

ExpiWell

ESM WORKSHOP SERIES · PART THREE

Manage the *Data*

Cleaning, missing data, and the problematic responses that survive both.

By Louis Tay and Marilyn Uy

expiwell.com

Manage the Data

1

Clean the data

Duplicate replies, backfilled responses, and the joins between surveys and sensors.

2

Missing data

Which kind of missingness you have decides what you are allowed to do about it.

3

Problematic data

Responses that are present, and still should not be trusted.

Four things to look for when cleaning

● Double replies

Multiple responses to the same question for the same time point

Often a sign the platform is not built for ESM, or lost a network round trip.

● Backfill

Responses filled in after the desired time window has closed

The better question is whether your software prevents backfill in the first place.

● Linking over time

Connecting a respondent's surveys across the study

ExpiWell links a participant's surveys automatically — no researcher-assigned ID question needed.

● Sensor joins

Connecting wearable and location data to survey responses

ExpiWell captures Fitbit, Apple Watch, SensorKit and GPS alongside responses.

Missing data

1

Types of missing data

MCAR, MAR and MNAR are not degrees of the same problem — they are different problems.

2

Addressing missing data

Most of the work happens before collection, not after it.

3

Recommended resources

Where to read further on estimation from incomplete data.

Three types of missing data

1

Missing Completely at Random

MCAR. Nothing systematic about which responses are missing, and estimates are not biased.

2

Missing at Random

MAR. Missingness is related to an observed variable rather than the missing one — e.g., if women are less likely to report weight.

3

Missing Not at Random

MNAR. Missingness is related to the value of the missing variable itself — e.g., if heavier respondents are less likely to report weight.

Most of this is preventable

1

Address it before it happens

Pilot the burden — too many surveys in a day produces missing data by design. Incentivize completeness, e.g., a bonus above 70%.

2

Calculate missing-data statistics per respondent

Percentage of missing responses, percentage of missing surveys, and whether particular individuals carry an inordinate share.

3

Report those statistics

Missingness is a result. Reporting it is what lets a reader judge everything else.

Best practices when missingness is not random

1

Examine demographic differences

Correlate demographics with level of missingness, or compare high-missing against low- or no-missing groups.

2

Do post-hoc follow-up on key variables

In person or by phone. Assuming you obtain full data, run the same comparisons against level of missingness.

3

Use estimators that tolerate unbalanced data

Multilevel mixed models handle different numbers of observations per participant without listwise deletion — assuming MAR.

Estimating from incomplete data

1

Multilevel mixed modeling

Handles missing and unbalanced data — different numbers of observations per participant — without dropping cases.

2

Maximum likelihood via the EM algorithm

Dempster et al. (1977), already implemented in most mixed-model software, e.g., Mplus.

3

Maximum likelihood and multiple imputation

Recommended for longitudinal data, and available in SAS, R and SPSS.

Recommended resources

1

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm.

Journal of the Royal Statistical Society, Series B, 1–38.

2

Newman, D. A. (2003). Longitudinal modeling with randomly and systematically missing data.

Organizational Research Methods, 6(3), 328–362.

3

Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art.

Psychological Methods, 7(2), 147.

Problematic data

1

Careless responding

Data can be present and still not usable — see Meade & Craig (2012), Psychological Methods.

2

Attention checks

Include them in ESM surveys (Keith, Tay, & Harms, 2017), e.g., “Select option ‘Neither Agree nor Disagree’.”

3

Separate carelessness from missingness

Attention checks are what tell you a low response rate is not simply a lack of care on the responses that did arrive.

Learning multilevel modeling



R

Finch, Bolin, & Kelley (2014)

Multilevel modeling using R.



Mplus and Stata

Byrne (2013) · Rabe-Hesketh & Skrondal (2008)

Structural equation modeling with Mplus; multilevel and longitudinal modeling using Stata.



HLM

Raudenbush & Bryk (2002)

Hierarchical linear models: applications and data analysis methods.



SPSS and SAS

Heck, Thomas, & Tabata (2013) · Littell et al. (2006)

Multilevel modeling with IBM SPSS; SAS system for mixed models.

What the platform does before cleaning starts

Three of the four cleaning problems above are decided by the platform, not the analyst.



Response timing

Every response carries when it was submitted

- Surveys are scheduled in the participant's own time zone
- Submission and response rates in the Participation Log



Automatic linking

A participant's surveys are joined across the study

- No researcher-assigned ID question needed
- Linking holds across multiple scheduled surveys



Export

Responses and sensor streams come out together

- Export participant responses from Insights
- GPS and wearable data download alongside them

ExpiWell

NEXT IN THE SERIES

Part Four: Analyze the Data

Activities, trajectories, dynamics, and multilevel structure.

By Louis Tay and Marilyn Uy

expiwell.com