~~About the log files

**1. Log files**

These are the two log files. Right now they live in chk\_lapses – where should they be saved?

**1) log\_correct\_ema:** This file documents data corrections to all EMA variables. These include corrections to lapse variables (lapse start and end times) as well as non-lapse variables (e.g. utc, the abstinence question, extra surveys). Lapses included in this log were either 1) added via a note in a data log or 2) added due to identification of a problematic/suspicious lapse in the chk\_lapse script (e.g. negative lapse length).

**3) log\_add\_lapses:** This file documents all lapses reported by the participant (either in audio messages, follow up study visits, etc) that are not tied to a specific ema report.

**1. Variables**

Variables to the left of and including the “notes” column are what will eventually remain in the file:

* subid and response id (for matching records)
* utc (NOT the true utc of the start date, leftover from previous Qualtrics settings, but is often how lapse reports are referenced in notes so we need it)
* var\_name (variable to be changed)
* old\_value (previous value of the variable to be corrected)
* new\_value (value you are changing the corrected variable to)
* log\_action (change indicates updating data, note does not alter data)
* log\_reason (the reason this case is in the log. Data log entry, negative lapse, date formatting, etc)
* ema\_type (morning vs later, to link to ema variables)
* log\_note (the data we found to support our log\_action)

There are extra variables in the log file right now to provide context for you while reviewing the lapses. These will be deleted after your review:

* proposed\_change: this clearly dictates what the change is from the notes column; megan and I kept this so that we were clear on our decisions
* flag: This identifies which cases need your review. We tried grouping review cases on common themes/decision to make (e.g. rounding decisions). The meaning of each of the flags are described in the last section of this document.
* Other lapse variables (ema start and end dates, lapse start and end dates, lapse length, report length) are included to the right of notes for you to look at while reviewing notes.

**2.** **Order of corrections**

The order of the data log corrections is intentional. Cases are organized by sub ID within log\_reason groups (since later log\_reasons are dependent on previous data cleaning, e.g. adding timezones). For each step, problematic cases are identified, and the corrected log for that section is read back into the script before moving onto the next step.

**1. Data log entry**: These cases were manually entered into log\_ema based on notes in the ema section of risk/analysis/notes/data\_cleaning\_log.docx as well as John’s ema cleaning in *cln.database.R* (lines 8 – 141) and *fun\_data.R* (lines 28 – 54). I did not incorporate anything before these steps.

**2. Unfinished lapse reports:** Like previous EMA cleaning, we discard lapse reports that do not provide any lapse date information (unless specified in a data log entry) and retain partial EMAs that have lapse date data (unless specified in data log to remove the lapse). These cases are not entered into log\_ema for examination.

**3. Bad date format:** Lapse dates with bad formatting: This code identifies lapse hours that are non-numeric or out of range, and lapse dates that are not in mm-dd-yyyy format. These cases are added to log\_ema for review.

**4. Multiple timezones:** This code identifies lapses that occurred on days with multiple timezones and adds them to log\_ema for review. Afterwards, each lapse has one timezone associated with it

**5. New timezone.** Any lapse that was not America/Chicago was reviewed in the log to confirm the new tz is correct.

**6. GPS > 1 day:** This code identifies lapses that did not have gps data on the day of their lapse and adds them to the log\_ema file.

**7. Non ema lapses:** Recorded in log\_add\_lapses.csv, these are lapses reported by the participant (in study follow up, audio, etc) that are not clearly linked to a specific ema lapse report. These will include subid, report\_date (day which they reported the lapse to us), start and end dates and times of their lapse, the source of the report, notes.

**8. Lapse/Report length:** This code locates cases that have problems with lapse\_length (the lapse end time – lapse start time), report length (start\_date – lapse start time). Negative values are unacceptable for lapse\_length and are tolerated up to -1 hrs for report length. We also examined cases with unusually long lengths (determined by histogram: >10 hrs lapse length, > 30 hrs).

**3. Data cleaning rules**

We always attempt to make our conclusions based on the most concrete, positive evidence. We will not allow lapse lengths to remain negative or for report lengths to be more negative than one hour since these types of reports are impossible.

Here are categories of data cleaning decisions we need to make that you will see in the “flag” variable. Ongoing categories mean we do not have an answer, Resolved list the data cleaning rule we have been using. These rules may not work for all cases but are just to help you review/make decisions.

***\*\*Ongoing\*\****

**ROUNDING:** ~~Participants were told by staff (no documentation) that they should round their lapse start and end times up to the next hour. Therefore, will allow lapses that are up to 1 hour later than the completion time of surveys they were reported on.~~ \*\*UPDATE\*\* After discussing with Megan, she does not believe participants were \*explicitly\* told to round up, but believes people did because we told them to report the same start and end time if they only had one drink. The only instructions in the sop are:

“The hour of the first drink simply means to give us an approximate time of when you had the drink. You will then be asked to indicate the date of the last drink that you have not yet reported, and the hour of the last drink that you have not yet reported. If you only had one drink, the date and hour for your first and last drink will be the same. Example; if you had a drink at 11:00 p.m., and no other drinks before completing a survey you would indicate the same date and time for your first and last drink.”

At least for me, that doesn’t indicate that people would round up a whole hour (e.g. they would report a 1 pm start time for a lapse at 12:03), but they could have. What is your recommended rounding rule?

**RANDOM TIME:** Some participant reports can recall the date but not the time of their past lapses. The data cleaning log used differing times to code lapses on these days (19:00, 12 – 13). When we truly have no other data to pull from (no audios, participant doesn’t indicate lapsing in the evening vs morning, etc) think our options would be:

1. Code the entire day as a lapse (e.g. 00:00– 24:00)
2. Code a static lapse at the participant’s most common lapse start time
3. Pick a standardized random time of day (but what?)

**NO EVIDENCE / STATIC LAPSE VS FULL EMA PERIOD:** Sometimes we have more information (at least a lapse start or end time, ema completion time) to help decide a more accurate time than a random time. This often comes down to wondering if the lapse should extend the maximum time possible (e.g. the full window up to an ema) or if we are better coding a static lapse (e.g. 13:00 – 13:00).

**DATA LOG ERROR:** Sometimes the data log suggests lapse times that are impossible (e.g. reporting a lapse into the future). Do we leave as corrected or override the data log?

**NON LAPSE EVIDENCE VS…**: I know we talked a little bit about not trusting negative data, but it is pretty telling if someone who does their audios every day stops doing them during a three day binge, etc. These types of flags are for when Megan and I weren’t sure which weaker evidence should take precedence.

***\*\*Resolved\*\****

**MULTIPLE LAPSES:** We talked about this one, but I left the flags in the file so you could see examples to see if you agree. If a participant reports multiple, separate lapses they were trying to report on a single lapse report, the lapse closest to the EMA report will be coded within the EMA response. The others will be entered as non-ema lapses in the add\_lapses log.