# ManyBabies1 Secondary Analysis

The effect of ‘Lab Factors’ on fussout rates/latencies and infant-level and laboratory-level effect sizes

STATUS: PREREGISTERED AS OF 6/8/18

# AsPredicted Format Preregistration

***1) Data collection.*** *Have any data been collected for this study already?*

Yes
No
X It's complicated. We have already collected some data but explain in Question 8 why readers may consider this a valid **pre**-registration nevertheless.

***2) Hypothesis.*** *What's the main question being asked or hypothesis being tested in this study?*

Because of the high cost of running infant participants and the small samples of individual labs, researchers in cognitive development often have the impression that ‘incidental’ factors may significantly impact whether infants (1) make it through the study (without fussing out) and (2) attend properly to stimuli/display an expected pattern of looking. These factors are numerous, and our impressions often vary from each other! In this project, we aim to make some initial estimates of the relative impact of some of these factors. The belief of the first author (MK) is that a subset of these effects will *not* have an appreciable impact on fuss-out rates; of those which possibly do have effects (and which we have sufficient evidence to evaluate), we will conduct a limited confirmatory analysis on a held-out portion of the ManyBabies 1 dataset.

***3) Dependent variable.*** *Describe the key dependent variable(s) specifying how they will be measured.*

There are two dependent variables - data ‘missingness’ (trial level variable) and effect size (participant-level variable).

***4) Conditions.*** *How many and which conditions will participants be assigned to?*

Participants are not randomly assigned to conditions for the purposes of these analyses; the ManyBabies1 dataset includes the key within-participant variable of ADS vs. IDS, and includes variation in methodology (grouped as:Headturn Preference Procedure, Central Fixation, Eye-tracking) and infant age, which are both likely to impact our dependent variables. They also vary along many other dimensions, which, while of considerably less scientific interest to the phenomena of IDS, may also impact our dependent variables.

***5) Analyses.*** *Specify exactly which analyses you will conduct to examine the main question/hypothesis.*

Because of the large number of IVs and exploratory nature of this project, we aim in this preregistration to (1) Clarify decision-making for including/excluding data and variables from these analyses (2) Lay out a general analytic approach that can be applied to each of the LabFactor variables individually and (3) Define all LabFactor variables to a sufficient extent that we don't have to make major decisions contingent on the data after having seen that data.

***6) Outliers and Exclusions.*** *Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.*

No outlier data will be excluded. Observations will be excluded from the dataset only if participants clearly come from outside the population we intended to sample or if individual scores are outside the expected range without explanation and therefore clearly spurious (e.g. a recorded measure of 19 s on a trial that lasts 18 s). Independent variables will be analyzed only if we believe them to have enough data to be meaningful. How we determine 'enough data to be meaningful' for the **ragged/comprehensive dataset** and the **high-density dataset** is described in the outline below.

***7) Sample Size.*** *How many observations will be collected or what will determine sample size?*

Sample size for this analysis is set by the MB1 project – we will analyze all data that was collected from this project (which has not yet been analyzed), believed to be approximately 1500 babies across 68 labs. As of 6/1/18, 30 labs had submitted MB1 data, 6/15/18 is the final deadline.

***8) Other.*** *Anything else you would like to pre-register?*

From #1 above, here's a clarification of the status of the data to be analyzed in this project. MB1 involves data collected at the level of individual participants and at the level of labs. The status of each is as follows:

*Individual participants*: All participants have finished being run for MB1. As of 6/1, many but not all labs have submitted their datasheets for this project (deadline 6/15). The data cleaning problem for this project is very hard, and we have not yet calculated summary statistics or availability for any of the variables described in this project.

*Labs*: Data at the lab level comes from a few sources:

*Pre-data questionnaire*: Already collected

*Video 'walkthrough'*: Labs have mostly already submitted these; they have not been viewed or coded by this team (or anyone). This doc will specify an initial summary of variables to be coded from this video; we expect to update this doc with a final list, *subsequent* to ICIS but *prior* to coding the video data.

*Post-data questionnaires 1 and 2*: In development, not yet collected. This preregistration will set the variables to be collected, and we will update this document with the final text of that survey once available. The fact that there are 2 separate surveys is a logistic rather than scientific decision! The first serves as a checklist for labs submitting their data and gathers feedback about the experience of participating in MB1, the second will have the majority of 'Lab Factor' questions.

***9) Name:***

ManyBabies 1 - LabFactors Project

## Outline Introduction/Project History

This is a preregistered secondary analysis that ‘piggybacks’ on MB1 (MB1 Main preregistration.) In this preregistation we aim to follow the best practices for secondary (confirmatory) analyses on existing data laid out in Weston et al. (2018).

Infant psychological data is precious and difficult to collect. Not only do we rely on a limited range of behaviors (looking times, heart rates, reaches for objects) from our pre-verbal participants, they are famously mercurial subjects, playing with their shoes and socially engaging with their caregivers rather than attending to experimental stimuli, and, in the limit, 'fussing out' of experiments – a crying baby rightly overrides our desires to collect data that day. While we have considerable anecdotal evidence and impressionistic opinions on how to manage this, hard data on the dynamics of this critical roadblock to cognitive development research remains relatively sparse. From meta-analyses of developmental studies, we know that there is substantial variance in drop-out rates even within age groups and methods, but we know very little about the sources of this variance. From 687 studies where dropout rates were recorded in MetaLab (data downloaded on 2018-05-19): Median is 25% (range: 0-75%); There is some evidence that methods differ in dropout rates, but even within methods, dropout rates vary greatly, and age seems to play only a small role (Bergmann et al, 2017; Table 3).

The ManyBabies 1 project, a large-scale collaboration of developmental psychology labs to replicate a classic finding – infants' preference for infant-directed over adult-directed speech – provides a unique opportunity not only to take stock of the field and discover how our methods and approaches differ, but to being to understand the factors that make these effects so difficult to measure. Here, we are preregistering a plan to analyze additional variables collected alongside the main MB1 project, consisting of a wide range of 'lab factors' that researchers believe may impact either whether a baby fusses out of a study, or whether they truly attend to stimuli (and thus produce an expected effect in the study.) This is an exploratory project, designed to guide further research, but because of the size, sprawl, and uniqueness of this dataset, we aim to use it as well as possible, by pre-registering our analytic plan and, where possible, holding out part of the dataset to allow for confirmation of exploratory effects.

A special note: This project has a strict intermediate deadline: we will do a preliminary presentation at ICIS in 6 weeks; this may not be enough time to process all variables as intended; we expect therefore to be unblinded and to present simpler analyses, but commit to presenting the plan documented here in the final presentation of this dataset.

Guiding principles for this project are as follows:

- Because of the extensive list of candidate variables, we expect to encounter unanticipated issues regarding the coding/representation of individual columns; thus, analyses will be initially developed/debugged on pilot, scrambled or dummy data, and run on the ‘real’ data only in their final form.
- For the same reason, we will specify a fairly detailed, but **general** plan for how we'll tackle each variable. (In short, we aim to learn something about the (potential) effect sizes of these lab factors.)
- Data is expected to be unevenly distributed – reporting this extensive list of 'extra' variables was optional for labs, and so we will be discovering how much information is available as we go.
- We do not include data from the MB1B (Bilingual) co-project in this preregistration because of timing (bilingual data collection will continue into summer), but point out its usefulness as an additional confirmatory dataset!
- Because much of our interest is in fuss-out rates, we will explicitly ensure that *all* tested babies (even those who produced no data for MB1 analyses) are included in this dataset. Note that while the bulk of this paper will simply treat data as ‘missing’ or ‘nonmissing’, we also can evaluate reasons for failure (fuss out, equipment failure, etc.) or limit analyses to data present and data *absent because the baby fussed out* (i.e., not including data missing for other reasons)

# Analysis Decision Tree

For each variable (See Variable Definitions, below), we'll conduct the following to decide whether and how to measure its potential impact on the MB1 participants.

1. Is there enough data to analyze this variable at all?
   a) *Lab level variable*: At least 10 observations to attempt analysis. There are 68 labs registered for MB1; we'll assume that if we have fewer than 10 labs we cannot say anything particularly meaningful about the variability on this axis. (For analyses we conduct, a valid outcome is ‘we don’t have the data to draw a conclusion’)
      (i) An issue identified: if there is serious imbalance in how variables were collected, a lab factor might appear to be conflated with method : our proposed solution to *watch out for* this but not exclude variables if they are mainly available from 1 method. (I suspect in many cases we will not have power to interpret any differences between methods…)
   b) *Participant level variable*: At least 100 total participants from at least 10 labs. These numbers are determined as follows: The minimum contribution from a lab is 16 babies, but some may not be appropriate to include in the fuss-out analysis (ie because they

are out of the intended population, e.g. discovered after testing to be non-monolingual), and some variables may not have been collected from all babies.

2. Separation of variables into 'ragged/comprehensive' and 'high-density' datasets
   a) Ragged/comprehensive: all variables that **don't** make it into the set below will be treated as fully exploratory analyses.
   b) High-density: We will aim to identify a sub-set of labs and variables such that 80% of babies in those labs have data for 80% of the variables. We will then determine whether this dataset is large enough to warrant creating 'hold out' subsets to confirm exploratory analyses on. If we are able to, this holdout set will be created by sampling a random 1/3 for exploratory analyses – for participant-level variables, split so that a random third from *each lab* is selected; for lab-level variables, so that a random third of labs reporting that variable are present.

# Dependent Variable Specification and Statistical Testing Plan

For variable modeling in general, we will *not* focus on interactions between implementation factors; we will follow the MB1 main analysis by using e.g. only age * native language (the current MB1 plan) as demographic/possible moderating factor; if the main analysis yields a different structure we'll follow that decision. Similarly, we will follow the same random effects structure as the main analysis, anticipating that we may need to prune additional random effects due to convergence issues.)

Because a goal of this analysis is to provide guidance for researchers, we will (in addition to this modeling) present descriptive results for all measurable variables, including both standardized effect sizes and in relevant units, in order to convey both our confidence in our estimates of these effect sizes and how they compare to the theoretical effect sizes we aim to measure.

## Dependent Variable 1 – 'Missing' (Measures trial-by-trial rate of data loss)

MISSING (T/F): Baby either did or didn't contribute a datapoint on each of 16 trials. True if a trial had no included looking time (e.g., no looking recorded, a look under 2 s, or no looking because the infant had already terminated the experiment). IMPORTANT NOTE: A baby who came to lab but then fussed out before contributing any data to the MB1 analyses is an important data point in this analysis! They have Missingness = TRUE for all 16 trials.

Because we have the complete observations from MB1, we are able to report whether a baby contributes data on every trial, rather than relying on coarser measurements like whether or not a baby made it to the end of a study. Neat! There are many reasons a trial may be missing (e.g. equipment failure, caregiver interference); however depending on how labs report data it may not be possible to know this **at the trial level**. We'll therefore plan to **not** differential by reason in the main analysis, but will be able to check whether results vary if we exclude babies who eventually terminated a study early for a non-fussout reason.

## Dependent Variable 2: Effect Size (Magnitude of the IDS>ADS preference per baby)

BABYEFFECTSIZE (Numerical) – Cohen's d, calculated for every baby completing at least two ADS trials and at least two IDS trials. We use Cohen's d (rather than eg. mean differences) because raw looking time values are expected to vary by method and by age.

This analysis is somewhat tied to the question of whether the dataset, as a whole, replicates the finding that babies attend more strongly to infant-directed than adult-directed speech, not yet known at the time of this preregistration. (MK believes that (1) in the world, babies probably **do** enjoy baby-talk but (2) it is possible – though maybe not probable – that a careful and preregisted implementation of this method may fail to stably measure a nonzero effect of this preference on babies' looking times.) In the event that this is **not** the case, we believe these analyses may be helpful for understanding that failure to replicate!

Note that in addition to using Cohen's d for the modeling portion of this project, we plan to also 'translate' our descriptive reporting of this variable back to seconds (mean & confidence intervals) for communication purposes.  Two additional thoughts on this variable to keep in mind when drawing conclusions. First, there are two ways that IVs might affect this variable. One is direct: a lab factor may may an infant's reactions less directly related to their underlying IDS/ADS manipulation. The other is indirect: if a lab factor leads to higher fuss0out rates, fewer trials mean less stable measurement. The second thought is that a larger effect size measurement does not mean a particular implementation/IV level is ‘better’; for instance larger effect size by graduate student testers could reflect lack of hypothesis blinding!

## Statistical Testing Plan

Throughout these analyses, *we will not attempt to measure interactions among lab-factor IVs* – while there are likely many unexpected confounds between these factors, we begin this project by estimating the effect of each variable alone. The plan below will be accompanied by descriptive statistics for all variables that have enough data (as defined above) that show effect sizes with confidence intervals; the modeling here will provide a way to quantify these effects, understand them in the context of the MB1 analyses, and guide further confirmatory research.

We plan to follow the general effects structure used for the model in MB1 (expecting that we may need to further prune the random effects structure to deal with nonconvergence), with the addition of including method (Central Fixation, Eyetracker, or Head-turn Preference). For the first lab-factor DV (MISSING), adaptation from the MB1 model is straightfoward – rather than a linear regression to predict looking time, logistic regression is used to predict whether the datapoint is present or absent. Then, following the 'moderator' strategy used in MB1, we add the target IV as a main effect, and an interaction between trial number and the target IV, accompanied by appropriate random slopes/intercepts nested within labs.

as an additional fixed effect to the final MB1 model used, e.g. “We fit the model specified above with the addition of a second-session main effect and trial type by second-session interaction (and with a second-session random slope and intercept nested within labs). ” Thus, our model specifications will be of the form:

```
MISSING ~ LABFACTORVAR * method * age_days + LABFACTORVAR * trial_num
      + age_days * trial_num + (trial_num | subid) +
      (trial_num * age_days * LABFACTORVAR * method | lab) + (age_days
      | item)
```

Note that we include random slope/intercept for method nested within lab, because some labs use more than one method; however most labs use only one method. As a final note, because our predictors (LABFACTORVAR) may not always have monotonic effects, we are prepared to use GAMM rather than linear predictors only for these variables.

For the second lab factor dependent variable, BABYEFFECTSIZE, we modify this framework to account for the fact that we are now modeling (summary score) data at the level of each baby. Thus, the general form of the analysis becomes

```
BABYEFFECTSIZE ~ LABFACTORVAR * method * age_days + (age_days *
      LABFACTORVAR * method| lab)
```

# Variable Specifications

This is a narrative/summary description of the variables we'll (attempt to) submit to the general analysis plan just described. You should take a look at the MB1 Data Dictionary as the definintive version of this list; the LabFactors project uses a subset of these variables. As part of this preregistration, the 'official' data dictionary is being harmonized to match the (expected) form of all the post-data collection variables to be collected as labs finish up submitting their MB1 datasets.

Below, the variables are organized as to their level (trial, participant or lab) and source- that is, in which data sheet(s) this variable should be found. We also star some additional variables which we will **not** attempt to analyze in our pipeline, but will report descriptively to undersand what labs are doing (e.g. recruiting methods.)

## Participant-Level Datasheets

second_session - Was this data collected as the child’s second study in the testing session? (Note that second-session data is dropped for the confirmatory dataset in the main MB1 paper)
caregiver_seat - Was the child in a caregiver’s lap or in a baby seat/highchair?
Optional_Beard - Did the RA have a beard or visible stubble *that day*? Answered by all RAs, not just males
Optional_RAType: Was the RA today an undergraduate, graduate student, postgrad, or other
Optional_Last_Feed_Minutes: How long, in minutes, since the last time the baby ate? (when the arrive)
Optional_Last_Sleep_Minutes: How long, in minutes, since the last time baby slept?

Optional_TOD: What time of day is it?
Optional_Season: What season is it (Spring, winter, summer fall)?
Optional_TermTime: Is your university “in session” right now?

## Lab-Level: Pre-Data Collection Questionnaire

(This is a subset of a longer survey which included lots of logistics about the running of MB1!)

**Recruitmentmethod - mail-outs? Hospital visits? Nursery visits? Facebook ads? Community event (e.g., Baby Expo), preschool, hospital, community organization (e.g., YWCA, Music Together), mailings, online ads (including facebook), emails (e.g., to the postdoc group on campus or whatever), word of mouth.
Compensation – What if anything do families receive as thank-you for participating?
Experimenter-Location - where is the online coder located?
SoundAttenuation - what steps are taken to mask external noise?
**Carseat – Does the lab use a carseat for testing infants?
**Headphones – What kind of headphones does the lab use?
**Testing, TestingMaskingMusic, TestingVisualDisplay – How often does the lab check that equipment is fully functional?
RAs – Who can run MB1? Undergrads, Ras, Grad students?
Training/Training criteria: what training must RAs have to run babies in this lab? (Moved ‘policy version’ here; we now ask about the principal RA(s) that *actually* ran MB1 in the post-data Q)

## Lab-Level: Main Post-Data Collection Questionnaire

Link to the qualtrics survey: https://umanitobapsych.az1.qualtrics.com/jfe/form/SV_cMiOtYCAOyhpQ9v
Note that we don't yet have access to how Qualtrics will actually format the variable names!

LanguageLab - Is this a language development lab? (1-7)
NewMethodology - to what extent was this study a new METHODOLOGICAL approach for you? (1-7)

## Lab-Level: Supplemental Post-Data Collection Questionnaire

Link to survey: https://umanitobapsych.az1.qualtrics.com/jfe/form/SV_3X9Cxm4Q6htwSnb
Same caveats about variable naming apply here!

RAKidExperience – For the main RA testing in your lab, how many participants under the age of 4 had they run prior to MB1?
RAInfantExperience For the main RA testing in your lab, how many INFANTS had they run prior to MB1?
RAHours – On averages, how many hours a week did this person work in the lab?
WalkingDistance – How far do families have to walk to get to the lab?
NumRAs - At a minimum, how many lab members are needed to run an MB1 session in your lab?

VisualClutterWarmup- In the space where Caregivers and families sit to fill out paperwork and 'warm up', how much visual clutter is there?
PeacefulWarmup- On a scale from 1 (very peaceful) to 7 (very stimulating), how peaceful/stimulating would you say your warmup room is?
WarmupToTestDistance- How far apart are the spaces where families fill out paperwork and 'warm up' and where the MB1 experiment happens?
VisitOrdering- Which of the following sequence matches the **typical** experience for families participating in MB1 in your lab?
WallColor - What colour(s) are the walls in your testing room or the booth you test children in?
BoothLighting - How dark is the lighting during testing?
BoothSize – How large in width/depth is the testing room (or booth/section, if applicable) where the baby & caregiver sit during MB1? (e.g. 4 feet by 5 feet)
VisualClutterBooth - In the space where MB1 experiments actually take place, how much visual clutter is there?
PeacefulBooth - On a scale from 1 (very peaceful) to 7 (very stimulating), how peaceful/stimulating would you say your testing area is?
Multipurpose - Is this room/space used for anything other than the MB1 experiment or similar setups
StimSize - How large is your visual display for MB1 (i.e. the checkerboard) in inches (e.g. 8 inches by 10 inches)
DisappearingRA – Did the infant see an RA who entered the room with the family and then went out of the child's view, and was that RA still partly visible?
CaregiverAudioBlinding - How, if at all, were caregivers participating in MB1 kept from **hearing** the experiment?
CaregiverVisualBlinding - How, if at all, were caregivers participating in MB1 kept from **seeing** the experiment?

## Video Walkthrough

After much discussion, we have decided to hold off deciding! Melanie Soderstrom is leading a student project that will do an initial qualitative analysis on variation between labs; we hope to be able to compare videos to self-reported practices as well.

# Author List

Melissa Kline (mekline@mit.edu)
Krista Byers-Heinlein (k.byers@concordia.ca)
Kiley Hamlin (kiley.hamlin@psych.ubc.ca)
Christina Bergmann (chbergma@gmail.com)
Melanie Soderstrom (M_Soderstrom@umanitoba.ca)
Casey Lew-Williams (caseylw@princeton.edu)
Elizabeth A. Simpson (simpsone@miami.edu)
Michael C. Frank (mcfrank@stanford.edu)
Stephanie Barbu (stephanie.barbu@univ-rennes1.fr)

Virginie Durier (virginie.durier@univ-rennes1.fr)
Jennifer Rennels (jennifer.rennels@unlv.edu)
Eon-Suk Ko (eonsukko@chosun.ac.kr)

## Relevant Docs

**The Mega-MB1 Data Dictionary <- should get added to the MB1 data folder too**

**TODO: Upload and Re-Link**

MB1 Main preregistration

MB1B doc (for comparison)

Sample data & data dictionary - Participant level

Sample data & data dictionary - Trial level

Participant questionnaire

Data entry form

MB1 pilot data - use for developing R code analyses

MB1 ‘Big Manual’

Walkthrough Video Instructions

Text from Laboratory Questionnaire

Weston et al. (2018): Best practices for secondary analysis preregistration