Hypothesis: 1

Analyst 2

Reviewer: Reviewer 3

Co-reviewer: Reviewer 2

Error check 1: Gross analytical errors?

Did you detect any gross or obvious errors in the final empirical model submitted by the analyst to test this hypothesis?

	No, I did not detect any gross or obvious errors in the final empirical model.	*Yes, I did detect any gross or obvious errors in the final empirical model*
Reviewer	**X**	
Co-reviewer	X	

Do you have any additional comments about your short assessment?

- No, I don't have any additional comments.

Co reviewer, do you have any additional comments?

- No additional comments

Error check 2: identifying analytical components final model

Number of observations used in final empirical model and analysis: n = 65

Females = 65

Males = 0

Unit of analysis: Combination of Threads (*ThreadId*) and female participants (*n_female*). However, the participant identifier *n_female* does not contain their *UserId* but just the order of their first post in a given thread.

1. High level statement of the analysis:

Analyst 2 has reported a main effect regression coefficient of value -1.3152 for Hypothesis 1, and a p value of 0.0429

2. Verbal interpretation of the result:

For a 1-unit change in “n_females” contributors, the “n_posts” variable decreases by 1.3152 units.

3. Data filters:

The analyst has used the edge dataset with filters applied to:

- Include only communication of type “conversation”
- “gender” identified as female
- “n_females” in a conversation block > 0 (Include only threads where there is more than one female)

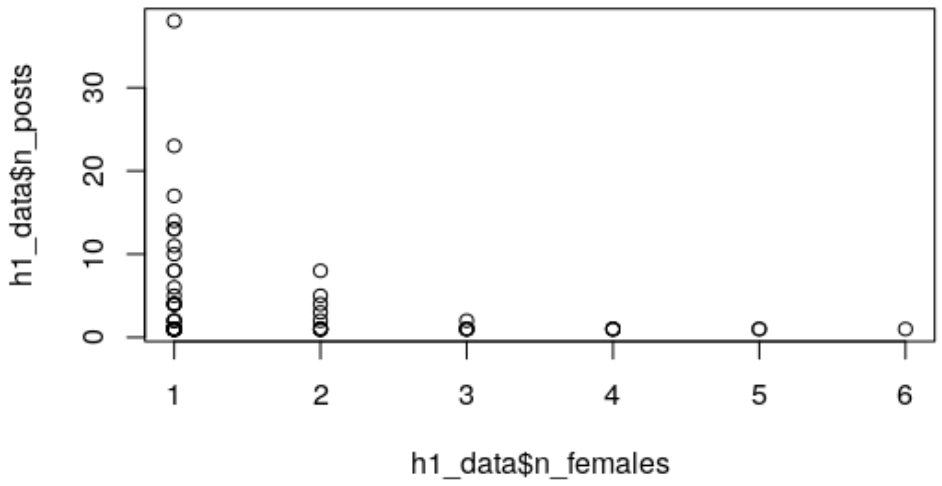
4. Dependent variable operationalization:

- The dependent variable “n_posts” is not an original variable from the Edge dataset.
- It represents the total number of posts each female contributor makes in each conversation.
- For example, conversation #99 might have
    - Female A contributing 5 times
    - Female B contributing 10 times
    - Female C contributing 4 times
    - Female D contributing 3 times
- Specifically, the analyst operationalizes it through the following 10 steps:
    1. The data is filtered such that communication type is conversation, gender is identified and gender is female.
    2. The data is sorted by (Year, ThreadId, Id_num, Order)
    3. The data is grouped by (Year, ThreadId, Id_num)
    4. A variable “first_post” is added, it equals “1” for the first entry in this grouped data, and zero otherwise. (Add one for the first post of each female in each thread, otherwise add zero)
    5. The data is then ungrouped and arranged again by (Year, ThreadId, Order)
    6. The data is re-grouped by (Year, ThreadId)

7. A variable “n_females” is added. It is defined as a cumulative sum of “first_post” variable, with a lag=1. (It indicates how many females have posted in the same thread before)
8. Replace missing values with zeros in “n_females”.
9. The data is ungrouped, and then re-grouped by: (Year, ThreadId, n_females)
10. “n_posts” is defined as the number of rows per group

5. Independent variables Operationalization:

- The independent variable “n_female” is not an original variable from the Edge dataset.
- It represents the number of female contributors organized by the order in which each female contributor posted in each conversation.
- For example, conversation #99 (see 4. Above for details on this) might show that
  - A was the 1st female to contribute (the dependent variable shows she made 5 contributions)
  - B was the 2nd female to contribute (the dependent variable shows she made 10 contributions)
  - C was the 3rd female to contribute (the dependent variable shows she made 4 contributions)
  - D was the 4th female to contribute (the dependent variable shows she made 3 contributions)
- Plotting the independent variable and the dependent variable results in a graph like this:



- Please note: within a conversation, the 1st (or 2nd or 3rd) female contributor may have contributed less than subsequent female contributors. The analyst only *assigns values for the independent variable based on when a female contributor first contributed*. A visual inspection of the plot seems to suggest that subsequent female contributors contribute less which would be in line with the analyst's results.

- Specifically, the analyst operationalizes it through steps 1-8 as described in the "Dependent variable operationalization" above.

Verbal summary of code submitted by analyst:

In #	Main Task	Var Name	Code Description
3 - 4	Import packages	-	Import two libraries
9	Read the data	edge	Read the CSV file "edge1.1"
25	Select relevant columns	h1data	Select these columns from the datafile (Year, Title, Type, ThreadId, Id, Id_num, Role, Female, Order).
26 - 30	Identifies women's order in conversation	h1_first_arrived	Filter the data where: (1) "Type" variable = 2 (2) No missing values in "Female" variable (3) "Female" variable = 1 "Filter where communication type is conversation, and gender is female"
31		h1_first_arrived	Arrange the data based on the columns: (Year, ThreadId, Order)
33-37	Restructure the dataset to get #female contributors and #posts in each "conversational block"	h1_data	Filter the data where: (1) "Type" variable = 2 (2) No missing values in "Female" variable (3) "Female" variable = 1
38		h1_data	Arrange the data based on the columns: (Year, ThreadId, Id_num, Order)
39		h1_data	Group it by: (Year, ThreadId, Id_num)
40		h1_data	Add first_post vector with "1" for the first entry and "0" otherwise (In each thread, mark the first post for each female participant)
41		h1_data	Ungroup (h1_data)
42		h1_data	Arrange the data based on these columns instead: (Year, ThreadId, Order)
43		h1_data	Group it by: (Year, ThreadId)
44		h1_data	Add a new variable "n_females" = lag(cumulative sum for "first_posts") (Count the number of females that posted in the thread before this entry)
45		h1_data	Replace the missing values in "n_females" with zero

47		h1_data	Ungroup (h1_data)
48		h1_data	Group it by: (Year, ThreadId, n_females)
49		h1_data	Get the number of posts by summing the number rows per group.
50		h1_data	Ungroup (h1_data)
51		h1_data	Filter the data where: “n_females” > 0 (Consider only threads with more than one female participants)
54	Fit linear regression model	h1_result	Fit linear regression model between “n_posts” and “n_females”, that are in (h1_data)
114	Print model summary		Print “h1_result” summary

Appendix 1: Minimal code to reproduce analyst’s result:

File: [H1 Analyst 2_minimal]

```
01: #added libraries used:
02: library(dplyr)
03: library(effsize)
04: # ------------------------------
05: # Data from edge, effectively a convience sample of "communication" or a cenus
of the site, depending perspective
06: edge <-
read.csv("~/Documents/work/CDA-2/Boba/MinimalCodes/Minimal-Code-Sample/edge1.1_anon
ymized.csv", stringsAsFactors = FALSE)
07: # ------------------------------
08: h1_data <- edge %>%
09:   filter(Type == 2, #communications of type conversation
10:         !is.na(Female), #gender is identified
11:         Female == 1 #gender is Female
12:   ) %>%
13:   arrange(Year, ThreadId, Id_num, Order) %>%
14:   group_by(Year, ThreadId, Id_num) %>% #for each person in thread in year in
order
15:   mutate(first_post = c(1, rep(0,n()-1)))  %>% # mark 1st post 1, others 0
16:   ungroup() %>%
17:   arrange(Year, ThreadId, Order) %>%
18:   group_by(Year, ThreadId) %>% #for each thread in each year in order
19:   mutate(n_females = lag(cumsum(first_post)), #the cumlative sum at the entry
```

```
above the
20:         n_females = ifelse(is.na(n_females), 0, n_females) #current one is the
number
21:  ) %>%   #of females at post time, so shifted down to current
22:  ungroup() %>%
23:   group_by(Year, ThreadId, n_females) %>% #for each sequence of the same number
of females
24:   summarise(n_posts= n()) %>% #calculate the amount of activity
25:   ungroup() %>%
26:   filter(n_females > 0) #posts with no females can only have one post from a
female before
27: # they become posts with one female, so discarding them as there is no
variation.
28: # with a range of numeric values and numeric outcomes, I'm doinga linear
regression
29: h1_result <- lm(n_posts ~ n_females, data=h1_data)
30: # ---------------------------------
31: ### reporting
32: # summary(h1_result)
```

Appendix 2: Raw analyst code:

File: [Analyst 2]

```
001: # Pretty final, well annotated script with just the needed parts in it
002: # ---------------------------------
003: # added libraries used:
004: library(dplyr)
005: library(effsize)
006: # ---------------------------------
007: # Data from edge, effectively a convience sample of "communication" or a cenus of
the site, depending perspective
008: # ---------------------------------
009: edge <- read.csv("edge1.1.csv", stringsAsFactors = FALSE)
010: # ---------------------------------
011: # Hypothesis 1: “A woman’s tendency to participate actively in the conversation
correlates positively with the number of females in the discussion.”
012: # For this problem I am treating "participate actively" as posting messages when I
waman has identified as present
013: # by posting at least once.
014: # I am treating this as a regression problem as of conversation segments defined by
an increase in the number of women present
```

```
# -------------------------------
# If a A woman’s tendency to participate actively in the conversation correlates positively with the number of females in the discussion,
# then conversational segments marked by having more females enter should have higher female activity
# I am using conversation segments as I do not think knowing the number of females at the end of the thread should be affecting the
# amount of discussion at the start
# -------------------------------
# to do this I need to use data for which the order indicates the order of arrival (not annual questions) and restucture
# the data set into conversational blocks rather than messages in conversations
# -------------------------------
# for each women, identifies the entry when they first arrive in a conversation
h1data <- edge %>% select(Year, Title, Type, ThreadId, Id, Id_num, Role, Female, Order)
h1_first_arrived <- edge %>% #start with edge data
 filter(Type == 2, #communications of type conversation
        !is.na(Female), #gender is identified
        Female == 1 #gender is Female
 ) %>%
 arrange(Year, ThreadId, Order)
# -------------------------------
h1_data <- edge %>%
  filter(Type == 2, #communications of type conversation
         !is.na(Female), #gender is identified
         Female == 1 #gender is Female
  ) %>%
  arrange(Year, ThreadId, Id_num, Order) %>%
  group_by(Year, ThreadId, Id_num) %>% #for each person in thread in year in order
  mutate(first_post = c(1, rep(0,n()-1)))  %>% # mark 1st post 1, others 0
  ungroup() %>%
  arrange(Year, ThreadId, Order) %>%
  group_by(Year, ThreadId) %>% #for each thread in each year in order
  mutate(n_females = lag(cumsum(first_post)), #the cumlative sum at the entry above the
         n_females = ifelse(is.na(n_females), 0, n_females) #current one is the number
 ) %>%                      #of females at post time, so shifted down to current
 ungroup() %>%
 group_by(Year, ThreadId, n_females) %>% #for each sequence of the same number of females
 summarise(n_posts= n()) %>% #calculate the amount of activity
```

```
050:  ungroup() %>%
051:  filter(n_females > 0) #posts with no females can only have one post from a female
before
052: # they become posts with one female, so discarding them as there is no
variation.
053: # with a range of numeric values and numeric outcomes, I'm doinga linear regression
054: h1_result <- lm(n_posts ~ n_females, data=h1_data)
055: # ------------------------------
056: ##########################################################
057 - 111: Hypothesis 2
112: # ------------------------------
113: ### reporting
114: summary(h1_result)
115 - 116: Hypothesis 2
```