ManyBabies Exploratory Analysis: Data Processing

Project start date: 2025-09-19
Project end date: 2025-12-12

Contributors: Alexandra Sarafoglou (alexandra.sarafoglou@gmail.com); Ingmar Visser (I.Visser@uva.nl); David Osten (david.leon.osten@gmail.com); Saulė Remeikaitė (saule.remeikaite@gmail.com)

Project aims:
Finding communalities between multiple ManyBabies studies.
Infant looking time (LT) as the main variable of interest across MB projects.
Focus on habituation that happens during LT in experimental / training trials.

Project objectives:

- Combining and standardizing MB1, MB3, MB4 datasets into one for easier further exploration
- Producing clear process and data documentation

Studies:

- **MB1**

https://manybabies.org/MB1/
Replication of infants' preference for infant-directed speech (IDS) over adult-directed speech (ADS). Measured with single-screen central fixation, eye tracking, and the head-turn preference procedure (HPP). Infants participated in 2 training trials (familiarization phase), followed by 16 test trials (habituation phase), and looking times generally declined across trials, showing habituation to the stimuli.
Results indicated that infants preferred to listen to IDS relative to ADS.

- **MB3**

https://manybabies.org/MB3/
Research investigating whether infants can learn abstract algebraic rules over patterned syllable sequences (ABA vs. ABB) and generalize those rules to novel syllable sequences. Infants participated in training trials (familiarization phase), followed by a maximum of 12 test trials (habituation phase).
Looking time analyses and habituation across trials are used to assess rule learning across the full set of test trials.

- **MB4**

https://manybabies.org/MB4/
Replication of the classic helper vs. hinderer infant paradigm testing whether infants prefer prosocial (helper) over antisocial (hinderer) agents. Social evaluation is

assessed via infants' responses to the test displays following the training sequence. Infants participated in a maximum of 14 training trials, during all of which habituation was happening.
Results indicated that infants did not show a reliable preference for helpers over hinderers overall.

Datafiles retrieved from OSF:

- MB1: 02_validated_output.csv
- MB3: 02_validated_merged_output.csv
- MB4:_clean_data.csv

Timeline:

1. Papers and datafiles analyzed; we chose datafiles that were preprocessed but still had the most amount of variables (see above) and started the process of identifying relevant variables
    - we excluded MB2 project because it did not contain LT data
    - raw datafiles were too big and messy to work with, so we decided to start with preprocessed datafiles
2. Visualization table for final datafile created (see *Table 1* below); relevant variables identified and defined, with more work to do on standardizing them
    - decided to structure the final datafile so that we have trials going down vertically (each infant has rows of trials; each trial of the same participant goes on a single row)
    - decided to have 3 columns (MB1/MB3/MB4) and add NA for data that are not from that specific project
3. Continued to work on variables from different projects and on combining them into final datafile
    - decided to create a new variable "Participant ID" = project + lab + original participant ID
    - decided to keep some variables that were excluded from original projects, since they did not interfere with studying habituation (MB4 "failed to make choice" ; MB1 "bilingualism" ; "age")
    - excluded some variables as they were excluded in original projects ("developmental disorder" ; "premature" ; "equipment failure" ; "session / trial error type")

MB4 is scarce in regards to demographic variables, but it is a possibility to reach out to the original team via Ingmar if they are needed later on.

4. Merged all 3 datafiles into one and documented it step-by-step
5. Uploaded code and files onto a private GitHub repository; accessible with admin permission later, if needed

Table 1

White: Variables were (1) relevant (2) common to all.

Pink: Variable has to be created but data for this is available.

Red: Data missing for some, but variable still created for continuity purposes.

VARIABLES COMPARISON		
MB1	MB3	MB4
Identifiers	**Identifiers**	**Identifiers**
project_id	project_id	project_id
lab	lab_id	lab_id
subid	participant_id	subj_id
unique_subject_identifier	unique_subject_identifier	unique_subject_identifier
Method	**Method**	**Method**
method	method	method
Age	**Age**	**Age**
age_days	participant_age_days	age_days
age_group	age_group	age_group
Participant sex	**Participant sex**	**Participant sex**
gender	participant_sex	participant_gender
Stimulus type	**Stimulus type**	**Stimulus type**
stimulus type (audio, visual, audiovisual, etc.)	stimulus type (audio, visual, audiovisual, etc.)	stimulus type (audio, visual, audiovisual, etc.)
Trial Number	**Trial Number**	**Trial Number**
trial_num	trial_num	trial_number
Trial Type (Condition)	**Trial Type (Condition)**	**Trial Type (Condition)**

trial_type	trial_type	condition
Habituation	**Habituation**	**Habituation**
habituation	habituation	habituation
Looking Time	**Looking Time**	**Looking Time**
looking_time	looking_time_seconds	lookingtime_videoevent
		lookingtime_freezeframe
Size of household		**Size of household**
household_size	household_size	not available
Participant Place of Birth	**Participant Place of Birth**	**Participant Place of Birth**
not available	participant_place_of_birth	not available
Parental education	**Parental education**	**Parental education**
parenta_education	caregiver1_education_construct_response	not available
parentb_education	caregiver2_education_construct_response	not available
Second session		**Second session**
second_session	second_session	second_session
Primary Caregiver Gender	**Primary Caregiver Gender**	**Primary Caregiver Gender**
parenta_gender	caregiver1_gender	primarycaregiver_gender
Language	**Language**	**Language**
lang_group	language_group	language_group
lang1	lang1_name	primary_language
lang1_exposure	lang1_exposure	percent_primarylanguage

Position	**Position**	**Position**
caregiver_seat	caregiver_seat	caregiver_seat