

Preprocessing ESM data:

A step-by-step framework, reporting templates, tutorials, and R code website

¹Research Group of Quantitative Psychology and Individual Differences, KU Leuven

²Center for Contextual Psychiatry, KU Leuven

³Family Lab, Ghent University

Collected data

Dyad	ID	Beepnr	Scheduled	Sent	Start	End	Item1	Item2	Item3	Item4	Item5
1	1	1	8:03	8:03	9:50	9:54	34	32	67	1	5
1	NA	NA	10:10	10:10	10:25	12:01	85	19	34	4	6
1	1	3	12:30	12:25	NA	NA	2	1	2	1	1
3	1	5	14:05	14:06	NA	NA	0	0	__na__	0	0
1	1	4	16:19	16:19	16:19	NA	83	38	17	2	3
1	1	6	18:03	18:03	18:05	18:05	101	130	-49	6	4
NA	1	7	20:13	20:15	20:15	20:15	8	10	37	3	6

NA = missing value

Collected data

Dyad	ID	Beepnr	Scheduled	Sent	Start	End	Item1	Item2	Item3	Item4	Item5
1	1	1	8:03	8:03	9:50	9:54	34	32	67	1	5
1	NA	NA	10:10	10:10	10:25	12:01	85	19	34	4	6
1	1	3	12:30	12:25	NA	NA	2	1	2	1	1
3	1	5	14:05	14:06	NA	NA	0	0	__na__	0	0
1	1	4	16:19	16:19	16:19	NA	83	38	17	2	3
1	1	6	18:03	18:03	18:05	18:05	101	130	-49	6	4
NA	1	7	20:13	20:15	20:15	20:15	8	10	37	3	6

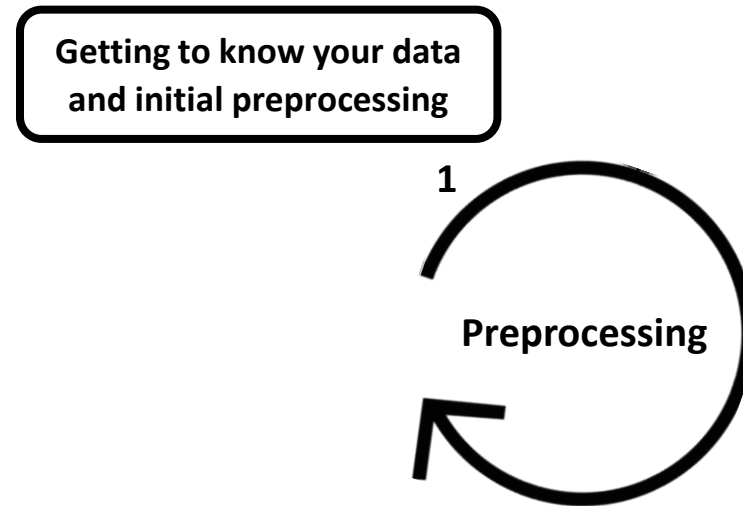
NA = missing value

- Preprocessing, two functions:
 - **Checking and solving issues**
 - **Data quality insight**
- No common framework



Step-by-step framework and Gallery of R functions

The framework



➔ From importing data to checking variable consistency

Step 1: Getting to know your data and initial preprocessing

Dyad	ID	Age	Beepnr	Start	End	Item1	Item2	Item3	Branch	ItemA	ItemB
1	1	21	1	9:50	9:54	34	32	67	1	3	NA
1	NA	21	NA	10:25	12:01	102	120	-34	2	NA	4
NA	1	21	3	NA	NA	0	0	0	0	0	0
...
5	10	30	14	16:19	16:25	83	38	17	1	NA	4
5	10	NA	13	18:05	18:05	1	30	49	1	4	NA
5	10	NA	13	18:05	18:05	1	30	49	1	4	NA

Consistency (bracketed over Dyad, ID, Age)

Out of range (arrow pointing to Item3 value -34)

Missing code (bracketed over the third row)

Duplication (bracketed over the last two rows)

Coherence in branching (bracketed under Branch, ItemA, ItemB)

ESM preprocessing gallery

- A gallery ([link](#)) that follows the preprocessing framework:
 - Tutorials
 - R code
 - Functions

ESM Preprocessing Gallery

Welcome to the R Gallery for preprocessing data from ESM studies! This website, based on Revol et al. (in preparation), is a comprehensive resource for those interested in preprocessing data from ESM studies. The purpose is to help researchers in navigating the ESM preprocessing steps (ESM preprocessing framework), by providing them with helpful tools and resources (R code and functions) and assisting them in reporting this critical step effectively. Basic knowledge of R, `dplyr`, and `ggplot2` is required, and some adaptation of the code to your specific situation may be necessary. Please remember to [cite us](#) if you find our framework and resources helpful in your study.

Variables and data summary

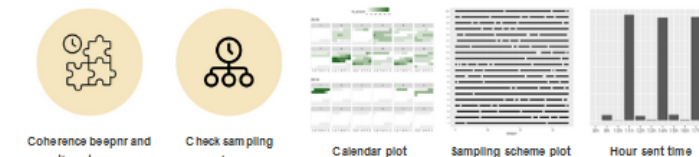
+ More

This section is dedicated to the first look and the first general preprocessing methods when you have imported your data in R. In the process, it helps to better understand the structure and the content of the dataset before going into more ESM specific preprocessing steps.

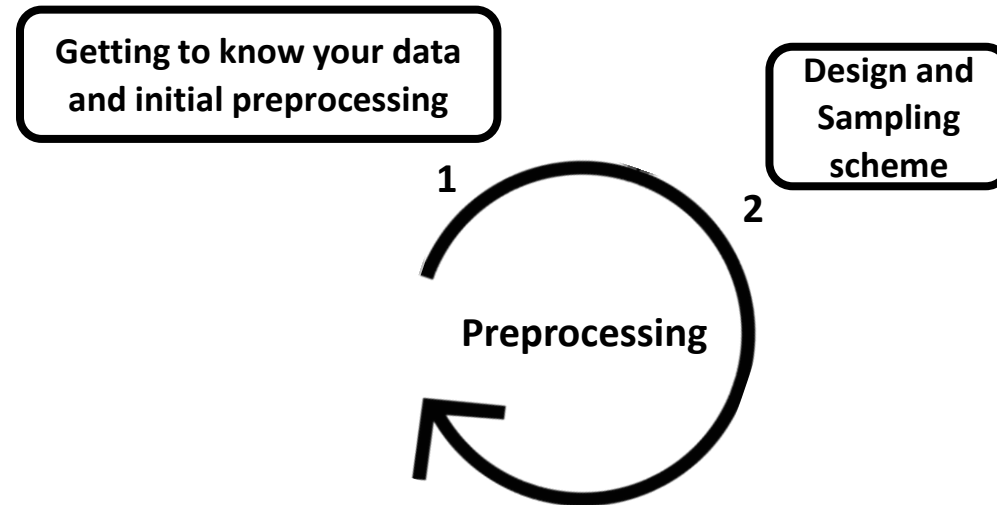


Design and Sampling scheme

This section is dedicated to checking if the design and the sampling scheme of the study have been well followed.

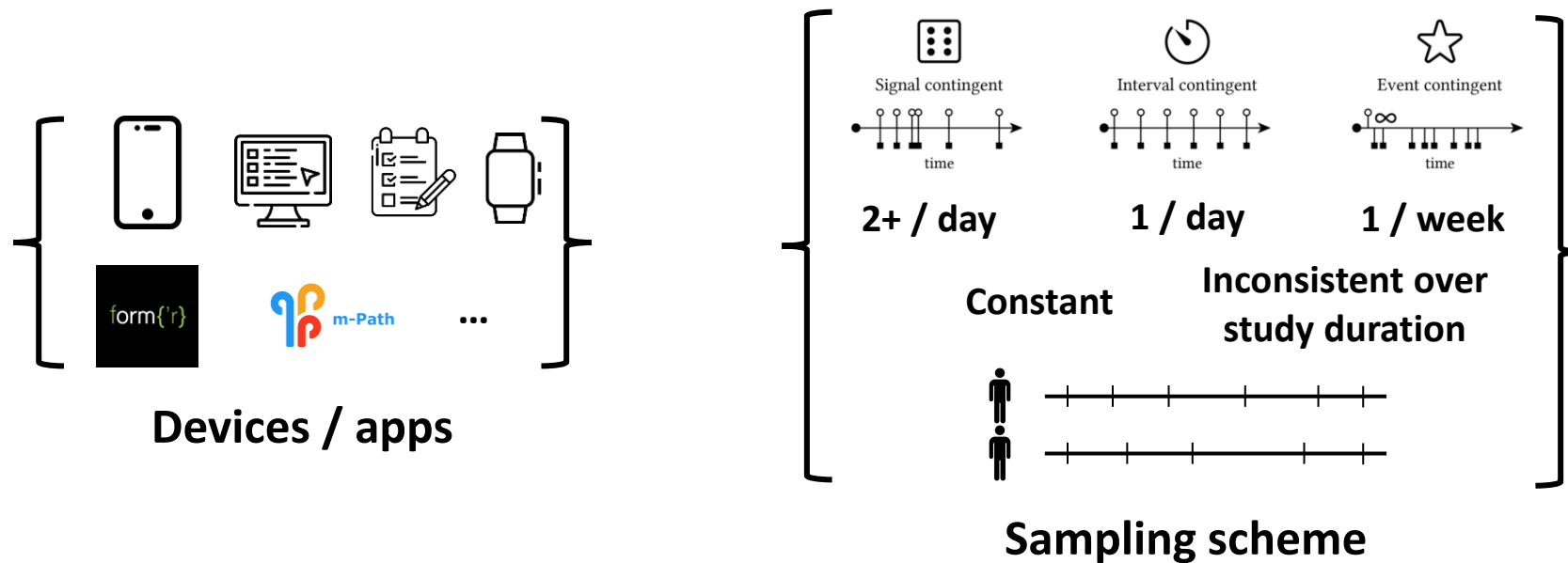


The framework



- ➔ Check if no mismatches between the sampling scheme determined by the study design and the data collected by the app/device.

Step 2: Design and Sampling scheme complexity



Step 2: Design and Sampling scheme complexity

Dyad	ID	Beepnr	Scheduled	Sent	Start	End	Item1	...	Item5
1	1	1	8:03	8:03	9:50	9:54	34		5
1	1	2	10:10	10:10	10:25	12:01	85		6
1	1	3	12:30	12:25	NA	NA	2		1
1	1	5	14:05	14:06	NA	NA	NA		NA
1	1	4	16:19	16:10	16:19	16:23	83		3
1	1	6	18:03	21:03	18:05	18:05	101		4
1	1	7	23:13	23:15	20:15	20:15	8		6

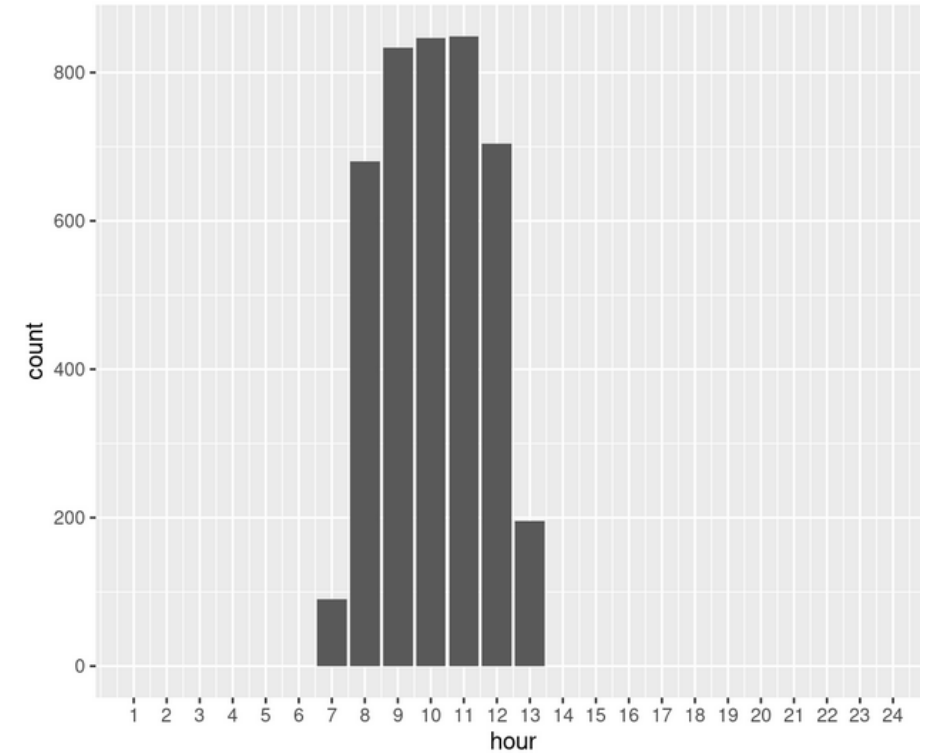
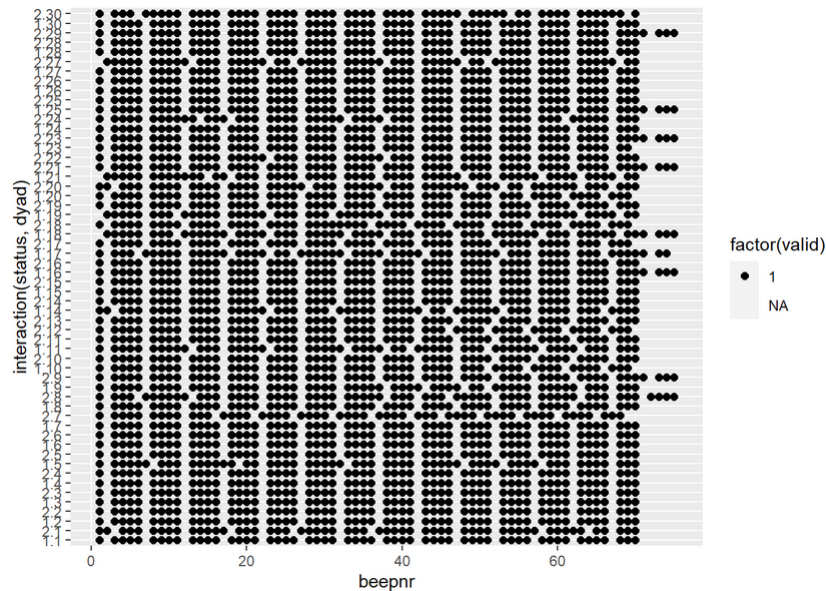
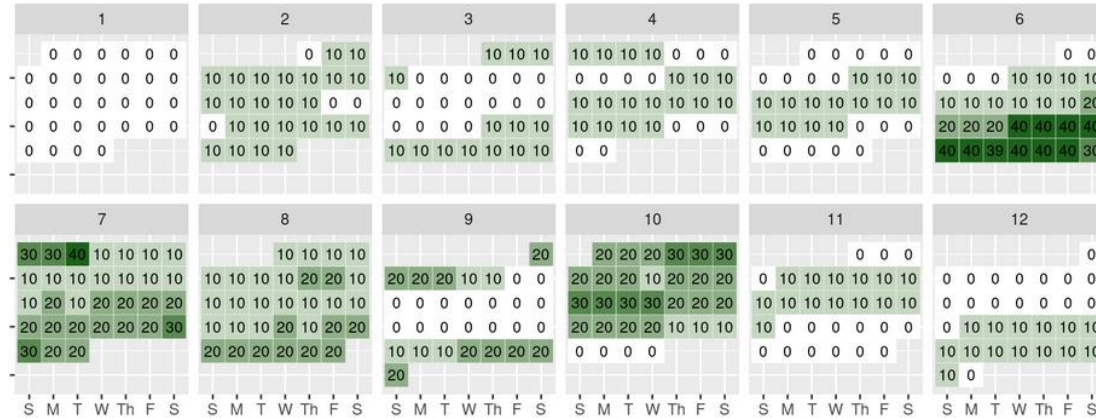
Coherence Time interval

Sampling scheme

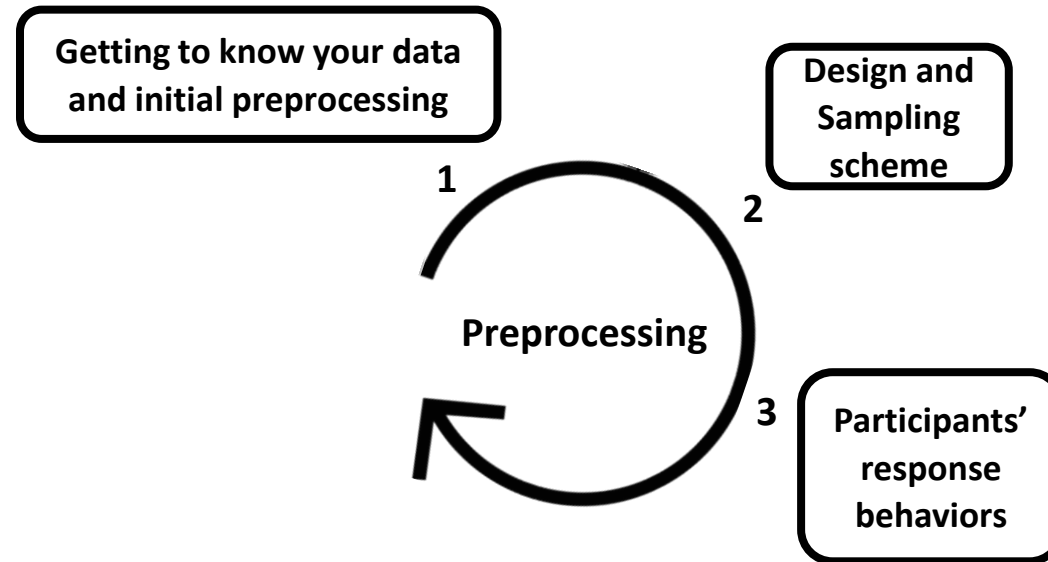
What time/date?

Step 2: Design and Sampling scheme complexity

2018

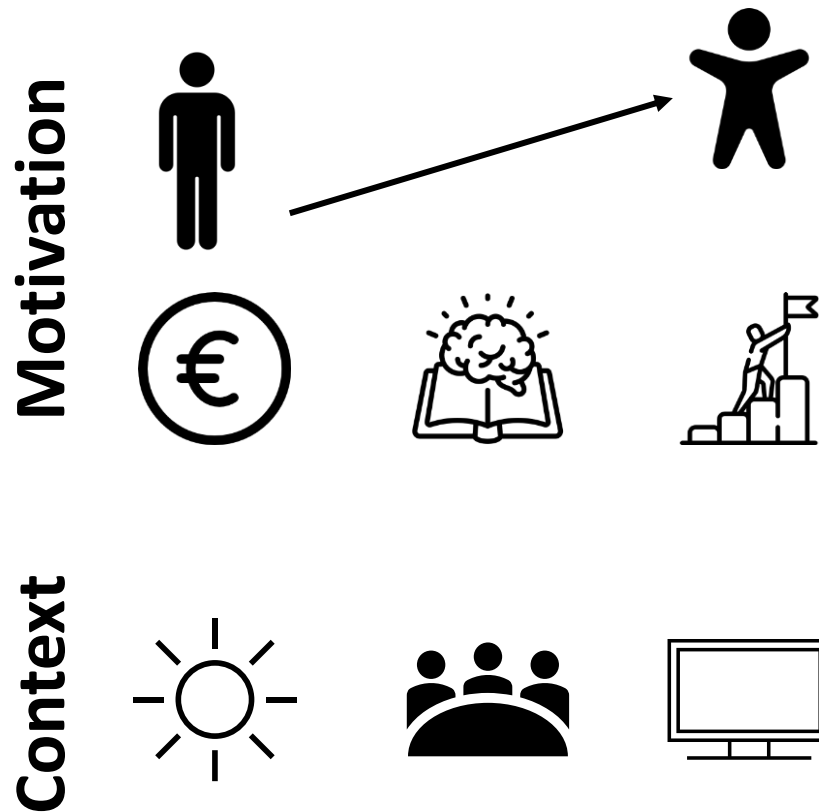


The framework



➔ Large variability in the participants' response behaviors, both between and within

Step 3: Participants' response behaviors



- Careless responding
- Compliance rate
- Pattern of responses

Explain behaviors
↓

Time intervals
┌ ──┐ ┌ ──┐
└ ──┘ └ ──┘

Explain behaviors
↓

Dyad	ID	Age	Beepnr	Valid	Sent	Start	End	Place	PA1	PA2	PA3	PA4	PA5
1	1	25	1	1	8:03	9:50	9:54	Home	34	32	67	1	5
...	↑ Gaps ↓
1	1	25	18	1	10:10	10:25	10:25	Work	99	100	97	98	99
1	1	25	19	1	12:25	12:30	14:31	Home	1	100	1	100	1
...
4	7	46	4	1	14:19	14:19	14:20	Work	83	38	17	2	3
4	7	46	5	0	16:19	NA	NA	Home	NA	NA	NA	NA	NA
4	7	46	6	0	18:03	NA	NA	Home	NA	NA	NA	NA	NA
4	7	46	7	0	20:15	23:30	23:36	Home	39	45	41	54	30

Careless

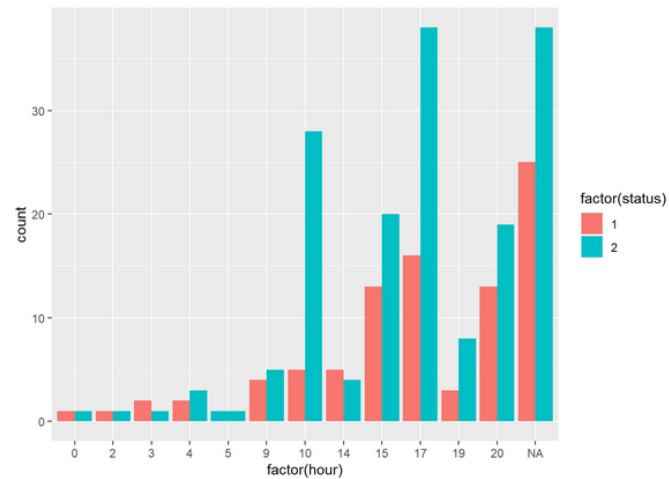
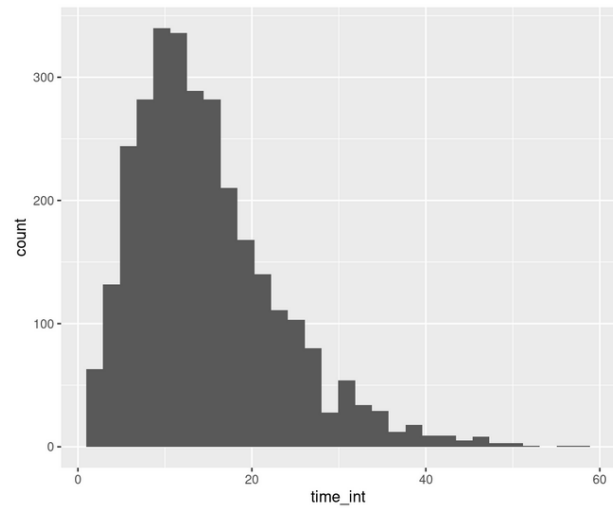
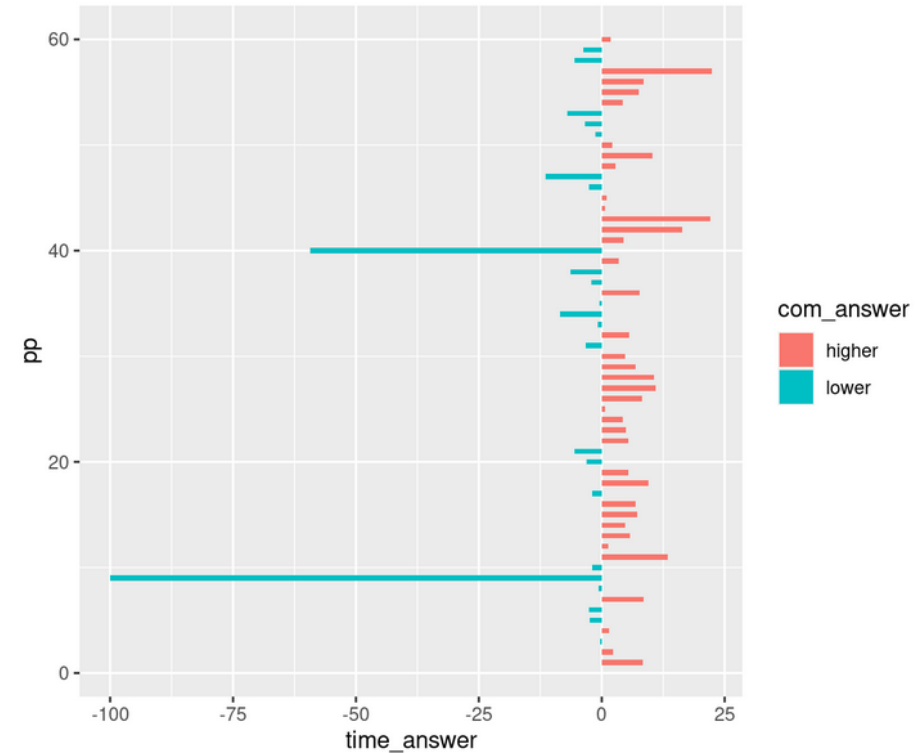
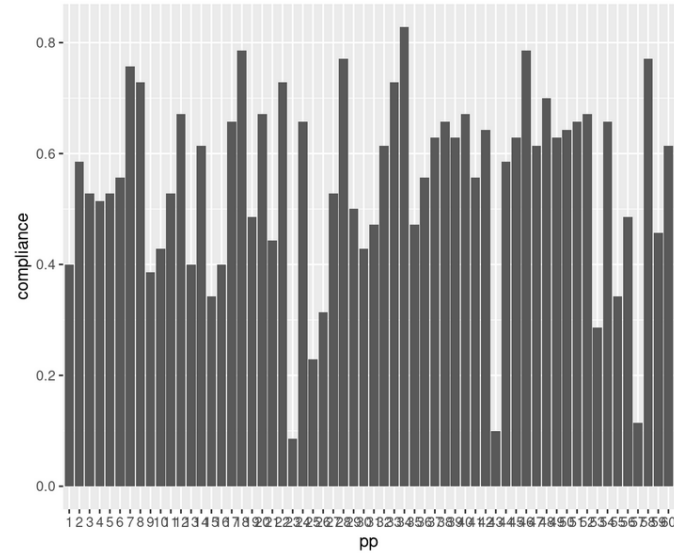
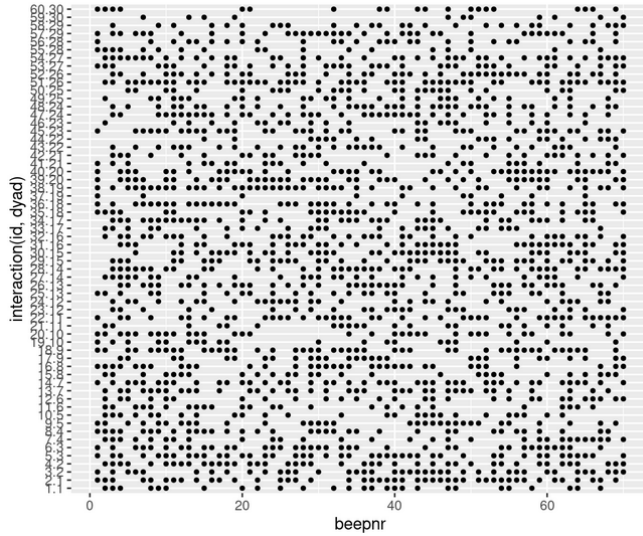
Pattern: missing at home

↑
Response rate
Compliance

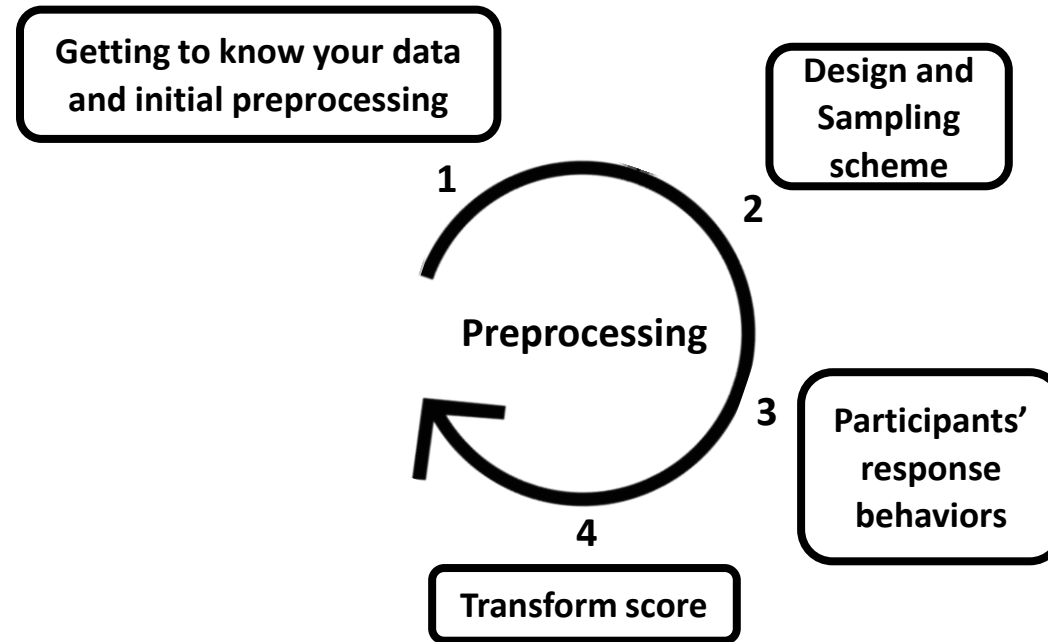
↑ ↑
Response time

↑
Missed time

Step 3: Participants' response behaviors

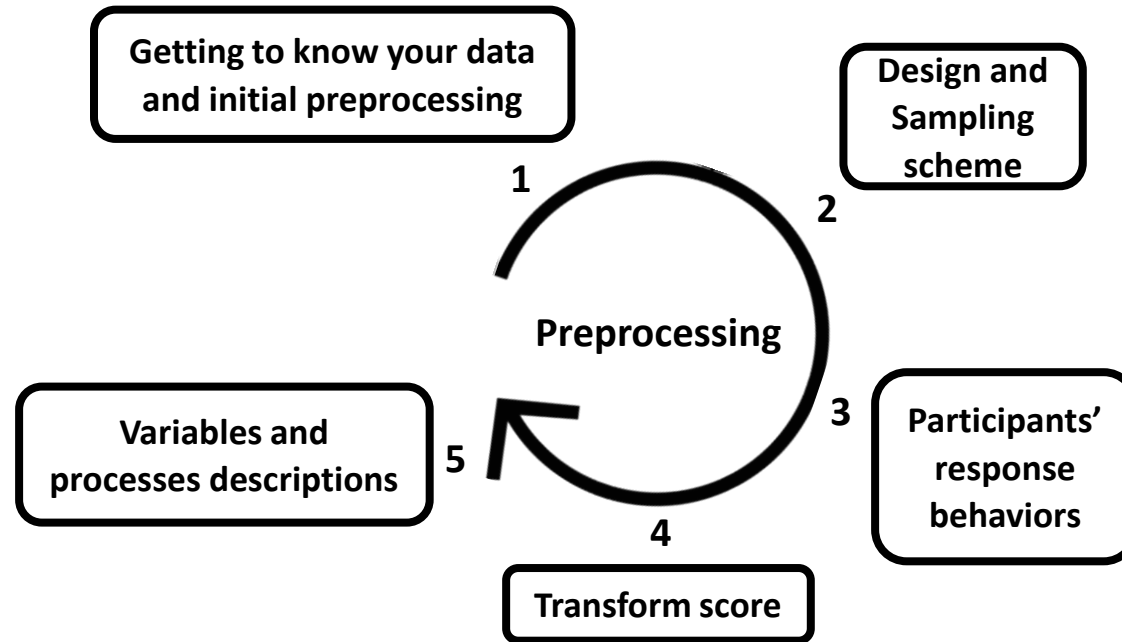


The framework



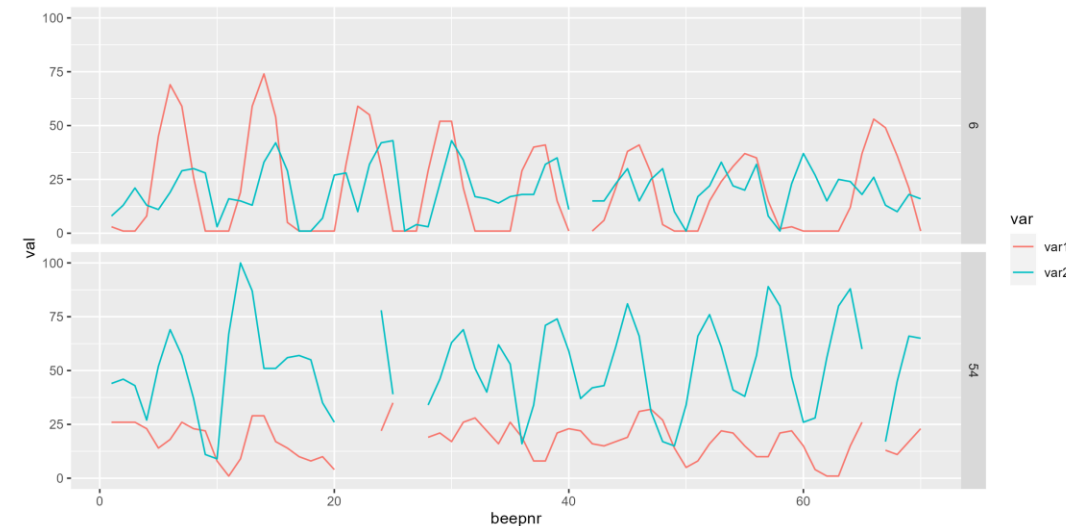
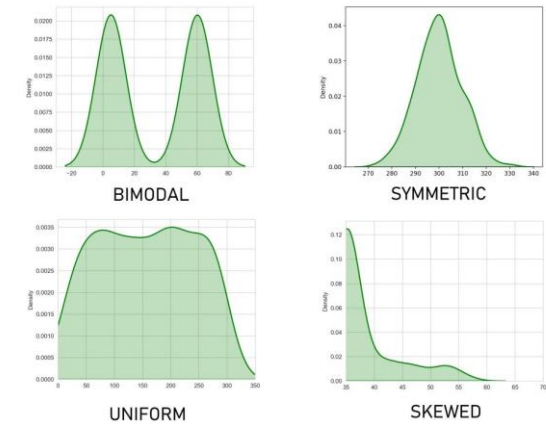
➔ Variety of variables/scores to compute
(e.g., time-invariant, time-variant, centering, window computing)

The framework

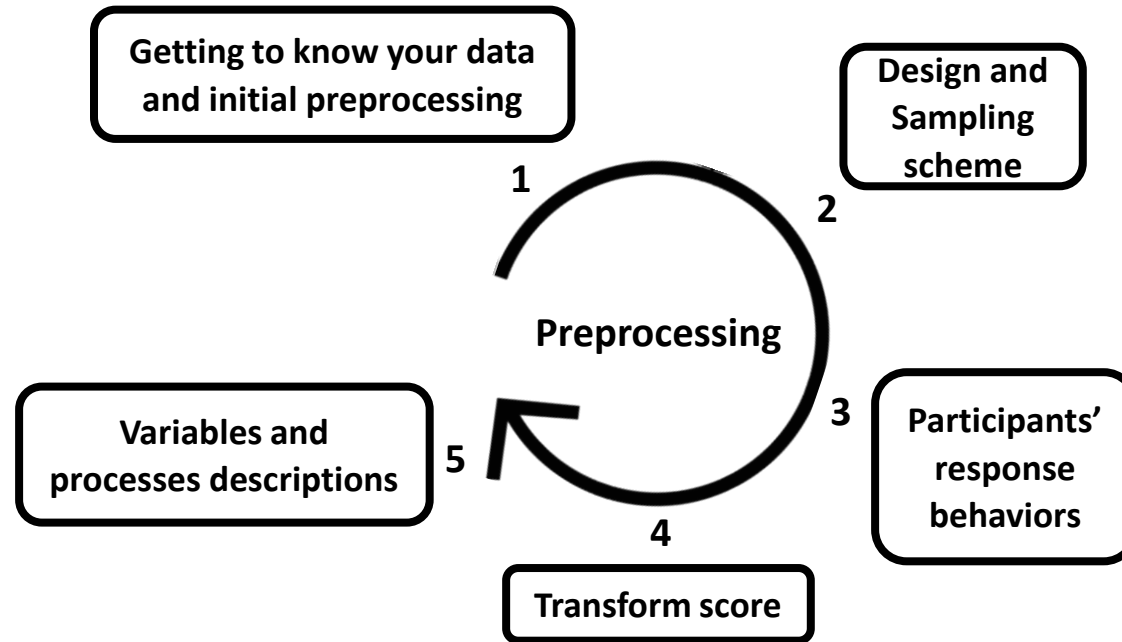


Step 6: Description of the variable of interest

- Variables:
 - Variables' distribution (e.g., floor effect, multimodality, skewness)
 - Implication for statistical model (e.g., semi-continuous variable)
- Insights on participants:
 - Exploring intra-individual processes and inter-individual differences



The framework



Steps	Main task	Composition
1 Getting to know your data and initial preprocessing	First insight on variables and data structure, basic preprocessing and checking variable consistency	<ul style="list-style-type: none"> • Import data and delete pilot/test cases • Check data structure and reformat dataset structure • Rename, relabel, and reformat variables • Check duplication (observation, timestamps, answers) • Check variables' type-values coherence (e.g., timestamps, identification, time-invariant variables) • Create time variables (e.g., beep number, continuous variables) • Overall missing values analysis • Common descriptive statistics (e.g., mean, range)
2 Design and Sampling scheme	Check if data collection design and sampling scheme have been followed well	<ul style="list-style-type: none"> • Overview of the actual sampling scheme • Check the sampling scheme (e.g., times beeps were scheduled and sent, observations outside of the sampling scheme, time to sent beeps, participant duration) • Check consistency of the observation/timestamps order (e.g., within participants and dyads) • Look for missing beeps (missed beeps should be recorded) • Descriptive statistics overall and per participant (e.g., number of beeps sent, duration)
3 Participants' response behaviors	Investigate how well participants engaged with the ESM study looking particularly for problematic patterns of behaviors	<ul style="list-style-type: none"> • Overview of how well participants followed the sampling scheme • Time the beeps were started or missed • Time intervals of responses (e.g., time to start, time per item, time interval between participant or dyad's observations) • Missingness correlates (e.g., time-related pattern) • Careless responding, capture and investigate over- and under-consistent responses • Compliance and response rate/frequency (e.g., over participants or dyads, lagging completion) • Descriptive statistics (e.g., number of beeps, duration, gaps length of missed observations)
4 Transform score	Compute and modify variables of interests	<ul style="list-style-type: none"> • Usual descriptive statistics (e.g., mean, standard deviation) • Lagging and centering • Compute special variables (e.g., systemic and dyadic variables, affect scores) • Check computation procedure and created variables
5 Variables and processes descriptions	Descriptive insights within and between participant processes and on the variables of interest themselves	<ul style="list-style-type: none"> • Variable descriptive statistics (e.g., summary tables) and visualizations (e.g., distributions, correlation plots) • Time series visualization (e.g., trends, intra- and inter-individual differences) • Participants' contexts and states visualizations

A large teal triangle is positioned on the left side of the slide, pointing towards the top-left corner.

Reporting

Transparency, replication and collaboration

Reporting

Preprocess report

Step 1 to 4

Step 1: Getting to know your data and initial preprocessing

Issue A: Spotted issue description



Data modification:

```
data[30:100, 3] = NA
```

...

Step 2: Design and Sampling scheme

...

→ What has been done?

Data description

On preprocessed data

Codebook:

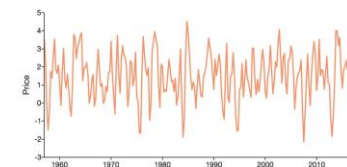
Variable	Position	Label	Measurement Level	Role	Column Width	Alignment	Print Format	Write Format	Missing Values
id	1	Employee Code	Scale	Input	8	Right	F4	F4	
gender	2	Gender	Nominal	Input	1	Left	A1	A1	
bdate	3	Date of Birth	Scale	Input	13	Right	ADATE10	ADATE10	
educ	4	Educational Level (years)	Ordinal	Input	8	Right	F2	F2	0
jobcat	5	Employment Category	Ordinal	Input	8	Right	F1	F1	0
salary	6	Current Salary	Scale	Input	8	Right	DOLLAR8	DOLLAR8	\$0
salbegin	7	Beginning Salary	Scale	Input	8	Right	DOLLAR8	DOLLAR8	\$0
jobtime	8	Months since hire	Scale	Input	8	Right	F2	F2	0
preexp	9	Previous Experience (months)	Scale	Input	8	Right	F6	F6	
minority	10	Minority Classification	Ordinal	Input	8	Right	F1	F1	9

Variables in the working file

Participant book:

Participant	min_date	max_date	nb_answer	compliance	var1_1_stats	var1_1_viz	var1_2_stats	var1_2_viz
1 1	2018-11-21 10:00:26	2018-12-04 20:00:34	1	0.01	mean = 23.17 sd = 24.35 n_length = 35 n_unique = 22		mean = 45.46 sd = 42.29 n_length = 35 n_unique = 18	
2 2	2018-11-21 00:30:31	2018-12-04 20:00:16	8	0.11	mean = 24.15 sd = 31.19 n_length = 47 n_unique = 21		mean = 30.4 sd = 44.93 n_length = 47 n_unique = 10	
3 3	2018-05-11 10:00:36	2018-05-24 20:00:37	5	0.07	mean = 41.38 sd = 18.01 n_length = 45 n_unique = 28		mean = 24.02 sd = 20.84 n_length = 45 n_unique = 20	
4 4	2018-05-11 15:00:25	2018-05-24 20:00:02	7	0.1	mean = 16.16 sd = 15.7 n_length = 44 n_unique = 20		mean = 18.41 sd = 20.05 n_length = 44 n_unique = 20	

Time series visualization:






...

→ What does the data look like?



Take home message

Take home message

- Preprocessing is
 - Time-consuming 
 - Challenging 
 - But primordial! 
- To support researchers, we are developing:
 - A Step-by-Step Framework
 - ESM Preprocessing Gallery ([link](#))
 - Templates to report preprocessing

Any questions?

@JordanRevol
jordan.revol@kuleuven.be