# Investigating the quality of digital trace data and data donation

Alexandru Cernat *(University of Manchester)*

Florian Keusch *(University of Mannheim)*
Ruben Bach *(University of Mannheim)*
Paulina K. Pankowska *(Utrecht University)*

www.alexcernat.com

@cernat_a

# Measuring online behaviour in social research

Digital behaviour increasingly important in the social world

Most studies rely on self-reports from surveys

# Surveys vs. digital trace data

| | **Surveys** |
|---|---|
| **Strengths** | - Probability samples<br>- Freedom of design<br>- Long term comparability |
| **Weaknesses** | - Fragmentary/discrete information<br>- High burden<br>- Measurement error |

# Surveys vs. digital trace data

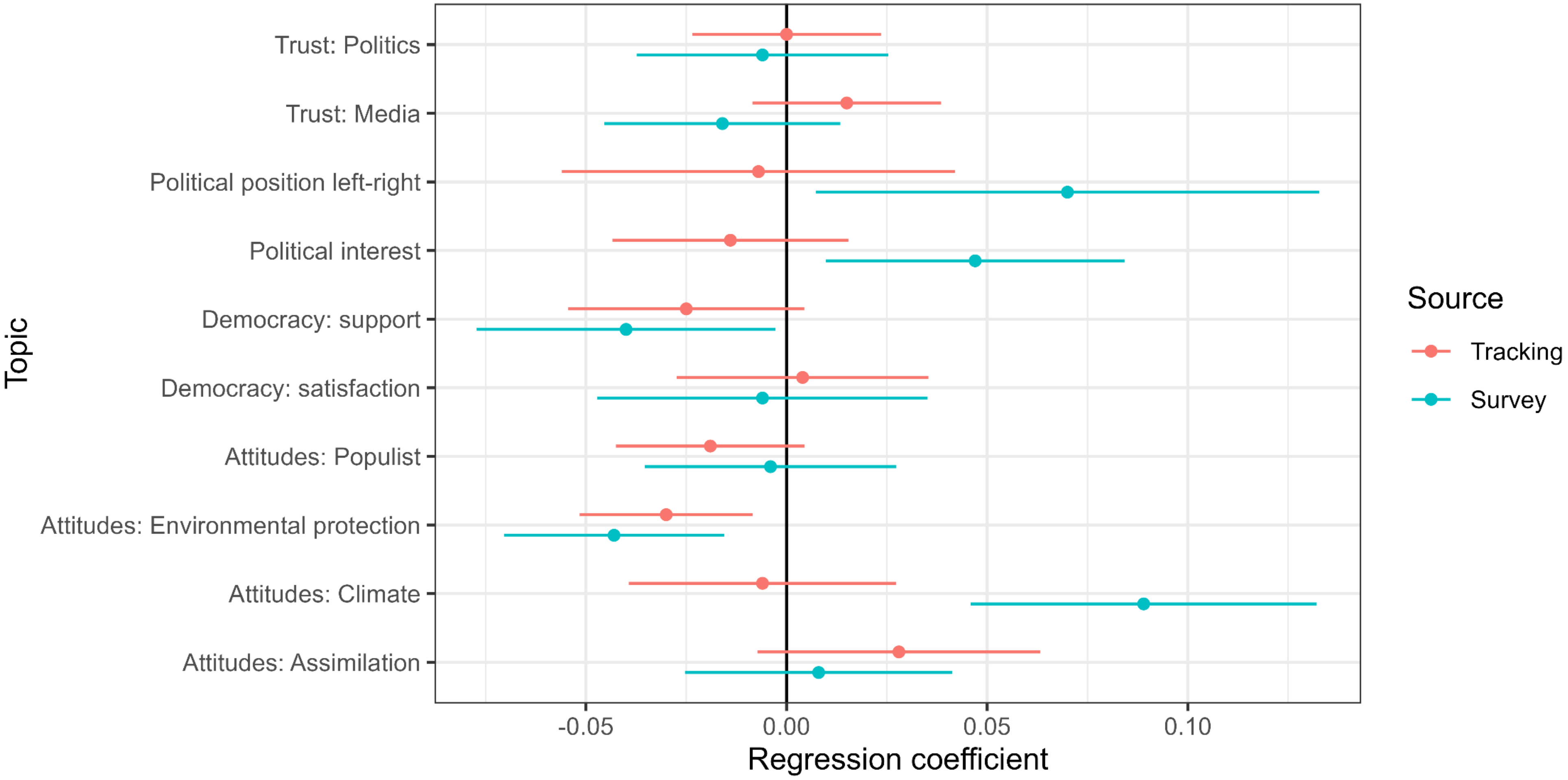
| | Surveys | Digital meter data |
|---|---|---|
| **Strengths** | - Probability samples<br>- Freedom of design<br>- Long term comparability | - Direct measurement<br>- Low burden<br>- Detailed/high frequency |
| **Weaknesses** | - Fragmentary/discrete information<br>- High burden<br>- Measurement error | - Selective/small samples<br>- Technology dependent<br>- Measurement error/missing data |

# Surveys vs. digital trace data

| | Surveys | Digital meter data | Data donation |
|---|---|---|---|
| **Strengths** | - Probability samples<br>- Freedom of design<br>- Long term comparability | - Direct measurement<br>- Low burden<br>- Detailed/high frequency | - Direct access to data<br>- Works with all platforms<br>- Users control info shared |
| **Weaknesses** | - Fragmentary/discrete information<br>- High burden<br>- Measurement error | - Selective/small samples<br>- Technology dependent<br>- Measurement error/missing data | - Convoluted process<br>- Linking with other data<br>- Separate process for each platform |

# Why does it matter?

*The effect of Facebook usage on…*

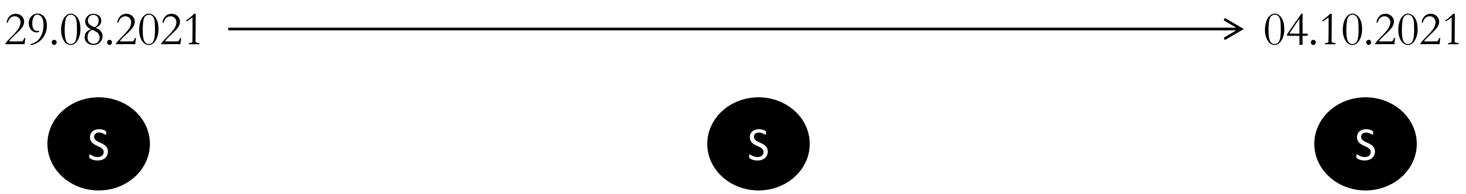


# Understanding the data quality in new forms of data

Understand the selection bias in data donation

Understand the measurement quality of digital trace data

# Our design



29.08.2021

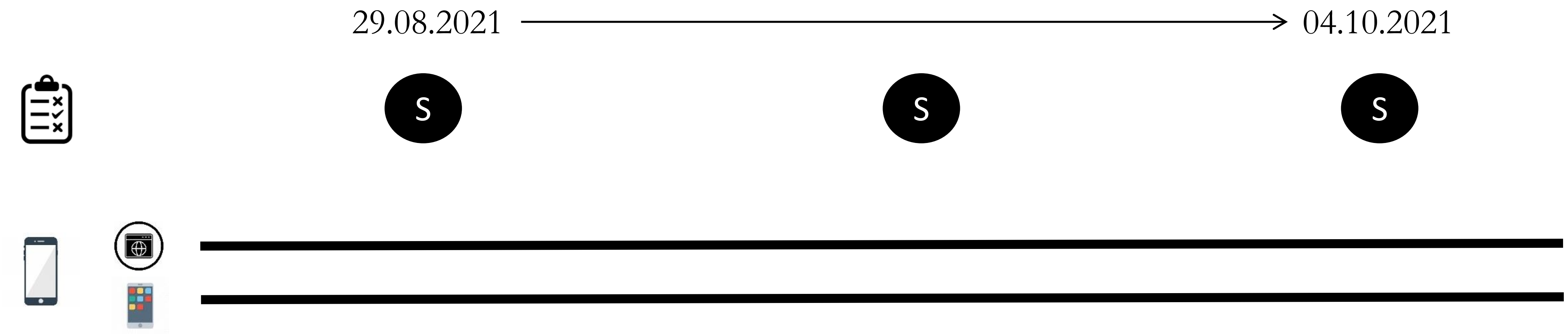
04.10.2021

S

S

S

# Our design



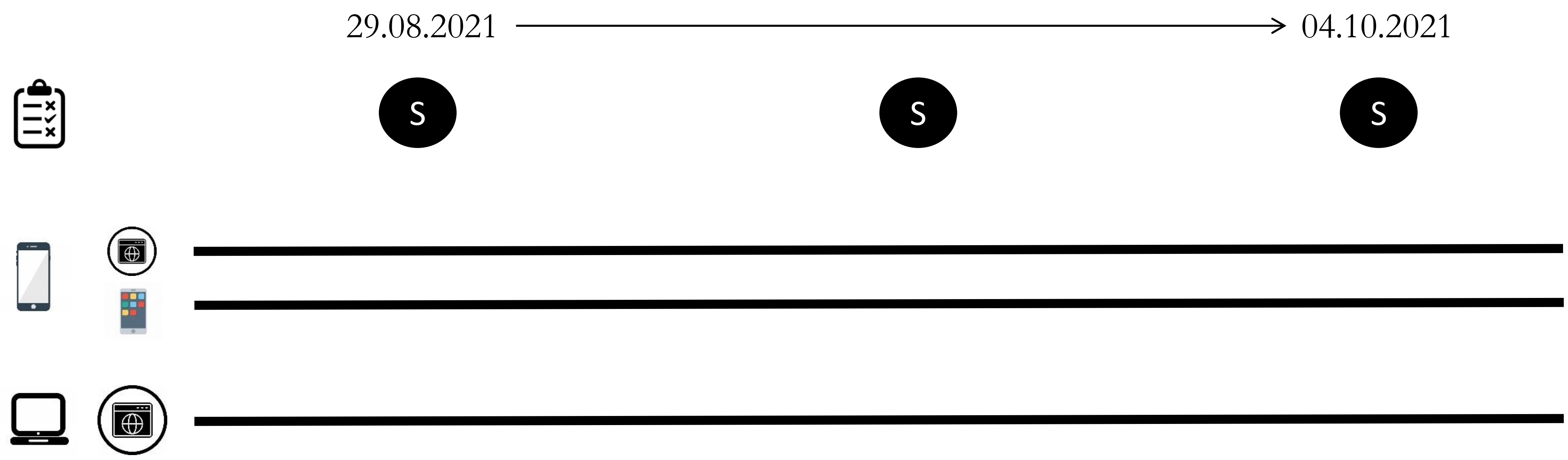
29.08.2021 → 04.10.2021

S

S

S

# Our design



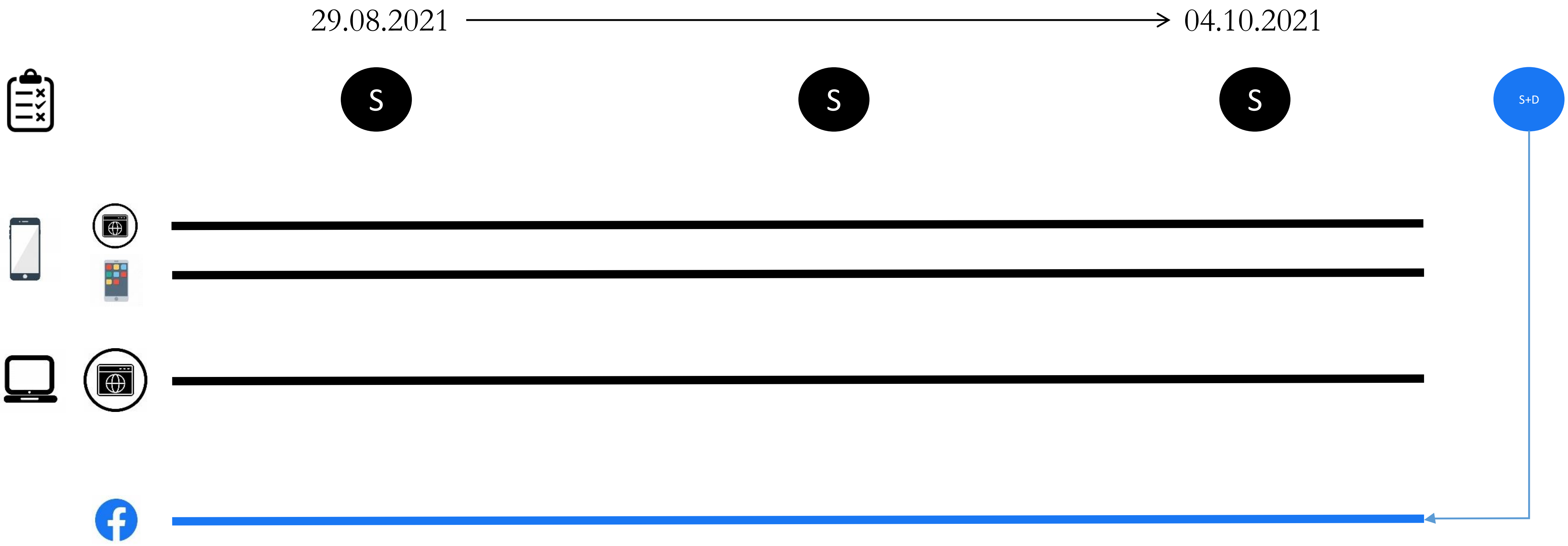
29.08.2021 → 04.10.2021

S

S

S

# Our design



29.08.2021 → 04.10.2021

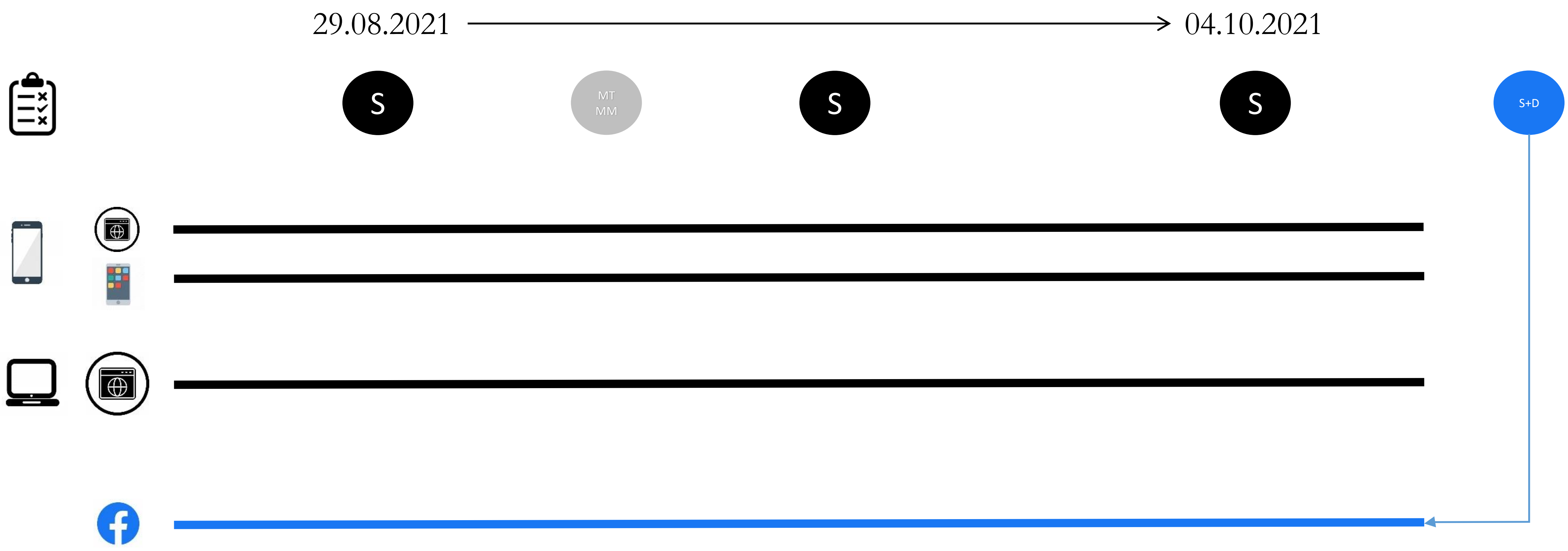
S

S

S

S+D

# Our design



29.08.2021 → 04.10.2021

S

MT
MM

S

S

S+D

# Study 1

Do you have two minutes to talk about your data? Willingness to participate and nonparticipation bias in Facebook data donation

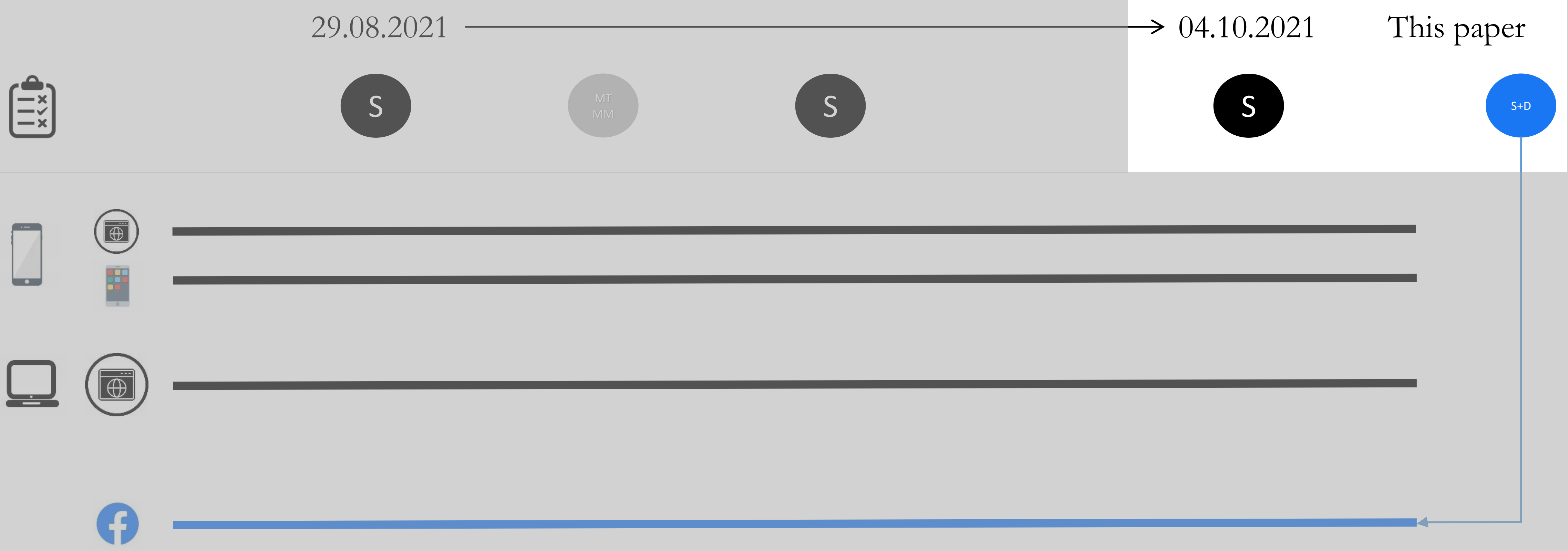
# How large is selection bias with data donation?

*How successful are Facebook users donating the data?*

- *What effect does the framing of the data donation request have on willingness to donate? (gain vs. loss)*

*What bias does arise from selective willingness to donate and successful donation of Facebook data?*

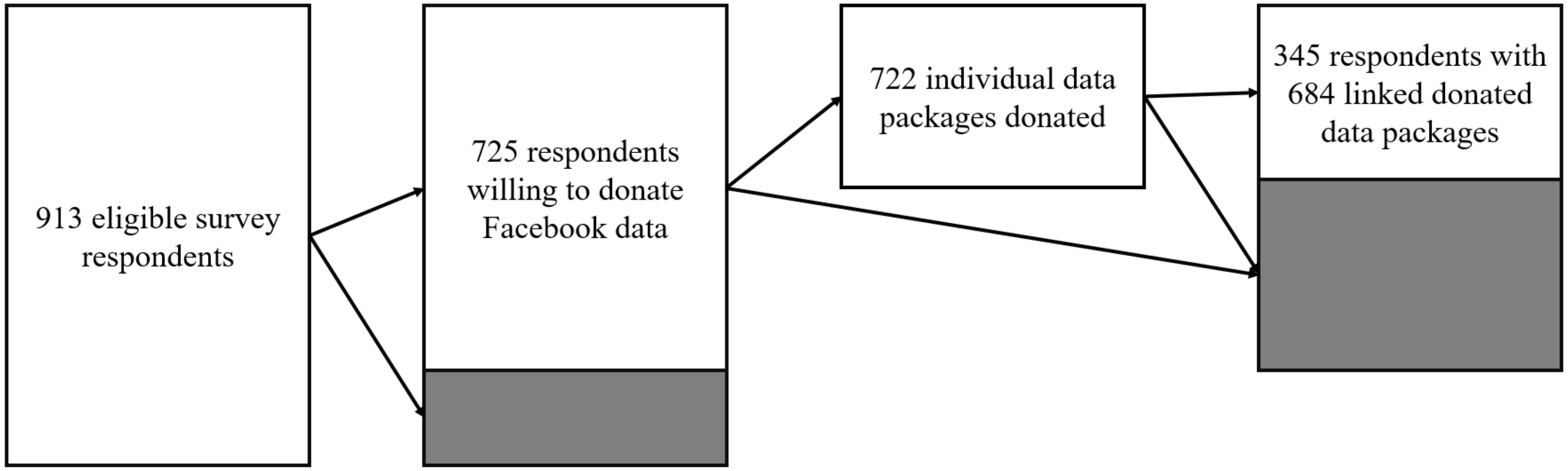
# The data



29.08.2021 → 04.10.2021

This paper

# Participation flowchart



913 eligible survey respondents

725 respondents willing to donate Facebook data

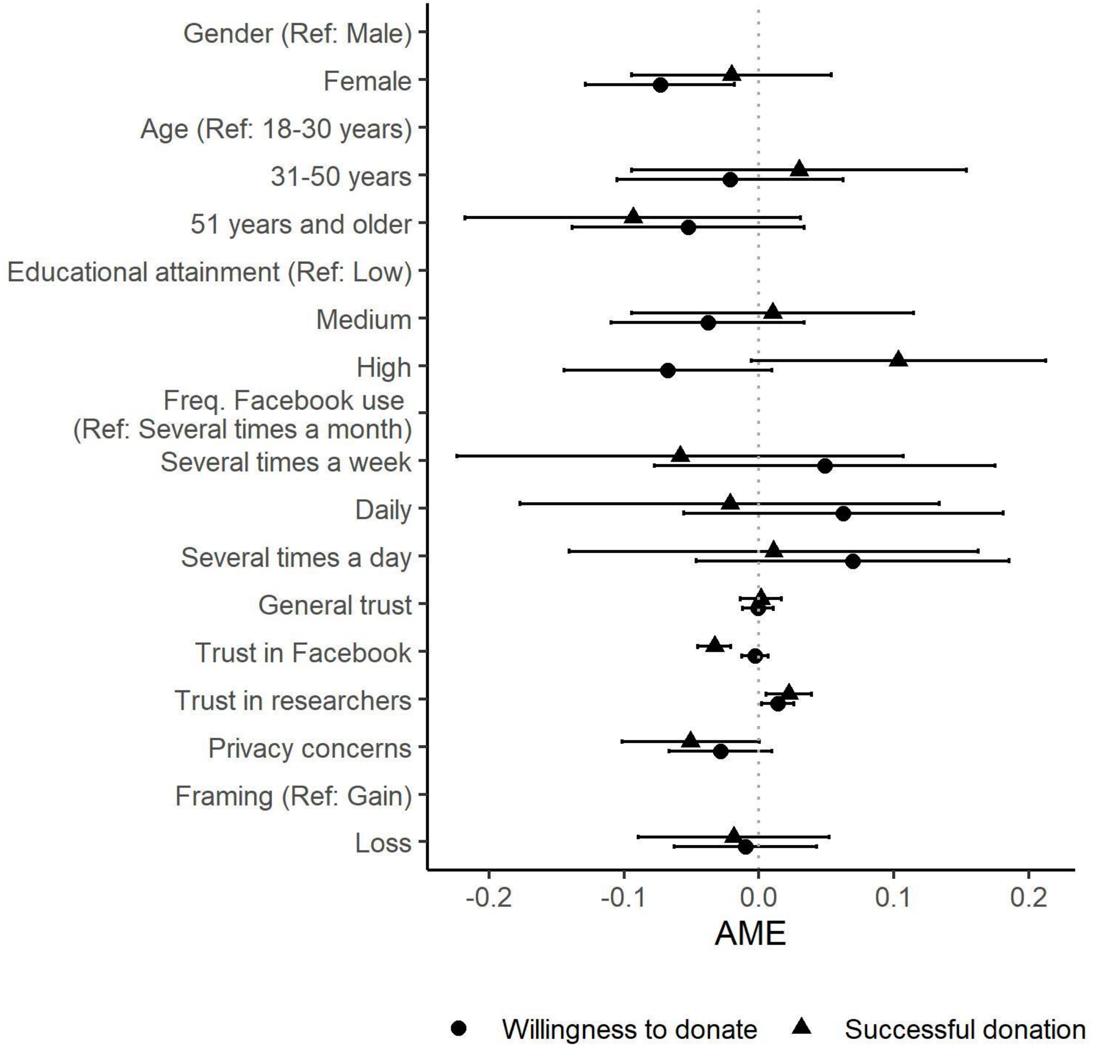
722 individual data packages donated

345 respondents with 684 linked donated data packages

**Willigness to Donate: 79%**

**Successful Donation: 48%**

# How large is selection bias with data donation?



# Conclusions

**Willingness to donate** Facebook data 79%

- gain or a loss framing in the data donation request did not make a difference
- privacy was a major driver for not being willing
- trust in researchers makes people more willing to donate

# Conclusions

**Willingness to donate** Facebook data 79%

- gain or a loss framing in the data donation request did not make a difference
- privacy was a major driver for not being willing
- trust in researchers makes people more willing to donate

**Data donation** rate was 48% (of those willing)

- technical issues with the data donation process
- individuals who expressed lower trust in Facebook were more successful in donating their data
- donors and non-donors did not differ in self-reported frequency of Facebook use, indicating no bias in this substantive measure

# Study 2

Estimating measurement quality in digital trace data and surveys using the MultiTrait MultiMethod model

# How best to measure online behaviours?

Digital trace data seen as a way to complement or replace survey data

Some researchers treat digital trace data as “gold standard” in terms of measurement

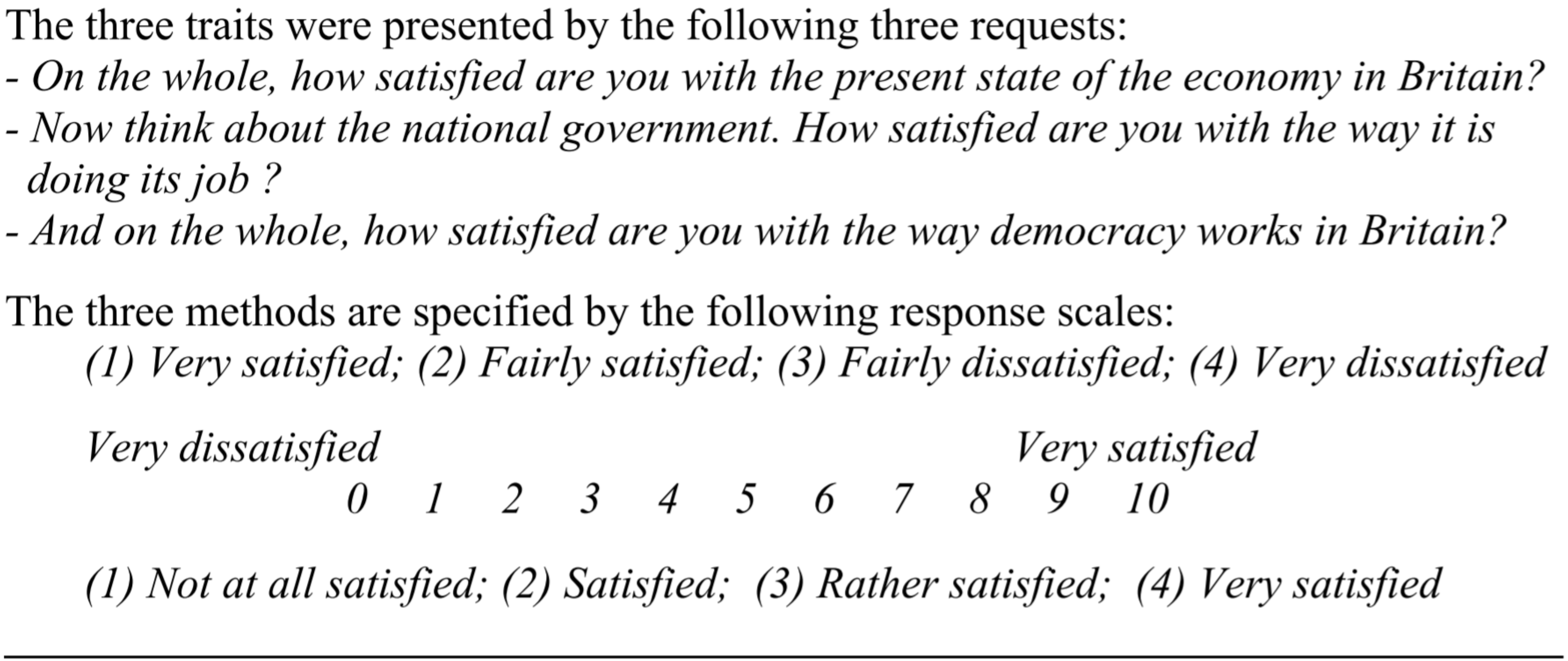
# Short intro to MultiTrait MultiMethod

A way to estimate:

- validity
- reliability
- random error

A **within experimental design**

# Example wording



The three traits were presented by the following three requests:

- *On the whole, how satisfied are you with the present state of the economy in Britain?*
- *Now think about the national government. How satisfied are you with the way it is doing its job ?*
- *And on the whole, how satisfied are you with the way democracy works in Britain?*

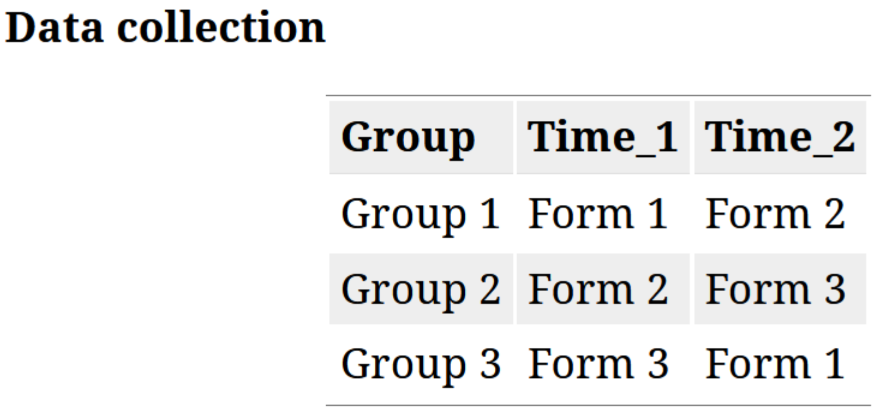
The three methods are specified by the following response scales:

*(1) Very satisfied; (2) Fairly satisfied; (3) Fairly dissatisfied; (4) Very dissatisfied*

*Very dissatisfied                    Very satisfied*
*0   1   2   3   4   5   6   7   8   9   10*

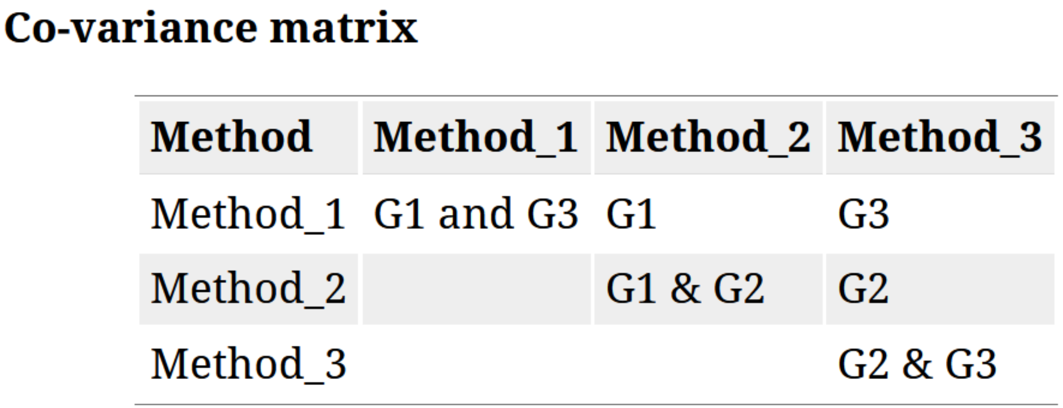
*(1) Not at all satisfied; (2) Satisfied;  (3) Rather satisfied;  (4) Very satisfied*

# Example split-ballot design



**Data collection**

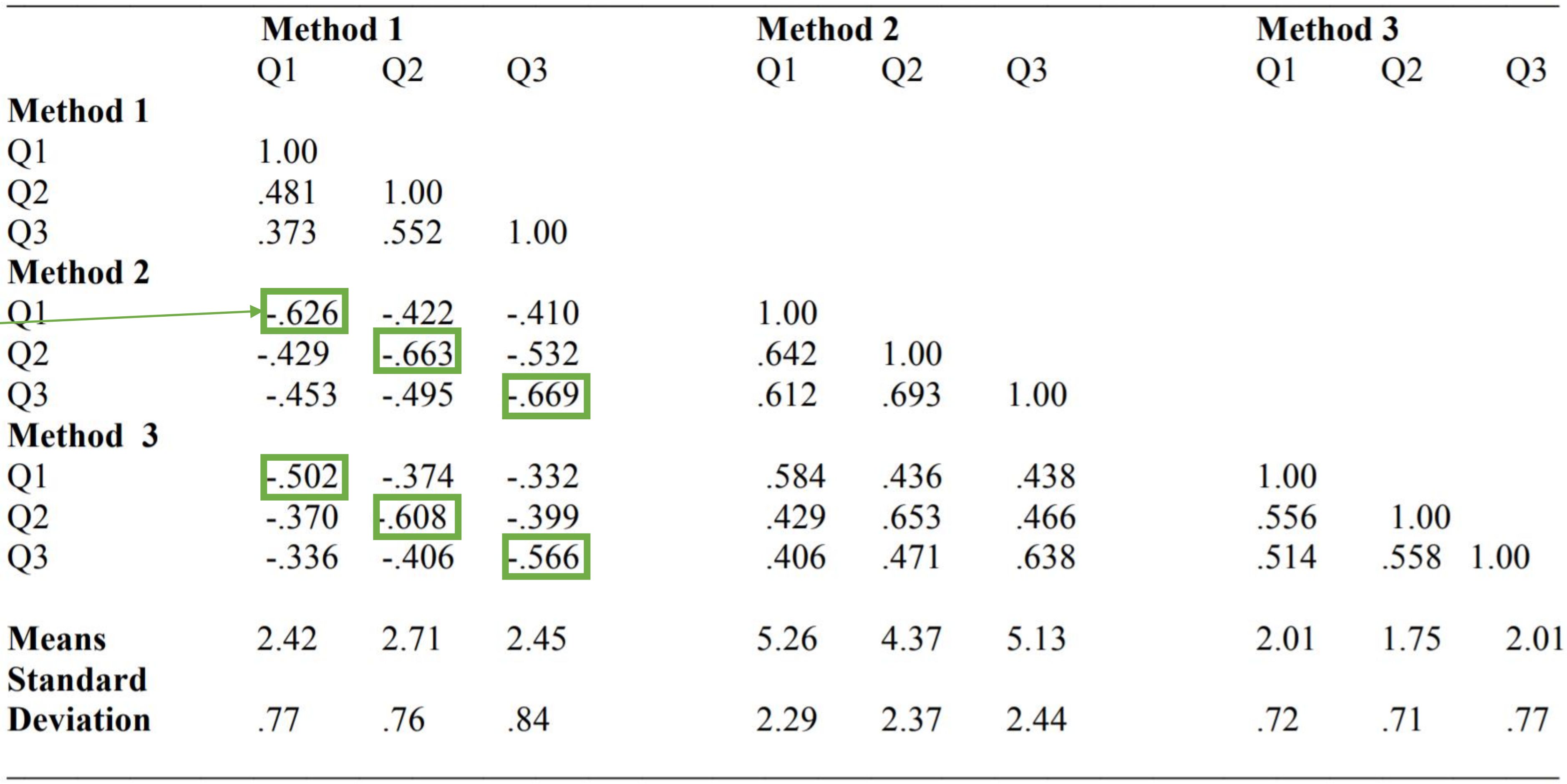
| Group | Time_1 | Time_2 |
|---|---|---|
| Group 1 | Form 1 | Form 2 |
| Group 2 | Form 2 | Form 3 |
| Group 3 | Form 3 | Form 1 |



**Co-variance matrix**

| Method | Method_1 | Method_2 | Method_3 |
|---|---|---|---|
| Method_1 | G1 and G3 | G1 | G3 |
| Method_2 |  | G1 & G2 | G2 |
| Method_3 |  |  | G2 & G3 |

# Example correlation matrix

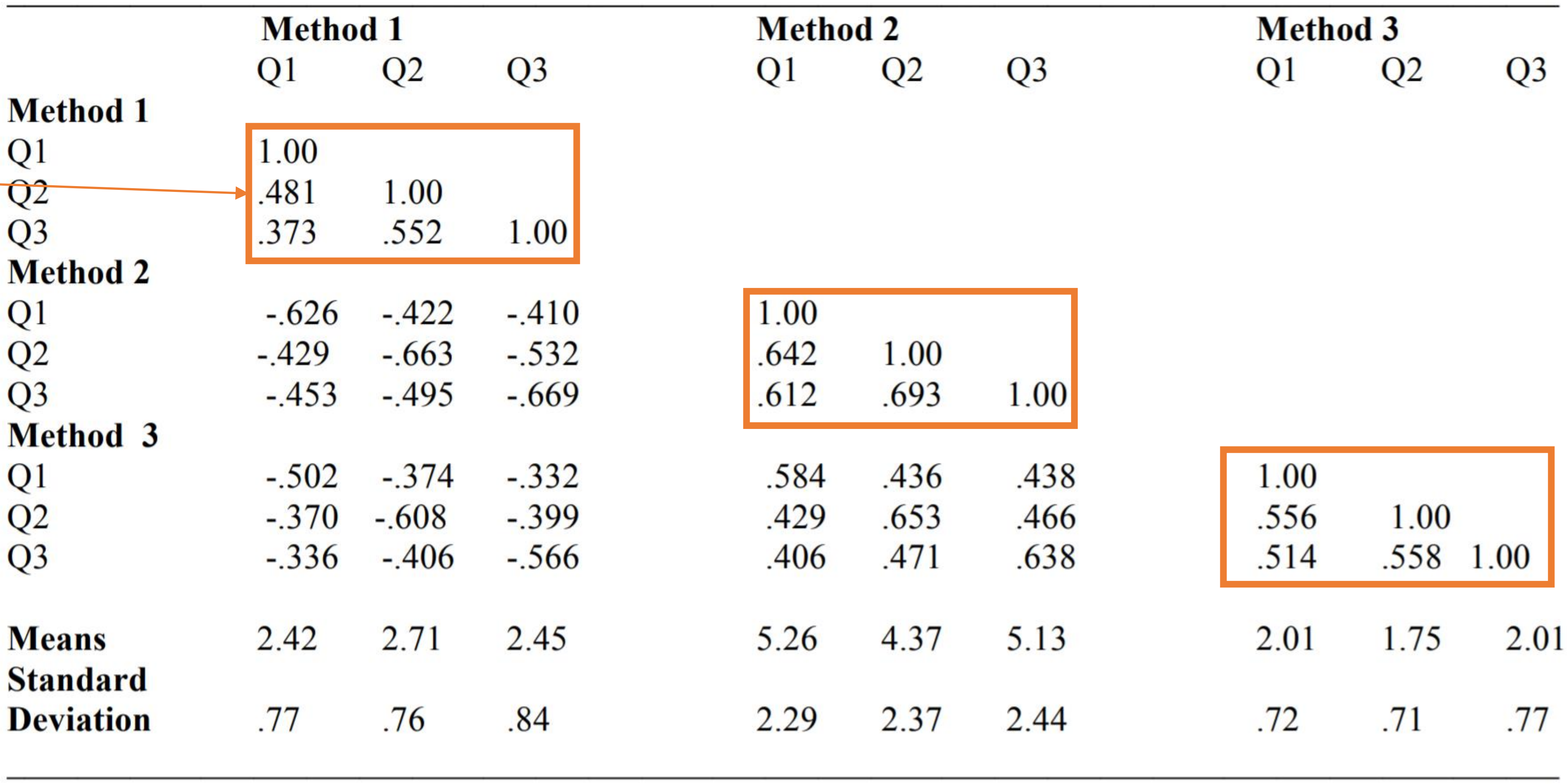


| | Method 1 | | | Method 2 | | | Method 3 | | |
|---|---|---|---|---|---|---|---|---|---|
| | Q1 | Q2 | Q3 | Q1 | Q2 | Q3 | Q1 | Q2 | Q3 |
| **Method 1** | | | | | | | | | |
| Q1 | 1.00 | | | | | | | | |
| Q2 | .481 | 1.00 | | | | | | | |
| Q3 | .373 | .552 | 1.00 | | | | | | |
| **Method 2** | | | | | | | | | |
| Q1 | -.626 | -.422 | -.410 | 1.00 | | | | | |
| Q2 | -.429 | -.663 | -.532 | .642 | 1.00 | | | | |
| Q3 | -.453 | -.495 | -.669 | .612 | .693 | 1.00 | | | |
| **Method 3** | | | | | | | | | |
| Q1 | -.502 | -.374 | -.332 | .584 | .436 | .438 | 1.00 | | |
| Q2 | -.370 | -.608 | -.399 | .429 | .653 | .466 | .556 | 1.00 | |
| Q3 | -.336 | -.406 | -.566 | .406 | .471 | .638 | .514 | .558 | 1.00 |
| **Means** | 2.42 | 2.71 | 2.45 | 5.26 | 4.37 | 5.13 | 2.01 | 1.75 | 2.01 |
| **Standard Deviation** | .77 | .76 | .84 | 2.29 | 2.37 | 2.44 | .72 | .71 | .77 |

Consistent variance

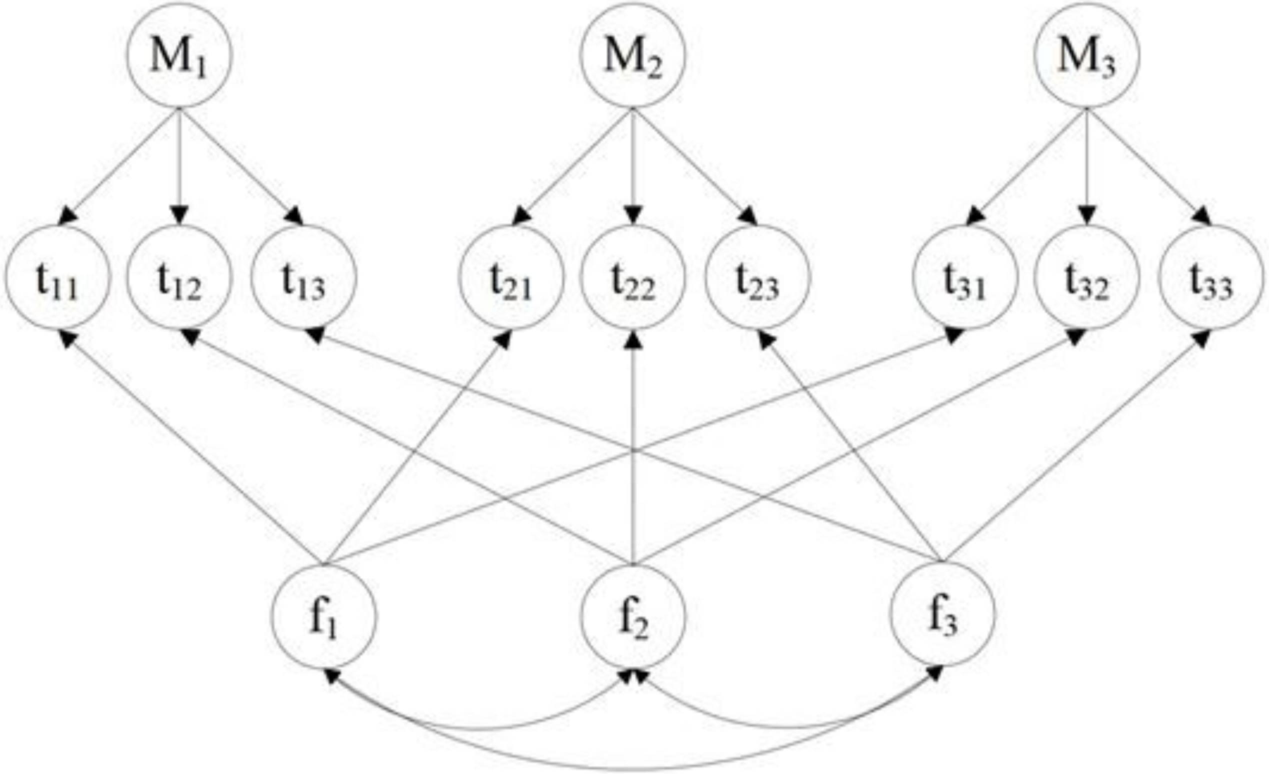
# Example correlation matrix

Method variance

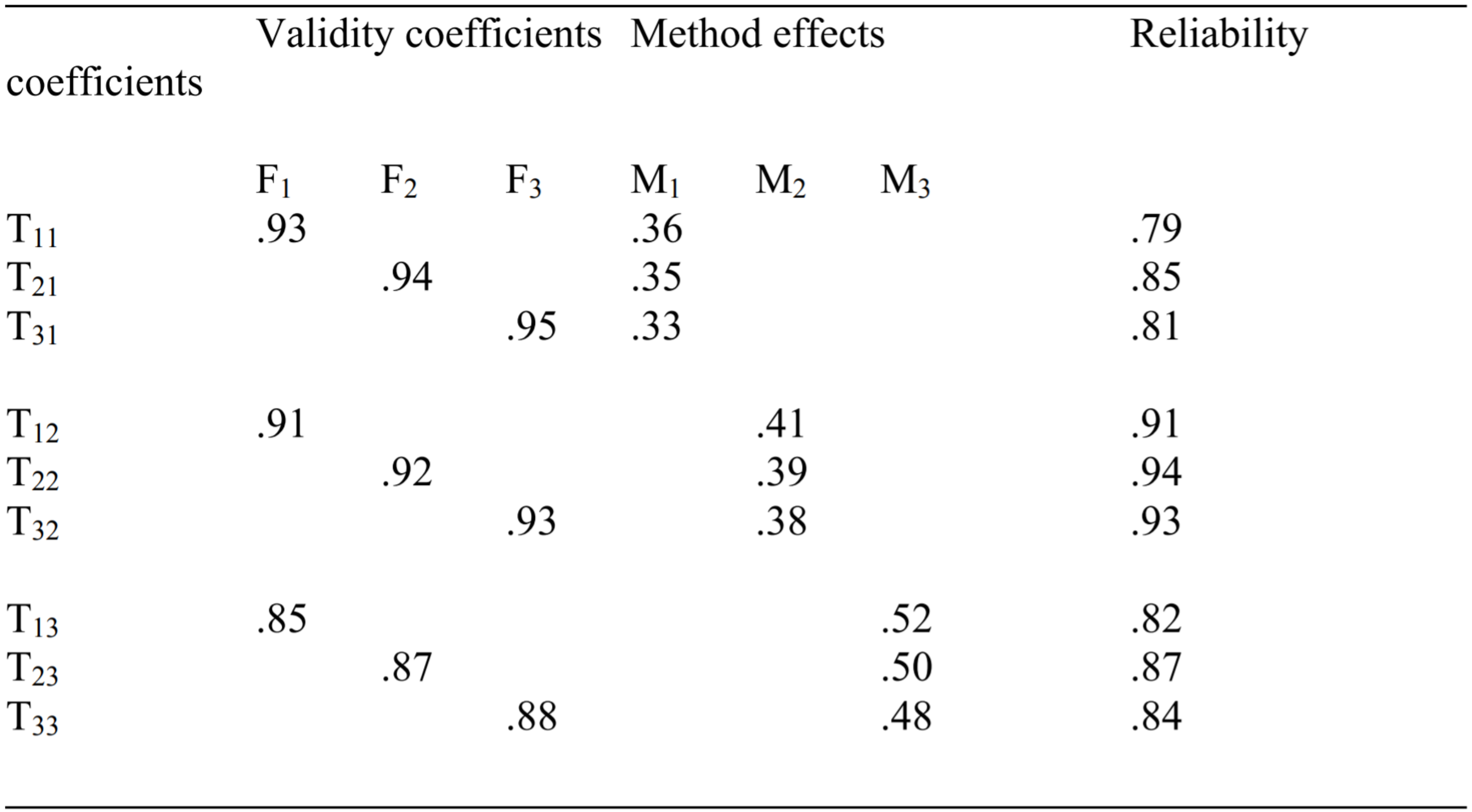


| | **Method 1** | | | **Method 2** | | | **Method 3** | | |
|---|---|---|---|---|---|---|---|---|---|
| | Q1 | Q2 | Q3 | Q1 | Q2 | Q3 | Q1 | Q2 | Q3 |
| **Method 1** | | | | | | | | | |
| Q1 | 1.00 | | | | | | | | |
| Q2 | .481 | 1.00 | | | | | | | |
| Q3 | .373 | .552 | 1.00 | | | | | | |
| **Method 2** | | | | | | | | | |
| Q1 | -.626 | -.422 | -.410 | 1.00 | | | | | |
| Q2 | -.429 | -.663 | -.532 | .642 | 1.00 | | | | |
| Q3 | -.453 | -.495 | -.669 | .612 | .693 | 1.00 | | | |
| **Method 3** | | | | | | | | | |
| Q1 | -.502 | -.374 | -.332 | .584 | .436 | .438 | 1.00 | | |
| Q2 | -.370 | -.608 | -.399 | .429 | .653 | .466 | .556 | 1.00 | |
| Q3 | -.336 | -.406 | -.566 | .406 | .471 | .638 | .514 | .558 | 1.00 |
| **Means** | 2.42 | 2.71 | 2.45 | 5.26 | 4.37 | 5.13 | 2.01 | 1.75 | 2.01 |
| **Standard Deviation** | .77 | .76 | .84 | 2.29 | 2.37 | 2.44 | .72 | .71 | .77 |

# Statistical model



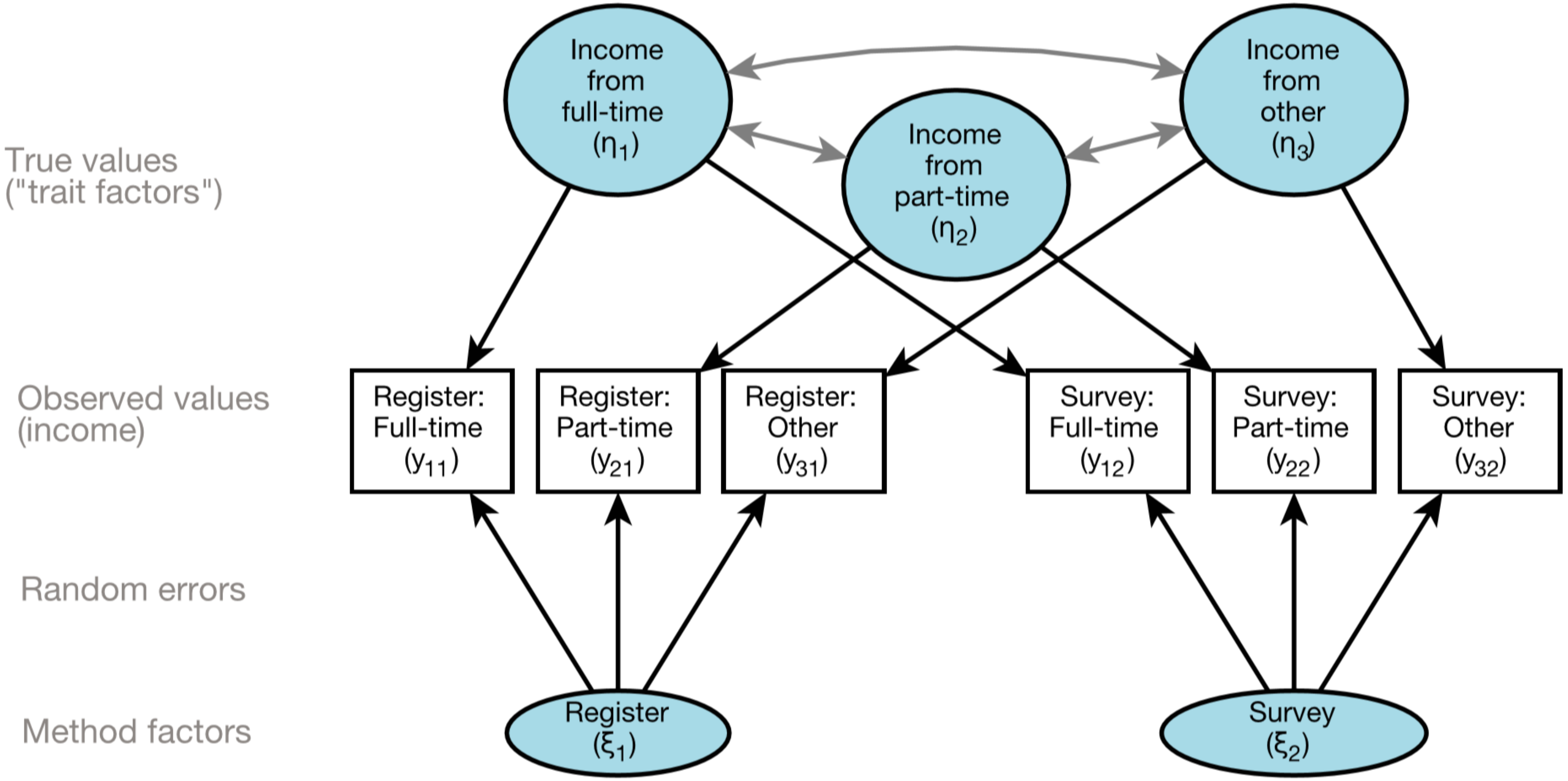
# Example of results



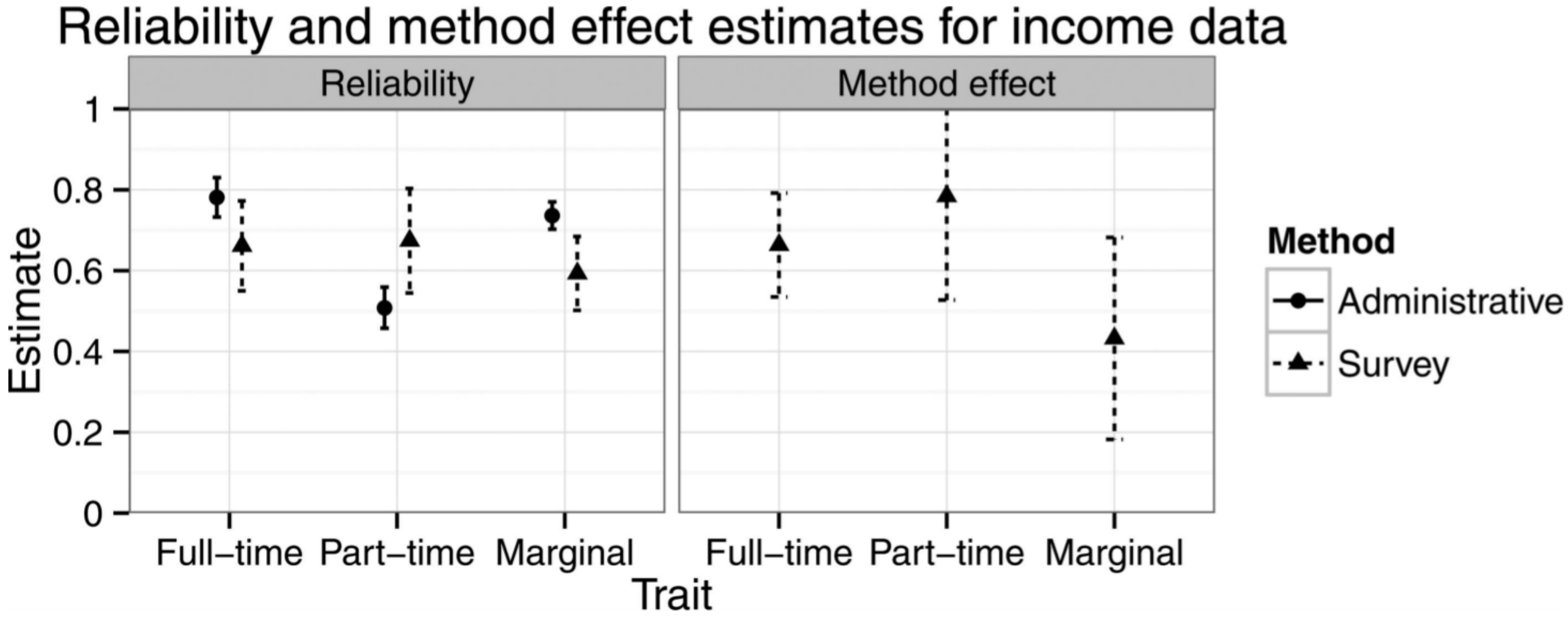
| | Validity coefficients | | | Method effects | | | Reliability |
|---|---|---|---|---|---|---|---|
| coefficients | | | | | | | |
| | $F_1$ | $F_2$ | $F_3$ | $M_1$ | $M_2$ | $M_3$ | |
| $T_{11}$ | .93 | | | .36 | | | .79 |
| $T_{21}$ | | .94 | | .35 | | | .85 |
| $T_{31}$ | | | .95 | .33 | | | .81 |
| $T_{12}$ | .91 | | | | .41 | | .91 |
| $T_{22}$ | | .92 | | | .39 | | .94 |
| $T_{32}$ | | | .93 | | .38 | | .93 |
| $T_{13}$ | .85 | | | | | .52 | .82 |
| $T_{23}$ | | .87 | | | | .50 | .87 |
| $T_{33}$ | | | .88 | | | .48 | .84 |

# Extending the MTMM to different data sources - model

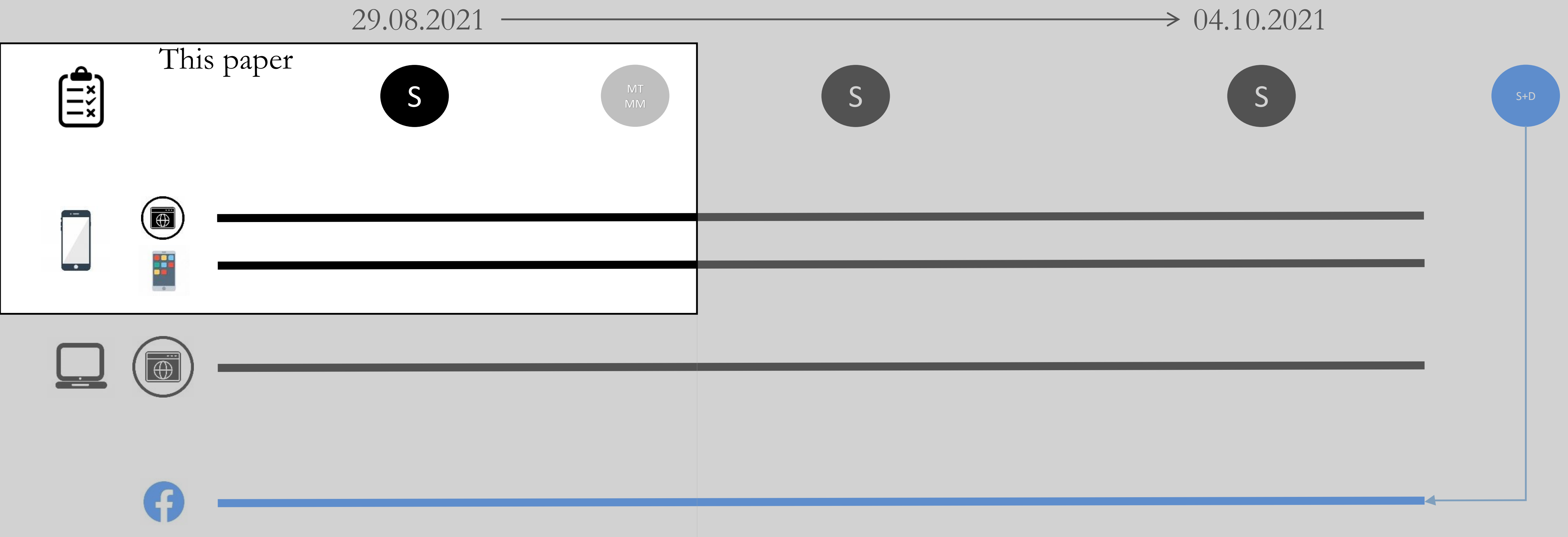
*(Oberski et al. 2017)*


Income from full-time ($\eta_1$)
Income from part-time ($\eta_2$)
Income from other ($\eta_3$)
True values ("trait factors")
Observed values (income)
Register: Full-time ($y_{11}$)
Register: Part-time ($y_{21}$)
Register: Other ($y_{31}$)
Survey: Full-time ($y_{12}$)
Survey: Part-time ($y_{22}$)
Survey: Other ($y_{32}$)
Random errors
Method factors
Register ($\xi_1$)
Survey ($\xi_2$)

# Extending the MTMM to different data sources - results *(Oberski et al. 2017)*



# Our MTMM design

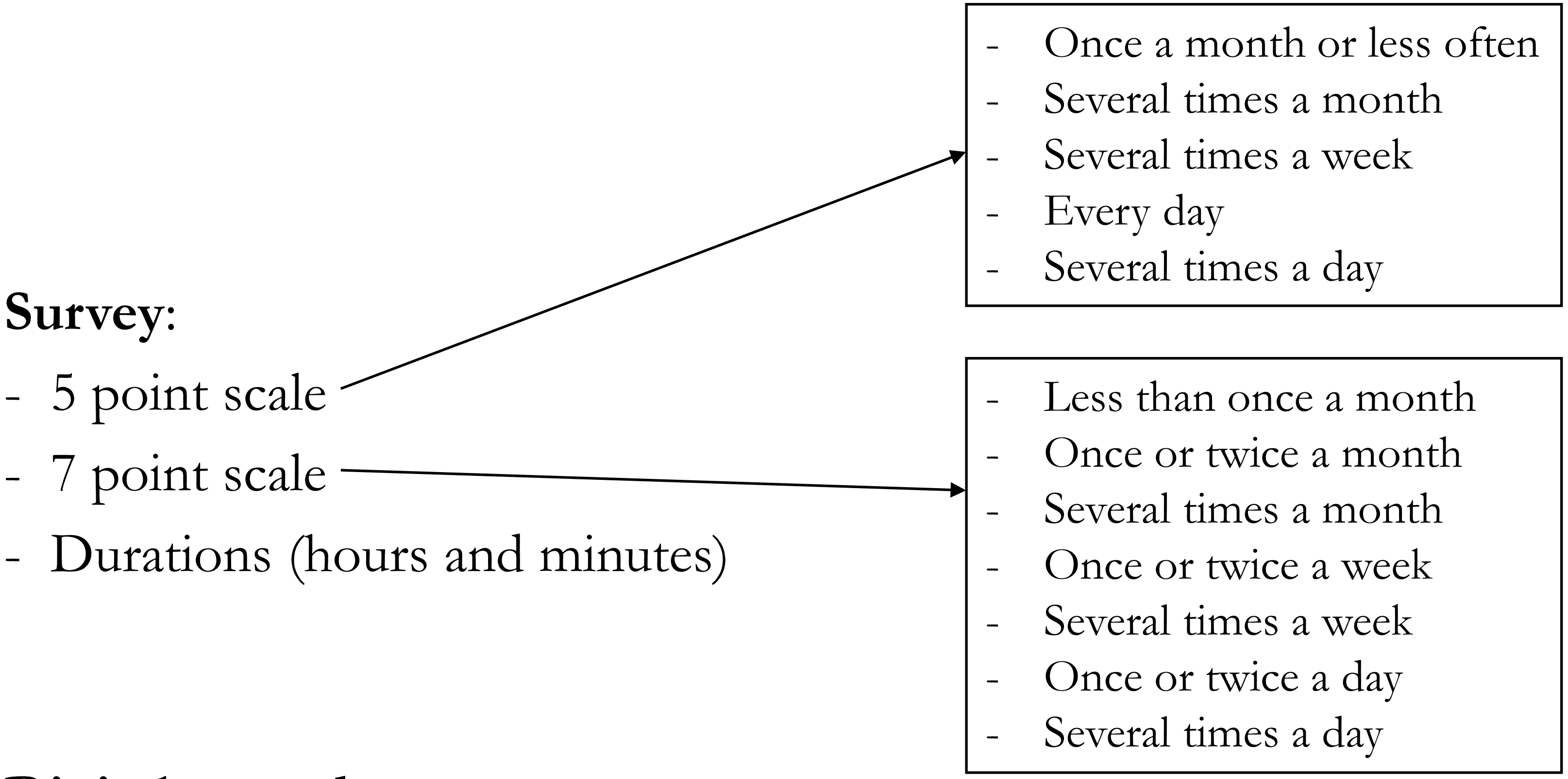


# How best to measure online behaviours?

Using the phone to:

- call
- write text message
- take photos
- social media
- web browsing

# Measurements



**Survey**:

- 5 point scale
- 7 point scale
- Durations (hours and minutes)

- Once a month or less often
- Several times a month
- Several times a week
- Every day
- Several times a day

- Less than once a month
- Once or twice a month
- Several times a month
- Once or twice a week
- Several times a week
- Once or twice a day
- Several times a day

**Digital trace data**:

- How many times they do the activity
- How much time they spend doing the activity

# Creating the indicators in digital trace data

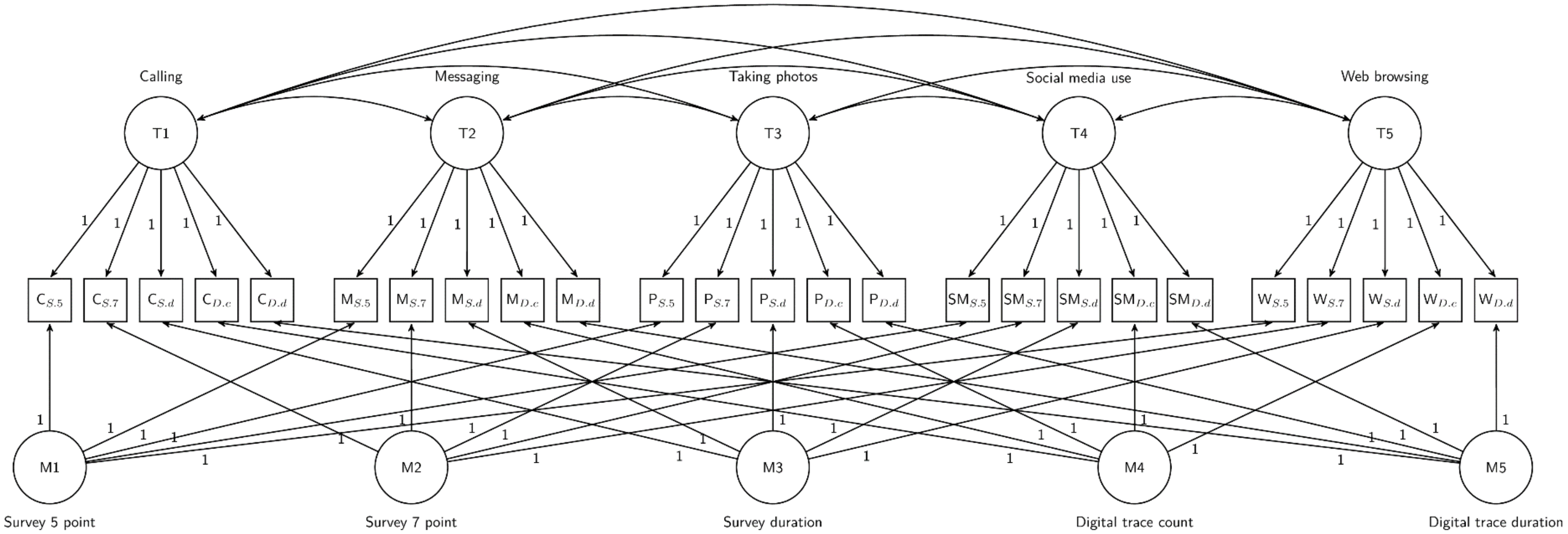
1. Use advanced searching to identify all relevant activities

2. Hand code long tail of ambiguous apps

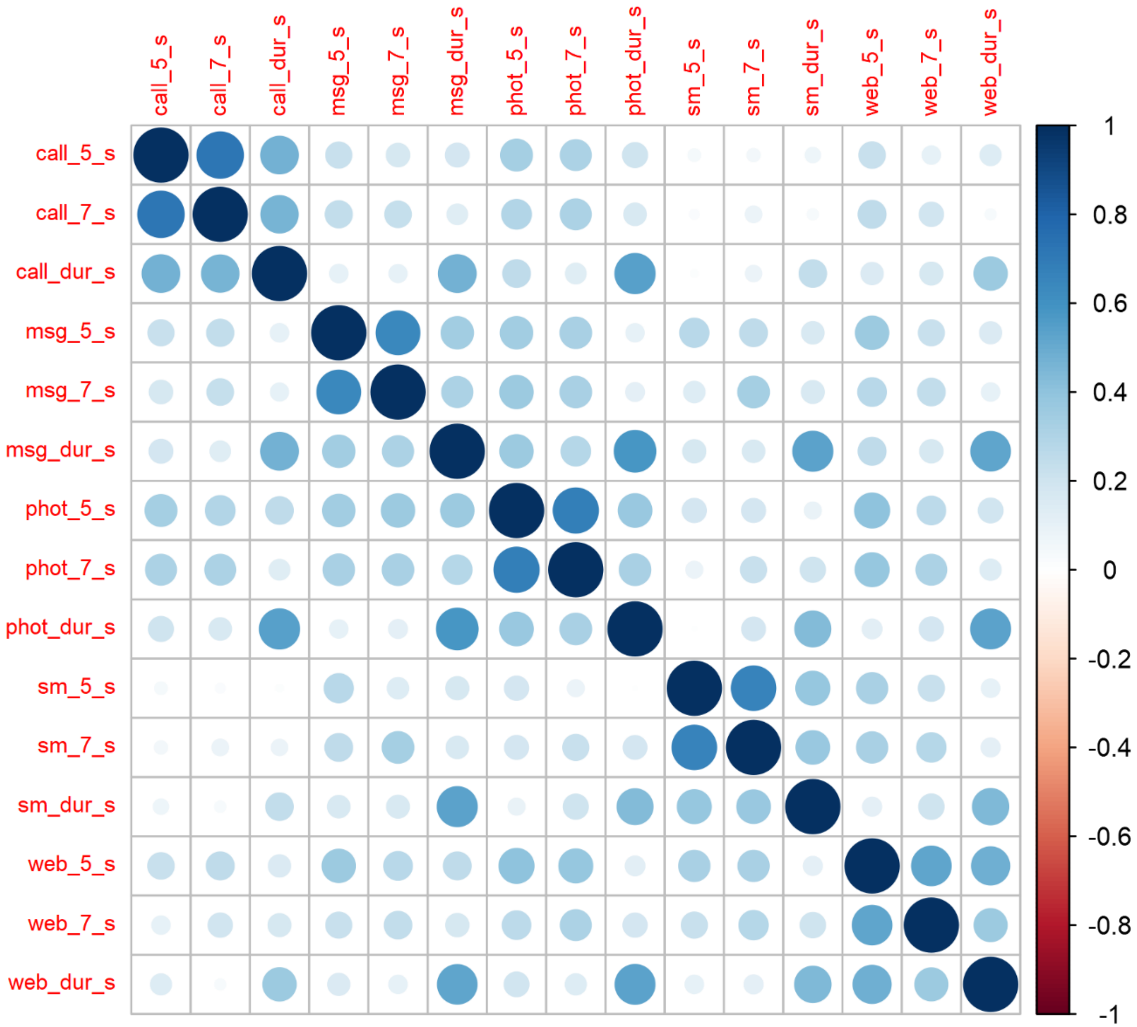
3. Calculate duration/count and aggregate

4. Take the log to deal with skewed distribution

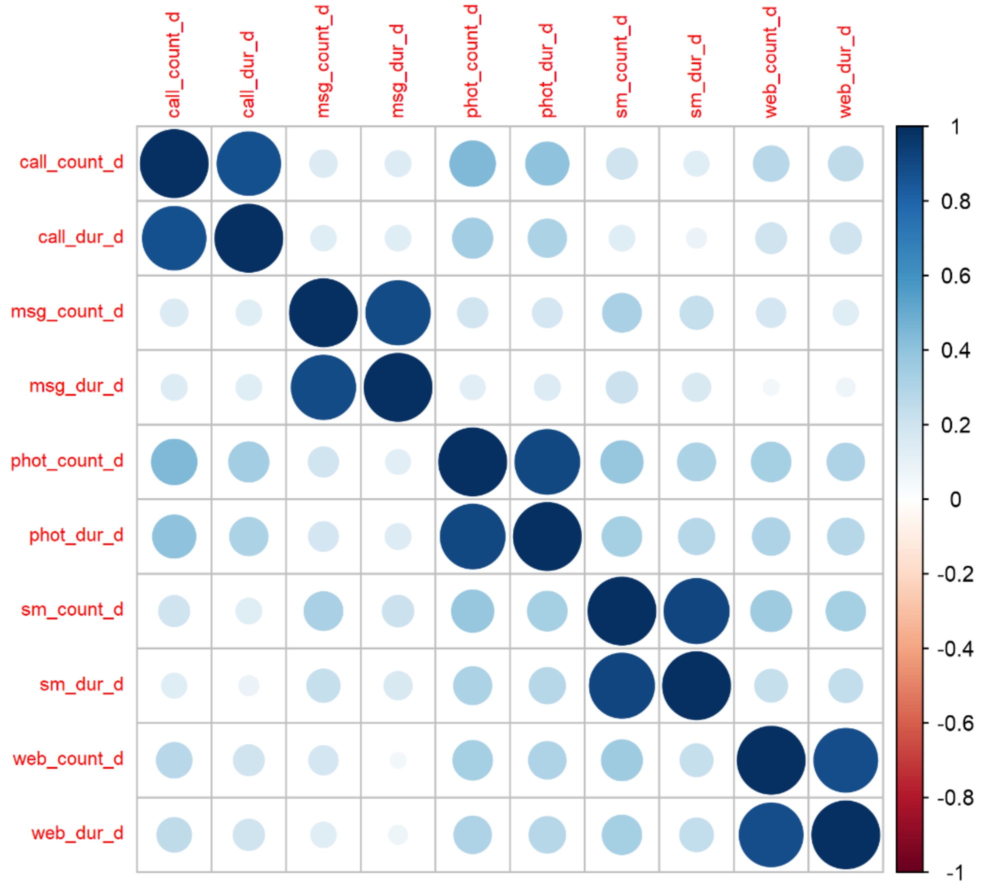
# MTMM model used



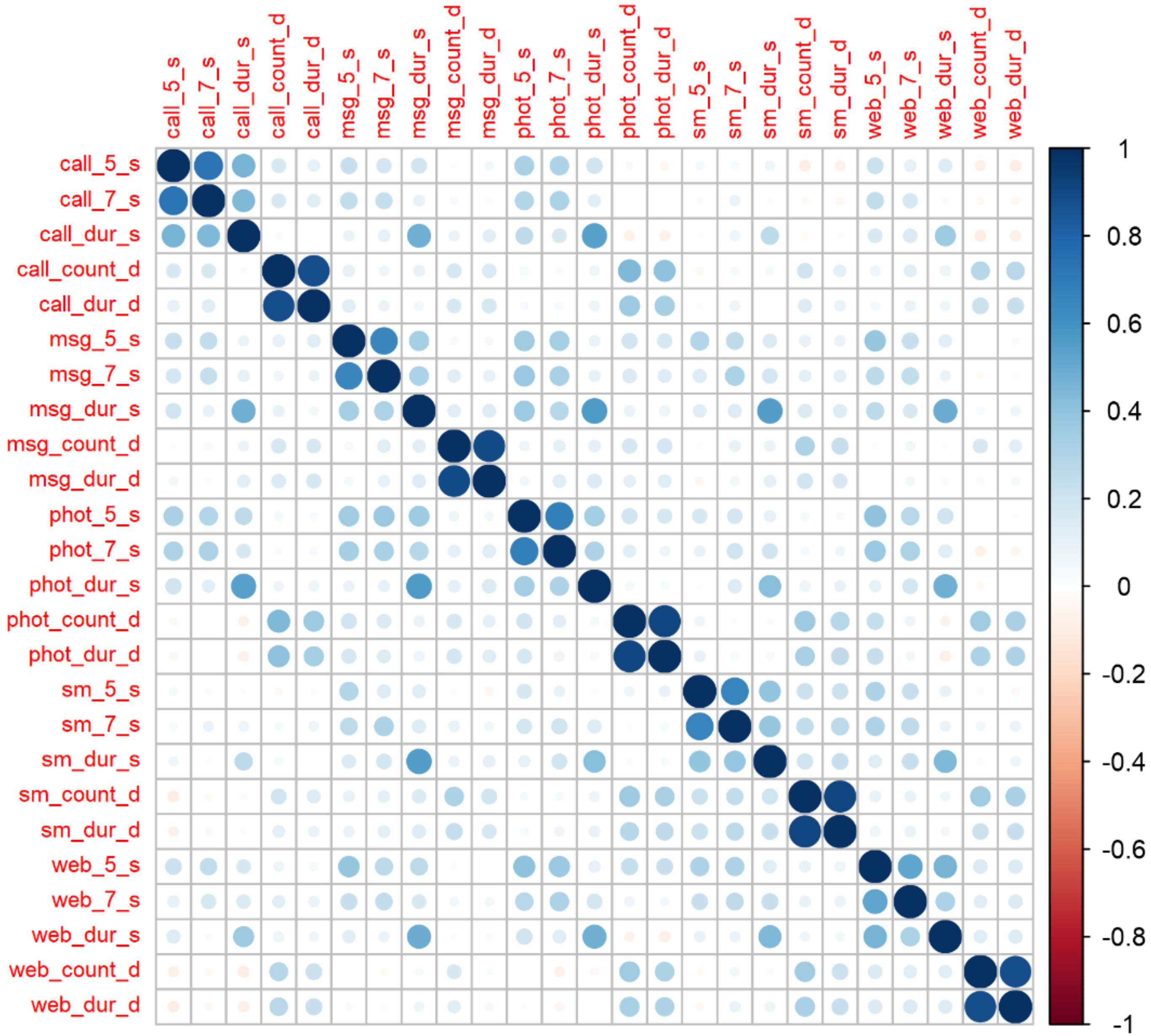
# Correlation matrix survey data



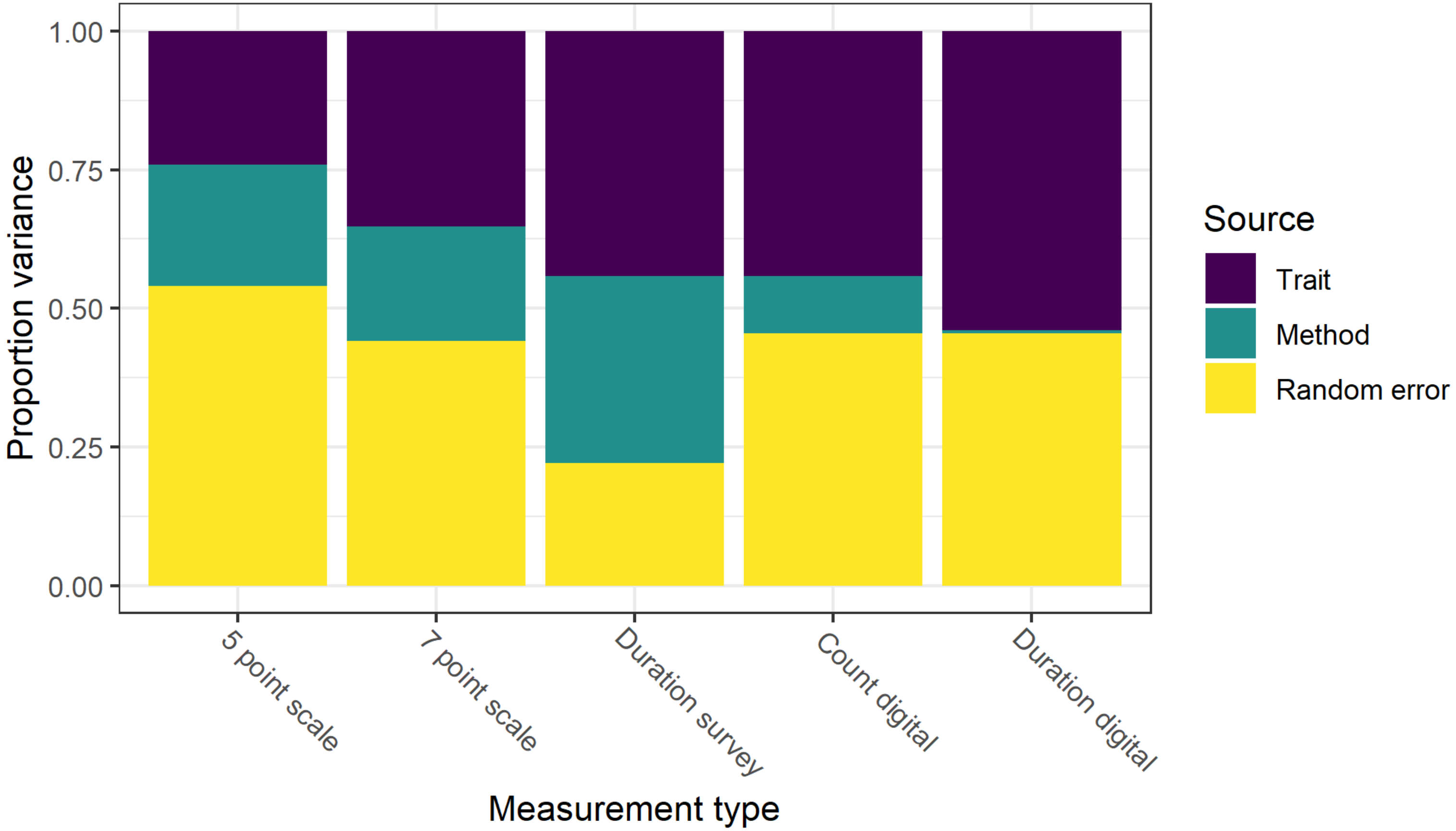
# Correlation matrix digital trace data



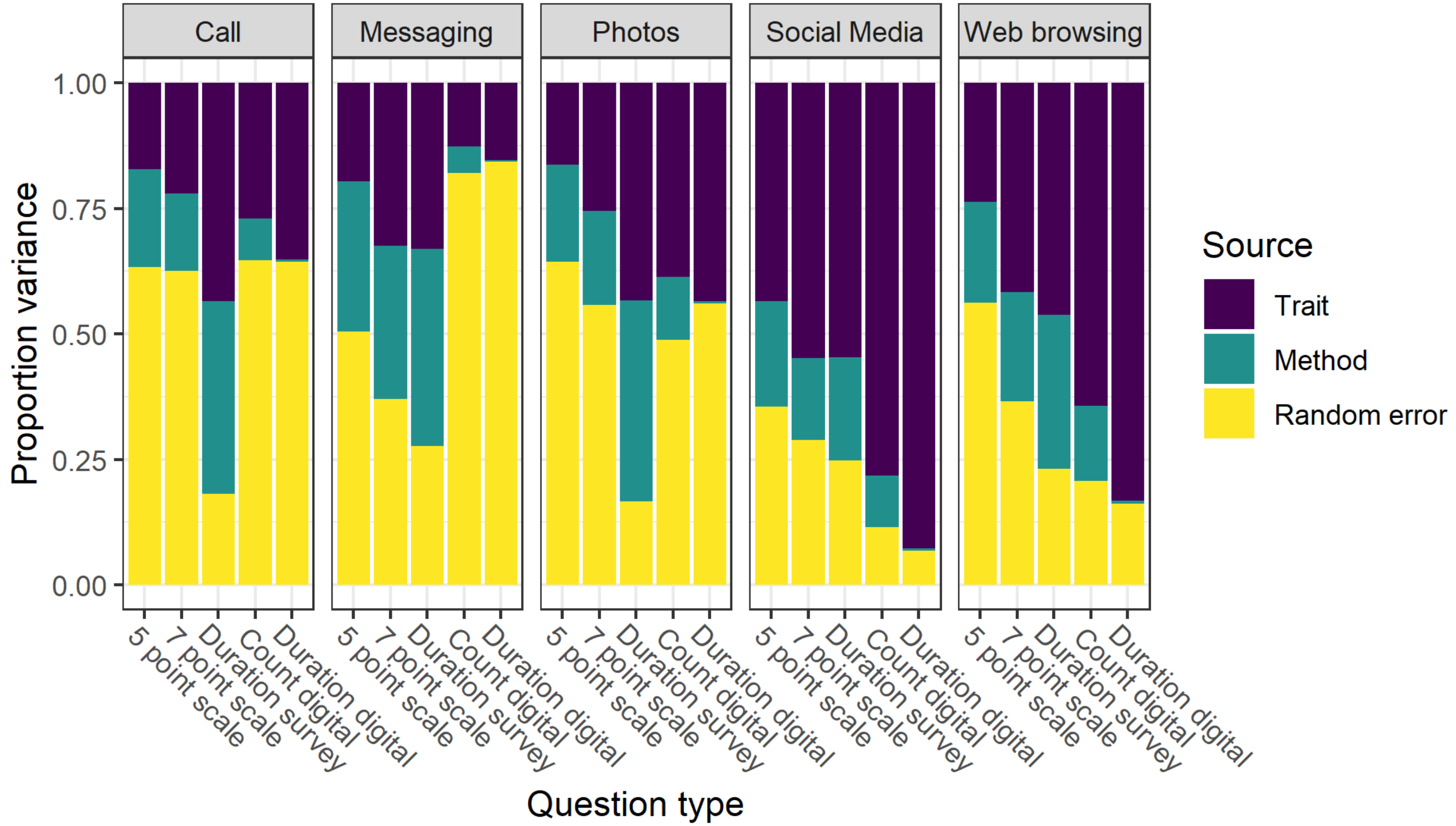
# Full correlation matrix



# MTMM variance decomposition - method



# Variance decomposition – method x trait



# Conclusions

- Digital trace seems more precise but far from perfect

- Further investigation in how text messaging is defined

- Investigate impact on substantive results

# Points for discussion

- Do different data sources measure the same concepts?
- How to decide which data source to use for what measures?
- Would combining measures from multiple sources improve measurement quality?

# An exploration of digital trace data quality

**Alexandru Cernat**

*University of Manchester*

*Social Statistics Department*

www.alexcernat.com

@cernat_a