



Applying multiverse analysis to experience sampling data: Investigating whether preprocessing choices affect robustness of conclusions

Jeroen Weermeijer¹ · Ginette Lafit^{1,2} · Glenn Kiekens^{1,3} · Martien Wampers¹ · Gudrun Eisele¹ · Zuzana Kasanova¹ · Thomas Vaessen¹ · Peter Kuppens² · Inez Myin-Germeys¹

Accepted: 16 December 2021
© The Psychonomic Society, Inc. 2022

Abstract

The experience sampling method (ESM) has revolutionized our ability to conduct psychological research in the natural environment. However, researchers have a large degree of freedom when preprocessing ESM data, which may hinder scientific progress. This study illustrates the use of multiverse analyses regarding preprocessing choices related to data exclusion (i.e., based on various levels of compliance and exclusion of the first assessment day) and the calculation of constructs (i.e., composite scores calculated as the mean, median, or mode) by reanalyzing established group differences in negative affect, stress reactivity, and emotional inertia between individuals with and without psychosis. Data came from five studies and included 233 individuals with psychosis and 223 healthy individuals (in total, 26,892 longitudinal assessments). Preprocessing choices related to data exclusion did not affect conclusions. For both stress reactivity and emotional inertia of negative affect, group differences were affected when negative affect was calculated as the mean compared to the median or mode. Further analyses revealed that this could be attributed to considerable differences in the within- and between-factor structure of negative affect. While these findings show that observed differences in affective processes between individuals with and without psychosis are robust to preprocessing choices related to data exclusion, we found disagreement in conclusions between different central tendency measures. Safeguarding the validity of future experience sampling research, scholars are advised to use multiverse analysis to evaluate the robustness of their conclusions across different preprocessing scenarios.

Keywords Preprocessing · Researcher degrees of freedom · Multiverse · Experience sampling methods · Ecological momentary assessment · Ambulatory assessment

Introduction

The experience sampling method (ESM, also called ecological momentary assessment or ambulatory assessment; Csikszentmihalyi et al., 1977) is an ecologically valid self-report diary technique used for capturing and quantifying data in daily life. It involves sampling individuals' mental states, symptoms, and context by prompting them to fill out a questionnaire multiple times per day for several consecutive days. In this way, data rich in moment-to-moment information is obtained, providing a unique and detailed insight into the individual's everyday life flow (Myin-germeys et al., 2018). While ESM is not new and has its roots in ecological psychology (Larson & Csikszentmihalyi, 2014), an increasing number of researchers are now turning to ESM to examine within-person psychological processes (Stange et al., 2019). ESM is, for instance, often

Ginette Lafit and Glenn Kiekens contributed equally to this work.

✉ Jeroen Weermeijer
Jeroen.Weermeijer@kuleuven.be

¹ Center for Contextual Psychiatry, Department of Neuroscience, KU Leuven, Herestraat 49 bus 1029, 3000 Leuven, Belgium

² Quantitative Psychology and Individual Differences, Department of Psychology and Education Sciences, KU Leuven, Leuven, Belgium

³ Research Group of Clinical Psychology, Faculty of Psychology and Educational Sciences, KU Leuven, Leuven, Belgium

used to study how people feel in daily life, including both negative (NA) and positive (PA) affective states. In addition, ESM can provide crucial insights into the dynamics of these affective processes, notably regarding how people respond emotionally to real-life events (most often studied in terms of stress reactivity; Lardinois et al., 2011) or, conversely, how their emotions have become decoupled from events (which is captured in the notion of emotional inertia; Kuppens et al., 2010, 2012).

These core characteristics of people's emotional functioning in daily life have proven crucial for understanding important differences in the experience of affect in people who do and do not suffer from mental disorders. For instance, NA and stress reactivity are consistently found to be elevated in individuals with psychosis (Lardinois et al., 2011; Myin-Germeys & van Os, 2007; Reininghaus et al., 2016b). Similarly, emotional inertia is increased in patients suffering from depression compared to healthy individuals (Kuppens et al., 2010), and may even prospectively predict the onset of depression (Kuppens et al., 2012). The potential clinical merit of ESM and these dynamic parameters as markers of psychopathology, as well as technological advances that make it possible to gather real-world data about the psychological state of individuals, has led to an increase in the use of ESM in recent years (van Berkel et al., 2017). The growth of ESM research has, however, brought with it an increasing heterogeneity in study designs and preprocessing strategies. While the study design (i.e., included items and sampling frequency) is decided upon based on a specific rationale, preprocessing strategies are typically made without a specific *a priori* justification. Studies frequently differ regarding the exclusion of participants based on a predefined compliance rate (Trull & Ebner-Priemer, 2020; i.e., the minimum number of completed assessments relative to assessment occasions), the exclusion of the first day of data collection (Stone et al., 2002), the grouping of ordinal variables into specific measures, and the centering of variables. However, out of all these preprocessing choices, only the effect of centering has been well studied (e.g., Hamaker & Grasman, 2014). For example, it is well known that person-mean centering is required when studying within-person associations (Enders & Tofghi, 2007).

Multiverse analysis (MA) is a statistical technique used to investigate the effect of different preprocessing choices on the statistical results and conclusions of an empirical study (Steege et al., 2016). Specifically, it involves creating a collection of datasets, with each dataset stemming from a different line of data preprocessing choices. The hypothesis is evaluated on each unique dataset, yielding a distribution of statistical results such as *p*-values. Using this distribution, one can then explore the impact of preprocessing choices on the robustness of conclusions by considering whether the same conclusion is reached under different scenarios (akin to

sensitivity analyses that investigate the impact of alternative modeling choices; Steege et al., 2016)

While MA has been used in the past to examine the robustness of findings in cross-sectional survey data (Steege et al., 2016; Stern et al., 2019), researchers only recently started applying MA to investigate the validity of conclusions in ESM research. For example, Dejonckheere et al. (2019) recently demonstrated that, irrespective of the selected mood items, an inverse NA–PA association becomes more potent when anticipating personally relevant events (e.g., the release of exam results for students). While the study of Dejonckheere et al. (2019) highlights how MA can be used to strengthen the validity of ESM research, many frequently made preprocessing choices have not yet been empirically investigated, cautioning against the specification of a particular combination of preprocessing choices (also known as knife-edge specification; Weston et al., 2019; Young & Holsteen, 2017).

One frequent preprocessing choice for which there exists substantial heterogeneity is the use of a compliance cutoff for analysis. Researchers often exclude participants from analyses based on low compliance. However, what constitutes low compliance varies widely across studies [e.g., 8% (Gaudiano et al., 2018), 20% (Edwards et al., 2018), 33% (Lataster et al., 2010), and 50% (Li et al., 2019)] and often goes unreported (Trull & Ebner-Priemer, 2020). Similarly, some studies have excluded the data of the first day of ESM for participants, labeling it a familiarization day (e.g., Stone et al., 2002). This is based on the assumption of an initial elevation bias affecting results (e.g., Shrout et al., 2018), although emerging evidence does not support this postulation (Arslan et al., 2020). Importantly, however, it remains unclear whether this choice affects conclusions. Using MA to provide greater clarity about the potential impact of these two preprocessing choices on often studied dynamic processes in ESM research would provide meaningful information to the scientific community.

Apart from investigating preprocessing choices that often vary across studies (i.e., compliance cutoff and exclusion of day 1), it is equally important to question preprocessing choices that appear undisputedly uniform. One such example is how researchers group items to form psychological constructs. A momentary psychological construct, such as NA, is typically defined as the mean value of a set of Likert-type items. However, this approach has been criticized (Jamieson, 2004), as data collected using Likert-type items are ordinal, which can be problematic given that arithmetic operations—such as calculating a mean—cannot be conducted on ordinal data, as intervals between values cannot be considered equal (Wu & Leung, 2017). However, Likert-type items responses are often treated as if they come from an interval scale (Wu & Leung, 2017). To justify this, the ordinal data must approach the properties of interval data (Chyung

et al., 2017). For example, respondents must share similar interpretations of data intervals (e.g., the interval between 1 and 2 must be considered equal to the interval between 4 and 5). Simulation research has demonstrated that the mean is biased when this is not the case (Lindstädt et al., 2020). As opposed to computing the mean, the median or mode values can be used (Jamieson, 2004; Lindstädt et al., 2020). However, the use of this alternative and its impact on statistical conclusions, is yet to be explored in ESM literature.

The current study

Using a large pooled ESM dataset of individuals with and without psychosis, the current study employs and demonstrates how MA can be used to examine the influence of preprocessing choices on the robustness of conclusions. More specifically, we look at choices related to (1) person exclusion (i.e., based on various levels of compliance), (2) data exclusion (i.e., exclusion of the first assessment day), and (3) the calculation of constructs (i.e., composite scores calculated as the mean, median, or mode). Building upon existing research, we evaluated the following three hypotheses: (1) NA is elevated in individuals with psychosis compared to healthy individuals (hypothesis 1; see also Blanchard et al., 1998); (2) momentary stress is associated with momentary NA, and this association (i.e., stress reactivity) is more potent in individuals with psychosis than healthy individuals for various types of momentary stress (i.e., social, activity, or event stress; hypotheses 2.1 to 2.3, respectively; see also Lataster et al., 2013; van Winkel et al., 2015; Reininghaus et al. 2016); and (3) emotional inertia (as a risk marker of depression) is elevated in individuals with psychosis (of which more than 50% report comorbid depression; Buckley et al., 2009) compared to healthy individuals (hypothesis 3; see also Kuppens et al., 2010). For each hypothesis, a multiverse of datasets was created based on various combinations of each of the three preprocessing choices discussed above. The analysis plan underlying these analyses was pre-registered and can be consulted online (https://osf.io/vefkg/?view_only=0d70ea4e0b8d4241901516131cc38cad).

Method

Sample and assessment procedure

This study combines data from methodologically similar experience sampling studies (Bak et al., 2001; Lataster et al., 2013; Myin-Germeys et al., 2001; Thewissen et al., 2008; Vaessen et al., 2018). All of these studies include patients with psychotic disorders and healthy individuals. General inclusion criteria were as follows: (1) age between 18 and 70 years, and (2) having sufficient proficiency of the Dutch

language to comprehend the content of questionnaires. For the clinical group (i.e., patients with psychosis), participants were required to have Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV) diagnoses of schizophrenia, schizoaffective disorder, or schizophreniform disorder. General exclusion criteria were (1) brain disease or (2) history of head injury with loss of consciousness. Participants presenting with a family history of schizophrenia spectrum disorder were excluded from participation. In these studies, participants were recruited from across Belgium and the Netherlands. The pooled sample comprised 456 participants, including 233 individuals with and 223 without psychosis (in total 26,892 longitudinal assessments). The mean age of the patient sample was 34.85 years ($SD = 11.08$), and 33.47% were female. The mean age of the sample of healthy individuals was 37.48 years ($SD = 12.56$), and 60.99% were female.

In each study, participants were asked to complete an ESM questionnaire ten times per day, for six consecutive days. Questionnaires were filled out using paper-and-pencil booklets. A signal-contingent semi-random sampling scheme was used in which ten prompts to fill out the questionnaire were delivered via a digital wristwatch in 90-minute blocks, between 7:30 AM and 10:30 PM (minimum 15 minutes apart from prior assessments). Questionnaire items consisted of items scored on seven-point Likert scales and several open-ended questions. For the current study, only a subset of the Likert scale items was used. A detailed overview is available on this project's preregistration webpage on the Open Science Framework (OSF) (https://osf.io/vefkg/?view_only=0d70ea4e0b8d4241901516131cc38cad).

Included baseline and ESM variables

Baseline variables included were age expressed in years, biological sex at birth (i.e., male or female), and group indicating a clinically diagnosed psychotic disorder. Age was measured as the absolute number of years since birth, whereas gender was measured as binary and referred to biological sex at birth (male, coded as 0; female, coded as 1). The binary group variable (i.e., patient vs. control) was coded "1" if a participant had been diagnosed with a psychotic disorder and "0" for healthy individuals.

ESM variables selected for this study were based on previous work (e.g., Myin-Germeys et al., 2003; Reininghaus et al., 2016a) and included momentary NA and momentary stress. *Momentary NA* was measured using six items: "I feel uncertain," "I am lonely," "I am anxious," "I feel irritated," "I feel sad," and "I feel guilty." *Momentary Stress* was conceptualized as using indices of momentary social stress, activity stress, and event stress. Social stress was assessed with the item "I prefer being alone." If participants were not in the presence of others, they were instructed to skip this

Table 1 Variable characteristics

Variable	ICC	Mean value	SD
I feel unsure [NA item]	0.49	1.67	1.61
I feel lonely [NA item]	0.56	1.58	1.65
I feel anxious [NA item]	0.51	1.49	1.13
I feel irritated [NA item]	0.36	1.72	1.79
I feel sad [NA item]	0.51	1.63	1.50
I feel guilty [NA item]	0.52	1.34	0.95
NA as mean	0.65	1.57	0.80
NA as median	0.60	1.42	0.90
NA as mode	0.56	1.38	0.94
I prefer being alone [SS item]	0.37	1.84	2.49
I would prefer doing something else [AS item]	0.19	2.59	3.61
It [the ongoing activity] costs much effort [AS item]	0.31	2.10	2.56
It [the ongoing activity] is challenging [AS item]	0.34	2.68	3.90
I am skilled to do this [AS item]	0.38	5.23	2.50
Activity stress as mean	0.32	3.15	0.90
Activity stress as median	0.30	2.77	2.10
Activity stress as mode	0.27	2.54	2.92
(Un)pleasantness of the current event [ES item]	0.18	1.41	2.78

question. Activity stress was framed as stress experienced in currently ongoing activities and was assessed using four items, of which the last item was reverse-coded (“I would prefer doing something else,” “It [the ongoing activity] costs much effort,” “It [the ongoing activity] is challenging,” and “I am skilled to do this”). Except for momentary event stress, each variable was measured using questions rated on 7-point Likert scales that ranged from 1 “not at all” to 7 “very much.” Event stress was defined as the unpleasantness of the most important event since the last beep and was measured using one item rating (un)pleasantness of that event, measured on a bipolar scale ranging from -3 “very unpleasant” to $+3$ “very pleasant.” Positive and neutral events (0 to $+3$) were coded 0 , whereas negative values were reverse coded (i.e., flipped from -1 , -2 , -3 to 1 , 2 , 3 , respectively). Finally, stress reactivity was operationalized as the effect of each momentary stress variable on momentary NA. Information on intraclass correlation coefficients (ICCs), means, and variances of all measured and calculated ESM variables can be found in table 1 below.

Preprocessing

Datasets for each unique combination of preprocessing choices were created. For hypotheses 1, 2.1, 2.3, and 3, this can be broken down as follows: exclusion based on different rates of compliance ranging from 0 to 50% (5% increments, 11 options), data of day 1 included/excluded (two options), and different central tendency measures to calculate NA (mean, median, mode, three options). Combining these

different choices results in 66 ($11 \times 2 \times 3$) unique datasets. For hypothesis 2.2, we have one additional choice dimension, as activity stress, unlike social (hypothesis 2.1) and event stress (hypothesis 2.3), was measured with multiple items. Therefore, the activity stress score can be computed as either the mean, mode, or median of these items (three extra options), resulting in 198 ($11 \times 2 \times 3 \times 3$) unique datasets.

Statistical analyses

Given that the obtained intensive longitudinal data have a hierarchical structure (i.e., repeated momentary measurements nested [level 1] within subjects [level 2]), linear mixed-effects models were used to test our hypotheses. In each model, age and gender were added as covariates. To avoid an overly small coefficient for the age covariate, we coded the age variable in line with recent work (Age-20)/10 (Rintala et al., 2019). ESM predictor variables were person mean-centered using the participant’s mean to allow interpretation of predictor variables at the within-person level in a relative fashion for each person (Enders & Tofighi, 2007; Hamaker & Grasman, 2014). Restricted maximum likelihood estimation was used to estimate variance components of our models (Lafit et al., 2021). A significant statistical result was obtained in scenarios where the p -value was below $.05$. For the first hypothesis, a linear mixed-effects model with a random intercept was estimated, with NA as the dependent variable and the diagnostic group (indicating psychosis) as a predictor variable. For hypotheses 2.1 to 2.3, we added momentary stress and the interaction with group

Table 2 Formulas for the models used in this study

Model	Level 1	Level 2	
		Random intercept	Random slope
Model 1	$NegativeAffect_{it} = \gamma_{0i} + \epsilon_{it}$	$\gamma_{0i} = \beta_{00} + \beta_{01}Group_i + \nu_{0i}$	—
Model 2.1-2.3	$NegativeAffect_{it} = \gamma_{0i} + \gamma_{1i}Stress^*_{it} + \epsilon_{it}$	$\gamma_{0i} = \beta_{00} + \beta_{01}Group_i + \nu_{0i}$	$\gamma_{1i} = \beta_{10} + \beta_{11}Group_i + \nu_{1i}$
Model 3	$NegativeAffect_{it} = \gamma_{0i} + \gamma_{1i}NA_{i-1} + \epsilon_{it}$	$\gamma_{0i} = \beta_{00} + \beta_{01}Group_i + \nu_{0i}$	$\gamma_{1i} = \beta_{10} + \beta_{11}Group_i + \nu_{1i}$

*Stress is defined as either social stress, activity stress, or event stress

to the model, including a random slope effect for the slope of momentary stress. For the third hypothesis, we also built upon the baseline model from hypothesis 1, including now the NA's autoregressive parameter (lagged within individuals and days, with the first beep of each day set to missing) as the predictor variable as well as the interaction with group. The random effect structure was again defined as having a random intercept and a random slope for lagged momentary NA. Except for this last model, all models included a first-order autoregressive [AR(1)] error structure. Mathematic formulas for the models can be found in Table 2. All analyses were performed in R, version 3.5.3 (2019), using the nlme package (Pinheiro et al., 2020).¹

Results

Based on the predefined compliance cutoff scores and data exclusion from day 1, the number of participants varied between 401 and 456. Table 3 provides a summary of the number of included participants under each exclusion criterion. The results concerning our first hypothesis (i.e., elevated momentary NA in individuals with psychosis compared to healthy individuals) indicated, as expected, significant differences in the mean level of NA between individuals with and without psychosis across all generated datasets. The estimated unstandardized coefficient value for group ranged from .53 to .66, with all corresponding *p*-values < .001 (Fig. 1, row 1).

Concerning the second hypothesis, we explored whether there was an association between momentary stress and momentary NA and whether this effect was stronger for individuals with than without psychosis. In line with our hypotheses, we found that social, event, and activity stress were each positively associated with NA across all generated datasets (Fig. 1, rows 2 to 4). The estimated unstandardized coefficient values ranged from .04 to .07 for social stress,

.06 to .11 for event stress, and .01 to .08 for activity stress (*p*-values < .01). The interaction term of diagnostic group with stress reactivity was also significant across all datasets for activity and event stress (*p* < .05). Unstandardized coefficient values ranged from .07 to .09 for activity stress and .03 to .07 for event stress, indicating a stronger association with momentary NA among individuals with psychosis than healthy individuals. In contrast, when considering social stress, we found that the interaction term became nonsignificant in those scenarios when NA was computed as the mean (i.e., 22 out of the 66 datasets; Fig. 1, row 2). Conversely, the interaction was consistently significant across all stress variables when NA was computed as the median or mode (*p* < .05).

Table 3 Number of included participants under different exclusion criteria

Compliance criterion	Data removed from day 1	Number of participants
0%	No	456
5%	No	454
10%	No	452
15%	No	452
20%	No	448
25%	No	444
30%	No	442
35%	No	437
40%	No	432
45%	No	419
50%	No	405
0%	Yes	452
5%	Yes	451
10%	Yes	449
15%	Yes	445
20%	Yes	444
25%	Yes	440
30%	Yes	437
35%	Yes	430
40%	Yes	423
45%	Yes	413
50%	Yes	401

¹ Interactive plots for viewing coefficient values per dataset are available in the HTML document provided on the preregistration of this project on the Open Science Framework (OSF) webpage (https://osf.io/vefkg/?view_only=0d70ea4e0b8d4241901516131cc38cad).

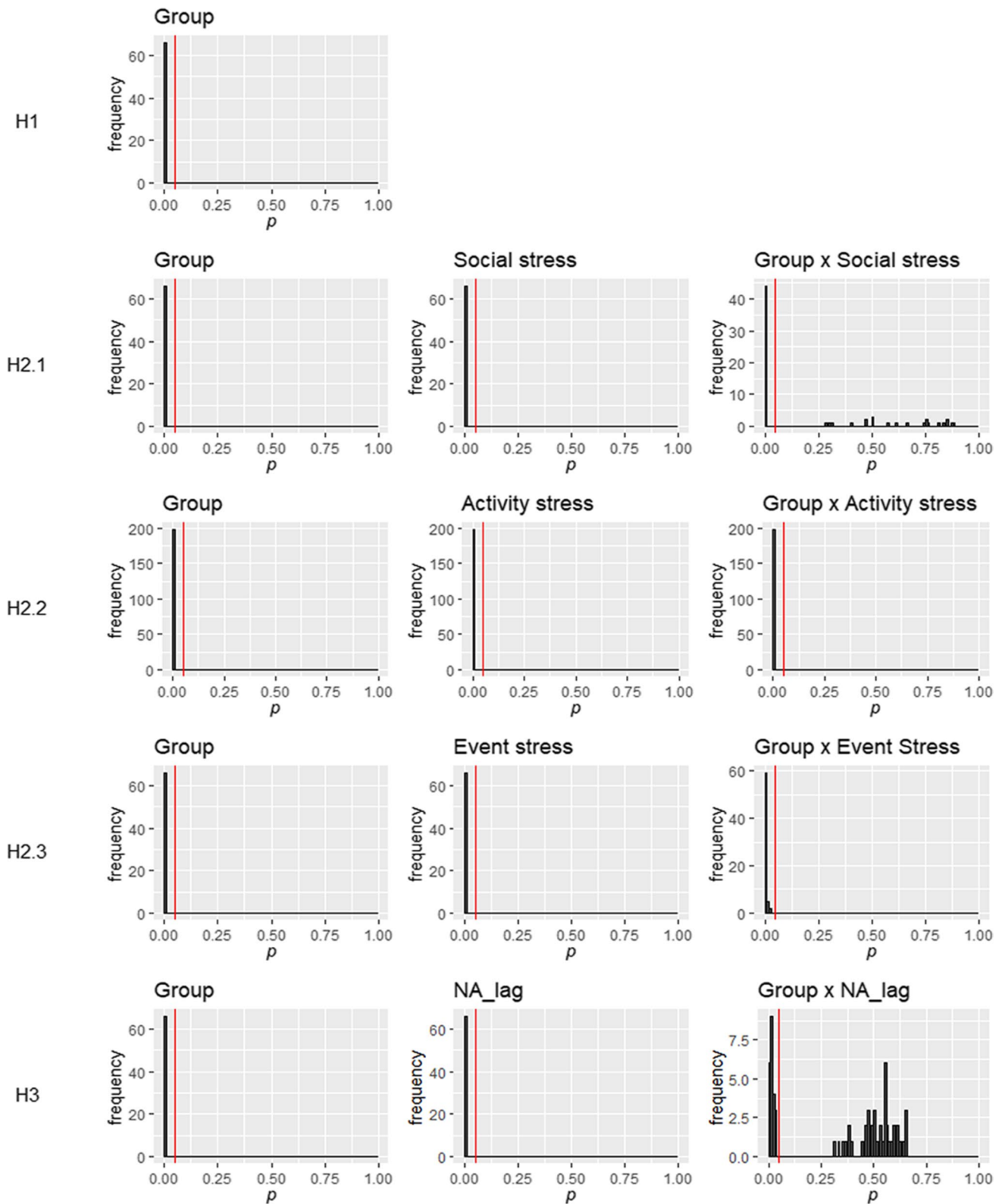


Fig. 1 Frequencies of p -values were obtained across generated datasets that were used for testing hypotheses 1 to 3. Each row represents one hypothesis. Hypothesis 1 = Momentary negative affect is elevated in individuals with psychosis as compared to controls; Hypotheses 2.1 to 2.3 = Momentary negative affect is predicted by either momentary social (2.1), activity (2.2), and event-related (2.3)

stress with group (individuals with psychosis vs. controls) moderating this effect. Hypothesis 3 = Momentary negative affect is predicted by momentary negative affect on the preceding beep with group (individuals with psychosis vs. control) moderating this effect. The red line indicates $p = .05$.

Table 4 Standardized factor loadings of the negative affect items at both the within- and between-person level for both patients with psychosis and healthy individuals

	I feel unsure	I am lonely	I am anxious	I feel irritated	I feel sad	I feel guilty
Patients with psychosis						
Within-person level	0.54	0.43	0.57	0.44	0.63	0.48
Between-person level	0.71	0.81	0.68	0.81	0.61	0.77
Healthy individuals						
Within-person level	0.54	0.36	0.56	0.42	0.53	0.41
Between-person level	0.95	0.59	0.89	0.69	0.79	0.76

Note. Standardized root mean square residuals for the within- and between-person covariance matrix were 0.036 and 0.022 for patients and 0.036 and 0.046 for controls, respectively. Other fit indices can be found in our online HTML file (https://osf.io/vefkg/?view_only=0d70ea4e0b8d4241901516131cc38cad).

For the third hypothesis, we investigated whether NA at time point t could be predicted by NA at the previous time point $t - 1$, with stronger anticipated effects for individuals with psychosis. The results only partially supported this (Fig. 1, row 5). While lagged NA was, as expected, a significant predictor of NA across all datasets ($p < .001$, unstandardized coefficient values ranging from .19 to .21), the interaction term (group \times lagged NA) was significant only in those scenarios when NA and lagged NA were computed as the mean ($p < .05$, unstandardized coefficient values ranging from .07 to .08). Conversely, nonsignificant results were observed when the median or mode ($.31 < p < .66$) were used, indicating that preprocessing choices regarding measures of central tendency impacted the conclusion of whether individuals with psychosis do (or do not) show meaningfully greater inertia of NA than healthy individuals.

Using the Lavaan R package (Rosseel, 2012), post hoc analyses were conducted to further explore this finding, in which we looked at NA's factor structure at the between- and within-person level. As shown in Table 4, NA's factor loadings were found considerably higher at the between- than the within-person level, suggesting that the negative emotions co-varied less within than across individuals. This finding was further supported by considerably lower internal consistencies at the within- ($\alpha = .67$) than the between-person level ($\alpha = .93$). Given that people may intensely experience a particular or specific constellation of negative emotions, we then used the leave-one-out approach (i.e., using only five out of the six NA items each time) to investigate whether the composition of NA was disproportionately affected by items considered in this study. Results revealed that conclusions differed as a function of the selected NA items (Table 5). Specifically, while excluding the feeling irritated emotion resulted in more datasets yielding a significant interaction term than the original MA (i.e., 55 instead of 44), the opposite effect was found for feeling lonely, anxious, sad, and guilty (Table 5). In contrast, no difference was observed for feeling unsure. Similarly, as the within- and between-person

factor loadings seemed to differ, we additionally investigated measurement invariance of the NA construct. To do this, we tested for configural invariance, metric invariance, and invariance of latent variances and covariances at the within-person and between-person level. Additionally, and specific for the between-person level, scalar invariance as well as invariance of latent means was investigated (see also Eisele et al., 2021; Lance et al., 2000; Ryu, 2014). Results show that, while there is configural invariance, and partial metric invariance (when freeing two equality constraints on NA items 1 and 6, 2 and 6, 3 and 6, 4 and 6), there was no invariance of latent variances and covariances at the within-person level. Our measurement invariance analyses at the between-person level showed that there was configural, metric, and scalar invariance as well as invariance for (co) variances and latent means. Results of the measurement invariance analysis are summarized in table 6.

Discussion

This study illustrated the use of MA in ESM research and reanalyzed established group differences in NA, stress reactivity, and emotional inertia between individuals with and without psychosis. Using a large pooled intensive longitudinal dataset, we investigated the robustness of statistical conclusions of preprocessing choices related to data exclusion (i.e., based on various levels of compliance and exclusion of the first assessment day) and the calculation of constructs (i.e., composite scores calculated as the mean, median, or mode). The findings revealed that while different data exclusion choices have no meaningful effect on conclusions, using the median or mode instead of mean values to compute NA can lead to different conclusions.

Prior research on NA, stress reactivity, and emotional inertia has consistently provided evidence for the existence of group differences between healthy individuals and those suffering from mental illnesses (Kuppens et al., 2010, 2012;

Table 5 Number of significant scenarios across the multiverse in a leave-one-out approach where each time one item was removed in the calculation of negative affect composite score

Hypothesis**	Predictor	Data used***						
		original	min NA1*	min NA2*	min NA3*	min NA4*	min NA5*	min NA6*
1	Group	66/66	66/66	66/66	66/66	66/66	66/66	66/66
2.1	Group	66/66	66/66	66/66	66/66	66/66	66/66	66/66
	Social stress	66/66	66/66	66/66	66/66	66/66	66/66	66/66
	Social stress × group	44/66	44/66	28/66	41/66	55/66	23/66	33/66
2.2	Group	198/198	198/198	198/198	198/198	198/198	198/198	198/198
	Activity stress	198/198	198/198	198/198	198/198	198/198	171/198	198/198
	Activity stress × group	198/198	198/198	198/198	198/198	198/198	198/198	198/198
2.3	Group	66/66	66/66	66/66	66/66	66/66	66/66	66/66
	Event stress	66/66	66/66	66/66	66/66	66/66	66/66	66/66
	Event stress × group	66/66	66/66	66/66	66/66	66/66	66/66	64/66
3	Group	66/66	66/66	66/66	66/66	66/66	66/66	66/66
	Lagged NA	66/66	66/66	66/66	66/66	66/66	66/66	66/66
	Lagged NA × group	22/66	22/66	11/66	16/66	22/66	22/66	22/66

*NA1 = I feel unsure, NA2 = I am lonely, NA3 = I am anxious, NA4 = I feel irritated, NA5 = I feel sad, NA6 = I feel guilty. Each row displays the number of scenarios out of the total number of possible scenarios in which a significant result ($p < .05$) was reached for the specified hypothesis and the predictor under investigation.

**Hypothesis 1 = Momentary negative affect is elevated in individuals with psychosis as compared to controls; Hypothesis 2.1 to 2.3 = Momentary negative affect is predicted by either momentary social (2.1), activity (2.2), and event-related (2.3) stress with group (individuals with psychosis vs. controls) moderating this effect. Hypothesis 3 = Momentary negative affect is predicted by momentary negative affect on the preceding beep with group (individuals with psychosis vs. control) moderating this effect

***Differences between the original multiverse and the leave-one-out multiverses could not be attributed to differences in the used central tendency measures

Table 6 Measurement invariance analyses of NA within-person structure

Model	CFI	RMSEA	df	AIC	BIC	χ^2	p	diff. χ^2	df diff.	P diff.	diff. RMSEA	diff. CFI
Configural	0.956	0.074	18	86872.57	87376.02	247.06	0					
Partial metric	0.953	0.070	21	86885.30	87369.38	265.79	0	18.73	3	0	0.004	0.003
Partial (co)variance	0.944	0.075	22	86931.14	87408.77	313.63	0	47.84	1	0	-0.005	0.009
<i>Measurement invariance of NA between-person structure</i>												
Configural	0.994	0.027	18	86674.67	87178.12	49.16	0					
Metric	0.993	0.026	23	86674.64	87145.81	59.13	0	9.97	5	0	0.001	0.001
Scalar	0.994	0.022	28	86665.98	87104.88	60.47	0	1.34	5	0	0.004	-0.001
(co)Variance	0.988	0.030	29	86695.03	87127.48	91.52	0	31.05	1	0	-0.008	0.006
Latent means	0.982	0.037	30	86726.26	87152.25	124.75	0	33.23	1	0	-0.007	0.006

Note. A significant difference between the different models was considered when two or more fit indices reached a significant cutoff, defined as $>.005$ for diff. CFI, $>.01$ for diff RMSEA, and $<.05$ for P diff.

Myin-Germeys & van Os, 2007; Reininghaus et al., 2016b). While our findings support these studies, it is essential to acknowledge that this was, in some instances, dependent upon specific preprocessing choices. Specifically, when NA was computed as the mean, as opposed to the median or mode, our results deviated from those in the literature (e.g., Reininghaus et al., 2016b; ultrahigh-risk and first-episode psychosis), suggesting that there are no differences concerning the degree to which social stress predicts NA in

individuals with psychosis compared to healthy individuals. Similarly, the strength of emotional inertia of NA differed between individuals with psychosis and healthy individuals only when NA was computed as the mean (e.g., comparable to previous work on patients suffering from depression; Kuppens et al., 2010) as opposed to the median or mode. Post hoc analyses suggest that this finding could be attributed to lower factor loadings of negative emotions at the within- than between-person level.

The observation of substantially different factor loadings for NA at the within- versus between-person level is not novel and has been reported previously (Eadeh et al., 2019; Möwisch et al., 2019; Vansteelandt et al., 2005; Zelenski & Larsen, 2000). Zelenski and Larsen (2000) found a similar factor as the one we observed and attributed this to affect measurements behaving differently at different levels of analyses. Specifically, these authors argued that emotions are more strongly intercorrelated across time between (e.g., someone who feels lonelier on average also feels sadder than the average person) than within individuals when the interrelation is assessed across short periods (e.g., someone who feels lonelier than usual may not feel sadder than usual in a specific situation). Our findings support this notion by finding lower within- than between-person factor loadings, suggesting that using mean, median, or mode values to compute scores may ignore the optimal weighting of items, which not only affects scale reliability, but also impacts conclusions in subsequent analysis (see also McNeish & Wolf, 2020). Therefore, it may be necessary to re-evaluate a traditional two-factor structure at the within-person level (Brose et al., 2015). Future work could investigate whether more than one lower-order factor (e.g., an internalizing and externalizing component) exists, and also explore the predictive utility of specific negative emotions at the individual level instead of the composite factor NA. In a similar vein, it would be necessary to determine whether the finding of differential factor loadings at the within- and between-person level found in this study can be generalized across patient populations. If this holds true, the implications are far-reaching, as it would reflect the need for operationalizing NA within individuals in a more granular way than is frequently done in ESM research. Furthermore, the finding of only partial metric invariance at the within-person level adds another level of complexity to ESM research investigating between-group differences in NA. More specifically, it suggests that the pattern of factor loadings may differ within individuals, irrespective of group, for several items. Therefore, future work is advised to evaluate measurement invariance before making substantive inferences based on ESM data.

While the sample size of the dataset ($N = 456$ and 26,892 longitudinal assessments) was considerably larger than previous ESM studies (N in the range of 42–99; e.g., Rauschenberg et al., 2017; van der Steen et al., 2017; Vasconcelos e Sa et al., 2016), four limitations should be considered in interpreting the findings. First, given that the operationalization of the ESM constructs used in this study varies across the literature (for examples of alternative conceptualization of NA-, social-, activity-, and event-related stress see Gerritsen et al., 2019; Lüdtke et al., 2017; and Sitko et al., 2016), it is unclear to what extent the current findings will be replicated under different operationalization protocols from the one used in this study. Second, the dataset used in this

study came from a series of studies that collected ESM data using a digital wristwatch and paper-and-pencil booklets. Importantly, however, prior studies consistently found that this method and full electronic data collection provide data that are equivalent psychometrically and in patterns of findings (Green et al., 2006; Gwaltney et al., 2008; Jacobs et al., 2005). Third, multiverse scenarios were constructed based on what we considered reasonable preprocessing choices. However, it is important to acknowledge that MA is an exploratory methodology (Simonsohn et al., 2015) and that opinions about which preprocessing choices need to be considered may vary among researchers and across labs. However, including many choices may make the multiverse of possible scenarios unmanageable, as the number of datasets to consider quickly becomes so large that it is challenging to disentangle each preprocessing choice (Del Giudice & Gangestad, 2020). Fourth, we only specified one model per hypothesis. In addition to the unique datasets considered for each hypothesis, a multiverse of possible statistical models exists. Future work could include alternatives such as ordinal variants of mixed-effects models (e.g., Bürkner & Vuorre, 2019) or the use of item response theory (IRT) models (e.g., Hedeker et al., 2006).

These limitations notwithstanding, the findings of this study have several implications for future research. First, we illustrated that MA is a useful and easy-to-use tool to investigate the impact of preprocessing decisions within experience sampling research. Second, our illustrative cases show that excluding data based on various compliance levels and excluding the first assessment day may be inadvisable. This reduces the power of analyses, and we found that applying different preprocessing choices related to data exclusion does not affect results. That said, we do recommend future work to explore whether this finding can be replicated across various sampling frequencies. Third, the finding that different methods for calculating (i.e., mean, median, mode) or conceptualizing NA (i.e., excluding or including items) affected conclusions in our sample illustrates the importance of exploring the effect of different but equally reasonable methods of operationalizing ESM constructs in future work. In doing so, researchers should be transparent about the choices made and the reasoning behind them, making this explicit during the preregistration process (Kirtley et al., 2019).

Conclusion

This study showed that MA could be used as a valuable data-analytical technique to investigate how different, but equally reasonable, data preprocessing choices affect ESM studies' statistical conclusions. We found that established group differences in NA, stress reactivity, and emotional inertia in

individuals with psychosis compared to healthy individuals were not affected by different data exclusion choices. In contrast, calculating NA as either the mean, median, or mode of items did affect the conclusions in our study. Additional analyses revealed that this was related to different factor loading patterns at the within and between person-level. These findings imply that considering only a single set of choices risks undermining the validity of conclusions. Therefore, the most significant contribution of this study is that it empirically illustrates the necessity of conducting MA to come to terms with different preprocessing choices in ESM research.

Funding This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement 777084 (DynaMORE project). Dr. Kiekens is supported by a Postdoctoral Fellowship from the Research Foundation Flanders (12ZZM21N).

Declarations

Conflict of Interest The authors declare that they have no conflict of interest.

References

- Arslan, R. C., Reitz, A. K., Driebe, J. C., Gerlach, T. M., & Penke, L. (2020). Routinely randomize potential sources of measurement reactivity to estimate and adjust for biases in subjective reports. *Psychological Methods*. Advance online publication. <https://doi.org/10.1037/met0000294>
- Bak, M., Van Der Spil, F., Gunther, N., Radstake, S., Delespaul, P., & Van Os, J. (2001). Maastricht Assessment of Coping Strategies (MACS-I): A brief instrument to assess coping with psychotic symptoms. *Acta Psychiatrica Scandinavica*, 103(6), 453–459. <https://doi.org/10.1034/j.1600-0447.2001.00324.x>
- Blanchard, J. J., Mueser, K. T., & Bellack, A. S. (1998). Anhedonia, positive and negative affect, and social functioning in schizophrenia. *Schizophrenia Bulletin*, 24(3), 413–424. <https://doi.org/10.1093/oxfordjournals.schbul.a033336>
- Brose, A., Voelkle, M. C., Lövdén, M., Lindenberger, U., & Schmiedek, F. (2015). Differences in the between–person and within–person structures of affect are a matter of degree. *European Journal of Personality*, 29(1), 55–71. <https://doi.org/10.1002/per.1961>
- Buckley, P. F., Miller, B. J., Lehrer, D. S., & Castle, D. J. (2009). Psychiatric comorbidities and schizophrenia. *Schizophrenia bulletin*, 35(2), 383–402. <https://doi.org/10.1093/schbul/sbn135>
- Bürkner, P. C., & Vuorre, M. (2019). Ordinal regression models in psychology: A tutorial. *Advances in Methods and Practices in Psychological Science*, 2(1), 77–101. <https://doi.org/10.1177/2515245918823199>
- Chyung, S. Y., Roberts, K., Swanson, I., & Hankinson, A. (2017). Evidence-based survey design: The use of a midpoint on the Likert scale. *Performance Improvement*, 56(10), 15–23. <https://doi.org/10.1002/pfi.21727>
- Csikszentmihalyi, M., Larson, R., & Prescott, S. (1977). The ecology of adolescent activity and experience. *Journal of Youth and Adolescence*, 6, 281–294. <https://doi.org/10.1007/BF02138940>
- Dejonckheere, E., Mestdagh, M., Houben, M., Rutten, I., Sels, L., Kuppens, P., & Tuerlinckx, F. (2019). Complex affect dynamics add limited information to the prediction of psychological well-being. *Nature Human Behaviour*, 3(5), 478–491. <https://doi.org/10.1038/s41562-019-0555-0>
- Del Giudice, M., & Gangestad, S. W. (2020). A traveler's guide to the multiverse: Promises, pitfalls, and a framework for the evaluation of analytic decisions. *Advances in Methods and Practices in Psychological Science* <https://doi.org/10.1177/2515245920954925>
- Eadeh, H. M., Breaux, R., Langberg, J. M., Nikolas, M. A., & Becker, S. P. (2019). Multigroup multilevel structure of the child and parent versions of the positive and negative affect schedule (PANAS) in adolescents with and without ADHD. *Psychological Assessment*, 32(4), 374–382. <https://doi.org/10.1037/pas0000796>
- Edwards, C. J., Cella, M., Emsley, R., Tarrier, N., & Wykes, T. (2018). Exploring the relationship between the anticipation and experience of pleasure in people with schizophrenia: An experience sampling study. *Schizophrenia Research*, 202, 72–79. <https://doi.org/10.1016/j.schres.2018.06.040>
- Eisele, G., Lafit, G., Vachon, H., Kuppens, P., Houben, M., Myin-Germeys, I., & Viechtbauer, W. (2021). Affective structure, measurement invariance, and reliability across different experience sampling protocols. *Journal of Research in Personality*, 92, 104094. <https://doi.org/10.1016/j.jrp.2021.104094>
- Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: a new look at an old issue. *Psychological Methods*, 12(2), 121. <https://doi.org/10.1037/2F1082-989X.12.2.121>
- Gaudio, B. A., Ellenberg, S., Price, L. H., & Moitra, E. (2018). Time-lagged predictors of daily medication nonadherence beliefs during the month post-hospital discharge in patients with psychotic-spectrum disorders. *Psychiatry Research*, 270, 253–256. <https://doi.org/10.1016/j.psychres.2018.09.048>
- Gerritsen, C., Bagby, R. M., Sanches, M., Kiang, M., Maheandiran, M., Preece, I., & Mizrahi, R. (2019). Stress precedes negative symptom exacerbations in clinical high risk and early psychosis: A time-lagged experience sampling study. *Schizophrenia Research*, 210, 52–58. <https://doi.org/10.1016/j.schres.2019.06.015>
- Green, A. S., Rafaeli, E., Bolger, N., Shrout, P. E., & Reis, H. T. (2006). Paper or plastic? Data equivalence in paper and electronic diaries. *Psychological Methods*, 11(1), 87. <https://doi.org/10.1037/1082-989X.11.1.87>
- Gwaltney, C. J., Shields, A. L., & Shiffman, S. (2008). Equivalence of electronic and paper-and-pencil administration of patient-reported outcome measures: A meta-analytic review. *Value in Health*, 11(2), 322–333. <https://doi.org/10.1111/j.1524-4733.2007.00231.x>
- Hamaker, E. L., & Grasman, R. P. P. P. (2014). To center or not to center? Investigating inertia with a multilevel autoregressive model. *Frontiers in Psychology*, 5(OCT), 1–15. <https://doi.org/10.3389/fpsyg.2014.01492>
- Hedeker, D., Mermelstein, R. J., & Flay, B. R. (2006). *Application of item response theory models for intensive longitudinal data* (pp. 84–108). New York, NY: Oxford University Press. Retrieved from: tinyurl.com/2asj5uey
- Jacobs, N., Nicolson, N. A., Derom, C., Delespaul, P. A. E. G., Van Os, J., & Myin-Germeys, I. (2005). Electronic monitoring of salivary cortisol sampling compliance in daily life. *Life Sciences*, 76(21), 2431–2443. <https://doi.org/10.1016/j.lfs.2004.10.045>
- Jamieson, S. (2004). Likert scales: How to (ab)use them. In *Medical Education*, 38(12), 1217–1218. <https://doi.org/10.1111/j.1365-2929.2004.02012.x>
- Kirtley, O. J., Lafit, G., Achterhof, R., Hiekkaranta, A. P., & Myin-Germeys, I. (2019). Making the black box transparent: A pre-registration template for studies using experience sampling

- methods. *Journal of Chemical Information and Modeling*, 53(9), 1689–1699. <https://doi.org/10.1017/CBO9781107415324.004>
- Kuppens, P., Allen, N. B., & Sheeber, L. B. (2010). Emotional inertia and psychological maladjustment. *Psychological Science*, 21(7), 984–991. <https://doi.org/10.1177/0956797610372634>
- Kuppens, P., Sheeber, L. B., Yap, M. B. H., Whittle, S., Simmons, J. G., & Allen, N. B. (2012). Emotional inertia prospectively predicts the onset of depressive disorder in adolescence. *Emotion*, 12(2), 283–289. <https://doi.org/10.1037/a0025046>
- Lafit, G., Adolf, J. K., Dejonckheere, E., Myin-Germeys, I., Viechtbauer, W., & Ceulemans, E. (2021). Selection of the number of participants in intensive longitudinal Studies: A user-friendly shiny app and tutorial for performing power analysis in multi-level regression models that account for temporal dependencies. *Advances in Methods and Practices in Psychological Science*, 4(1), 1–24. <https://doi.org/10.1177/2F2515245920978738>
- Lance, C. E., Vandenberg, R. J., & Self, R. M. (2000). Latent growth models of individual change: The case of newcomer adjustment. *Organizational Behavior and Human Decision Processes*, 83(1), 107–140. <https://doi.org/10.1177/2F109442810031002>
- Lardinois, M., Lataster, T., Mengelers, R., Van Os, J., & Myin-Germeys, I. (2011). Childhood trauma and increased stress sensitivity in psychosis. *Acta Psychiatrica Scandinavica*, 123(1), 28–35. <https://doi.org/10.1111/j.1600-0447.2010.01594.x>
- Larson, R., & Csikszentmihalyi, M. (2014). The experience sampling method. In *Flow and the foundations of positive psychology* (pp. 21–34). Springer, Dordrecht. https://doi.org/10.1007/978-94-017-9088-8_2
- Lataster, T., Collip, D., Lardinois, M., Van Os, J., & Myin-Germeys, I. (2010). Evidence for a familial correlation between increased reactivity to stress and positive psychotic symptoms. *Acta Psychiatrica Scandinavica*, 122(5), 395–404. <https://doi.org/10.1111/j.1600-0447.2010.01566.x>
- Lataster, T., Valmaggia, L., Lardinois, M., Van Os, J., & Myin-Germeys, I. (2013). Increased stress reactivity: A mechanism specifically associated with the positive symptoms of psychotic disorder. *Psychological Medicine*, 43(7), 1389–1400. <https://doi.org/10.1017/S0033291712002279>
- Li, X., Zhang, Y. T., Huang, Z. J., Chen, X. L., Yuan, F. H., & Sun, X. J. (2019). Diminished anticipatory and consummatory pleasure in dysphoria: Evidence from an experience sampling study. *Frontiers in Psychology*, 10, 1–10. <https://doi.org/10.3389/fpsyg.2019.02124>
- Lindstädt, R., Proksch, S. O., & Slapin, J. B. (2020). When experts disagree: Response aggregation and its consequences in expert surveys. *Political Science Research and Methods*, 8(3), 580–588. <https://doi.org/10.1017/psrm.2018.52>
- Lüdtke, T., Kriston, L., Schröder, J., Lincoln, T. M., & Moritz, S. (2017). Negative affect and a fluctuating jumping to conclusions bias predict subsequent paranoia in daily life: An online experience sampling study. *Journal of Behavior Therapy and Experimental Psychiatry*, 56, 106–112. <https://doi.org/10.1016/j.jbtep.2016.08.014>
- McNeish, D., & Wolf, M. G. (2020). Thinking twice about sum scores. *Behavior research methods*, 52(6), 2287–2305. <https://doi.org/10.3758/s13428-020-01398-0>
- Möwisch, D., Schmiedek, F., Richter, D., & Brose, A. (2019). Capturing affective well-being in daily life with the day reconstruction method: A refined view on positive and negative affect. *Journal of Happiness Studies*, 20(2), 641–663. <https://doi.org/10.1007/s10902-018-9965-3>
- Myin-germeys, I., Kasanova, Z., Vaessen, T., Vachon, H., Kirtley, O., Viechtbauer, W., & Reininghaus, U. (2018). Experience sampling methodology in mental health research: new insights and technical developments. *World Psychiatry*, 17, 123–132. <https://doi.org/10.1002/wps.20513>
- Myin-Germeys, I., Peeters, F., Havermans, R., Nicolson, N. A., DeVries, M. W., Delespaul, P., & Van Os, J. (2003). Emotional reactivity to daily life stress in psychosis and affective disorder: An experience sampling study. *Acta Psychiatrica Scandinavica*, 107(2), 124–131. <https://doi.org/10.1034/j.1600-0447.2003.02025.x>
- Myin-Germeys, I., & van Os, J. (2007). Stress-reactivity in psychosis: Evidence for an affective pathway to psychosis. *Clinical Psychology Review*, 27(4), 409–424. <https://doi.org/10.1016/j.cpr.2006.09.005>
- Myin-Germeys, I., Van Os, J., Schwartz, J. E., Stone, A. A., & Delespaul, P. A. (2001). Emotional reactivity to daily life stress in psychosis. *Archives of General Psychiatry*, 58(12), 1137–1144. <https://doi.org/10.1001/archpsyc.58.12.1137>
- Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R-core-team. (2020). *nlme: Linear and Nonlinear Mixed Effects Models* (R package version 3.1-144). <https://cran.r-project.org/package=nlme>
- Rauschenberg, C., van Os, J., Cremers, D., Goedhart, M., Schieveld, J. N. M., & Reininghaus, U. (2017). Stress sensitivity as a putative mechanism linking childhood trauma and psychopathology in youth's daily life. *Acta Psychiatrica Scandinavica*, 136(4), 373–388. <https://doi.org/10.1111/acps.12775>
- Reininghaus, U., Gayer-Anderson, C., Valmaggia, L., Kempton, M. J., Calem, M., Onyejiaka, A., Hubbard, K., Dazzan, P., Beards, S., Fisher, H. L., Mills, J. G., McGuire, P., Craig, T. K. J., Garety, P., Van Os, J., Murray, R. M., Wykes, T., Myin-Germeys, I., & Morgan, C. (2016a). Psychological processes underlying the association between childhood trauma and psychosis in daily life: An experience sampling study. *Psychological Medicine*, 46(13), 2799–2813. <https://doi.org/10.1017/S003329171600146X>
- Reininghaus, U., Kempton, M. J., Valmaggia, L., Craig, T. K. J., Garety, P., Onyejiaka, A., Gayer-Anderson, C., So, S. H., Hubbard, K., Beards, S., Dazzan, P., Pariante, C., Mondelli, V., Fisher, H. L., Mills, J. G., Viechtbauer, W., McGuire, P., Van Os, J., Murray, R. M., ... Morgan, C. (2016b). Stress sensitivity, aberrant salience, and threat anticipation in early psychosis: An experience sampling study. *Schizophrenia Bulletin*, 42(3), 712–722. <https://doi.org/10.1093/schbul/sbv190>
- Rintala, A., Wampers, M., Myin-Germeys, I., & Viechtbauer, W. (2019). Response compliance and predictors thereof in studies using the experience sampling method. *Psychological Assessment*, 31(2), 226–235. <https://doi.org/10.1037/pas0000662>
- Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. R package version 0.5-15 <http://lavaan.org>. *Journal of Statistical Software*, 48(2), 1–36. https://econpapers.repec.org/article/jssjstsof/v_3a048_3ai02.htm%0A
- Ryu, E. (2014). Factorial invariance in multilevel confirmatory factor analysis. *British Journal of Mathematical and Statistical Psychology*, 67(1), 172–194. <https://doi.org/10.1111/bmsp.12014>
- Shrout, P. E., Stadler, G., Lane, S. P., McClure, M. J., Jackson, G. L., Clavel, F. D., ... & Bolger, N. (2018). Initial elevation bias in subjective reports. *Proceedings of the National Academy of Sciences*, 115(1), 15–23. <https://doi.org/10.1073/pnas.1712277115>
- Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). Specification curve: Descriptive and inferential statistics on all reasonable specifications. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.2694998>
- Sitko, K., Varese, F., Sellwood, W., Hammond, A., & Bentall, R. (2016). The dynamics of attachment insecurity and paranoid thoughts: An experience sampling study. *Psychiatry Research*, 246, 32–38. <https://doi.org/10.1016/j.psychres.2016.08.057>
- Stange, J. P., Kleiman, E. M., Mermelstein, R. J., & Trull, T. J. (2019). Using ambulatory assessment to measure dynamic risk processes in affective disorders. *Journal of Affective Disorders*, 259, 325–336. <https://doi.org/10.1016/j.jad.2019.08.060>

- Steegeen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, 11(5), 702–712. <https://doi.org/10.1177/1745691616658637>
- Stern, J., Arslan, R. C., Gerlach, T. M., & Penke, L. (2019). No robust evidence for cycle shifts in preferences for men's bodies in a multiverse analysis: A response to Gangestad et al. (2019). *Evolution and Human Behavior*, 1–27. <https://doi.org/10.1016/j.evolhumbehav.2019.08.005.No>
- Stone, A. A., Shiffman, S., Schwartz, J. E., Broderick, J. E., & Hufford, M. R. (2002). Patient non-compliance with paper diaries. *British Medical Journal*, 324(7347), 1193–1194. <https://doi.org/10.1136/bmj.324.7347.1193>
- Team, R. C. (2019). *A Language and Environment for Statistical Computing* (3.5.3.). R Foundation for Statistical Computing. <https://www.r-project.org/>
- Thewissen, V., Bentall, R. P., Lecomte, T., van Os, J., & Myin-Germeys, I. (2008). Fluctuations in self-esteem and paranoia in the context of daily life. *Journal of Abnormal Psychology*, 117(1), 143–153. <https://doi.org/10.1037/0021-843X.117.1.143>
- Trull, T. J., & Ebner-Priemer, U. W. (2020). Ambulatory assessment in psychopathology research: A review of recommended reporting guidelines and current practices. *Journal of Abnormal Psychology*, 129(1), 56–63. <https://doi.org/10.1037/abn0000473>
- Vaessen, T., Kasanova, Z., Hernaus, D., Lataster, J., Collip, D., van Nierop, M., & Myin-Germeys, I. (2018). Overall cortisol, diurnal slope, and stress reactivity in psychosis: An experience sampling approach. *Psychoneuroendocrinology*, 96, 61–68. <https://doi.org/10.1016/j.psyneuen.2018.06.007>
- van Berkel, N., Ferreira, D., & Kostakos, V. (2017). The experience sampling method on mobile devices. *ACM Computing Surveys*, 50(6), 1–40. <https://doi.org/10.1145/3123988>
- van der Steen, Y., Gimpel-Drees, J., Lataster, T., Viechtbauer, W., Simons, C. J. P., Lardinois, M., Michel, T. M., Janssen, B., Bechdolf, A., Wagner, M., & Myin-Germeys, I. (2017). Clinical high risk for psychosis: the association between momentary stress, affective and psychotic symptoms. *Acta Psychiatrica Scandinavica*, 136(1), 63–73. <https://doi.org/10.1111/acps.12714>
- van Winkel, M., Nicolson, N. A., Wichers, M., Viechtbauer, W., Myin-Germeys, I., & Peeters, F. (2015). Daily life stress reactivity in remitted versus non-remitted depressed individuals. *European Psychiatry*, 30(4), 441–447. <https://doi.org/10.1016/j.eurpsy.2015.02.011>
- Vansteelandt, K., Van Mechelen, I., & Nezelek, J. B. (2005). The occurrence of emotions in daily life: A multilevel approach. *Journal of Research in Personality*, 39(3), 325–335. <https://doi.org/10.1016/j.jrp.2004.05.006>
- Vasconcelos e Sa, D., Wearden, A., Hartley, S., Emsley, R., & Barrowclough, C. (2016). Expressed Emotion and behaviourally controlling interactions in the daily life of dyads experiencing psychosis. *Psychiatry Research*, 245, 406–413. <https://doi.org/10.1016/j.psychres.2016.08.060>
- Weston, S. J., Ritchie, S. J., Rohrer, J. M., & Przybylski, A. K. (2019). Recommendations for increasing the transparency of analysis of pre-existing data sets. *Advances in Methods and Practices in Psychological Science*, 2(3), 214–227. <https://doi.org/10.1177/2515245919848684>
- Wu, H., & Leung, S. O. (2017). Can Likert scales be treated as interval scales? A simulation study. *Journal of Social Service Research*, 43(4), 527–532. <https://doi.org/10.1080/01488376.2017.1329775>
- Young, C., & Holsteen, K. (2017). Model uncertainty and robustness: A computational framework for multimodel analysis. In *Sociological Methods and Research*, 46(1), 3–40. <https://doi.org/10.1177/0049124115610347>
- Zelenski, J. M., & Larsen, R. J. (2000). The distribution of basic emotions in everyday life: A state and trait perspective from experience sampling data. *Journal of Research in Personality*, 34(2), 178–197. <https://doi.org/10.1006/jrpe.1999.2275>

Open Practices Statement

This paper was preregistered. Details on the preregistration, such as code used for our analysis, are available at https://osf.io/vefkg/?view_only=0d70ea4e0b8d4241901516131cc38cad. Data used in this study are not publicly available, but can be requested by contacting the first or last author.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.