Data Request form YOUth (version 3.0, September 10, 2019)

Introduction

The information you provide here will be used by the YOUth Data Management Committee to evaluate your data request. Details on this evaluation procedure can be found in the Data Access Protocol.

Moreover, your data request will be stored in an online repository available to all researchers who submit or have submitted a data request. The aim of this repository is to provide a searchable overview of past, current, and pending data requests. By default, we will publish the following information from your request on our researcher's website:

- After submission of a data request: the names and institutions of the contact person and participating researchers (**Section 1**) and the research context (**Section 2**).
- After approval of a data request: the complete request (**Section 1-5**).
  Exception: If you believe that publishing the complete request could do harm (e.g. when you propose to use a novel analysis technique) you can object to publishing the complete request. This should be indicated on the data request form with a rationale (**Section 5**). The YOUth Data Management Committee will review your matter and advise the YOUth Executive Board whether or not to publish the complete request. If you do not agree with the YOUth Data Management Committee about publishing the complete request, you have the possibility to withdraw your data request.

Section 1: Researchers

In this section, please provide information about the researchers involved with this data request.

- Name, affiliation and contact information of the contact person
- Name and details of participating researchers (e.g. intended co-authors)
- Name and details of the contact person within YOUth

Contact person for the proposed study:	
Name:	Bauke van der Velde
Institution:	Utrecht University
Department:	Depts of experimental and developmental psychology
Address:	Martinus J. Langeveldgebouw H037; Heidelberglaan 1; 3584 CS Utrecht
Email:	b.vandervelde@uu.nl
Phone:	+31 302533340

+ for other participating researchers

Contact person for the proposed study:	
Name:	Caroline Junge
Institution:	Utrecht University
Department:	Depts of experimental and developmental psychology
Address:	Martinus J. Langeveldgebouw H066; Heidelberglaan 1; 3584 CS Utrecht
Email:	c.m.m.junge@uu.nl
Phone:	+31 302531266

Contact person in YOUth Data Management Committee:	
Name:	Chantal Kemner
Institution:	Utrecht University
Department:	Depts of experimental and developmental psychology
Email:	c.kemner@uu.nl

Section 2: Research context

In this section, please briefly describe the context for your research plans. This section should logically introduce the next section (hypotheses). As mentioned, please note that this section will be made publicly available on our researcher's website after submission of your request.
Please provide:

- The title of your research plan
- A very brief background for the topic of your research plan
- The rationale for and relevance of your specific research plan
- The specific research question(s) or aim(s) of your research (Please also provide a brief specification)
- A short description of the data you request

References can be added at the end of this section (optional).

Title of the study
EEG recordings in developmental cognitive neuroscience: examining possible sources of data loss

Background of the topic of your research plan, rationale, relevance (max. 500 words)
Because EEG does not require children to make overt responses, its key appeal to developmental cognitive neuroscientists is that one can use the same passive task to gauge cognitive development across ages. EEG is therefore a popular testing method for infant studies, and indeed one of the main methods in the YOUth study to test development from infancy to early childhood. While its main benefit is its relative non-invasiveness, which allows children to move (relatively) freely while being measured, on the downside is that EEG studies with infants typically have large drop-out rates, ranging from 20-60%. There are multiple reasons why infants do not make it to the final sample or contribute fewer trials than administered. Some do not tolerate the cap, while others move too much during EEG recordings which contaminates recordings. Indeed, even though EEG allows for movement, noise in EEG data is far more common in infants/toddlers compared to adults. Also, their shorter attention spans for testing might lead to suboptimal testing times. Moreover, even data loss for infants who remain in the sample might lead to suboptimal data from which wrong conclusions could be drawn, which hampers comparisons to adult research and could negatively influence reliability in child research. Given the typical small sample sizes in infant research, such data loss is usually accepted prima facie, and considered inherent to infant studies. But testing infants is time-consuming, costly, and requires the support of parents and children. This begs a better understanding of the external reasons that contributes to data loss. To our knowledge, there is no study that examined the extent of such factors in contributing to data loss in EEG studies. The current paper addresses this: our main aim is to describe how a range of external factors potentially impacts EEG data quality in a large population-based longitudinal study. We first study how external effects concerning task factors (whether children are sitting on parent's lap or in high chair; length of experiment) and non-technical factors (research assistant; time of day; presence of additional measurements preceding data collection; age of child) influence data quality (i.e., data loss) in individual sessions. Secondly, we take advantage of the fact that YOUth is a longitudinal study, and we examine whether there are individual differences in testability: we ask whether some infants are more prone than others to data loss across sessions. To generalize our findings, we will carry out both analyses for the event-related potential task (face-house) as well as for the EEG-coherence task. Both tasks differ in their requirements in artefact-free data.

In sum, our analyses will further our understanding which factors contribute to data quality in infancy and toddlerhood. We will end with recommendations to prevent data loss in EEG recordings as much as possible.

The specific research question(s) or aim(s) of your research
1) Which (external) factors influence EEG data quality in single sessions? a) For ERP-facehouse? b) For EEG-coherence? 2) How does EEG data quality develop over sessions in longitudinal studies? a) For ERP-facehouse b) For EEG-coherence?

Summary of the data requested for your project: Please indicate which data you request to answer your research question.
Entire tested population so far (5mo/10mo/3yo). We need as large a subset as possible to make better substantiated conclusions regarding the influence of longitudinal study design on EEG data quality.

References (optional)
Hessels, R. S., & Hooge, I. T. (2019). Eye tracking in developmental cognitive neuroscience–the good, the bad and the ugly. *Developmental cognitive neuroscience*.

Section 3: Hypotheses

In this section, please provide your research hypotheses. For each hypothesis:

- Be as specific as possible
- Provide the anticipated outcomes for accepting and/or rejecting a hypothesis (or explain why this does not apply to your project, e.g. when using Bayesian statistics)

Exception: if you plan a hypotheses-free project, please use this section to explain why you don't formulate specific hypotheses.

Hypotheses
It is a descriptive study. Data loss will be related to the following external factors to determine the influence of these factors on data quality on single sessions for both tasks separately: - Research assistant (we expect that some researcher assistants are more capable than others; cf. for a similar case with eye tracking Hessels & Hooge, 2019) - Time of day (Morning vs afternoon; we expect that data loss is less severe in the morning) - Experiment order (whether EEG-session was first, or later in the whole procedure; we expect that data loss is less severe when we started with EEG session) - Length of experiment (chance of data quality loss vs time; we expect that data loss increases over time) - Electrode set (no hypothesis here; explorative; it should not matter) - How the child is seated (we expect that 10 month-olds are more likely to sit in a high chair; whether or not this affects data loss we have no hypothesis) - age (we expect less data loss in 3-yr-olds than in infants because older children can be better instructed and have longer attention spans; with respect to 5-month-olds vs 10-month-olds

– we expect more severe data loss to movement in the 10-month-olds, whereas 5-month-olds might be more likely to fall asleep)

Comparison between ERP and EEG coherence: we expect that external factors show similar patterns of data loss, albeit less extremely for the EEG coherence task because coherence analysis poses less strict constraints on artefact-free data.

Longitudinal aspect: - it is an empirical question whether or not the same infants contribute equal amounts of data loss across sessions, or are just as likely to tolerate capping.

Section 4: Methods

In this section, you should make clear how the hypotheses are tested. Be as specific as possible. Please describe:

- The study design and study population (Which data do you require from which subjects?)
- The general processing steps (to prepare the data for analysis)
- The analysis steps (How are the data analysed to address the hypotheses? If possible, link each description to a specific hypothesis)
- Any additional aspects that need to be described to clarify the methodological approach (optional)

Study design, study population and sample size (e.g. cross-sectional or longitudinal; entire population or a subset; substantiate your choices)

For the first set of descriptive analyses, we treat data cross-sectionally, and describe per age rage the extent of data loss, and how a range of factors might affect data loss. This means that for each task (facehouse and eegcoherence) we need as many participants as possible per age. Although both tasks are usually administered in one take, it is not necessary that all subjects must have data for both – we consider all data, even when only one of the tasks is (partly) administered. Missing data is an important metric in this study, as it is part of the aims to describe how much missing can occur

For the longitudinal aspect of the study, it is necessary to have subjects who participated at 5 months as well as 10 months; ideally also at 3 yrs.

Specific processing and analysis steps

Trial creation differs between experiments. For the ERP experiment, the presence of the stimulus determines the start and end of the trial. We will follow the same procedure as Di Lorenzo (submitted) in trial acceptance (except checks for infants' attending the screen). For the coherence experiment, the total length of the videos (60s) is divided into 60t trials of 1 second (for each video (6)). Artifacts will be defined as absence of signal, clipping, muscle artifacts and excessive noise. Data loss will be defined as the percentage trials with artifact (or missing) data compared to the total expected amount of data for each participant.

General processing steps to prepare the data for analysis

First, we make an inventory for how many subjects per age per task we have collected EEG data, (and logbook data available). EEG data will be analyzed exclusively using Matlab, by means of the FieldTrip toolbox (Oostenveld et al., 2011). The original 2048 Hz data will be down sampled to 512 Hz, using chip interpolation .

Section 5: Data request

In this section, please specify as detailed as possible which data (and from which subjects) you request. Include information regarding:

- Which wave(s)
- Which experiments, questionnaires, etc.
- How many sets (sample-size)
- Purpose of your data request
- Other aspects relevant to your data request (optional).

Select the appropriate wave(s) (more options are possible):

☐ Rondom zw – 20 weeks
☐ Rondom zw – 30 weeks
☒ Rondom 0 – 5 mo
☒ Rondom 0 – 10 mo
☒ Rondom 3 (data collection has started)
☐ Rondom 6 (not available yet)
☐ Rondom 9
☐ Rondom 12 (not available yet)
☐ Rondom 15 (not available yet)

Experiments and number of sets you request
Entire tested population (5mo/10mo/3yo) that participated in the facehouse and/or facecoherence. We need as large a subset as possible to make better substantiated conclusions regarding the influence of longitudinal study design on EEG data quality. With these sets we also need their logbooks, and information regarding general information (sex; age at test).

Other aspects relevant to your data request (optional)
Missing data is an important metric in this study, as it is part of the aims to describe how much missing can occur. The only kind of missing data that cannot be detected is when no EEG measurement was conducted for a participant in the YOUth cohort. Given that the emphasis is on EEG data quality metrics, we still need to know why such data is missing. We therefore need to consider those children who according to logbook did not start recording. Perhaps willingness to adhere to capping procedure is dependent on time of day or is relevant information for subsequent testing in later waves. **(if logbook says data is recorded, but there is no data – we consider this TRUE missing data – these cases cannot be not examined).**

Data request for the purpose of:

☒ Analyses in order to publish:
- ☒ Article
- ☐ Report
- ☒ Thesis
- ☐ Other. Please specify

☐ Analyses for data quality control only (data will not be published)

☐ Analyses for descriptive data only, e.g. in order to determine good datasets (data will not be published

DISCLAIMER DATA ACCESS QUALITY CONTROL AND DESCRIPTIVE DATA: These data can only be used for data quality control analyses or descriptive data analyses only and may not be made public, for example by publishing them or otherwise making them available to others. If you want to use data for disclosure, permission of the YOUth data committee is required, and this data request protocol must be followed for analyses in order to publish.

Would you like to be notified when a new data lock is available?
In principle, data will be made available in data locks twice a year. This means that twice a year, the data is locked on a specific date and that all approved data request projects will receive the same locked data set.

☐ Yes
☒ No

Do you agree with publishing the complete request on our researcher's website after it is approved (by default)?

☒ Yes
☐ No. Please provide a rationale below.

NOTE 1: Please fill out the 'Form contributions to YOUth data collection' in Annex 1 to specify your contribution to YOUth in order to gain access to the requested data.
NOTE 2: Please fill out the 'Data Selection Template' (.xlsm) to specify the sort of data you request.

This Annex 1 together with the Data Selection Template and this Request Form should be sent to the Secretary of the Scientific Director (I.Bleeker@uu.nl).