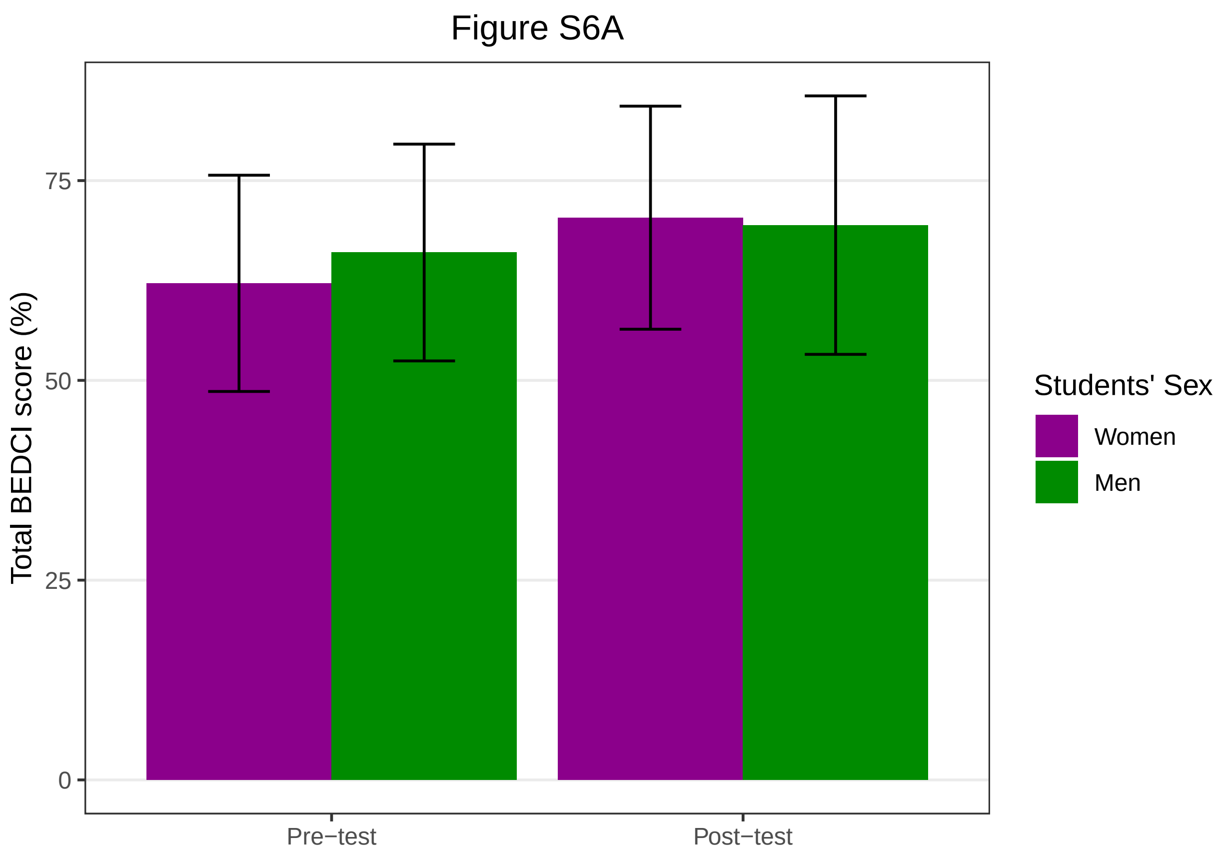
```
# Split data by sex, counting students
numFStudents_2017 = sum(data_2017$Sex == "Female")
numMStudents_2017 = sum(data_2017$Sex == "Male")

# Perform t-tests for pairwise combinations of tests and genders
fpre_mpre = t.test(x = (data_2017[data_2017$Sex == "Female", "Pre_CITotal"] / 14) * 100, y = (data_2017
fpre_fpost = t.test(x = (data_2017[data_2017$Sex == "Female", "Pre_CITotal"] / 14) * 100, y = (data_201
fpost_mpost = t.test(x = (data_2017[data_2017$Sex == "Female", "Post_CITotal"] / 14) * 100, y = (data_2
mpre_mpost = t.test(x = (data_2017[data_2017$Sex == "Male", "Post_CITotal"] / 14) * 100, y = (data_2017

# Create variables to hold results
sigOut = data.frame(X = c("Female, Pre-test", "Female, Pre-test", "Female, Post-test", "Male, Pre-test"

# Enter test results into data frame
sigOut$t[1] = fpre_mpre$statistic # Female, pre-test v male, pre-test
sigOut$pValue[1] = fpre_mpre$p.value
```

```r
sigOut$t[2] = fpre_fpost$statistic # Female, pre-test v female, post-test
sigOut$pValue[2] = fpre_fpost$p.value

sigOut$t[3] = fpost_mpost$statistic # Female, post-test v male, post-test
sigOut$pValue[3] = fpost_mpost$p.value

sigOut$t[4] = mpre_mpost$statistic # Male, pre-test v male, post-test
sigOut$pValue[4] = mpre_mpost$p.value

for (i in 1:4) {
  if (sigOut$pValue[i] < 0.001) {
    sigOut$Significance[i] = "***"
  } else if (sigOut$pValue[i] < 0.01) {
    sigOut$Significance[i] = "**"
  } else if (sigOut$pValue[i] < 0.05) {
    sigOut$Significance[i] = "*"
  }
}

# Print statistical test results
print(sigOut)
```

```
##                   X                 Y          t     pValue Significance
## 1  Female, Pre-test    Male, Pre-test -0.9490637 0.34816263
## 2  Female, Pre-test Female, Post-test -1.8895807 0.06646876
## 3 Female, Post-test   Male, Post-test  0.2065868 0.83731411
## 4    Male, Pre-test   Male, Post-test  0.0000000 1.00000000
```

```r
# Calculate means and standard deviations for pre- and post-test by sex
gSE = data.frame(Test = rep(c("Pre", "Post"), 2), Gender = rep(c("Female", "Male"), each = 2), Mean = re

gSE$Mean[1] = mean((data_2017[data_2017$Sex == "Female", "Pre_CITotal"] / 14) * 100)
gSE$SD[1] = sd((data_2017[data_2017$Sex == "Female", "Pre_CITotal"] / 14) * 100)

gSE$Mean[2] = mean((data_2017[data_2017$Sex == "Female", "Post_CITotal"] / 14) * 100)
gSE$SD[2] = sd((data_2017[data_2017$Sex == "Female", "Post_CITotal"] / 14) * 100)

gSE$Mean[3] = mean((data_2017[data_2017$Sex == "Male", "Pre_CITotal"] / 14) * 100)
gSE$SD[3] = sd((data_2017[data_2017$Sex == "Male", "Pre_CITotal"] /  14) * 100)

gSE$Mean[4] = mean((data_2017[data_2017$Sex == "Male", "Post_CITotal"] / 14) * 100)
gSE$SD[4] = sd((data_2017[data_2017$Sex == "Male", "Post_CITotal"] / 14) * 100)

# Order variables for plotting
gSE$Test <- factor(gSE$Test, levels = c("Pre", "Post"), ordered = TRUE)

figS6a <- ggplot(gSE, aes(x = Test, y = Mean, fill = Gender)) +
  geom_bar(position = position_dodge(), stat = "identity") +
  geom_errorbar(aes(ymin = Mean - SD, ymax = Mean + SD), width = 0.3, position = position_dodge(0.9)) +
  scale_fill_manual("Students' Sex", values = c("magenta4", "green4"), labels = c("Women", "Men")) +
  scale_x_discrete(breaks = c("Pre", "Post"), labels = c("Pre-test", "Post-test")) +
  ylab("Total BEDCI score (%)") +
  theme_bw() +
```

```
  theme(panel.grid.minor = element_blank(), panel.grid.major.x = element_blank(), axis.title.x = elemen
  ggtitle("Figure S6A") + theme(plot.title = element_text(hjust = 0.5))

plot(figS6a)
```

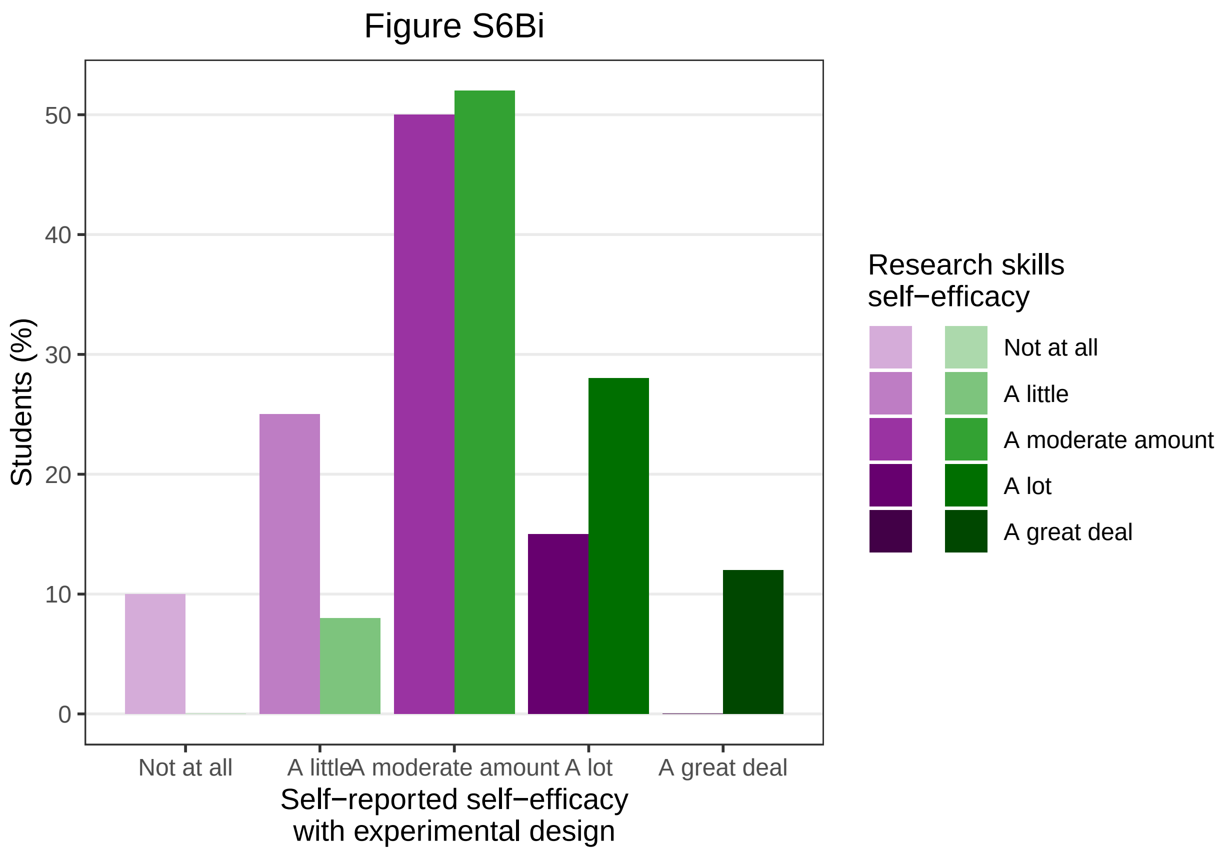



```
ggsave(figS6a, file="./Figures/FigS6A.svg", width = 3, height = 5)
```

Figure S6Bi

```
# Getresponses for each of the five Likert scale responses for Question 5 on the self-efficacy survey (
fQ5_pre = (as.vector(table(factor(data_2017[(data_2017$Sex == "Female"),"Pre_Q5"], levels = 1:5))) / nr
mQ5_pre = (as.vector(table(factor(data_2017[(data_2017$Sex == "Male"),"Pre_Q5"], levels = 1:5))) / nrow
Q5_pre_freq = c(fQ5_pre, mQ5_pre)

# Create data frame with frequency of responses (pre-test values)
Q5_pre = data.frame(lev = rep(1:5, 2), value = Q5_pre_freq, gender = rep(c("Female", "Male"), each = 5)
Q5_pre$col <- factor(Q5_pre$col , levels = c("1", "2", "3", "4", "5", "6", "7", "8", "9", "10"), ordere
Q5_pre$gender <- factor(Q5_pre$gender, levels = c("Female", "Male"), ordered = TRUE)

# Get responses for each of the five Likert scale responses for Question 5 on the self-efficacy survey
fQ5_post = (as.vector(table(factor(data_2017[(data_2017$Sex == "Female"),"Post_Q5"], levels = 1:5))) / 
mQ5_post = (as.vector(table(factor(data_2017[(data_2017$Sex == "Male"),"Post_Q5"], levels = 1:5))) / nr
Q5_post_freq = c(fQ5_post, mQ5_post)
```

```
# Create data frame with frequency of responses (post-test values)
Q5_post = data.frame(lev = rep(1:5, 2), value = Q5_post_freq, gender = rep(c("Female", "Male"), each = 5
Q5_post$col <- factor(Q5_post$col , levels = c("1", "2", "3", "4", "5", "6", "7", "8", "9", "10"), orde
Q5_post$gender <- factor(Q5_post$gender, levels = c("Female", "Male"), ordered = TRUE)

# Histogram
figS6bi <- ggplot(Q5_pre, aes(x=lev, y=value)) +
  geom_bar(aes(fill = col), stat = "identity", position = position_dodge()) +
  theme_bw() +
  scale_fill_manual("Research skills\nself-efficacy", values = c("1" = "#D5ACD9", "2" = "#BE7DC4", "3" =
  guides(fill=guide_legend(ncol = 2)) +
  scale_x_continuous("Self-reported self-efficacy\nwith experimental design", breaks=c(1:5) ,limits=c(0
  ylab("Students (%)") +
  theme(panel.grid.minor = element_blank(), panel.grid.major.x = element_blank()) +
  ggtitle("Figure S6Bi") + theme(plot.title = element_text(hjust = 0.5))

plot(figS6bi)
```




```
ggsave(figS6bi, file="./Figures/FigS6Bi.svg", width = 5, height = 3)
```

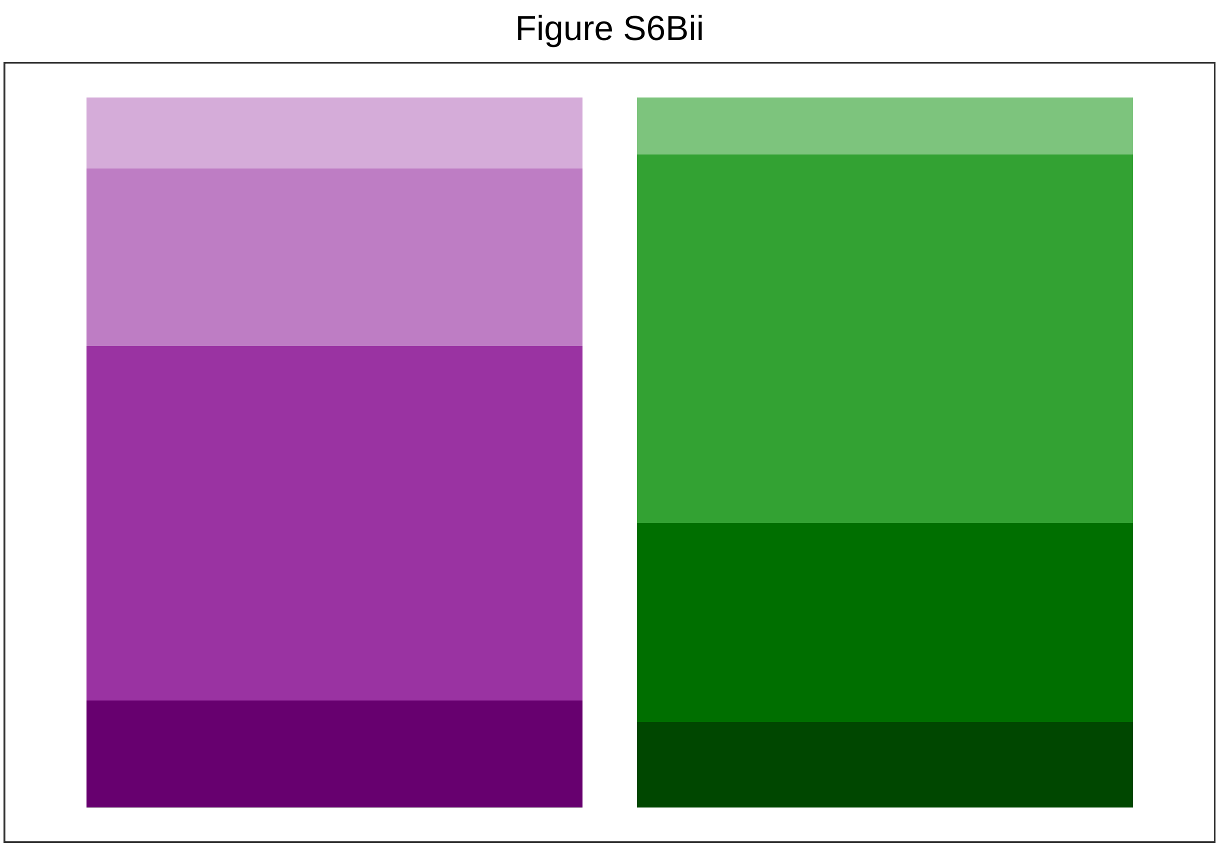
Figure S6Bii

```
# Statistical test
ordLogRegTest = as.vector(ordLogReg(data_2017, 5, "pre")$coefficients)[c(6,8)] # t value and p value
```

```r
# Significance
sig1 = "is NOT a statistically significant difference"
sig2 = ""

if (ordLogRegTest[2] < 0.001) {
  sig1 = "IS a statistically significant difference"
  sig2 = "***"
} else if (ordLogRegTest[2] < 0.01) {
  sig1 = "IS a statistically significant difference"
  sig2 = "**"
} else if (ordLogRegTest[2] < 0.05) {
  sig1 = "IS a statistically significant difference"
  sig2 = "*"
}

# Print output
cat(paste("By ordinal logistic regression, there ", sig1, " between male and female student responses o
```

```
## By ordinal logistic regression, there IS a statistically significant difference between male and fema
```

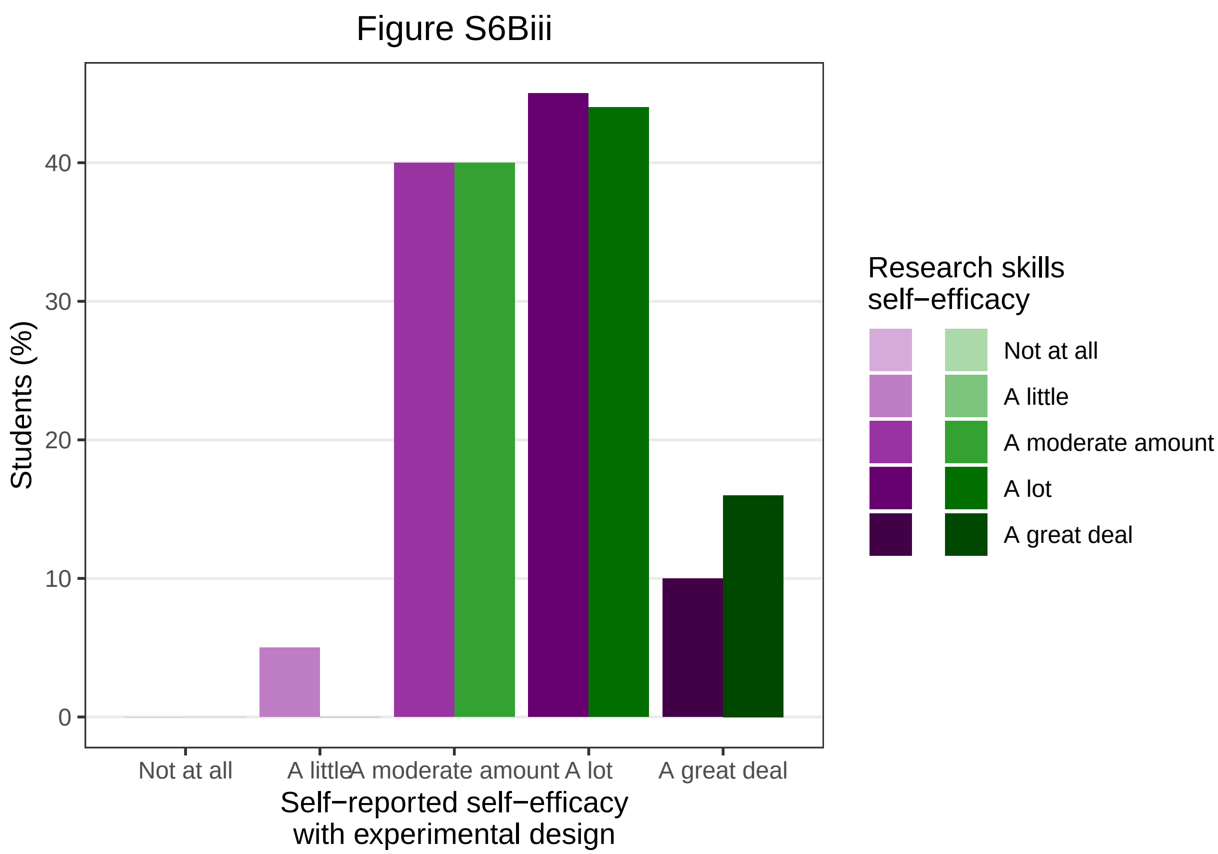
```r
# Stacked bar chart
figS6bii <- ggplot(Q5_pre, aes(x=gender, y=value)) +
  geom_bar(aes(fill=col), stat = "identity", position = "fill") +
  theme_bw() +
  scale_fill_manual("legend", values = c("1" = "#D5ACD9", "2" = "#BE7DC4", "3" = "#9A33A2", "4" = "#670
  theme(panel.grid.minor = element_blank(), panel.grid.major = element_blank(), axis.title.x=element_bl
  ggtitle("Figure S6Bii") + theme(plot.title = element_text(hjust = 0.5))

plot(figS6bii)
```




```
ggsave(figS6bii, file="./Figures/FigS6Bii.svg", width = 1, height = 5)
```

Figure S6Biii

```
# Histogram
figS6biii <- ggplot(Q5_post, aes(x=lev, y=value)) +
  geom_bar(aes(fill = col), stat = "identity", position = position_dodge()) +
  theme_bw() +
  scale_fill_manual("Research skills\nself-efficacy", values = c("1" = "#D5ACD9", "2" = "#BE7DC4", "3" =
  guides(fill=guide_legend(ncol = 2)) +
  scale_x_continuous("Self-reported self-efficacy\nwith experimental design", breaks=c(1:5) ,limits=c(0
  ylab("Students (%)") +
  theme(panel.grid.minor = element_blank(), panel.grid.major.x = element_blank()) +
  ggtitle("Figure S6Biii") + theme(plot.title = element_text(hjust = 0.5))

plot(figS6biii)
```




```
ggsave(figS6biii, file="./Figures/FigS6Biii.svg", width = 5, height = 3)
```

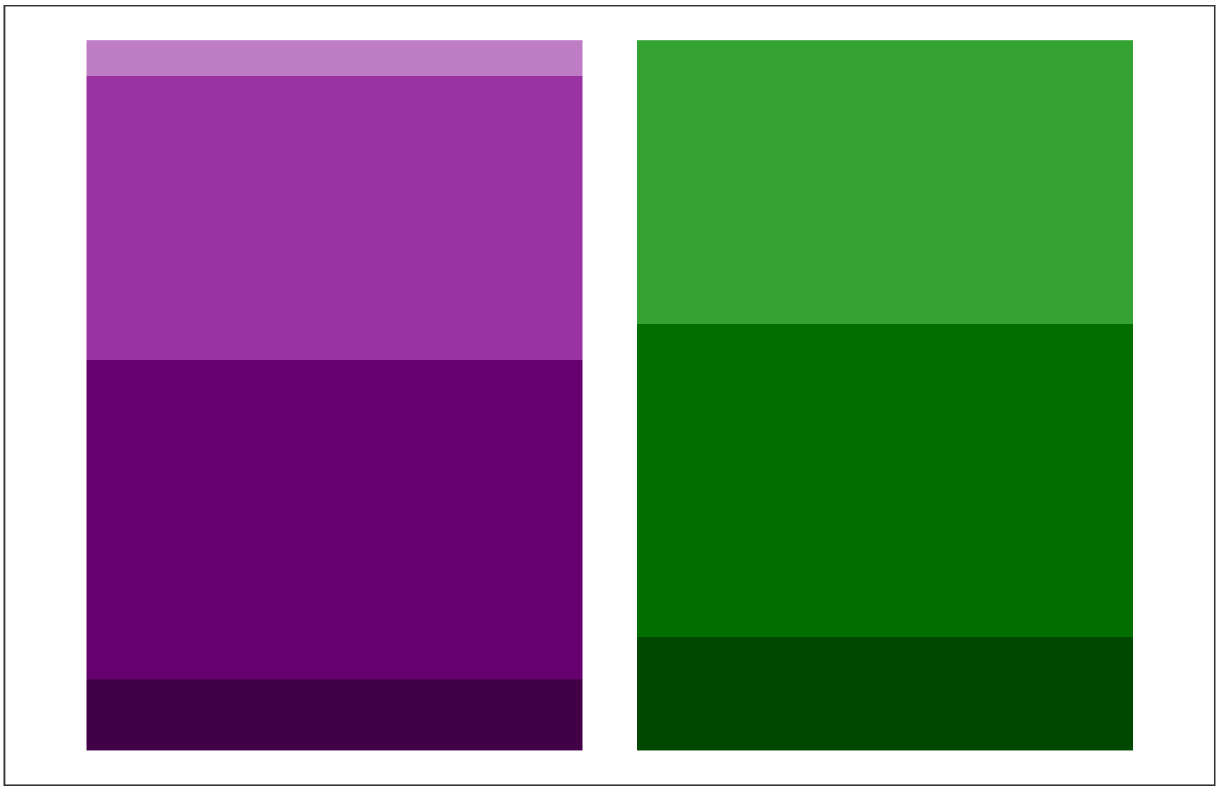
Figure S6Biv

```
# Statistical test
ordLogRegTest = as.vector(ordLogReg(data_2017, 5, "post")$coefficients)[c(6,8)] # t value and p value

# Significance
sig1 = "is NOT a statistically significant difference"
sig2 = ""

if (ordLogRegTest[2] < 0.001) {
  sig1 = "IS a statistically significant difference"
  sig2 = "***"
} else if (ordLogRegTest[2] < 0.01) {
  sig1 = "IS a statistically significant difference"
  sig2 = "**"
} else if (ordLogRegTest[2] < 0.05) {
  sig1 = "IS a statistically significant difference"
  sig2 = "*"
}

# Print output
cat(paste("By ordinal logistic regression, there ", sig1, " between male and female student responses o
```

```
## By ordinal logistic regression, there is NOT a statistically significant difference between male and
```

```
# Stacked bar chart
figS6biv <- ggplot(Q5_post, aes(x=gender, y=value)) +
  geom_bar(aes(fill=col), stat = "identity", position = "fill") +
  theme_bw() +
  scale_fill_manual("legend", values = c("1" = "#D5ACD9", "2" = "#BE7DC4", "3" = "#9A33A2", "4" = "#670
  theme(panel.grid.minor = element_blank(), panel.grid.major = element_blank(), axis.title.x=element_bla
  ggtitle("Figure S6Biv") + theme(plot.title = element_text(hjust = 0.5))

plot(figS6biv)
```

Figure S6Biv



```
ggsave(figS6biv, file="./Figures/FigS6Biv.svg", width = 1, height = 5)
```

Table S1

Demographics of student participants in the study from 2017 and 2018.

```
# Gender
mat = matrix(c(data3['Women', 'Year2017'], data3['Women', 'Year2018'], data3['Men', 'Year2017'], data3[
cat("GENDER\n")
```

```
## GENDER
```

```r
prop.test(t(mat), correct=FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  t(mat)
## X-squared = 1.8713, df = 1, p-value = 0.1713
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.32975834  0.05732244
## sample estimates:
##    prop 1    prop 2
## 0.3846154 0.5208333
```

```r
# Race
mat = matrix(c(data3['URM', 'Year2017'], data3['URM', 'Year2018'], data3['NonURM', 'Year2017'], data3['
cat("RACE\n")
```

```
## RACE
```

```r
prop.test(t(mat), correct=FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  t(mat)
## X-squared = 0.16267, df = 1, p-value = 0.6867
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.3437991  0.2262256
## sample estimates:
##    prop 1    prop 2
## 0.4615385 0.5203252
```

```r
# Previous years of research lab experience
mat = matrix(c(as.matrix(table(data_2017$LabExp)), as.matrix(table(factor(data_2018$LabExp, levels = 1:
cat("LAB EXPERIENCE\n")
```

```
## LAB EXPERIENCE
```

```r
chisq.test(mat)
```

```
## Warning in chisq.test(mat): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  mat
## X-squared = 12.59, df = 6, p-value = 0.05002
```

```r
# Previous degree subject
mat = matrix(c(colSums(data_2017[,3:9]), colSums(data_2018[,3:9])), nrow = 7, dimnames = list("Degree" =
cat("DEGREE\n")
```

```
## DEGREE
```

```r
chisq.test(mat)
```

```
## Warning in chisq.test(mat): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  mat
## X-squared = 6.0568, df = 6, p-value = 0.4169
```

```r
# Program
mat = matrix(c(data3['BSPH', 'Year2017'], data3['BIG', 'Year2017'], data3['BBS', 'Year2017'], data3['Bi
cat("PROGRAM\n")
```

```
## PROGRAM
```

```r
chisq.test(mat)
```

```
## Warning in chisq.test(mat): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  mat
## X-squared = 13.209, df = 6, p-value = 0.03984
```

```r
# Experience with experimental design
cat("EXPERIENCE\n")
```

```
## EXPERIENCE
```

```r
wilcox.test(x = data_2017$ExperimetalDeignExperience, y = data_2018$ExperimetalDeignExperience, alterna
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  data_2017$ExperimetalDeignExperience and data_2018$ExperimetalDeignExperience
## W = 1277, p-value = 0.849
## alternative hypothesis: true location shift is not equal to 0
```

```r
# Comfort with experimental design
cat("COMFORT\n")
```

```
## COMFORT
```

```r
wilcox.test(x = data_2017$ExperimetalDeignComfort, y = data_2018$ExperimetalDeignComfort, alternative =
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  data_2017$ExperimetalDeignComfort and data_2018$ExperimetalDeignComfort
## W = 1306.5, p-value = 0.9946
## alternative hypothesis: true location shift is not equal to 0
```

```r
# Pre-test total self-efficacy in research skills
cat("PRE SELF-EFFICACY\n")
```

```
## PRE SELF-EFFICACY
```

```r
wilcox.test(x = melt(data_2017[,14:27])$value, y = melt(data_2018[,14:27])$value, alternative = "two.si
```

```
## No id variables; using all as measure variables
## No id variables; using all as measure variables
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  melt(data_2017[, 14:27])$value and melt(data_2018[, 14:27])$value
## W = 249748, p-value = 0.4159
## alternative hypothesis: true location shift is not equal to 0
```

```r
# Post-test total self-efficacy in research skills
cat("POST SELF-EFFICACY\n")
```

```
## POST SELF-EFFICACY
```

```r
wilcox.test(x = melt(data_2017[,28:41])$value, y = melt(data_2018[,28:41])$value, alternative = "two.si
```

```
## No id variables; using all as measure variables
## No id variables; using all as measure variables
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  melt(data_2017[, 28:41])$value and melt(data_2018[, 28:41])$value
## W = 255055, p-value = 0.9221
## alternative hypothesis: true location shift is not equal to 0
```

```r
# Net change in research skills self-efficacy
cat("DELTA SELF-EFFICACY\n")
```

```
## DELTA SELF-EFFICACY
```

```
wilcox.test(x = (melt(data_2017[,28:41])$value - melt(data_2017[,14:27])$value), y = (melt(data_2018[,28
```

```
## No id variables; using all as measure variables
## No id variables; using all as measure variables
## No id variables; using all as measure variables
## No id variables; using all as measure variables
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  (melt(data_2017[, 28:41])$value - melt(data_2017[, 14:27])$value) and (melt(data_2018[, 28:41]
## W = 262874, p-value = 0.3343
## alternative hypothesis: true location shift is not equal to 0
```

Table S2

Students significantly improved in many aspects of research skills self-efficacy during their first semester of graduate school.

```
# Note: this code was already included for the creation of Figure 3B
# Calculate the p-value for each question and save to an output table
Items = c("Understand contemporary concepts in your field", "Make use of the primary scientific research
sigOut = data.frame(Item = Items, Question = 1:14, V = rep(0, 14), pValue = rep(0, 14), Significance = r
for (i in 1:14) {
  wTest = wilcoxonSignedRankTest(data, i)
  sigOut$V[i] = wTest$statistic
  sigOut$pValue[i] = round(wTest$p.value, 4)
}

wTest_Total = wilcox.test(melt(SE_pre)$value, melt(SE_post)$value, paired=TRUE, alternative = "two.side
```

```
## Using Sex as id variables
## Using Sex as id variables
```

```
tmp = data.frame(Item = "Total", Question = 15, V = wTest_Total$statistic, pValue = round(wTest_Total$p
sigOut[15,] = tmp

pValues = p.adjust(sigOut$pValue, method = "fdr")
for (i in 1:15) {
  sigOut$pAdj[i] = round(pValues[i], 4)
  if (sigOut$pAdj[i] < 0.001) {
    sigOut$Significance[i] = "***"
  } else if (sigOut$pAdj[i] < 0.01) {
    sigOut$Significance[i] = "**"
  } else if (sigOut$pAdj[i] < 0.05) {
    sigOut$Significance[i] = "*"
  }
}

# Print signficance table
print(sigOut)
```

```
##                                                                                         Item
## 1                                                   Understand contemporary concepts in your field
## 2  Make use of the primary scientific research literature in your field (e.g., journal articles)
## 3            Identify a specific question for investigation based on the research in your field
## 4                                         Formulate a research hypothesis based on a specific question
## 5                                         Design an experiment or theoretical test of the hypothesis
## 6                                                  Understand the importance of 'controls' in research
## 7                                                                              Observe and collect data
## 8                                                                             Statistically analyze data
## 9                                         Interpret data by relating results to the original hypothesis
## 10                                        Reformulate your original research hypothesis (as appropriate)
## 11                                                Relate your results to the 'bigger picture' in your field
## 12                                                    Orally communicate the results of research projects
## 13                                                              Write a research paper for publication
## 14                                                                                 Think independently
## V                                                                                                Total
##    Question       V pValue Significance   pAdj
## 1         1   524.5 0.0071            * 0.0118
## 2         2   488.0 0.0784              0.0905
## 3         3   677.0 0.0164            * 0.0224
## 4         4   429.5 0.0000          *** 0.0000
## 5         5   339.5 0.0000          *** 0.0000
## 6         6   389.0 0.0033           ** 0.0062
## 7         7   607.5 0.3072              0.3072
## 8         8   498.0 0.0004           ** 0.0012
## 9         9   385.5 0.0028           ** 0.0060
## 10       10   286.0 0.0000          *** 0.0000
## 11       11   506.0 0.1198              0.1284
## 12       12   413.0 0.0024           ** 0.0060
## 13       13   411.0 0.0208            * 0.0260
## 14       14   619.0 0.0114            * 0.0171
## V        15 31223.5 0.0000          *** 0.0000
```

Table S3

Students significantly improved in concept inventory questions relating to controls, hypotheses, biological variation, and accuracy.

```
# Note: this code was already included for the creation of Figure 2C
# Calculate the p-value for each question and save to an output table
coreConcepts = c("Controls", "", "Hypotheses", "", "Biological variation", "", "Accuracy", "Extraneous 
questions = c(1, 5, 2, 9, 3, 10, 4, 6, 14, 7, 12, 8, 13, 11) # Questions ordered by subject
sigOut = data.frame(CoreConcept = coreConcepts, Question = questions, Chi = rep(0, 14), pValue = rep(0,
i = 1 # Iterate over every row of output
for (q in questions) {
  mnTest = mcNemarChiSquaredTest(data_2017, q)
  sigOut$Chi[i] = round(mnTest$statistic, 2)
  sigOut$pValue[i] = round(mnTest$p.value, 4)
  i = i + 1
}

pValues = p.adjust(sigOut$pValue, method = "fdr")
```

```
for (i in 1:14) {
  sigOut$pAdj[i] = round(pValues[i], 4)
  if (sigOut$pAdj[i] < 0.001) {
    sigOut$Significance[i] = "***"
  } else if (sigOut$pAdj[i] < 0.01) {
    sigOut$Significance[i] = "**"
  } else if (sigOut$pAdj[i] < 0.05) {
    sigOut$Significance[i] = "*"
  }
}

# Print signficance table
print(sigOut) # Also Table S3
```

```
##                CoreConcept Question  Chi pValue Significance   pAdj
## 1                 Controls        1 4.00 0.0455              0.1592
## 2                                 5 0.10 0.7518              1.0000
## 3               Hypotheses        2 5.88 0.0153              0.1592
## 4                                 9 0.00 1.0000              1.0000
## 5     Biological variation        3 0.00 1.0000              1.0000
## 6                                10 4.00 0.0455              0.1592
## 7                 Accuracy        4 4.92 0.0265              0.1592
## 8       Extraneous factors        6 0.12 0.7237              1.0000
## 9                                14 0.12 0.7237              1.0000
## 10    Independent sampling        7 0.00 1.0000              1.0000
## 11                               12 0.75 0.3865              1.0000
## 12         Random sampling        8 0.00 1.0000              1.0000
## 13                               13 0.00 1.0000              1.0000
## 14 Purpose of experiments       11 0.08 0.7728              1.0000
```

Table S4

Backgrounds of women and men participants in the study.

```
# Split data by sex
data_women = data[data$Sex == "Female",]
data_women = data_women[!is.na(data_women$Sex),]
data_men = data[data$Sex == "Male",]
data_men = data_men[!is.na(data_men$Sex),]

# Race
mat = matrix(c(data4['URM', 'Women'], data4['URM', 'Men'], data4['NonURM', 'Women'], data4['NonURM', 'Me
cat("RACE\n")
```

```
## RACE
```

```
prop.test(t(mat), correct=FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
```

```
##  correction
##
## data:  t(mat)
## X-squared = 0.00016644, df = 1, p-value = 0.9897
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.2868378  0.2830854
## sample estimates:
##    prop 1    prop 2
## 0.4615385 0.4634146
```

```
# Previous years of research lab experience
mat = matrix(c(as.matrix(table(data_women$LabExp)), as.matrix(table(factor(data_women$LabExp, levels = 
cat("LAB EXPERIENCE\n")
```

```
## LAB EXPERIENCE
```

```
chisq.test(mat)
```

```
## Warning in chisq.test(mat): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  mat
## X-squared = 14.11, df = 6, p-value = 0.02844
```

```
# Previous degree subject
mat = matrix(c(colSums(data_women[,3:9]), colSums(data_men[,3:9])), nrow = 7, dimnames = list("Degree" 
cat("DEGREE\n")
```

```
## DEGREE
```

```
chisq.test(mat)
```

```
## Warning in chisq.test(mat): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  mat
## X-squared = 12.209, df = 6, p-value = 0.05746
```

```
# Program
mat = matrix(c(data4['BSPH', 'Women'], data4['BIG', 'Women'], data4['BBS', 'Women'], data4['BioP', 'Wom
cat("PROGRAM\n")
```

```
## PROGRAM
```

```r
chisq.test(mat)
```

```
## Warning in chisq.test(mat): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  mat
## X-squared = 83.836, df = 6, p-value = 5.751e-16
```

```r
# Experience with experimental design
cat("EXPERIENCE\n")
```

```
## EXPERIENCE
```

```r
wilcox.test(x = data_women$ExperimetalDeignExperience, y = data_men$ExperimetalDeignExperience, alterna
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  data_women$ExperimetalDeignExperience and data_men$ExperimetalDeignExperience
## W = 1174.5, p-value = 0.5997
## alternative hypothesis: true location shift is not equal to 0
```

```r
# Comfort with experimental design
cat("COMFORT\n")
```

```
## COMFORT
```

```r
wilcox.test(x = data_women$ExperimetalDeignComfort, y = data_men$ExperimetalDeignComfort, alternative =
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  data_women$ExperimetalDeignComfort and data_men$ExperimetalDeignComfort
## W = 1217, p-value = 0.8288
## alternative hypothesis: true location shift is not equal to 0
```

```r
# Pre-test total self-efficacy in research skills
cat("PRE SELF-EFFICACY\n")
```

```
## PRE SELF-EFFICACY
```

```r
wilcox.test(x = melt(data_women[,14:27])$value, y = melt(data_women[,14:27])$value, alternative = "two.
```

```
## No id variables; using all as measure variables
## No id variables; using all as measure variables
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  melt(data_women[, 14:27])$value and melt(data_women[, 14:27])$value
## W = 264992, p-value = 1
## alternative hypothesis: true location shift is not equal to 0
```

```
# Post-test total self-efficacy in research skills
cat("POST SELF-EFFICACY\n")
```

```
## POST SELF-EFFICACY
```

```
wilcox.test(x = melt(data_men[,28:41])$value, y = melt(data_men[,28:41])$value, alternative = "two.side
```

```
## No id variables; using all as measure variables
## No id variables; using all as measure variables
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  melt(data_men[, 28:41])$value and melt(data_men[, 28:41])$value
## W = 225792, p-value = 1
## alternative hypothesis: true location shift is not equal to 0
```

```
# Net change in research skills self-efficacy
cat("DELTA SELF-EFFICACY\n")
```

```
## DELTA SELF-EFFICACY
```

```
wilcox.test(x = (melt(data_women[,28:41])$value - melt(data_women[,14:27])$value), y = (melt(data_men[,2
```

```
## No id variables; using all as measure variables
## No id variables; using all as measure variables
## No id variables; using all as measure variables
## No id variables; using all as measure variables
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  (melt(data_women[, 28:41])$value - melt(data_women[, 14:27])$value) and (melt(data_men[, 28:4
## W = 256806, p-value = 0.08496
## alternative hypothesis: true location shift is not equal to 0
```