Data Sets for Many-Analyses

Julia Haaf

2/28/2020

This document takes the original data and subsets it using different exclusion criteria. It is therefore a documentation of all data sets used in the many-analyses.

```
merged <- readRDS("data/merged_deidentified_full.rds")

# head(merged)
nrow(merged)
```

```
## [1] 2281
```

```
# table(merged$race)
# sum(is.na(merged$race))
# sum(is.na(merged$countryofbirth))

levels(merged$countryofbirth)
```

```
## [1] ""  "1" "2" "4" "6"
```

```
# merged$usborn <- ifelse(merged$countryofbirth %in% c(1, "united states", "United State", "United Stat
#                                                      , "us", "US", "usa", "USA", "U.S.A", "U.S."
#                                                      , "The United States of America.", "America"), 1
# not needed anymore)

levels(merged$race)
```

```
## [1] "1" "2" "3" "4" "5" "6"
```

```
# merged$iswhite <- ifelse(merged$race %in% c(1, "White/Caucasian ", "White/Caucasian/Welsh,German,Ital
#                                             , "1 AND 2", "1 AND 4 AND 5", "1 AND 6", "1,2", "1,2,4",
#                                             , "1,4,6", "1,5", "1,6", "white/ native american", "Afric
#                                             , "Caucasion & Hispanic/Latina", "Thai and Caucasian"), 1
# not needed anymore


dates <- case_when(merged$date == "4/11" ~ "4/11/2017",
                   merged$date == "4/12" ~ "4/12/2017",
                   merged$date == "4/18" ~ "4/18/2017",
                   merged$date == "4/19" ~ "4/19/2017",
                   merged$date == "4/20" ~ "4/20/2017",
                   merged$date == "4/21" ~ "4/21/2017",
                   merged$date == "4/25" ~ "4/25/2017",
                   merged$date == "4/26" ~ "4/26/2017",
                   merged$date == "4/28" ~ "4/28/2017",
                   merged$date == "4/13/2107" ~ "4/13/2017",
                   TRUE ~ merged$date)
```

```
mdy <- mdy(dates) # automatically convert dates in the month-day-year format
```

```
## Warning: 168 failed to parse.
```

```
ymd <- ymd(dates) # automatically convert dates in the year-month-day format
```

```
## Warning: 1193 failed to parse.
```

```
mdy[is.na(mdy)] <- ymd[is.na(mdy)] # combine

merged$date_formated <- mdy

# rules for N-based site exclusions
nbysource <- with(merged, tapply(dv_order, list(source), length))
include2 <- names(which(nbysource >= 60))
include3 <- names(which(nbysource >= 80))

# Create variables, indexes, and exclusion rules ----
# compute exclusion rules
merged <- merged %>% group_by(source) %>% mutate(row_id = row_number()) %>% ungroup() %>%
  mutate(merged,
         # Participant-level exclusion rules
         # make sure that all IH participants are included in the ML-based ERs
         pass_ER1 = ifelse(expert==0, TRUE, pass_ER1),
         pass_ER2 = ifelse(expert==0, TRUE, pass_ER2),
         pass_ER3 = ifelse(expert==0, TRUE, pass_ER3),
         # In addition to the ML4 application of the exclusion rules (i.e., keep all In-House participa
         # we can also apply them ourselved to our best interpretation
         # Exclusion rule 1:
         #1. Wrote something for both writing prompts
         #2. Completed all six items evaluating the essay authors)
         pass_AR1 = (msincomplete == 0 | is.na(msincomplete)) & # completed both prompts
           !is.na(prous3) & !is.na(prous4) & !is.na(prous5) &  # P provided all 3 ratings of pro-us ess
           !is.na(antius3) & !is.na(antius4) & !is.na(antius5),# P provided all 3 ratings of anti-us
         # Exclusion rule 2:
         # as above, plus
         #3. Identify as White (race == 1)
         #4. Born in USA (countryofbirth == 1)
         pass_AR2 = pass_AR1 &
           (race == "1") & # white ps, NA race discarded
           (countryofbirth == "1"),
         # Exclusion rule 3:
         # as above, plus
         # 5. Score a 7 or higher on the American Identity item
         pass_AR3 = pass_AR2 &
           (americanid >= 7 & !is.na(americanid)), # strongly ID as american, NAs discarded
         ## N-based exclusions
         pass_NR1 = TRUE, #include all labs
         pass_NR2 = source %in% include2, # only labs with more than 60 participants (as preregistered)
         pass_NR3 = source %in% include3, # only labs with more than 80 participants (the target, as us
         ## Protocol-based exclusions
         pass_PR1 = TRUE, #include all labs
         pass_PR2 = expert == 1, # only AA protocols
         ## Timing-based exclusions
         # Exclude data collected before the pre-reg was registered (February 15, 2017)
         # for some labs, the data is NA, so for those we just went with what the ML4 authors
         # wrote on the how many first participants should be excluded.
         # This applies to: ufl: 181 participants excluded
         pass_TR1 = TRUE, # include all participants
         pass_TR2 = (date_formated > "2017-02-15" | is.na(date_formated)) & !(source=="ufl" & row_id < 
)
merged$row_id <- NULL #remove redundant column
```

Differences with ML4

In ML4, the participant-level exclusion criteria are only applied to the AA protocols, even though many are also available for the IH protocols (e.g., completeness of the measures).

In ML4 `NA`'s are excluded for ER1. I don't think it's justified to exclude all `NA`'s for ER1 because that would exclude an additional 300 participants (2 entire labs who only have `NA`'s for the completeness of the writing prompts). For the expert lab that has the same, the authors assume that the prompts were fine (based on detailed notes not reporting any abnormalities).

Exclusion Criteria

Person-level exclusion (Taken from Klein et al., 2019):

Note that these exclusion criteria *only applied to AA sites*!

1. All participants who did not complete the materials.
2. 1 + All participants who do not identify as white and who were born outside of the United States.
3. 2 + All participants who responded lower than 7 on an American Identity item.

Study-level exclusion based on N (from the preregistration):

1. No study-level exclusion.
2. All studies that have fewer than 60 participants collected.
3. All studies that have fewer than 40 participants per cell. –> this one I can't find again, only that they would exclude labs that tested fewer than 80 participants in total. In this case, it just comes down to the same anyways.

Study-level exclusion based on expert advise (from the comment):

1. No exclusion based on expert advise.
2. In-house studies are excluded.

Participant-level exclusion based on timing: 1. No exclusion based on time of data collection 2. All participants whose data was collected before the lead team's analysis plan was preregistered.

We include this as a separate dimension because Klein et al. always apply this criterion, whereas Chatard et al. don't and we also think one shouldn't.

Application of participant-based exclusion criteria for IH participants: 1. apply participant-based exclusion criteria *only* to AA studies, keep all participants from IH studies (Klein et al., Chatard et al.) 2. apply participant-based exclusion criteria to all participants, excluding participants for whom the relevant information is missing (us).

All exclusion criteria are crossed in the following way:

```
person.ex <- 1:3
n.ex <- 1:3
expert.ex <- 1:2
time.ex <- 1:2
```

```r
ih.ex <- 1:2

crit <- expand.grid(person.ex, n.ex, expert.ex, time.ex,ih.ex)
nrow(crit)
```

```
## [1] 72
```

```r
# we have 72 potential combinations of exclusion criteria settings
# let's see if we can combine them and generate datasets automatically

apply_excl = function(set,data){
  excl_rules = function(x){
    out = list()
    out[[ifelse(x[5]==1, paste0("pass_ER",x[1]), paste0("pass_AR",x[1]))]] <- TRUE
    out[[paste0("pass_NR",x[2])]] <- TRUE
    out[[paste0("pass_PR",x[3])]] <- TRUE
    out[[paste0("pass_TR",x[4])]] <- TRUE
    return(out)
  }
  rules <- excl_rules(as.numeric(set))
  apply_rules = function(data, rules){
    rules <- rules[names(data)]
    idx <- Map(`%in%`, data, rules)
    idx[is.na(names(rules))] <- list(rep(TRUE, nrow(data)))
    data[Reduce(`&`, idx), ]
  }
  dat <- apply_rules(data,rules)
  # also remove NA  is dv
  dat <- dat[!is.na(dat$pro_minus_anti),]

  metaset <- run.analysis(dat)
  write.csv(metaset, file=paste0("data/update/metaset_",do.call(paste0, as.list(set)),".csv"), row.names
  write.csv(dat, file=paste0("data/update/reanalysis_",do.call(paste0, as.list(set)),".csv"), row.names
}

apply(crit,1, function(f) apply_excl(f,merged))
```

```
## NULL
```

Inclusive Analysis

We would propose to include the most inclusive set and vary theory-based participant-level criteria. No time-based: does not make sense, teams had no access to each other's data, lead time did not access prior to preregistration. No study-level based: only lowers power if applied.

Original Analysis

We aim at reanalyzing the key findings of Klein et al (2019) using the exclusion settings (1,2,1,2,1), (2,2,1,2,1), and (3,2,1,2,1). This means excluding data collected before February 15, 2017 and using the N=60 per site as the minimum inclusion. The last two criteria were added from the preprint to the published version.

Analysis for the Main Claim of the Comment

We aim at reanalyzing the main claim of Chatard et al (2020) using the exclusion settings (1,3,2,1,1), (2,3,2,1,1), and (3,3,2,1,1). Note that the study-level exclusions are based on the number of participants for exclusion criterion 1.