Norming WebCDI

Benny deMayo

6/15/2020

```
#first we read in the WG data that is just from the SES pilot projects
all_wg_raw <- readInWebCDI(fb_wg_directory)

facebook_ws_raw <-
  readInWebCDI(fb_ws_directory) %>%
  select( #drop a bunch of columns that were screwing up the merge with prolific data
    -opt_out,
    -country,
    -sibling_boolean,
    -sibling_data,
    -sibling_count,
    -caregiver_other
  )

prolific_raw <-
  readInWebCDI(prolific_data_directory) %>%
  select(
    colnames(facebook_ws_raw), #drop columns that are prolific specific
    -opt_out, #drop a bunch of columns that were screwing up the merge with facebook data
    -country,
    -sibling_boolean,
    -sibling_data,
    -sibling_count,
    -caregiver_other
  )

#all of the ws data collected
all_ws_raw <-
  facebook_ws_raw %>%
  bind_rows(prolific_raw) %>%
  mutate(completed = case_when(
    stringr::str_to_lower(completed) == "true" ~ TRUE,
    stringr::str_to_lower(completed) == "false" ~ FALSE
  ))
```

```
#we have to read in the full dataset and select out the manually-coded columns
wg_exclusion_info <-
  read_csv(all_data_wg_path) %>%
  select(
    study_name,
    subject_id,
    hearing_exclude,
```

```
    vision_exclude,
    illnesses_exclude
  ) %>%
  mutate(subject_id = as.character(subject_id))

wg <-
  all_wg_raw %>%
  filter(completed == TRUE) %>% #only take completed administrations
  #join it with the exclusion data we have from the screened files
  left_join(
    wg_exclusion_info,
    by = c("study_name", "subject_id")
  ) %>%
  filter(repeat_num == "1") %>% #done
  filterBirthweight() %>% #done
  filterMultilingual() %>% #done
  filterIllnesses() %>%
  filterVision() %>%
  filterHearing() %>%
  getCompletionInterval() %>%
  getEthnicities() %>%
  getMaternalEd() %>%
  filter(completion_time >= min_completion_time) %>%
  filter_age_wg()

ws1_exclusion_info <-
  read_csv(all_data_ws1_path) %>%
   select(
    study_name,
    subject_id,
    hearing_exclude,
    vision_exclude,
    illnesses_exclude
  ) %>%
  mutate(subject_id = as.character(subject_id))

ws2_exclusion_info <-
  read_csv(all_data_ws2_path) %>%
   select(
    study_name,
    subject_id,
    hearing_exclude,
    vision_exclude,
    illnesses_exclude
  ) %>%
  mutate(subject_id = as.character(subject_id))

ws_all_exclusion_info <-
  bind_rows(ws1_exclusion_info, ws2_exclusion_info)

ws <-
  all_ws_raw %>%
  filter(completed == TRUE) %>%
```

```
  #join it with the exclusion data we have from the screened files
  left_join(
    ws_all_exclusion_info,
    by = c("study_name", "subject_id")
  ) %>%
  filter(repeat_num == "1") %>%
  filterBirthweight() %>%
  filterMultilingual() %>%
  filterIllnesses() %>%
  filterVision() %>%
  filterHearing() %>%
  getCompletionInterval() %>%
  getEthnicities() %>%
  getMaternalEd() %>%
  filter(completion_time >= min_completion_time) %>%
  filter_age_ws()

prolific_filtered_n <-
  ws %>%
  filter(str_detect(study_name, "prolific")) %>%
  nrow()
```

```
#Calculating exclusions

completed_wg <-
  all_wg_raw %>%
  filter(completed == TRUE) %>%
  left_join(
    wg_exclusion_info,
    by = c("study_name", "subject_id")
  ) %>%
  getCompletionInterval()


completed_ws <-
  all_ws_raw %>%
  filter(completed == TRUE) %>%
  left_join(
    ws_all_exclusion_info,
    by = c("study_name", "subject_id")
  ) %>%
  getCompletionInterval()
```

More fine-grained exclusion information.

```
#Disclaimer that this is some of the worst code I've ever written. Sorry everyone.

n_total_wg <- nrow(completed_wg)
n_total_ws <- nrow(completed_ws)

excl_col_names <-
  c(
    "Exclusion",
```

```
    "WG exclusions",
    "% of full WG sample excluded",
    "WS exclusions",
    "% of full WS sample excluded"
  )


#First take away kids who have done the survey more than once.
wg_minus_repeats <-
  completed_wg %>%
  filter(repeat_num == "1")

wg_repeats_n <- n_total_wg - nrow(wg_minus_repeats)

ws_minus_repeats <-
  completed_ws %>%
  filter(repeat_num == "1")

ws_repeats_n <- n_total_ws - nrow(ws_minus_repeats)

repeat_admins <-
  c(
    "Not first administration",
    wg_repeats_n,
    percent(wg_repeats_n / n_total_wg),
    ws_repeats_n,
    percent(ws_repeats_n / n_total_ws)
  )

names(repeat_admins) <- excl_col_names

#Next take away kids born pre-term or with low birthweight.

wg_minus_premie <-
  wg_minus_repeats %>%
  filterBirthweight()

wg_premie_n <- nrow(wg_minus_repeats) - nrow(wg_minus_premie)

ws_minus_premie <-
  ws_minus_repeats %>%
  filterBirthweight()

ws_premie_n <- nrow(ws_minus_repeats) - nrow(ws_minus_premie)

premies <-
  c(
    "Premature or low birthweight",
    wg_premie_n,
    percent(wg_premie_n / n_total_wg),
    ws_premie_n,
    percent(ws_premie_n / n_total_ws)
  )
```

```r
names(premies) <- excl_col_names

#Next take away kids with multilingual exposure

wg_minus_multiling <-
  wg_minus_premie %>%
  filterMultilingual()

wg_multiling_n <- nrow(wg_minus_premie) - nrow(wg_minus_multiling)

ws_minus_multiling <-
  ws_minus_premie %>%
  filterMultilingual()

ws_multiling_n <- nrow(ws_minus_premie) - nrow(ws_minus_multiling)

multiling <-
  c(
    "Multilingual exposure",
    wg_multiling_n,
    percent(wg_multiling_n / n_total_wg),
    ws_multiling_n,
    percent(ws_multiling_n / n_total_ws)
  )

names(multiling) <- excl_col_names

#Next exclude kids with problems of illness, vision, or hearing
wg_minus_health <-
  wg_minus_multiling %>%
  filterIllnesses() %>%
  filterVision() %>%
  filterHearing()

wg_health_n <- nrow(wg_minus_multiling) - nrow(wg_minus_health)

ws_minus_health <-
  ws_minus_multiling %>%
  filterIllnesses() %>%
  filterVision() %>%
  filterHearing()

ws_health_n <- nrow(ws_minus_multiling) - nrow(ws_minus_health)

health <-
  c(
    "Illnesses/Vision/Hearing",
    wg_health_n,
    percent(wg_health_n / n_total_wg),
    ws_health_n,
    percent(ws_health_n / n_total_ws)
  )
```

```r
names(health) <- excl_col_names

#Now filter out kids who are the wrong age
wg_minus_age <-
  wg_minus_health %>%
  filter_age_wg()

wg_age_n <- nrow(wg_minus_health) - nrow(wg_minus_age)

ws_minus_age <-
  ws_minus_health %>%
  filter_age_ws()

ws_age_n <- nrow(ws_minus_health) - nrow(ws_minus_age)

age <-
  c(
    "Out of age range",
    wg_age_n,
    percent(wg_age_n / n_total_wg),
    ws_age_n,
    percent(ws_age_n / n_total_ws)
  )

names(age) <- excl_col_names

#Now we need to get rid of people who did the survey too fast

wg_minus_fakes <-
  wg_minus_age %>%
  filter(completion_time >= min_completion_time)

wg_fake_n <- nrow(wg_minus_age) - nrow(wg_minus_fakes)

ws_minus_fakes <-
  ws_minus_age %>%
  filter(completion_time >= min_completion_time)

ws_fake_n <- nrow(ws_minus_age) - nrow(ws_minus_fakes)

fakes <-
  c(
    "Completed survey too quickly",
    wg_fake_n,
    percent(wg_fake_n / n_total_wg),
    ws_fake_n,
    percent(ws_fake_n / n_total_ws)
  )

names(fakes) <- excl_col_names

#calculate total amount of WG exclusions
total_wg_exclusions <-
```

```
  wg_repeats_n +
  wg_premie_n +
  wg_multiling_n +
  wg_health_n +
  wg_age_n +
  wg_fake_n

#calculate total amount of WS exclusions
total_ws_exclusions <-
  ws_repeats_n +
  ws_premie_n +
  ws_multiling_n +
  ws_health_n +
  ws_age_n +
  ws_fake_n

#make a row in the table for this
totals <-
  c(
    "Total exclusions",
    total_wg_exclusions,
    percent(total_wg_exclusions / n_total_wg),
    total_ws_exclusions,
    percent(total_ws_exclusions / n_total_ws)
  )

names(totals) <- excl_col_names

#now make the table
exclusion_tbl <-
  bind_rows(repeat_admins, premies, multiling, health, age, fakes, totals)

knitr::kable(exclusion_tbl)
```

Exclusion	WG exclusions	% of full WG sample excluded	WS exclusions	% of full WS sample excluded
Not first administration	0	0%	0	0%
Premature or low birthweight	7	3%	1	0%
Multilingual exposure	18	6%	23	8%
Illnesses/Vision/Hearing	4	1%	4	1%
Out of age range	1	0%	26	9%
Completed survey too quickly	132	48%	122	40%
Total exclusions	162	58%	176	58%

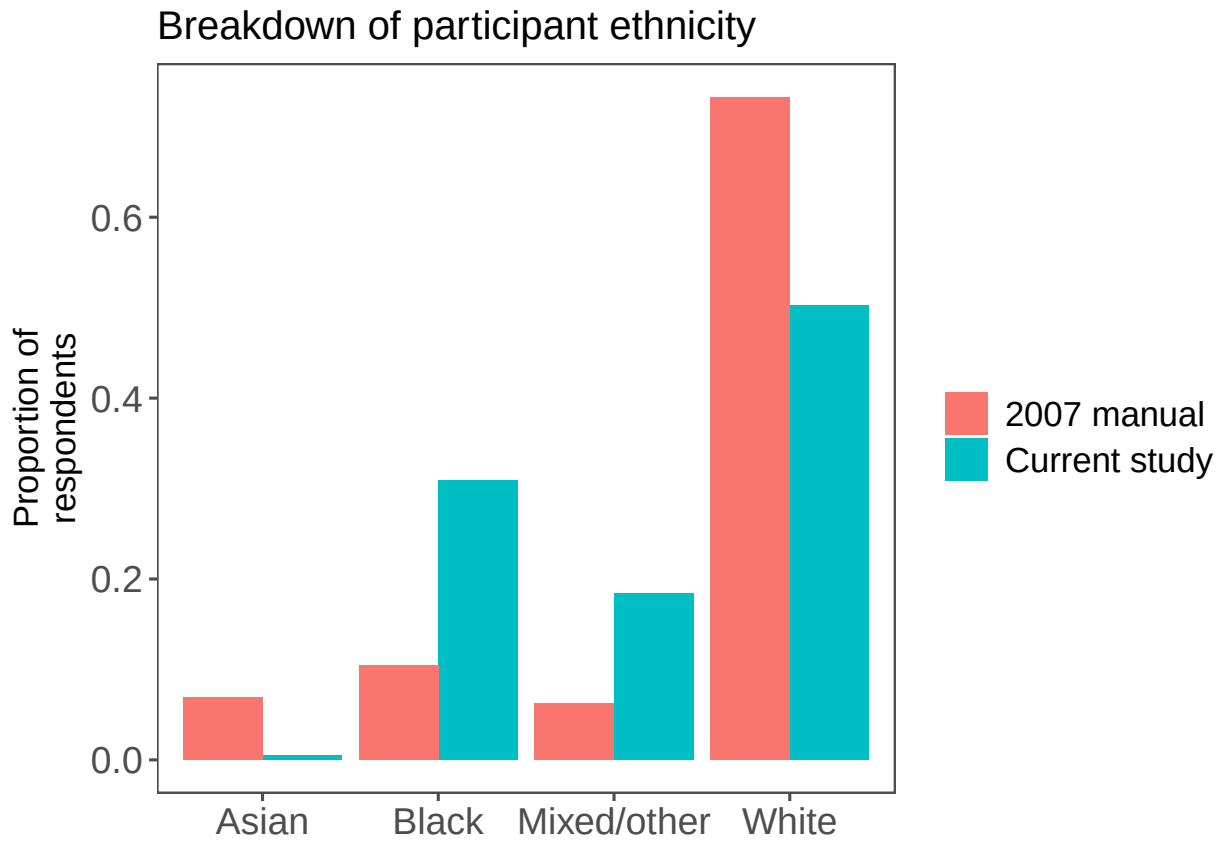
```
total_admin <- nrow(ws) + nrow(wg)
```

```
#combine relevant demographic information from both WS and WG
demographics_df <-
```

```
  bind_rows(
    wg %>%
      select(
        study_name,
        subject_id,
        age,
        ethnicity,
        maternal_ed,
        words_produced = `Words Produced`
      ),
    ws %>%
      select(
        study_name,
        subject_id,
        age,
        ethnicity,
        maternal_ed,
        words_produced = `Total Produced`
      )
  )

total_n <- nrow(demographics_df)
ethnicity_na_n <-
  nrow(demographics_df %>% filter(ethnicity == "No ethnicity reported"))
ethnicity_total_n <- total_n - ethnicity_na_n
maternal_ed_na_n <- nrow(demographics_df %>% filter(maternal_ed == "Not reported"))
maternal_ed_total_n <- total_n - maternal_ed_na_n
```

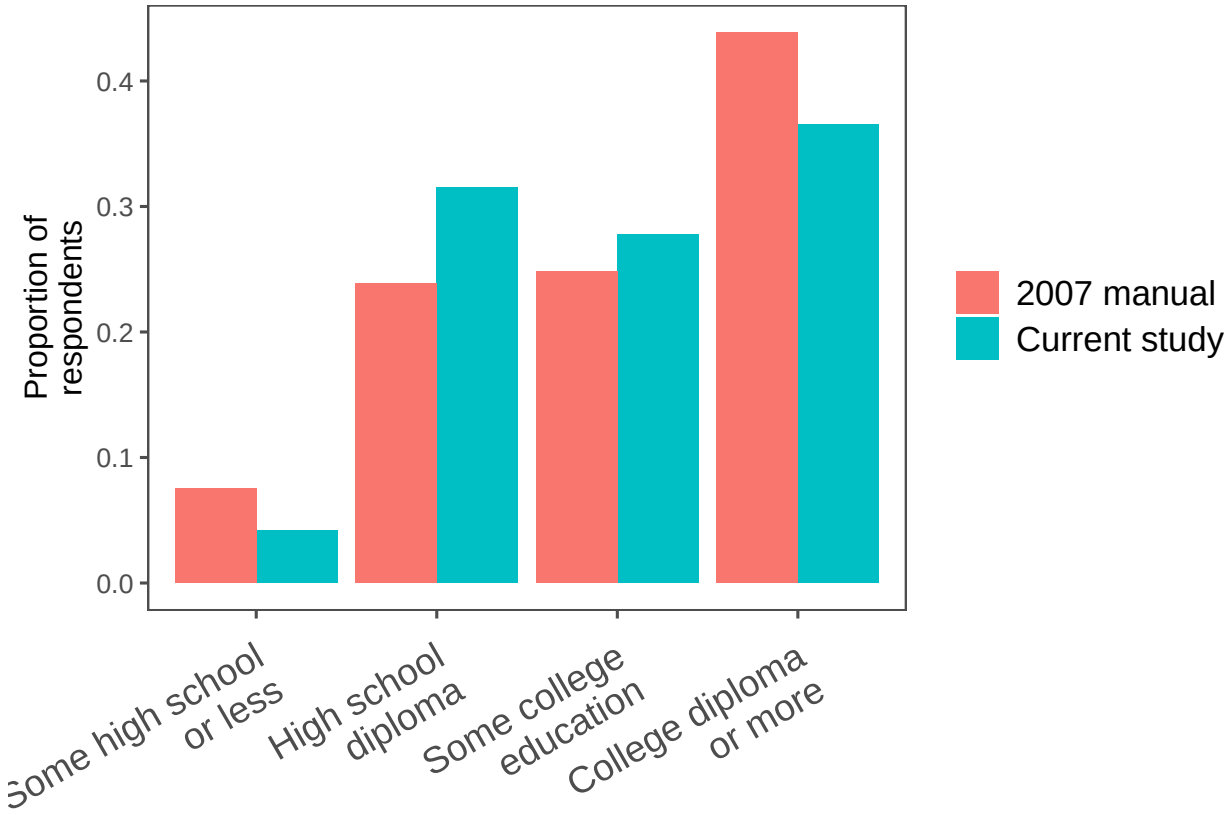
```
#Ethnicity plot creation
ethnicity_plot_df <-
  demographics_df %>%
  getEthnicitySummary() %>%
  filter(
    !is.na(ethnicity),
    ethnicity != "No ethnicity reported"
  ) %>%
  mutate(`Current study` = prop.table(n)) %>%
  left_join(old_ethnicity_numbers, by = "ethnicity") %>%
  select(-n) %>%
  pivot_longer(
    cols = c(`Current study`, `2007 manual`),
    names_to = "study",
    values_to = "proportion"
  )

ethnicity_plot <-
  ethnicity_plot_df %>%
  ggplot(aes(ethnicity, proportion, fill = study)) +
  geom_col(position = "dodge") +
  labs(
    y = "Proportion of\nrespondents",
    title = "Breakdown of participant ethnicity"
    # caption = str_c(
```

```
    #   "A total of ",
    #   total_admin,
    #   " CDI's were collected. ",
    #   total_exclusions,
    #   " did not meet inclusion criteria\n(multilingual status, vision/hearing impairments, premature
    #   min_completion_time,
    #   " minutes); ",
    #   ethnicity_na_n,
    #   " did not report ethnicity, leaving a final N = ",
    #   ethnicity_total_n,
    #   "."
    # )
  ) +
  theme_few() +
  theme(
    legend.title = element_blank(),
    axis.text = element_text(size = 14),
    axis.title = element_text(size = 13),
    legend.text = element_text(size = 13),
    axis.title.x = element_blank(),
    plot.title = element_text(size = 15),
    plot.caption = element_text(hjust = 0)
  )

ethnicity_plot
```



```r
#Maternal education plots

maternal_ed_plot_df <-
  demographics_df %>%
  count(maternal_ed) %>%
  mutate(`Current study` = prop.table(n)) %>%
  left_join(old_momed_numbers, by = "maternal_ed") %>%
  select(-n) %>%
  pivot_longer(
    cols = c(`Current study`, `2007 manual`),
    names_to = "study",
    values_to = "proportion"
  ) %>%
  mutate(
    maternal_ed = fct_relevel(
      maternal_ed,
      "Some high school or less",
      "High school diploma",
      "Some college education",
      "College diploma or more"
    )
  )

x_axis_labs <- c(
  "Some high school\n or less",
  "High school\ndiploma",
  "Some college\neducation",
  "College diploma\nor more"
)

maternal_ed_plot <-
  maternal_ed_plot_df %>%
  ggplot(aes(maternal_ed, proportion, fill = study)) +
  geom_col(position = "dodge") +
  labs(
    x = "Education Level",
    y = "Proportion of\nrespondents"
    # title = "Maternal education",
    # caption = str_c(
    #   "A total of ",
    #   total_admin,
    #   " CDI's were collected. ",
    #   total_exclusions,
    #   " did not meet inclusion\ncriteria",
    #   " (multilingual status, vision/hearing impairments, premature\nbirth, completion time under ",
    #   min_completion_time,
    #   " minutes); leaving a final N = ",
    #   total_n,
    #   "."
    # )
  ) +
  theme_few() +
  theme(
```

```
    legend.title = element_blank(),
    axis.text.x = element_text(angle = 30, vjust = 0.9, hjust = 1, size = 13.5),
    axis.title.x = element_blank(),
    legend.text = element_text(size = 13),
    plot.caption = element_text(hjust = 0)
  ) +
  scale_x_discrete(labels = x_axis_labs)

maternal_ed_plot
```




```
momed_ses_df <-
  demographics_df %>%
  mutate(highschool = case_when(
    maternal_ed == "Some high school or less" ~ "high_school",
    maternal_ed == "High school diploma" ~ "high_school",
    maternal_ed == "Some college education" ~ "college",
    maternal_ed == "College diploma or more" ~ "college"
  ))

momed_ses_df %>%
  filter(
    age <= 30
  ) %>%
  ggplot(aes(age, words_produced, color = highschool)) +
  geom_point() +
```

```
geom_smooth(method = "lm")
```