

Understanding Ecological-Momentary-Assessment Data: A Tutorial on Exploring Item Performance in Ecological-Momentary-Assessment Data



Björn S. Siepe¹ , **Carlotta L. Rieble²**, **Rayyan Tutunji²**,
Aljoscha Rimpler³, **Julius März⁴**, **Ricarda K. K. Proppert²**,
and **Eiko I. Fried²** 

¹Psychological Methods Lab, Department of Psychology, University of Marburg, Marburg, Germany; ²Department of Clinical Psychology, Leiden University, Leiden, Netherlands; ³Department of Psychometrics and Statistics, University of Groningen, Groningen, Netherlands; and ⁴Department of Child and Adolescence Psychiatry, Erasmus Medical Centre, Rotterdam, Netherlands

Abstract

The use of ecological-momentary-assessment (EMA) data to study individuals in their everyday lives is popular in many areas of social and life sciences. At the same time, EMA data sets are complex, the psychometric properties of EMA items are often not investigated systematically, and scales are often neither standardized nor validated beyond their face validity. Here, we present different descriptive statistics and data-visualization techniques to increase the understanding of the performance of EMA items. We apply these techniques to a wide range of items used in a large EMA data set (599 participants, 360 time points) collected in the WARN-D study to investigate their distributions, contextual influences, change over time, sources of variability, and relationship with classical static measures of psychopathology. We discuss the theoretical and substantive implications of our findings and provide researchers with R code that they can adapt to their own EMA data, as well as literature recommendations for each topic. We hope to inspire more researchers to share in-depth descriptive summaries of their experience-sampling data such that the field can move forward in understanding the performance of EMA measures across contexts.

Keywords

ecological momentary assessment, data visualization, time-series analysis, intensive longitudinal data, tutorial, open materials

Received 1/25/24; Revision accepted 9/9/24

The use of ecological momentary assessment (EMA) to study individuals in their everyday lives has become widespread in many areas of social sciences. “EMA” refers to various methods that include repeated assessments of individuals in their day-to-day lives (Shiffman et al., 2008). EMA data collection, often carried out via smartphone apps, comes with several benefits, such as increased ecological validity, the possibility to investigate contextual influences, and the opportunity to use a within-persons perspective to explore how concepts of interest develop, interact, and change over time (Trull &

Ebner-Priemer, 2009). To facilitate such rich insights, EMA data sets tend to be relatively complex. For example, EMA data contain a large number of measurement points nested within individuals, usually with multiple assessments per day (Wrzus & Neubauer, 2023). The data frequently feature many different constructs of interest,

Corresponding Author:

Björn S. Siepe, Psychological Methods Lab, Department of Psychology, University of Marburg, Marburg, Germany
Email: bjoern.siepe@uni-marburg.de



commonly using only one item per construct to minimize burden for participants (Dejonckheere et al., 2022). In addition, different sampling schemes are often used for different items, such as asking about sleep quality only in the morning versus measuring current affect multiple times per day.

These and other common features of EMA data, such as attrition and interindividual and intraindividual variation, pose two core challenges for assessment, analysis, and valid inference (Hall et al., 2021). First, common descriptive measures and visualizations, such as sample averages or aggregate histograms of items, fail to convey all relevant information about the highly multivariable and dynamic nature of the data. Given that a thorough description and visualization of EMA data is a crucial requirement for choosing appropriate statistical models and gaining a better theoretical understanding of the constructs under study, there is an urgent need to tackle this challenge. Second, classical psychometric strategies using latent-variable models to assess the quality of measures cannot easily be applied to the many single-item measures common in EMA (M. S. Allen et al., 2022), leaving open the question of how well EMA items perform or function and, as a subsequent step, whether they can be considered valid. To put it differently, the EMA literature is currently in a bit of a free-for-all situation when it comes to assessment because it is largely unclear what “good” items should look like and which psychometric standards this evaluation should be based on.

To tackle these two challenges, in this article, we showcase different descriptive statistics and data-visualization techniques for EMA data to increase the understanding of the performance of individual items. With “item performance,” we loosely refer to item distributions and other relevant properties across individuals, contexts, and time. We consider obtaining a deeper understanding of EMA data a necessary precursor for discussions about the validity of EMA data for two reasons. First, understanding data better is a necessary first step for selecting appropriate statistical models. For example, statistical models often assume normally distributed and unimodal data, and violating these assumptions may have important ramifications for the quality of estimates (Haslbeck et al., 2023; von Klipstein et al., 2023). Second, it is a crucial element for epistemic iteration, that is, for updating theories about constructs based on an improved understanding of measurements and then improving measurements based on updated theories (Chang, 2004). To do so, we stay at the descriptive level of individual items and their bivariate associations and thus do not consider multivariate measurement models, such as reflective latent variables, commonly used in psychometrics.

Our line of reasoning is based on a long tradition of arguing for the importance of data visualization before

model fitting (e.g., Anscombe, 1973; Tukey, 1977; Wainer & Thissen, 1981), a call that is arguably even more relevant with increasing complexity of data for which researchers are no longer interested in just sample averages but also dynamic aspects of the data and interindividual and intraindividual variation. We argue that many key features of data may be obscured if researchers focus solely on the results of modeling techniques and neglect the informational content of raw data. While many researchers may already thoroughly explore and visualize their raw data, this information is usually not available in publications. In that sense, the intensive use of (visual) exploratory data analysis can be tied to theory building in psychology (Haig, 2013). Establishing robust, that is, replicable, phenomena, which the field aims to explain with theories, means finding generalizable patterns in data. This can often be achieved via relatively simple methods, such as visualizations or basic correlations, but requires this information to be available to build a cumulative psychological science. This way, insights generated from individual work can then serve as building blocks to generate and refine explanations and analysis techniques.

Below, we provide a tutorial for researchers interested in better understanding their EMA data focused on the following topics: (a) distributions, (b) context, (c) temporal (in)stability, (d) disentangling variability sources, and (e) measuring across different time scales. We provide R code suitable for standard EMA data formats so that researchers can adopt our visualizations and analyses for their data. In contrast to existing visualization frameworks for EMA data (e.g., Bringmann et al., 2020; Rimpler et al., 2024), which were developed to provide feedback to participants or clinicians, our core goal is to provide a tool for the research community. As a supplement to the article, we further provide a “full report,” which contains multiple additional analyses referenced throughout the article, and a “tutorial” file containing the R code to reproduce all figures on synthetic data (available at OSF, <https://osf.io/yf3up/>).¹

For the tutorial, we use data from the WARN-D project, which aims to build a personalized early warning system for depression. The 3-month EMA data—with around 40 EMA items, around 360 measurement points, and several hundred participants—are representative of the complex nature of EMA data, featuring different sampling frequencies, interindividual and intraindividual variability, state- and trait-like variables, missing data, attrition, and other aspects we discuss below (Fried et al., 2023).

After a brief introduction to data and software, the tutorial is structured into five sections. In the first three sections, we cover insights into individual EMA items based on (a) distributions, (b) context, and (c) time. In the fourth section, we expand on these insights by

(d) disentangling different sources of variability in our data. Finally, we move beyond the univariable case to (e) assess the association of EMA items with more static measures, such as a validated depression questionnaire. For each section, we present several analyses to facilitate a better understanding of an EMA data set and highlight their implications from both a theoretical and substantive standpoint. Analyses are accompanied by visualization examples, our theoretical and statistical interpretation of the results in the WARN-D data, and resources for further reading on the respective topic and potential modeling strategies. We finish the tutorial with an outlook on where researchers might go next after having described and visualized their data.

Method

WARN-D

The overarching goal of the WARN-D project is to build a personalized early warning system for depression in higher-education students. To do so, around 2,000 students, split into four cohorts of around 500 students each, are followed over 2 years. Participating students had to be enrolled in a Dutch institution of higher education, and all surveys were available in both English and Dutch. Participants completed a baseline survey and then underwent 3 months of daily data collection by completing questionnaires and wearing a smartwatch to track digital phenotyping parameters, such as activity, sleep, and heart rate. After this period, students complete follow-ups every 3 months for 2 years. In this article, we use data from a subset of participants from the first two cohorts for illustrative purposes. EMA data collection has been completed for all cohorts, but follow-up assessments are ongoing at the time of writing. The project is funded by the European Research Council in the Horizon 2020 research and innovation program (Grant 949059), and data collection was approved by the Leiden University Research Ethics Committee (2021-09-06-E.I.FriedV2-3406). An in-depth description of WARN-D, including an extensive codebook of all measures and information on reimbursement and data-collection procedures, is available in the protocol article by Fried et al. (2023). Because one of the primary aims of the WARN-D project is creating a prediction system, we excluded a subset of the whole sample for cross-validation techniques in future projects (see the associated preregistration at Tutunji et al., 2023), reducing the sample size available at the time of writing this article from 865 to 599.

During the 3 months of daily data collection, participants received prompts four times per day at semirandom times. In addition, participants received an extra survey on Sundays that included questions about the past week. The full list of items used here is provided

in the online supplement. Unless stated otherwise, all EMA items were assessed on a Likert scale from 1 (*not at all*) to 7 (*very much*). On average, participants missed 45.85% of prompts ($Mdn = 44.12$, $SD = 28.15$). The average age of participants was 22.42 years ($SD = 3.95$). Of the participants, 80.70% identified as women, 16.61% identified as men, and 2.68% indicated another gender identity. More than half of the participants (54.77%) were of Dutch/Belgian/German nationality, and 35.51% had another European nationality; the rest indicated a non-European nationality.

Unless stated otherwise, we exclude participants from analyses if they responded to fewer than 30% of measurement points, after which, $n = 499$ participants remained. We do not investigate missing data in-depth but provide some ideas on how to explore missing data in the full report and link further resources there.

Software and visualization

We use the statistical programming language R (Version 4.3.1; R Core Team, 2023) for all visualizations and analyses. Details, such as specific R packages used, are available in the supplementary materials (<https://osf.io/yf3up/>). Throughout the tutorial, we follow guiding principles for informative visualizations (see e.g., Midway, 2020) when possible, including making heterogeneity across participants, items, and time explicit; using color to highlight differences while remaining colorblind-friendly; and plotting raw data points and/or uncertainty when presenting aggregates or estimates. With these principles in mind and with the overarching goal of gaining a better understanding of our data, the figures and analyses presented below were created and conducted in an iterative process. We recognize that many alternative visualizations of EMA data may be equally or more useful for other research projects. Rather than providing a single set of recommended figures, we aim to showcase our visualization workflow and inspire researchers to adapt and improve on it.

Our commented R code is intended for a long-data format in which each row is a single response of an individual at a certain time point. For a more general introduction to using R and RStudio for data visualization and an explanation for how to restructure data into a long format, see Nordmann et al. (2022). For further resources on data visualization, we refer readers to Hehman and Xie (2021). WARN-D data collection is still ongoing, and we want to avoid having different small parts of the data shared across many different articles. We will make data available on the WARN-D project hub (<https://osf.io/frqdv/>) when all data are collected, cleaned, and checked (excluding potentially identifiable data). To make this article reproducible in the future, we share the exact participant IDs we used for this

Table 1. Means (*SDs*) of Different Individual Summary Statistics

Item	iMean	iMedian	iSD	iSkew	iRMSSD
Cheerful	4.34 (0.86)	4.42 (1.04)	1.18 (0.31)	-0.27 (0.57)	1.39 (0.37)
Irritable	1.98 (0.78)	1.61 (0.88)	1.08 (0.41)	2.09 (1.60)	1.36 (0.51)
Motivated	3.76 (0.87)	3.77 (1.11)	1.26 (0.34)	-0.10 (0.62)	1.53 (0.45)
Nervous	2.22 (0.91)	1.89 (1.09)	1.11 (0.41)	1.69 (1.62)	1.30 (0.49)
Overwhelmed	2.39 (1.02)	2.09 (1.23)	1.18 (0.44)	1.53 (1.86)	1.36 (0.52)
Relaxed	4.34 (0.80)	4.48 (0.99)	1.27 (0.33)	-0.34 (0.54)	1.53 (0.43)
Ruminate	2.01 (0.90)	1.70 (1.01)	1.00 (0.45)	2.26 (2.22)	1.18 (0.52)
Sad	1.88 (0.77)	1.55 (0.83)	1.00 (0.42)	2.51 (2.10)	1.15 (0.47)
Stressed	2.58 (0.94)	2.28 (1.17)	1.25 (0.39)	1.07 (1.36)	1.42 (0.47)
Tired	3.73 (0.94)	3.67 (1.22)	1.41 (0.34)	0.14 (0.61)	1.73 (0.45)

Note: We did not exclude participants because of missingness here. All items ranged from 1 (*not at all*) to 7 (*very much*). Starting letter “i” refers to individual summary statistics. Skewness was calculated in the default way implemented in the R package *e1071* (Meyer et al., 2023). RMSSD = root mean square of successive difference.

article in the supplementary materials. We have also created a mock data set that allows researchers to understand the data structure that we are using, available on OSF (<https://osf.io/yf3up/>). It has a similar structure as the real data but does not mimic any associations or patterns therein and is used within the tutorial code supplement. We recommend researchers wishing to adopt our analyses start with that document.

Results

In the following sections, we introduce statistical and visualization techniques to summarize information that we believe is critical to a better understanding of the EMA data. We provide an extensive online supplementary document, which we call “full report” and includes code to reproduce all analyses in this article and several additional analyses for each of the sections below. In each of the following five sections, we first explain the importance of the topic, perform example analyses and visualizations in the WARN-D data, and then interpret them. At the end of each section, we provide a “Further reading” section that contains literature relevant to each topic.

Distributions

Summary statistics. We start by displaying a range of summary statistics for specific items to gain a first understanding of individual distributions. Table 1 contains means, medians, standard deviations, skew, and the root mean square of successive differences (RMSSDs; for the original publication on the mean square of successive differences, see von Neumann et al., 1941), which has been suggested as a measure of the (in)stability of psychological constructs, such as affect (Schoevers et al., 2021). We

show the RMSSD as an example of a time-series variability measure that has gained popularity in the context of EMA and refer to the further reading section below for more information. Instead of providing a between-subjects average, we calculate each statistic for the whole time series of every item for each participant first. We then calculate the means and standard deviations of the individual statistics across people. For example, suppose we have time series of three individuals A, B, and C, each with 100 time points. We first calculate the individual mean for each person in that person’s time series. If we obtained the means of 1, 4, and 7 for individuals A, B, and C, respectively, we could then calculate the mean across these three mean values as 4. We do the same for all individual summary statistics below—first aggregating within individuals and then across individuals.

We use the starting letter “i” to denote individual summary statistics. For example, the iMedian for cheerful is the mean of all person-specific medians for the item. Table 1 shows, for example, that participants are on average not very irritable, that aggregated means and medians of motivated are around the center of the scale (3.5), that items such as cheerful and relaxed are left-skewed, and that multiple negative-affect items are right-skewed (i.e., exhibited floor effects). The RMSSD is similar across items but shows a large interindividual variation in every item.

Although these statistics provide some first information about our data, the raincloud plots in Figure 1 facilitate a better understanding of the distributions of individual means. These plots include individual observations, density estimates of the distributions, and box plots to show central tendencies and quantiles of the distribution, thus improving over traditional bar or box plot only (M. Allen et al., 2021; Hehman & Xie, 2021).

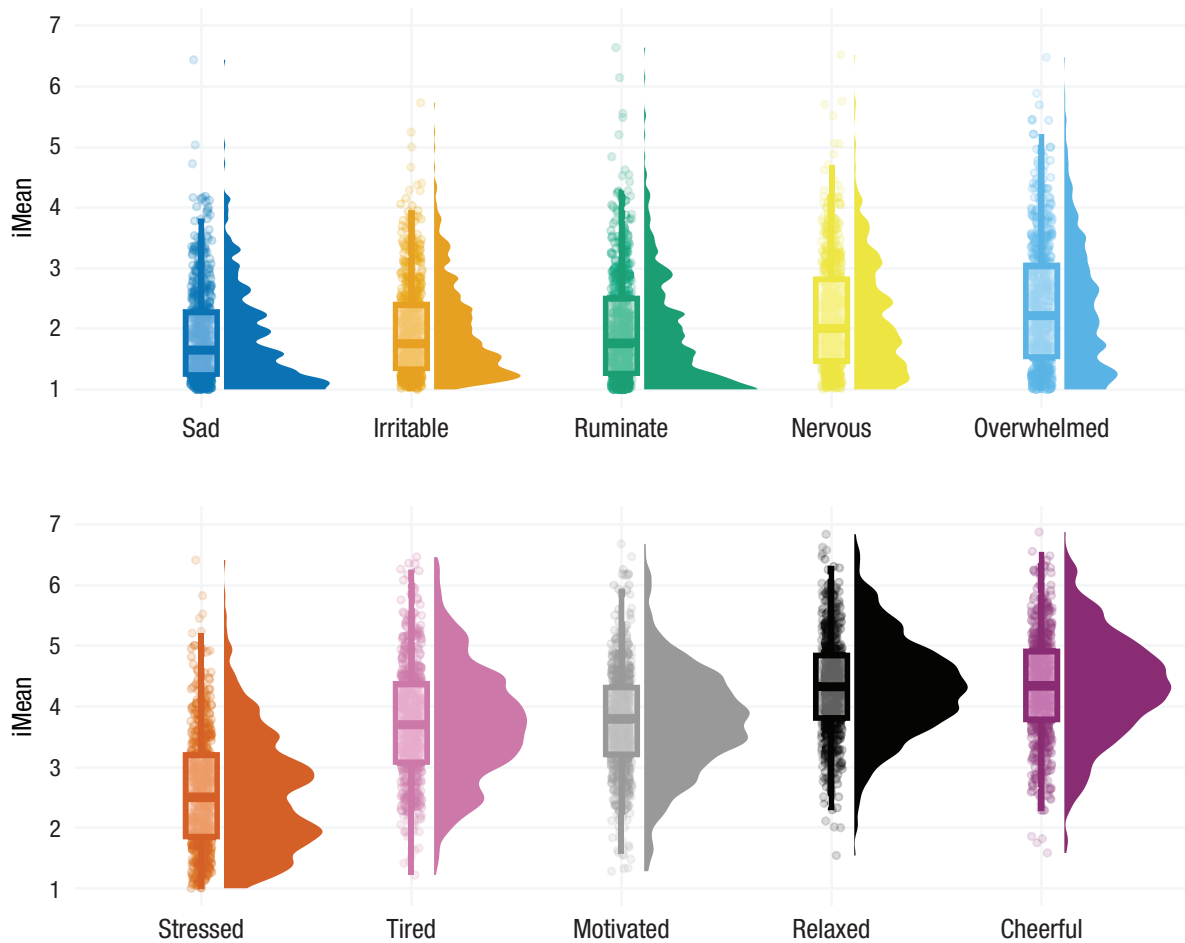


Fig. 1. Raincloud plots for example items. The box plot displays the first and third quartiles of the distribution as upper and lower hinges and the median as a horizontal bar. The whiskers extend to 1.5 times the interquartile range. Dots represent individual means.

In this example, we found both large heterogeneity in individual means across all items and marked differences in the distribution of negative affect (floor effects at the aggregate level for some items) compared with positive ones. Some item distributions also seem to exhibit multiple modes instead of having just one clear peak. Although informative about the heterogeneity in average responses, the summary statistics presented in Table 1 may miss crucial distributional features of individual data. For example, two individuals may have the same individual means, but their distributions may look very different because of multimodality, which we explain further in the following paragraph.

Modality. One particular feature of interest is whether individual distributions are unimodal, that is, have one clear peak, or exhibit multimodality. This may be masked when calculating summary statistics: Even if the distribution of an item was strongly bimodal in every person, the aggregate distribution of all person-specific means of that

item could look perfectly normal. This is important for two reasons. First, the form of person-specific distributions may contain important information about the individuals and constructs that we are studying. For example, a multimodal distribution of an affect item could point to a specific response process in which an individual does not use certain regions of the response scale because of a tendency to choose extreme or center values only (Van Vaerenbergh & Thomas, 2013). It could also hint at phenomena such as individuals switching between multiple stable states, for example, a state of low negative affect and one with high negative affect, which may correspond to better and worse states of their overall mental health (Haslbeck et al., 2023). Second, typical summary statistics, such as the standard deviation, cease to be meaningful descriptors in the presence of multimodality (Smaldino, 2013), and many statistical models applied to time-series data assume unimodal (and symmetric) distributions; violations of these assumptions may have important ramifications for the quality of estimates (Haslbeck et al., 2023).

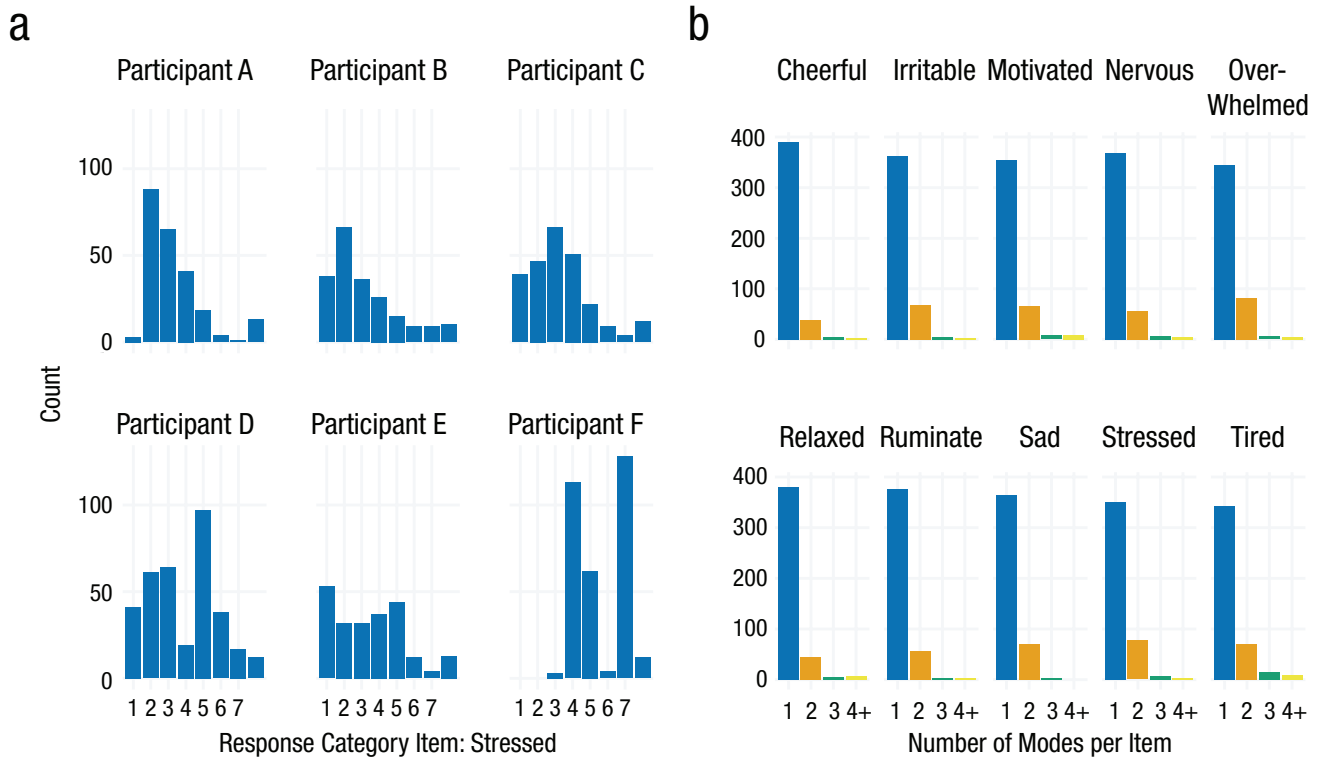


Fig. 2. Modality of ecological-momentary-assessment items. (a) Distributions for unimodal (first row) and bimodal (second row) distributions for the item stressed for six example participants. The x -axis denotes the response options; the height of a chart on the y -axis denotes the count of the respective response. (b) Results for the modality estimation for different items. The x -axis denotes the number of modes; the y -axis denotes the count of participants having a certain number of modes in their individual distribution.

One way to investigate modality is the visual inspection of marginal distributions (Haslbeck et al., 2023), achieved by creating histograms for every item and participant (for an example for the item stressed in six individuals, see Fig. 2a). The example of participant E illustrates that the number of modes is often somewhat ambiguous. For larger data sets, the modality estimation technique by Haslbeck et al. (2023) can be used (we present these results in the full report). We show the number of estimated modes for multiple items in Figure 2b. Although we use Likert scales with a limited range of 1 to 7, which makes it more difficult to identify bimodality, we still find considerable multimodality in some items. For example, around 20% of the response distributions for the items stressed and overwhelmed are estimated to be multimodal, which can be seen by comparing the height of the leftmost blue bar with the height of the other bars in Figure 2b.

Floor effects. Next to the presence of multimodality, floor and (less commonly) ceiling effects can occur in person-specific distributions in EMA data (Mestdagh & Dejonckheere, 2021). Floor effects occur when the average score is low and the response distribution is skewed and exhibits limited variability. This can violate standard assumptions of linear models and make the interpretation

of their results problematic, leading to incorrect conclusions and findings based on statistical artifacts (Terluin et al., 2016; von Klipstein et al., 2023). Floor effects may also help with understanding items and the underlying response processes. For example, one can speculate that floor effects could mean that symptoms or events are not sufficiently relevant or common for many participants in a study. Alternatively, some individuals may not conceptualize the rating of emotions or other constructs on a continuous (e.g., 1–7) scale; instead, they may first evaluate whether the emotion is present and only then rate response options higher than 1. Studying such response processes via cognitive interviews is urgently needed to get a better understanding of why EMA item distributions look like they do (see e.g., Zumbo & Hubley, 2017). We assess floor effects by calculating the proportion of participants that choose the lowest response category more than 80% of the time.² For several items, we find a substantive amount of individuals showing clear evidence of floor effects, leading to some of the group-level distributions shown in Figure 1. For example, for sad or ruminant, 30.47% and 28.04% of individuals, respectively, endorse the lowest response option more than 80% of the time. Visualizations of the floor effect and the limited use of some response options are presented in the full report in Section F3.

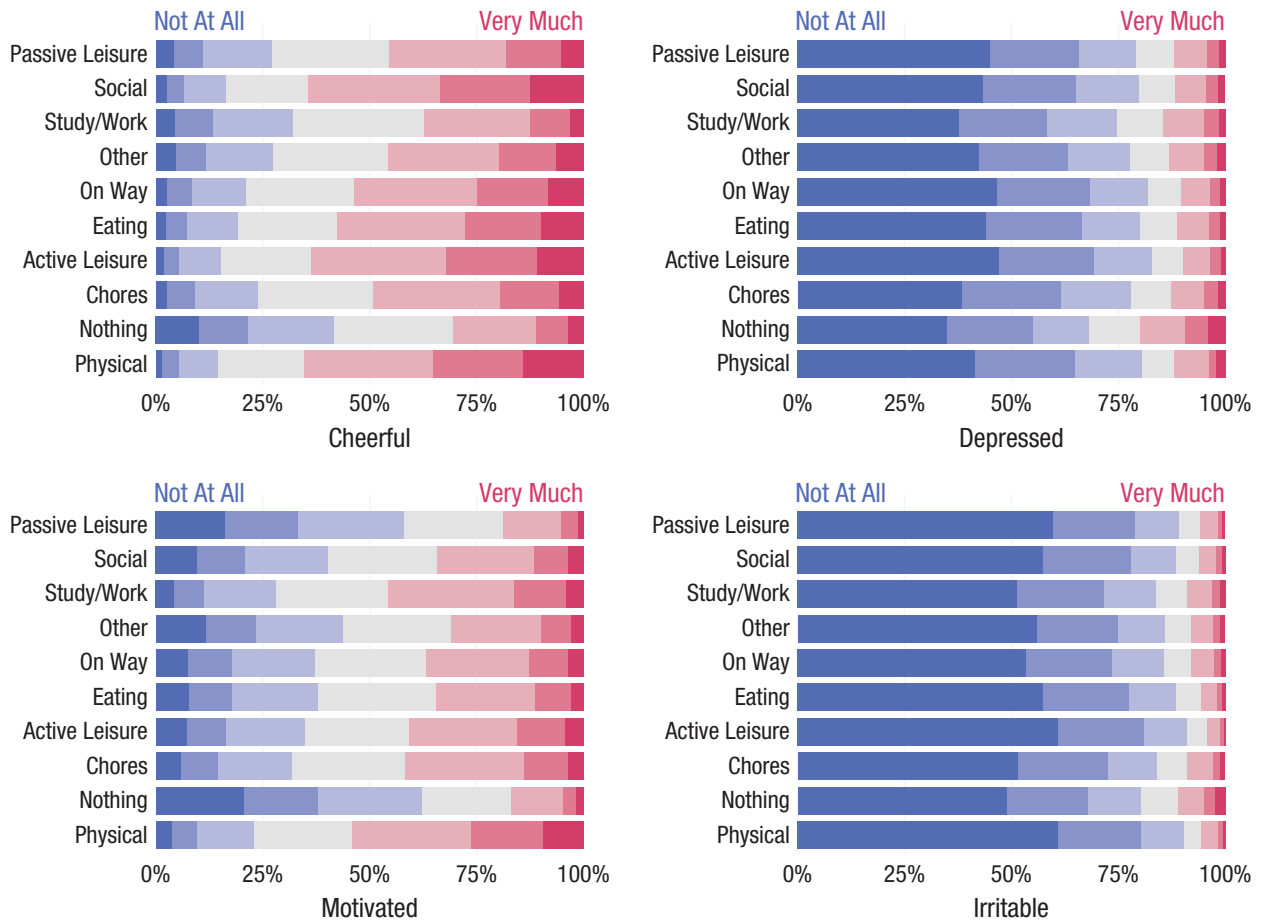


Fig. 3. Answers across activities. This figure displays the relative frequencies of answer options (x-axis) depending on the concurrent activity (y-axis) for four example items. Activities are ordered by their relative frequencies from top to bottom.

Further reading. In general, Revol et al. (2023) provided useful tools for the preprocessing and investigation of EMA data. For a nuanced perspective on calculating time-series summary statistics that aim to capture (in)stability, see Jahng et al. (2008). Haslbeck et al. (2023) investigated modality and skewness in various EMA studies in depth, whereas Haslbeck and Ryan (2022) probed the performance of common time-series models for bimodal response distributions. For some example modeling options for such data, see Cui et al. (2023) and Hamaker et al. (2010). Terluin et al. (2016) and von Klipstein et al. (2023) showed examples of when an ignored floor effect may have affected the conclusions of studies. Alternative time-series models better suited for heavily skewed or zero-inflated distributions have so far been rarely discussed in psychology. They are, for example, covered in Haqiqatkah et al. (2024) and in Ruf et al. (2021).

Context

One of the key promises of EMA studies is to “captur[e] life as it is lived” (Bolger et al., 2003), which includes variation across different contexts, situations, or activities.

Therefore, in this section, we demonstrate ways in which several contextual factors may be associated with response behavior. The question of the extent to which behavior or experiences are stable across different contexts is reminiscent of the well-established person-situation debate in personality psychology (Beck & Jackson, 2022). If EMA data are to advance understanding of such questions, items that measure constructs that likely vary across different contexts should indeed show such variation in empirical data. Which items researchers may expect to vary to which degree across different contexts is necessarily dependent on the population and constructs they are studying. For example, researchers may expect different levels of variability of psychopathological symptoms in a clinical sample compared with a student sample. Here, we aggregate across time and individuals and present two example items that illustrate different levels of stability across contexts.

Item response across activities. In Figure 3, we present the distribution of the answer options for the items cheerful, depressed, motivated, and irritable across different activities that participants reported when answering

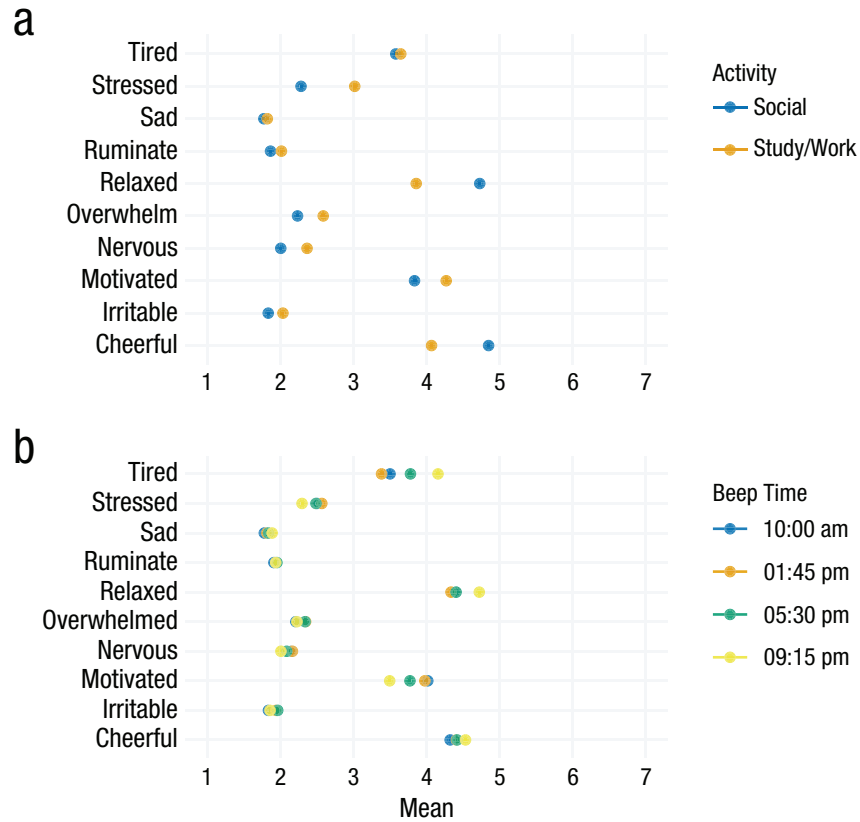


Fig. 4. Activity and time-of-day effects. (a) Item means across all participants when participants either indicated that they were in a “social” activity or a “study/work” activity. (b) Item means across the four daily beeps/notifications. The hours of the day indicated in the legend for Fig. 4b are approximate because the exact timing of beeps included some planned randomness (Fried et al., 2023). Horizontal lines around the points would indicate the standard error of the estimated means, but they are barely visible here because of our sample size.

prompts. The relative frequency (in percentages) of an answer option is displayed on the x -axis. For example, when individuals are in a social activity, they indicate that they are very cheerful (dark red) more often than during other activities. Likewise, participants indicate higher levels of motivation during study/work activities than during leisure. Such pronounced effects do not occur for depressed and irritable, for which responses seem to be more stable across contexts.

The difference in stability across contexts fits the different nature of the items. Whereas the depressed item was collected in the evening only and inquires about a summary of the whole day, the item cheerful aims to capture momentary affect (“How cheerful are you right now?”) and should be more responsive to different contexts given that it was collected four times a day. From a statistical perspective, we can learn that the inclusion of contextual variables may provide relevant additional information when modeling or predicting certain constructs that fluctuate considerably across contexts.

Furthermore, floor effects in items such as irritable impede their informativeness on an aggregate level because they show little insight into interindividual differences, and these items are potentially harder to include in typical statistical models. We take a deeper look comparing the two example activities study/work versus social activities in Figure 4a. We can again see that for some items, such as relaxed and cheerful, responses differ quite strongly across these activities, whereas responses for tired or sad are very similar for these different activities.

Time-of-day and weekday effects. The time when people answer surveys provides important situational context that may give rise to recurring patterns in EMA data that can be of interest for various research questions (e.g., Helliwell & Wang, 2014; Smith et al., 2018). Time-of-day effects, such as lower motivation in the evenings, can be indicative of natural circadian rhythms or external influences, such as work or university courses, and the daily

experiences of participants on a weekday compared with the weekend may be markedly different (Koudela-Hamila et al., 2019). This, in turn, can have implications for how to best measure EMA items because researchers might miss important information by infrequent sampling. Likewise, too-frequent sampling can also lead to a loss of information. In Cohort 1 of WARN-D, we queried participants about daily drug use in the evening prompt around 9:30 p.m., which likely misses some behavior later in the night. In Cohort 2, we adapted a less frequent but more functional sampling schedule and asked the item retrospectively on the weekend for the whole week.

We find the most pronounced time-of-day effects for the item tired, which increases by about 0.77 points on the 7-point Likert scale from around noon compared with the evening. Other items, such as sad or ruminate, stay mostly stable during the day.

Regarding weekend effects, the largest differences between weekdays and weekends are found for the items useful (0.26 points lower on weekends), relaxed (0.24 points higher on weekends), and stressed (0.24 points lower on weekends). In contrast, responses to items tired, sad, or irritable appear to be very similar on the weekend. The full report includes our calculations, more information on this, and visualizations for time-of-day/beep and weekday effects in Section F7. By “beep effects,” we mean potential systematic differences in responses across the different surveys per day or, in other words, across the time of day. Furthermore, we present additional contextual analyses of EMA items based on specific activities, the rating of negative events, and other variables in Section F6.

Further reading. Mestdagh and Dejonckheere (2021) discussed shortcomings of previous research and ways forward for the investigation of contextual influences in EMA. Beyond self-reported context, the abundance of GPS data has allowed the study of the influence of environmental context on mood measures (de Vries et al., 2021). Langener et al. (2023) reviewed how the social environment, including activity, is captured in experience-sampling and passive-sensing studies. Recent methodological innovations extend dynamical structural equation modeling to be able to include person-situation interactions (Castro-Alvarez et al., 2022). Adolf et al. (2017) showed how context-related changes in parameter estimates can be modeled for time-series analyses. Beck and Jackson (2022) displayed how information about situations can be leveraged for individual prediction of behaviors and experiences and discussed the relevance of situation information for personality research, and Cloos et al. (2022) used the sensitivity to emotionally relevant events as a quality criterion for EMA items. Both weekend effects and time-of-day effects have been studied across a range of

constructs (e.g., for suicide and mood, see Freichel and O’Shea, 2023; for positive affect, see Egloff et al., 1995). Gabriel et al. (2019) offered a discussion on time trends and how to model them, and Zhang and Volkow (2023) discussed the role of seasonality and circadian rhythms in psychiatric disorders.

Across time

Changes over time. Change over time is one of the main reasons why collecting longitudinal data is of interest to researchers. This includes, for example, intervention research; research into biological rhythms, such as menstrual cycles (Beddig et al., 2020); and research following participants during periods of stressful events, such as students over a semester with multiple exam periods. But other factors, such as shifting response styles of participants, may also cause observed changes. For example, initial elevation bias has been observed in multiple intensive longitudinal studies, meaning that individuals report higher state values at the beginning of data collection (Shrout et al., 2018). Yet summary statistics, such as those presented in the first two sections of this article, give very limited insight into stability or change over time. Potential changes are also important because time-series models commonly assume no (lasting) changes in means or variances over time (Ryan et al., 2023). In the following, we present tools that help highlight stability and changes over time. We first take a look at aggregate changes of multiple items throughout data collection, followed by the temporal stability of a single item at the group level. We then highlight interindividual differences across people researchers may miss when using group-level approaches.

Investigating item changes at the group level. To visualize how the daily averages of responses change over time, we visualize the average daily EMA responses over time (solid lines) and their overall mean across all time points (dotted line) in Figure 5. For both cohorts, we show that the reports of stressed were above the overall mean at the beginning of data collection and that the reverse was true for happy and relaxed. There are several explanations for this. For example, for Cohort 1, we know that there was an intense exam period until around Day 20. We further observed a clear periodicity in which higher negative affect and lower positive affect were being reported on Mondays or, in other words, at the vertical grid lines. Because WARN-D follows a sample from a general student population and does not include any intervention, overall mean changes over time were generally small. Using such a visualization may bring additional insight when studying distinct groups, for example, diagnostic conditions or treatment groups. In these cases, distinct change patterns between groups can often be expected, which is easy to see in plots.

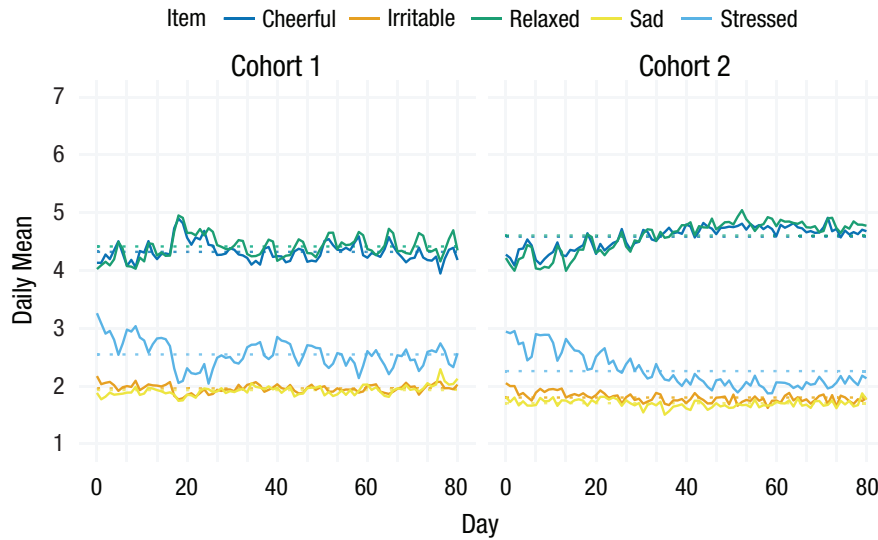


Fig. 5. Group-level changes of daily means of example items. This figure shows daily means for five example items, split by Cohort 1 and 2. Vertical grid lines represent Mondays. Solid lines indicate average daily ecological-momentary-assessment responses across all individuals. Dotted lines indicate the overall items' means across all time points and individuals.

Investigating item stability at the individual level. Beyond the aggregate change of mean levels of items over time, the stability or fluctuation within individual items over time is of central relevance for most EMA research. One of the most common indicators of temporal dependencies in time series is the autocorrelation. For a lag size of 1, the autocorrelation quantifies the linear dependency between the current and the previous time point. Autocorrelations are often used to infer the optimal lag size in time-series modeling. Beyond modeling decisions, the inspection of autocorrelations can help understand time dependence in data. Broadly, autocorrelations indicate the extent to which states persist over time and are resistant to change (Kuppens et al., 2010). Such effects are often called “inertia” and also play a role in the literature on identifying early warning signals for mental disorders (Helmich et al., 2021). The concept of long-range autocorrelations has also been studied as the “memory” of a time series in complexity science, indicating that the current state of a time series may be dependent on its state a long time ago (Olthof et al., 2020).

In Figure 6, we plot all individual autocorrelations for the item stressed across nine different lags, that is, beeps, while excluding overnight effects. The number 9 was selected here for ease of plotting. Depending on the specific item and context, a smaller or larger range of lags may be of interest. Each point reflects the autocorrelation of one individual. Note that we detrend the data by removing a linear trend of time.

The plot in Figure 6 indicates substantial heterogeneity across people at every lag size. Still, the association with the previous time point is the strongest overall. In

addition, autocorrelations seem to be elevated around Lags 4 and 5, which indicates some stability of the item stressed (measured 4 times a day) across roughly a 24-hr interval. Other items not shown here show slightly different patterns. For example, autocorrelations of the item cheerful are generally slightly lower compared with the item stressed, and they decrease somewhat faster with higher lag size. Theoretically, this informs about the tendency of certain item responses to persist over time and the interindividual differences therein. For modeling, significantly elevated autocorrelations beyond Lag 1 can indicate that for certain purposes and individuals, researchers should consider using models that take into account more than the first lag typical for psychological research.

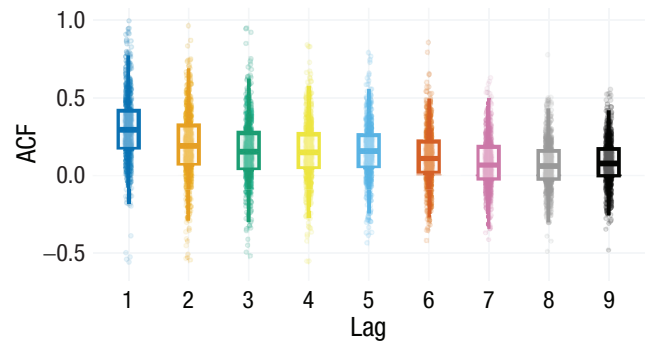


Fig. 6. Autocorrelations of the item stressed. The individual data points in this figure represent individual autocorrelation function (ACF) values. A linear trend was detrended before calculation, and overnight effects were not included. The ACF was calculated only for individuals with more than 100 data points.

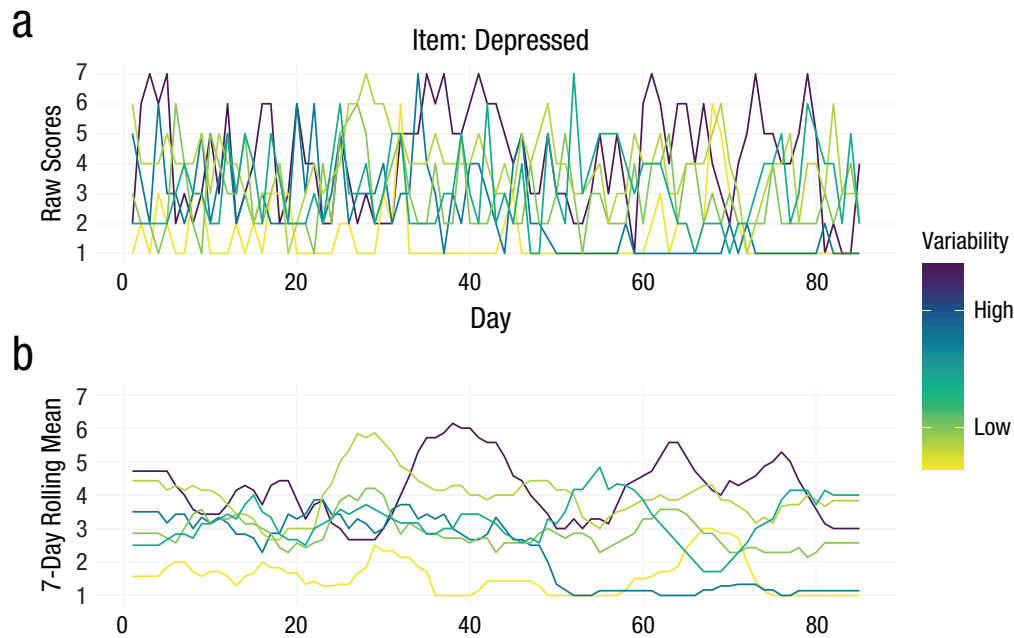


Fig. 7. Development of the item depressed for six participants. (a) Raw values of item responses. (b) Seven-day rolling means are visualized as an intuitive example of an aggregation window spanning a week. The x -axis denotes the day of the study, and the y -axis denotes the response value. Time series with higher variability (operationalized as standard deviation) are colored in darker colors, and time series with lower variability are in lighter colors.

Investigating item changes at the individual level. Plotting time-series data can inform about the nature and degree of changes over time. However, with many time points, variables, and individuals, such visualizations are often quite messy. This can make it hard to see certain changes over time and features of the individual response distribution. We therefore zoom into six participants for the remainder of this section.

Figure 7a shows the raw time-series data for the item depressed of these six participants. In this figure, it is very hard to distinguish individual lines and to recognize overall trend patterns. Figure 7b depicts the same data using a moving-window technique (Fig. 7a) in which (in this case) 7 days of data are averaged for each time point (e.g., Time Point 40 includes the average of Time Points 37–43; Time Point 41 includes the average of Time Points 38–44). Such visualizations can be flexibly applied in varying window sizes to summarize different statistics, such as the mean or variance, to simplify the visual interpretation of patterns over time. We additionally color the time series based on their variability over time such that time series with a higher fluctuation are shaded in darker colors and time series with less variability are colored in lighter colors. Overall, we show that the responses to depressed fluctuate highly and change abruptly for some participants, while they are more stable or gradually changing for others. This interindividual heterogeneity in change patterns is not directly

apparent from the other analyses in this article because they often condense the time-series information into a single summary statistic. This could imply, for example, that the stability of depressive symptoms may vary between individuals and that symptoms may change faster than what is often assumed (Fried et al., 2022). Statistically, it may be relevant to use methods that can account for or explicitly model such changes over time, which we sketch in the “Further reading” section.

Figure 8 illustrates how time series for one item (at the individual or group level) can be made more informative by displaying the marginal distribution of the item on the right side of the plot. Here, we show that responses of one individual to the items sad and stressed tend to fall in either the middle or the extreme values of the scale, which is not easily visible from the long time-series visualization alone.³ Creating individual time-series plots, such as in Figures 7 and 8, can be challenging in large samples. In such cases, we recommend looking into a random subsample of participants to obtain insight into potential patterns that may warrant further investigation.

Further reading. The analysis of lagged dependencies is explained in many standard time-series textbooks, such as Box et al. (2008). We chose the simple approach of calculating autocorrelation in a univariate and basic way, but other approaches allow one to handle the possible

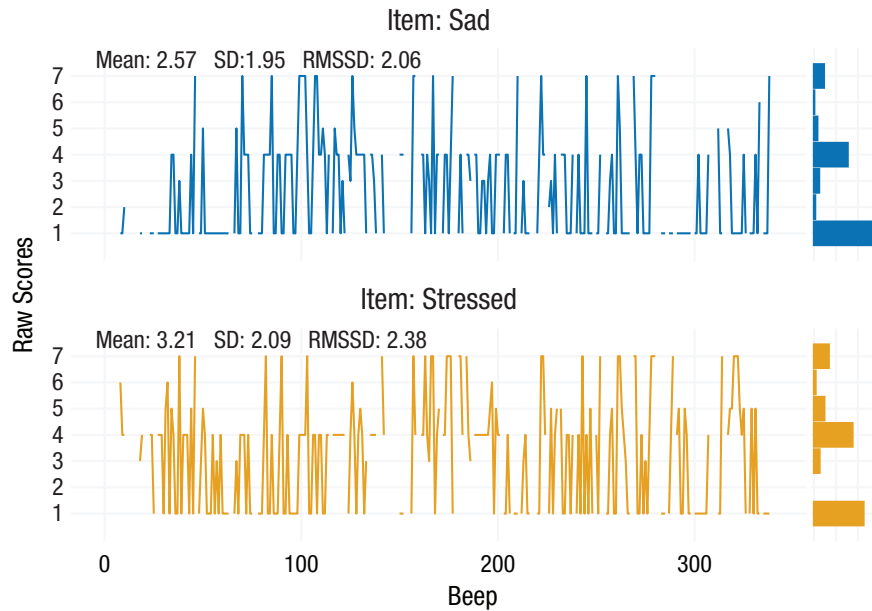


Fig. 8. Raw time series with marginal distribution. The figure shows time series for items sad and stressed for an example participant. The x -axis denotes the day of the study, and the y -axis denotes the response value.

nonstationarity of the data and the influence of other variables on a given item. Concerning the former, moving-window calculations of autocorrelations can also be used to deal with nonstationarity (see e.g., Olthof et al., 2020). Inertia is also often operationalized as an autoregressive effect in multivariable statistical models, such as multilevel models (Hamaker & Wichers, 2017; Jongerling et al., 2015), in which the influence of other variables and inter-individual differences can be modeled. The analysis of autocorrelations, as shown here, can be the first step in understanding the lag structure. Beyond that, for example, Jacobson et al. (2019) developed a tool for exploratory and confirmatory lag diagnostics in typical psychological time-series data. Change-point detection techniques also commonly use moving-window approaches (see e.g., Cabrieto et al., 2018). To explicitly model changes over time in intensive longitudinal data, readers may refer to Boker et al. (2002) for an example of changing associations between two time series or Haslbeck et al. (2021) for more complex multivariate methods.

Disentangling variability sources

Up to this point, we have looked at sources of variability, such as changes over time and contextual factors, mostly separately. Now we can try to synthesize them and try to understand where most of the variance in our data set comes from. This helps us answer questions such as the following: Is there more between-days variability than within-days variability? Are there strong changes in

responses over time that are consistent across individuals? And what is the ratio of within-individuals variability to between-individuals variability? This last question occurs regularly in the context of multilevel modeling and is commonly answered by calculating the intraclass correlation (ICC), which quantifies the proportion of variance because of stable between-persons differences (Hamaker, 2024). Expanding on this perspective, generalizability theory (Cranford et al., 2006; Schönbrodt et al., 2022) is concerned with decomposing variance at multiple hierarchical structures, such as beeps, days, individuals, or items in the data set. This can help researchers obtain a first insight into the structure of variation in their data. Here, we investigate the influence of time and inter-individual variability for an example item.

First, we estimate the ICC for every Likert-scaled EMA item in the data set. For more details about its calculation, see the full report. The mean of ICCs across items is 0.32 ($SD = 0.05$); useful (evening item: “Today, I felt productive/useful”) has the lowest ICC of 0.22, and ruminate (four times per day item: “I am experiencing negative thoughts right now”) has the highest one with 0.42. This means that more than half of the variance for all items appears to occur at the within-persons level (but note that within-persons variability can be overestimated in the presence of measurement error; Wilms et al., 2020). For more detailed results, see the full report.

We can extend this and decompose the variance of moments/beeps nested in days nested in individuals in

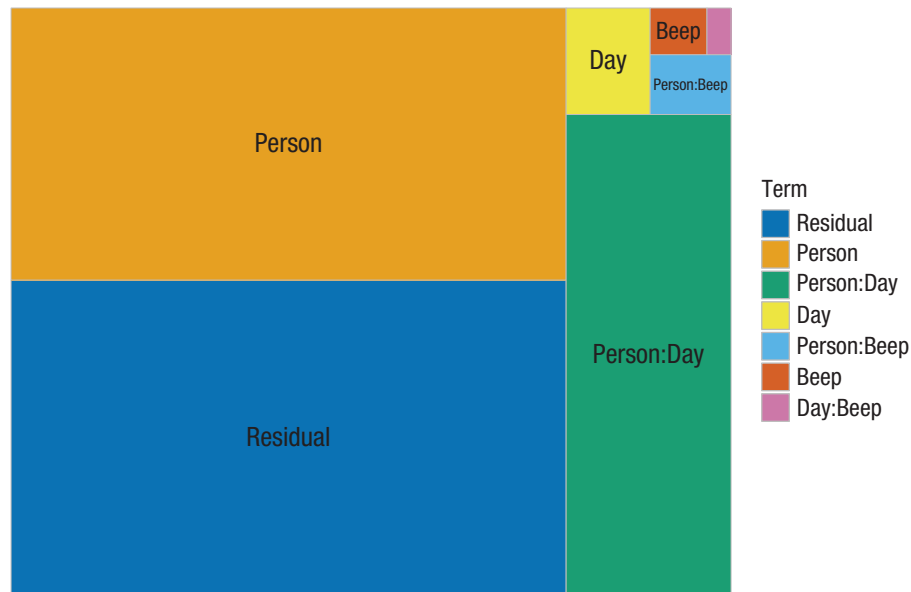


Fig. 9. Variance decomposition for the item stressed. The figure illustrates different variance components. The proportion of variance components is represented by their tile size.

our data set. The following description is, in parts, adapted from Schönbrodt et al. (2022). Technically, we estimate an intercept-only multilevel model including a random intercept variance for all factors and allocate the variance of item responses across multiple levels and factors without assuming linear relationships. This is summarized in Figure 9, which shows the proportion of variance components across these factors. The size of a tile represents the relative proportion of total variance that is attributable to a certain factor. “Person” contains between-persons differences, “Person:Day” contains variance between days (in which each day of each person is a unique element), “Person:Prompt” contains time-of-day effects for some individuals, “Prompt” contains general time-of-day effects, “Day” contains general effects of study day, “Day:Prompt” contains effects of certain prompts on certain days, and the residual contains the remaining residual variance left after all other factors have been taken into account. For code for this plot, see Section F7 in the full report.

The variance decomposition in Figure 9 shows the large extent of residual variance, which here represents within-persons variation in the item stressed that is not explained by the other factors. A substantial amount of variance can also be attributed to between-persons differences, as indicated by the size of Person in the plot. The interactions of Person with beep and day show between-individuals variability in time-of-day and day-of-study effects, respectively. The latter can be interpreted as different trajectories of stress between different participants. The general effect of the day of the study is small, which is to be expected given that we follow

students during their normal lives and do not expect particularly impactful events on any specific day. In addition, our sample comprised two cohorts in which a given day of the study refers to different dates. Note that the decomposition here builds on classical test theory and thereby comes with some strong assumptions, including the absence of autoregressive effects or the independence of different variance components (Schönbrodt et al., 2022). We sketch some potential solutions for these shortcomings in the “Further reading” section below.

Further reading. Various reliability indices could be computed from variance decomposition analyses (Schönbrodt et al., 2022). How to incorporate autocorrelation structures into variance decomposition for EMA is shown in Vansteelandt and Verbeke (2016). A more flexible approach to calculate reliability in longitudinal data with multiple indicators per construct is the use of dynamic-factor models (Fuller-Tyszkiewicz et al., 2017), and Schuurman et al. (2015) showed how to calculate between- and within-persons reliability with single-indicator measures. General recommendations regarding reliability in single and multiitem EMA measures are provided in Vogelsmeier et al. (2023).

Measuring different time scales

One additional source to better understand EMA items is to look at how they are associated with other features in the data, such as EMA items assessed weekly or psychometric scales assessed at baseline before the EMA

		Baseline				
		PHQ-9	GAD-7	PSS-10	SCOFF	ASRM
EMA	Anhedonia	0.52	0.4	0.48	0.16	0.05
	Irritable	0.38	0.33	0.46	0.14	0.08
	Depressed	0.46	0.43	0.49	0.19	0.07
	Emo_reg	0.44	0.45	0.48	0.2	0.09
	Ruminate	0.4	0.39	0.42	0.2	0.04
	Nervous	0.45	0.48	0.5	0.24	0.11
	Overwhelmed	0.39	0.4	0.47	0.18	0.12
	Sad	0.4	0.38	0.46	0.16	0.07
	Stressed	0.38	0.41	0.47	0.17	0.04
	Tired	0.37	0.37	0.35	0.17	-0.07
	Concentrate	-0.45	-0.33	-0.47	-0.14	0.04
	Connected	-0.38	-0.19	-0.33	-0.18	0.04
	Coping	-0.43	-0.31	-0.49	-0.15	0.05
	Useful	-0.4	-0.23	-0.39	-0.13	0.08
	Cheerful	-0.4	-0.25	-0.37	-0.17	0.12
	Motivated	-0.31	-0.21	-0.3	-0.11	0.1
	Outlook	-0.43	-0.29	-0.39	-0.18	0.13
	Relaxed	-0.38	-0.32	-0.41	-0.17	0.09

Fig. 10. Correlation of ecological momentary assessment (EMA) with baseline. This figure shows the correlation of EMA means with baseline items. The y -axis contains the mean of different EMA variables, aggregated to individual means across the whole study duration. The x -axis contains different baseline questionnaires. Emo_reg refers to the item, “Today, it was difficult to cope with my emotions.” For more information about all items and their phrasing, see the supplementary material on OSF.

stage. This allows researchers to better understand the association between dynamic momentary items and slower moving or trait-like measures. In traditional validation work, such analyses may be used to establish predictive (How well does a daily item predict later measures?) or concurrent (How well does a daily item overlap with less frequent assessments of the same construct?) validity (M. S. Allen et al., 2022). Here, we aggregate across time and individuals, calculating between-persons correlations for multiple EMA items and baseline questionnaires.

EMA and baseline items. Figure 10 depicts these correlations, with a focus on questionnaires assessing mental-health-related constructs, shown on the x -axis, including the Altman Self-Rating Mania Scale (ASRM; Altman et al., 1997), assessing manic symptoms; the Generalized Anxiety

Disorder (GAD-7; Spitzer et al., 2006) scale, assessing anxiety-disorder symptoms; an adapted form of the Patient Health Questionnaire-9 (PHQ-9; Kroenke et al., 2001), assessing depressive symptoms; and the 10-item version of the Perceived Stress Scale (PSS-10; Cohen & Williamson, 1988), and the SCOFF (acronym explained in Morgan et al., 1999) assessing eating disorder symptoms. We use sum-scoring for these questionnaires to obtain a single score per individual on the x -axis. On the y -axis, we include the mean of multiple EMA items over the whole data-collection period.

Correlations between EMA means and the depression (PHQ-9), anxiety (GAD-7), and stress (PSS-10) questionnaires appear fairly similar, indicating that EMA means do not differentiate well between the baseline questionnaires. Most items are only weakly connected to the eating-disorder (SCOFF) and mania (ASRM) questionnaires.

Daily and weekly items. We can also look into the relations of EMA items collected at different frequencies, such as multiple times per day, daily, or weekly. Specifically, here we investigate how validated retrospective reports over a longer period relate to aggregates of daily reports, such as comparing a validated scale for a construct assessed on the weekend (“Last week, I felt . . .”) with daily scores obtained every evening of the week (“Today, I felt . . .”). If items with a higher sampling frequency deliver additional information, this points to the added value of intensive daily versus less frequent assessments and can inform the frequency of assessment in future studies. In the WARN-D data, we investigate this question by looking at the association of the weekly means of a daily depression question (“Today, I felt down or depressed”) with various individual depressive symptoms assessed retrospectively once a week with the PHQ-9 (“Last week, I felt . . .”). We do not visualize these calculations here but rather show more detailed quantitative results in the full report in Section F9.

Overall, the correlations range from .17 to .57. Mean daily depression (averaged across the week) and a weekly retrospective depression item (“Feeling down or depressed”) correlated with .57 (95% confidence interval [CI] = [.55, .59]); the lowest correlation obtained was .17 between mean daily depression (across the week) and both motor retardation and overeating (95% CI = [.14, .20] for both). The correlation of the mean daily depression item with the sum score of the weekly PHQ-9 (Kroenke et al., 2001) was .56 (95% CI = [.54, .58]).⁴ In summary, the weekly measurement appears to (at least on average in this group of people) approximate the daily scores fairly well and certainly better than we had anticipated. In the end, the decision at what frequency to sample will depend on many other factors as well, including the specific research question.

Further reading. Recent work has compared momentary versus recalled affect (Greene et al., 2022; Leertouwer et al., 2022) and daily versus biweekly depressive symptoms (Horwitz et al., 2023). For example ways to study the predictive and concurrent validity for single-item measures, see Song et al. (2023). A spectrum of techniques that, among other advantages, allow modeling using items assessed at different sampling frequencies is subsumed under continuous-time modeling (van Montfort et al., 2018).

Discussion

EMA data are commonly collected across a variety of disciplines, and there are many rich data sets available for psychological researchers to analyze. However, EMA data tend to be quite complex, and in our tutorial, we aim to provide ideas on how researchers can better understand such data. To introduce and demonstrate

useful analytic and data-visualization tools, we have presented the properties and performance of various items in the WARN-D study. Most of these insights would not have been gained by applying only typical time-series models that capture multivariate associations between variables. We have emphasized the relevance of the information researchers can gain for both theoretical and statistical considerations. We hope that this tutorial and the associated R code will help researchers explore items in their EMA data in more detail but also share the results of such explorations in their articles or supplementary materials to build a basis for establishing robust phenomena. Making full use of all the information hiding in EMA data sets by employing the full breadth of available data-visualization and modeling tools—and sharing the results (possibly in longer online supplements, such as the full report that we provided)—will not only benefit the quality and transparency of analyses but also improve the understanding of the individuals and constructs under study. We believe that these tools can be helpful in a variety of contexts, from pilot studies with new questionnaires to the description and analysis of large samples with well-known items.

In our application of these techniques for the WARN-D data, we uncovered threats to the integrity of potential modeling strategies because of, for example, floor effects and bimodality. At the same time, we also learned more about the variability and possible contextual influences on the responses of participants. We have gained further insights into the potential added value of EMA above and beyond cross-sectional measures, for example, by investigating the change of item responses over time or assessing the overlap between EMA responses and typical trait-like questionnaires. Finally, we hope that sharing these results of our EMA data in the full report will help other researchers build on our EMA item battery and choose or modify items, for example, to achieve more desirable distributions.

We also note that this tutorial is aimed to provide a first step for researchers to better understand their EMA data. We see our focus on individual items and descriptive statistics as an important precursor to tackling more complex challenges, three of which we highlight below. First, we have not yet discussed multiitem constructs and measurement models, which are the default in psychometrics for cross-sectional data. Such models assume that each item of a scale measuring a purported construct, such as depression, neuroticism, or mathematical ability, is error-prone and that measuring multiple related items and then estimating the construct based on the shared variance of items improves measurement precision. For EMA data, there is a fast-growing literature on measurement models that can capture interindividual differences (McNeish et al., 2021), changes over time (Vogelsmeier et al., 2019), and nonlinear and state-switching properties of measurement models (Kelava &

Brandt, 2019). Second, readers may have noticed that we used terms such as item “functioning” and “performance” and refrained from terms such as “validity” or “validation.” Existing validity frameworks, such as the Standards for Educational and Psychological Testing (American Educational Research Association, 2014), do provide crucial opportunities in understanding psychological constructs (see e.g., Fried et al., 2022), but how to apply them to EMA data is not quite established. Furthermore, and related to our first point, validity terminology in psychology is traditionally associated with the evaluation of measurement models in the context of multiple indicators for a construct instead of focusing on individual items. New methods are emerging for marrying EMA data and notions of validity, such as investigating single-item reliability (Dejonckheere et al., 2022; Schuurman et al., 2015) and validity (M. S. Allen et al., 2022) in longitudinal studies. Third, establishing how our items perform in data is helpful to guide questions into an underappreciated aspect of the question of validity: understanding the response processes, that is, the cognitive processes underlying people’s item responses. Choosing the highest answer category in the item sad, for example, may have substantively different meanings across individuals and within individuals over time. What response processes underlie EMA item answers is largely unknown and can be tackled by conducting cognitive interviews in which participants are queried to make explicit how and why they answer items in the way they answer them (Stone et al., 2023). Understanding this process is crucial to understanding if the constructs researchers use actually measure what they intend to measure (Borsboom et al., 2004).

Transparency

Action Editor: Katie Corker

Editor: David A. Sbarra

Author Contributions

Björn S. Siepe: Conceptualization; Formal analysis; Methodology; Software; Visualization; Writing – original draft; Writing – review & editing.

Carlotta L. Rieble: Data curation; Investigation; Project administration; Validation; Writing – review & editing.

Rayyan Tutunji: Data curation; Investigation; Project administration; Validation; Writing – review & editing.

Aljoscha Rimpler: Data curation; Investigation; Writing – review & editing.

Julius März: Data curation; Investigation; Writing – review & editing.

Ricarda K. K. Proppert: Data curation; Investigation; Project administration; Validation; Writing – review & editing.

Eiko I. Fried: Conceptualization; Data curation; Funding acquisition; Investigation; Methodology; Project administration; Supervision; Visualization; Writing – original draft; Writing – review & editing.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

C. L. Rieble, R. Tutunji, R. K. K. Proppert, and E. I. Fried are supported by funding from the European Research Council under the European Union’s Horizon 2020 research and innovation program (Grant 949059).


Open Practices

This article has received the badge for Open Materials. More information about the Open Practices badges can be found at <http://www.psychologicalscience.org/publications/badges>.



ORCID iDs

Björn S. Siepe  <https://orcid.org/0000-0002-9558-4648>

Eiko I. Fried  <https://orcid.org/0000-0001-7469-594X>

Notes

1. The version of the supplementary files associated with this article is available in the folder “Version 1.1.”
2. This cutoff is arbitrary; lower proportions of choosing the lowest response may also pose difficulties for statistical models.
3. These figures were inspired by Haslbeck et al. (2023).
4. We calculated the PHQ-9 sum score by retransforming the relevant PHQ-15 items into the PHQ-9 by taking the higher value (e.g., the higher score of the two disaggregated PHQ-15 sleep symptoms “insomnia” and “hypersomnia” was used to reconstruct the PHQ-9 “insomnia or hypersomnia” item).

References

- Adolf, J. K., Voelkle, M. C., Brose, A., & Schmiedek, F. (2017). Capturing context-related change in emotional dynamics via fixed moderated time series analysis. *Multivariate Behavioral Research, 52*(4), 499–531. <https://doi.org/10.1080/00273171.2017.1321978>
- Allen, M., Poggiali, D., Whitaker, K., Marshall, T. R., van Langen, J., & Kievit, R. A. (2021). Raincloud plots: A multi-platform tool for robust data visualization. *Wellcome Open Research, 4*, Article 63. <https://doi.org/10.12688/wellcomeopenres.15191.2>
- Allen, M. S., Iliescu, D., & Greiff, S. (2022). Single item measures in psychological science. *European Journal of Psychological Assessment, 38*(1), 1–5. <https://doi.org/10.1027/1015-5759/a000699>
- Altman, E. G., Hedeker, D., Peterson, J. L., & Davis, J. M. (1997). The Altman self-rating mania scale. *Biological Psychiatry, 42*(10), 948–955.
- American Educational Research Association. (Ed.). (2014). *Standards for educational and psychological testing*.
- Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician, 27*(1), 17–21.

- Beck, E. D., & Jackson, J. J. (2022). Personalized prediction of behaviors and experiences: An idiographic person-situation test. *Psychological Science*, *33*(10), 1767–1782. <https://doi.org/10.1177/09567976221093307>
- Beddig, T., Reinhard, I., Ebner-Priemer, U., & Kuehner, C. (2020). Reciprocal effects between cognitive and affective states in women with premenstrual dysphoric disorder: An ecological momentary assessment study. *Behaviour Research and Therapy*, *131*, Article 103613. <https://doi.org/10.1016/j.brat.2020.103613>
- Boker, S. M., Rotondo, J. L., Xu, M., & King, K. (2002). Windowed cross-correlation and peak picking for the analysis of variability in the association between behavioral time series. *Psychological Methods*, *7*(3), 338–355. <https://doi.org/10.1037/1082-989X.7.3.338>
- Bolger, N., Davis, A., & Rafaeli, E. (2003). Diary methods: Capturing life as it is lived. *Annual Review of Psychology*, *54*, 579–616. <https://doi.org/10.1146/annurev.psych.54.101601.145030>
- Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*, 1061–1071. <https://doi.org/10.1037/0033-295X.111.4.1061>
- Box, G. E., Jenkins, G. M., & Reinsel, G. C. (2008). *Time series analysis* (4th ed.). John Wiley & Sons. <https://doi.org/10.1002/9781118619193>
- Bringmann, L. F., van der Veen, D. C., Wichers, M., Riese, H., & Stulp, G. (2020). ESMvis: A tool for visualizing individual Experience Sampling Method (ESM) data. *Quality of Life Research*, *30*, 3179–3188. <https://doi.org/10.1007/s11136-020-02701-4>
- Cabrieto, J., Tuerlinckx, F., Kuppens, P., Hunyadi, B., & Ceulemans, E. (2018). Testing for the presence of correlation changes in a multivariate time series: A permutation based approach. *Scientific Reports*, *8*, Article 769. <https://doi.org/10.1038/s41598-017-19067-2>
- Castro-Alvarez, S., Tendeiro, J. N., de Jonge, P., Meijer, R. R., & Bringmann, L. F. (2022). Mixed-effects trait-state-occasion model: Studying the psychometric properties and the person-situation interactions of psychological dynamics. *Structural Equation Modeling: A Multidisciplinary Journal*, *29*(3), 438–451. <https://doi.org/10.1080/10705511.2021.1961587>
- Chang, H. (2004). *Inventing temperature: Measurement and scientific progress*. Oxford University Press.
- Cloos, L., Kuppens, P., & Ceulemans, E. (2022). *Development, validation, and comparison of self-report measures for positive and negative affect in intensive longitudinal research*. PsyArXiv. <https://doi.org/10.31234/osf.io/5j7c6>
- Cohen, S., & Williamson, G. (1988). Perceived stress in a probability sample of the United States. In S. Spacapan & S. Oskamp (Eds.), *The social psychology of health* (pp. 31–67). Sage.
- Cranford, J. A., Shrout, P. E., Iida, M., Rafaeli, E., Yip, T., & Bolger, N. (2006). A procedure for evaluating sensitivity to within-person change: Can mood measures in diary studies detect change reliably? *Personality & Social Psychology Bulletin*, *32*(7), 917–929. <https://doi.org/10.1177/0146167206287721>
- Cui, J., Hasselman, F., & Lichtwarck-Aschoff, A. (2023). Unlocking nonlinear dynamics and multistability from intensive longitudinal data: A novel method. *Psychological Methods*. Advance online publication. <https://doi.org/10.1037/met0000623>
- de Vries, S., Nieuwenhuizen, W., Farjon, H., van Hinsberg, A., & Dirckx, J. (2021). In which natural environments are people happiest? Large-scale experience sampling in the Netherlands. *Landscape and Urban Planning*, *205*, Article 103972. <https://doi.org/10.1016/j.landurbplan.2020.103972>
- Dejonckheere, E., Demeyer, F., Geusens, B., Piot, M., Tuerlinckx, F., Verdonck, S., & Mestdagh, M. (2022). Assessing the reliability of single-item momentary affective measurements in experience sampling. *Psychological Assessment*, *34*(12), 1138–1154. <https://doi.org/10.1037/pas0001178>
- Egloff, B., Tausch, A., Kohlmann, C.-W., & Krohne, H. W. (1995). Relationships between time of day, day of the week, and positive mood: Exploring the role of the mood measure. *Motivation and Emotion*, *19*(2), 99–110. <https://doi.org/10.1007/BF02250565>
- Freichel, R., & O'Shea, B. A. (2023). Suicidality and mood: The impact of trends, seasons, day of the week, and time of day on explicit and implicit cognitions among an online community sample. *Translational Psychiatry*, *13*, Article 157. <https://doi.org/10.1038/s41398-023-02434-1>
- Fried, E. I., Flake, J. K., & Robinaugh, D. J. (2022). Revisiting the theoretical and methodological foundations of depression measurement. *Nature Reviews Psychology*, *1*(6), 358–368. <https://doi.org/10.1038/s44159-022-00050-2>
- Fried, E. I., Proppert, R. K. K., & Rieble, C. L. (2023). Building an early warning system for depression: Rationale, objectives, and methods of the WARN-D study. *Clinical Psychology in Europe*, *5*(3), 1–25. <https://doi.org/10.32872/cpe.10075>
- Fuller-Tyszkiewicz, M., Hartley-Clark, L., Cummins, R. A., Tomin, A. J., Weinberg, M. K., & Richardson, B. (2017). Using dynamic factor analysis to provide insights into data reliability in experience sampling studies. *Psychological Assessment*, *29*(9), 1120–1128. <https://doi.org/10.1037/pas0000411>
- Gabriel, A. S., Podsakoff, N. P., Beal, D. J., Scott, B. A., Sonnentag, S., Trougakos, J. P., & Butts, M. M. (2019). Experience sampling methods: A discussion of critical trends and considerations for scholarly advancement. *Organizational Research Methods*, *22*(4), 969–1006. <https://doi.org/10.1177/1094428118802626>
- Greene, T., Sznitman, S., Contractor, A. A., Prakash, K., Fried, E. I., & Gelkopf, M. (2022). The memory-experience gap for PTSD symptoms: The correspondence between experience sampling and past month retrospective reports of traumatic stress symptoms. *Psychiatry Research*, *307*, Article 114315. <https://doi.org/10.1016/j.psychres.2021.114315>
- Haig, B. D. (2013). Detecting psychological phenomena: Taking bottom-up research seriously. *The American Journal of Psychology*, *126*(2), 135–153. <https://doi.org/10.5406/amerjpsyc.126.2.0135>
- Hall, M., Scherner, P. V., Kreidel, Y., & Rubel, J. A. (2021). A systematic review of momentary assessment designs for mood and anxiety symptoms. *Frontiers in Psychology*, *12*, Article 1716. <https://doi.org/10.3389/fpsyg.2021.642044>
- Hamaker, E. L. (2024). The curious case of the cross-sectional correlation. *Multivariate Behavioral Research*, *59*, 1111–1122. <https://doi.org/10.1080/00273171.2022.2155930>

- Hamaker, E. L., Grasman, R. P. P., & Kamphuis, J. H. (2010). Regime-switching models to study psychological processes. In P. C. M. Molenaar & K. M. Newell (Eds.), *Individual pathways of change: Statistical models for analyzing learning and development* (pp. 155–168). American Psychological Association. <https://doi.org/10.1037/12140-009>
- Hamaker, E. L., & Wichers, M. (2017). No time like the present: Discovering the hidden dynamics in intensive longitudinal data. *Current Directions in Psychological Science*, *26*(1), 10–15. <https://doi.org/10.1177/0963721416666518>
- Haqiqatkah, M. M., Ryan, O., & Hamaker, E. L. (2024). Skewness and staging: Does the floor effect induce bias in multilevel AR(1) models? *Multivariate Behavioral Research*, *59*(2), 289–319. <https://doi.org/10.1080/00273171.2023.2254769>
- Haslbeck, J. M. B., Bringmann, L. F., & Waldorp, L. J. (2021). A tutorial on estimating time-varying vector autoregressive models. *Multivariate Behavioral Research*, *56*(1), 120–149. <https://doi.org/10.1080/00273171.2020.1743630>
- Haslbeck, J. M. B., & Ryan, O. (2022). Recovering within-person dynamics from psychological time series. *Multivariate Behavioral Research*, *57*(5), 735–766. <https://doi.org/10.1080/00273171.2021.1896353>
- Haslbeck, J. M. B., Ryan, O., & Dablander, F. (2023). Multimodality and skewness in emotion time series. *Emotion*, *23*(8), 2117–2141. <https://doi.org/10.1037/emo0001218>
- Helman, E., & Xie, S. Y. (2021). Doing better data visualization. *Advances in Methods and Practices in Psychological Science*, *4*(4). <https://doi.org/10.1177/251524592111045334>
- Helliwell, J. F., & Wang, S. (2014). Weekends and subjective well-being. *Social Indicators Research*, *116*(2), 389–407. <https://doi.org/10.1007/s11205-013-0306-y>
- Helmich, M. A., Olthof, M., Oldehinkel, A. J., Wichers, M., Bringmann, L. F., & Smit, A. C. (2021). Early warning signals and critical transitions in psychopathology: Challenges and recommendations. *Current Opinion in Psychology*, *41*, 51–58. <https://doi.org/10.1016/j.copsyc.2021.02.008>
- Horwitz, A. G., Zhao, Z., & Sen, S. (2023). Peak-end bias in retrospective recall of depressive symptoms on the PHQ-9. *Psychological Assessment*, *35*(4), 378–381. <https://doi.org/10.1037/pas0001219>
- Jacobson, N. C., Chow, S.-M., & Newman, M. G. (2019). The differential time-varying effect model (dtvem): A tool for diagnosing and modeling time lags in intensive longitudinal data. *Behavior Research Methods*, *51*(1), 295–315. <https://doi.org/10.3758/s13428-018-1101-0>
- Jahng, S., Wood, P. K., & Trull, T. J. (2008). Analysis of affective instability in ecological momentary assessment: Indices using successive difference and group comparison via multilevel modeling. *Psychological Methods*, *13*(4), 354–375. <https://doi.org/10.1037/a0014173>
- Jongerling, J., Laurenceau, J.-P., & Hamaker, E. L. (2015). A multilevel AR(1) model: Allowing for inter-individual differences in trait-scores, inertia, and innovation variance. *Multivariate Behavioral Research*, *50*(3), 334–349. <https://doi.org/10.1080/00273171.2014.1003772>
- Kelava, A., & Brandt, H. (2019). A nonlinear dynamic latent class structural equation model. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(4), 509–528. <https://doi.org/10.1080/10705511.2018.1555692>
- Koudela-Hamila, S., Grund, A., Santangelo, P., & Ebner-Priemer, U. W. (2019). Valence and motivation as predictors of student time use in everyday life: An experience sampling study. *Frontiers in Psychology*, *10*, Article 1430. <https://doi.org/10.3389/fpsyg.2019.01430>
- Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine*, *16*(9), 606–613. <https://doi.org/10.1046/j.1525-1497.2001.016009606.x>
- 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>
- Langener, A. M., Stulp, G., Kas, M. J., & Bringmann, L. F. (2023). Capturing the dynamics of the social environment through experience sampling methods, passive sensing, and ego-centric networks: Scoping review. *JMIR Mental Health*, *10*(1), Article e42646. <https://doi.org/10.2196/42646>
- Leertouwer, I. J., Schuurman, N. K., & Vermunt, J. K. (2022). Are retrospective assessments means of people's experiences? *Journal for Person-Oriented Research*, *8*(2), 52–70. <https://doi.org/10.17505/jpor.2022.24855>
- McNeish, D., Mackinnon, D. P., Marsch, L. A., & Poldrack, R. A. (2021). Measurement in intensive longitudinal data. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(5), 807–822. <https://doi.org/10.1080/10705511.2021.1915788>
- Mestdagh, M., & Dejonckheere, E. (2021). Ambulatory assessment in psychopathology research: Current achievements and future ambitions. *Current Opinion in Psychology*, *41*, 1–8. <https://doi.org/10.1016/j.copsyc.2021.01.004>
- Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2023, February 1). *E1071: Misc functions of the Department of Statistics, Probability Theory Group (formerly: E1071), TU Wien* (Version 1.7-13). <https://cran.r-project.org/web/packages/e1071/index.html>
- Midway, S. R. (2020). Principles of effective data visualization. *Patterns*, *1*(9), Article 100141. <https://doi.org/10.1016/j.patter.2020.100141>
- Morgan, J. F., Reid, F., & Lacey, J. H. (1999). The SCOFF questionnaire: Assessment of a new screening tool for eating disorders. *BMJ*, *319*(7223), 1467–1468. <https://doi.org/10.1136/bmj.319.7223.1467>
- Nordmann, E., McAleer, P., Toivo, W., Paterson, H., & DeBruine, L. M. (2022). Data visualization using R for researchers who do not use R. *Advances in Methods and Practices in Psychological Science*, *5*(2). <https://doi.org/10.1177/25152459221074654>
- Olthof, M., Hasselman, F., & Lichtwarck-Aschoff, A. (2020). Complexity in psychological self-ratings: Implications for research and practice. *BMC Medicine*, *18*(1), Article 317. <https://doi.org/10.1186/s12916-020-01727-2>
- R Core Team. (2023). *R: A language and environment for statistical computing* (Version 4.3.2). <https://www.R-project.org/>
- Revol, J., Carlier, C., Lafit, G., Verhees, M., Sels, L., & Ceulemans, E. (2023). *Preprocessing ESM data: A step-by-*

- step framework, tutorial website, R package, and reporting templates. PsyArXiv. <https://doi.org/10.31234/osf.io/hnu2t>
- Rimpler, A., Siepe, B. S., Rieble, C. L., Proppert, R. K. K., & Fried, E. I. (2024). Introducing FRED: Software for generating feedback reports for ecological momentary assessment data. *Administration and Policy in Mental Health and Mental Health Services Research*, *51*, 490–500. <https://doi.org/10.1007/s10488-023-01324-4>
- Ruf, A., Neubauer, A. B., Ebner-Priemer, U., Reif, A., & Matura, S. (2021). Studying dietary intake in daily life through multilevel two-part modelling: A novel analytical approach and its practical application. *International Journal of Behavioral Nutrition and Physical Activity*, *18*, Article 130. <https://doi.org/10.1186/s12966-021-01187-8>
- Ryan, O., Haslbeck, J. M., & Waldorp, L. (2023). *Non-stationarity in time-series analysis: Modeling stochastic and deterministic trends*. PsyArXiv. <https://doi.org/10.31234/osf.io/z7ja2>
- Schoevers, R. A., Borkulo, C. D., van Lamers, F., Servaas, M. N., Bastiaansen, J. A., Beekman, A. T. F., Hemert, A. M., van Smit, J. H., Penninx, B. W. J. H., & Riese, H. (2021). Affect fluctuations examined with ecological momentary assessment in patients with current or remitted depression and anxiety disorders. *Psychological Medicine*, *51*(11), 1906–1915. <https://doi.org/10.1017/S0033291720000689>
- Schönbrodt, F. D., Zygarr-Hoffmann, C., Nestler, S., Pusch, S., & Hagemeyer, B. (2022). Measuring motivational relationship processes in experience sampling: A reliability model for moments, days, and persons nested in couples. *Behavior Research Methods*, *54*(4), 1869–1888. <https://doi.org/10.3758/s13428-021-01701-7>
- Schuurman, N. K., Houtveen, J. H., & Hamaker, E. L. (2015). Incorporating measurement error in $n = 1$ psychological autoregressive modeling. *Frontiers in Psychology*, *6*, Article 1038. <https://doi.org/10.3389/fpsyg.2015.01038>
- Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological momentary assessment. *Annual Review of Clinical Psychology*, *4*, 1–32. <https://doi.org/10.1146/annurev.clinpsy.3.022806.091415>
- Shrout, P. E., Stadler, G., Lane, S. P., McClure, M. J., Jackson, G. L., Clavé, F. D., Iida, M., Gleason, M. E. J., Xu, J. H., & Bolger, N. (2018). Initial elevation bias in subjective reports. *Proceedings of the National Academy of Sciences, USA*, *115*(1), E15–E23. <https://doi.org/10.1073/pnas.1712277115>
- Smaldino, P. E. (2013). Measures of individual uncertainty for ecological models: Variance and entropy. *Ecological Modelling*, *254*, 50–53. <https://doi.org/10.1016/j.ecolmodel.2013.01.015>
- Smith, G. K., Mills, C., Paxton, A., & Christoff, K. (2018). Mind-wandering rates fluctuate across the day: Evidence from an experience-sampling study. *Cognitive Research: Principles and Implications*, *3*, Article 54. <https://doi.org/10.1186/s41235-018-0141-4>
- Song, J., Howe, E., Oltmanns, J. R., & Fisher, A. J. (2023). Examining the concurrent and predictive validity of single items in ecological momentary assessments. *Assessment*, *30*(5), 1662–1671. <https://doi.org/10.1177/10731911221113563>
- Spitzer, R. L., Kroenke, K., Williams, J. B., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: The GAD-7. *Archives of Internal Medicine*, *166*(10), 1092–1097.
- Stone, A. A., Schneider, S., & Smyth, J. M. (2023). Evaluation of pressing issues in ecological momentary assessment. *Annual Review of Clinical Psychology*, *19*, 107–131. <https://doi.org/10.1146/annurev-clinpsy-080921-083128>
- Terluin, B., de Boer, M. R., & de Vet, H. C. W. (2016). Differences in connection strength between mental symptoms might be explained by differences in variance: Reanalysis of network data did not confirm staging. *PLOS ONE*, *11*(11), Article e0155205. <https://doi.org/10.1371/journal.pone.0155205>
- Trull, T. J., & Ebner-Priemer, U. W. (2009). Using experience sampling methods/ecological momentary assessment (ESN/EMA) in clinical assessment and clinical research: Introduction to the special section. *Psychological Assessment*, *21*(4), 457–462. <https://doi.org/10.1037/a0017653>
- Tukey, J. W. (1977). *Exploratory data analysis* (Vol. 2). Addison-Wesley.
- Tutunji, R., Proppert, R. K. K., Rieble, C. L., & Fried, E. I. (2023). *Defining a generic holdout sample for combined exploratory and predictive analyses in the WARN-D dataset*. OSF. <https://doi.org/10.17605/OSF.IO/W9NXY>
- van Montfort, K., Oud, J. H. L., & Voelkle, M. C. (Eds.). (2018). *Continuous time modeling in the behavioral and related sciences*. Springer. <https://doi.org/10.1007/978-3-319-77219-6>
- Van Vaerenbergh, Y., & Thomas, T. D. (2013). Response styles in survey research: A literature review of antecedents, consequences, and remedies. *International Journal of Public Opinion Research*, *25*(2), 195–217. <https://doi.org/10.1093/ijpor/eds021>
- Vansteelandt, K., & Verbeke, G. (2016). A mixed model to disentangle variance and serial autocorrelation in affective instability using ecological momentary assessment data. *Multivariate Behavioral Research*, *51*(4), 446–465. <https://doi.org/10.1080/00273171.2016.1159177>
- Vogelsmeier, L. V. D. E., Jongerling, J., & Maassen, E. (2023). *Assessing and accounting for measurement in intensive longitudinal studies: Current practices, considerations, and avenues for improvement*. PsyArXiv. <https://doi.org/10.31234/osf.io/uat5r>
- Vogelsmeier, L. V. D. E., Vermunt, J. K., van Roekel, E., & De Roover, K. (2019). Latent Markov factor analysis for exploring measurement model changes in time-intensive longitudinal studies. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(4), 557–575. <https://doi.org/10.1080/10705511.2018.1554445>
- von Klipstein, L., Servaas, M. N., Lamers, F., Schoevers, R. A., Wardenaar, K. J., & Riese, H. (2023). Increased affective reactivity among depressed individuals can be explained by floor effects: An experience sampling study. *Journal of Affective Disorders*, *334*, 370–381. <https://doi.org/10.1016/j.jad.2023.04.118>
- von Neumann, J., Kent, R. H., Bellinson, H. R., & Hart, B. I. (1941). The mean square successive difference. *The Annals of Mathematical Statistics*, *12*(2), 153–162. <https://www.jstor.org/stable/2235765>

- Wainer, H., & Thissen, D. (1981). Graphical data analysis. *Annual Review of Psychology, 32*, 191–241.
- Wilms, R., Lanwehr, R., & Kastenmüller, A. (2020). Do we overestimate the within-variability? The impact of measurement error on intraclass coefficient estimation. *Frontiers in Psychology, 11*, Article 825. <https://doi.org/10.3389/fpsyg.2020.00825>
- Wrzus, C., & Neubauer, A. B. (2023). Ecological momentary assessment: A meta-analysis on designs, samples, and compliance across research fields. *Assessment, 30*(3), 825–846. <https://doi.org/10.1177/10731911211067538>
- Zhang, R., & Volkow, N. D. (2023). Seasonality of brain function: Role in psychiatric disorders. *Translational Psychiatry, 13*, Article 65. <https://doi.org/10.1038/s41398-023-02365-x>
- Zumbo, B. D., & Hubley, A. M. (Eds.). (2017). *Understanding and investigating response processes in validation research*. Springer. <https://doi.org/10.1007/978-3-319-56129-5>