

# Associations Between Ecological Momentary Assessment and Passive Sensor Data in a Large Student Sample

Björn S. Siepe<sup>1</sup>, Rayyan Tutunji<sup>2</sup>, Carlotta L. Rieble<sup>2</sup>, Ricarda K. K. Proppert<sup>2</sup>, and Eiko I. Fried<sup>2</sup>  
<sup>1</sup>Psychological Methods Laboratory, Department of Psychology, University of Marburg  
<sup>2</sup>Department of Clinical Psychology, Leiden University

Ecological momentary assessment (EMA) increases ecological validity but can be burdensome. To reduce this burden and to better understand psychological constructs in daily life, a growing chorus of voices has called for augmenting or replacing EMA data with data passively collected from wearable devices. It is thus critical to investigate the quality of wearable data and its overlap with typical self-report measures. Here we compared results from passive sensing and EMA data from a project to build a warning system for depression in students (WARN-D) in a large sample of 781 students. For 3 months, participants wore a Garmin VivoSmart 4 watch and answered EMA surveys (up to 352 observations). We investigated whether and to what extent passive sensor metrics were concurrently associated with different self-report measures purportedly measuring the same constructs. We focused on stress, tiredness, and sleep, all of which have transdiagnostic relevance to mental health and can arguably be assessed with self-report and physiological measures. We used longitudinal mixed-effects models to estimate average momentary associations and their interindividual heterogeneity. Self-report and wearable measures of sleep-related variables showed robust associations, while associations were weaker for tiredness, and measures of stress did not overlap for most individuals. These findings suggest that wearable data and their corresponding self-report measures may not necessarily measure similar constructs. We provide several explanations for this result, including semantic differences and measurement issues, and offer insights and ways forward for mental health research combining wearable and self-report data.

## General Scientific Summary

We investigated the concurrent overlap between self-report and wearable sensor data measuring stress, tiredness, and sleep. For the majority of individuals in our sample, we found that self-report and physiological measures of stress show very weak to no associations. These results raise several questions about differences between data sources and potential measurement issues.

**Keywords:** ecological momentary assessment, digital phenotyping, passive sensors, measurement, stress

Collecting high-resolution mental health data via ambulatory assessment has become a dominant theme in psychopathology research in recent years (Hamaker & Wichers, 2017). In particular, collecting self-report data in participants' daily lives, often referred

to as experience sampling, allows for new insights into inter and intraindividual variation (Hamaker & Wichers, 2017; Mestdagh & Dejonckheere, 2021; Siepe, Rieble, et al., 2025). However, collecting self-reported data over an extended period, possibly multiple

This article was published Online First July 31, 2025.

Daniel Fulford served as action editor.

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

Rayyan Tutunji  <https://orcid.org/0000-0002-3537-9888>

Carlotta L. Rieble  <https://orcid.org/0000-0002-4764-3906>

Ricarda K. K. Proppert  <https://orcid.org/0000-0002-4225-2439>

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

This research was funded in whole, or in part, by the Horizon Europe Framework Programme (Grant 949059). For the purpose of open access, the author has applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission.

Rayyan Tutunji, Carlotta L. Rieble, Ricarda K. K. Proppert, and Eiko I. Fried are supported by funding from the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Programme (Grant 949059). The authors declare that there were no conflicts of interest with respect to the authorship or the publication of this article. The additional online materials for this article are available at <https://osf.io/txgnq/>. This exploratory work was not pre-registered. This is the third version of this

article (March 12, 2025).

Björn S. Siepe served as lead for formal analysis, methodology, software, writing—original draft and contributed equally to conceptualization and writing—review and editing. Rayyan Tutunji contributed equally to data curation, investigation, project administration, validation, and writing—review and editing. Carlotta L. Rieble contributed equally to data curation, investigation, project administration, validation, and writing—review and editing. Ricarda K. K. Proppert contributed equally to data curation, investigation, project administration, validation, and writing—review and editing. Eiko I. Fried served as a lead for funding acquisition and supervision, contributed equally to data curation, funding acquisition, investigation, methodology, project administration, and writing—review and editing, and served in a supporting role for conceptualization, visualization, writing—original draft.

Correspondence concerning this article should be addressed to Björn S. Siepe, Psychological Methods Laboratory, Department of Psychology, University of Marburg, Gutenbergstraße 18, 35037 Marburg, Germany. Email: [bjorn.siepe@uni-marburg.de](mailto:bjorn.siepe@uni-marburg.de)

times per day, can be burdensome for participants. Data collected via passive sensors are increasingly being used to augment or to replace self-report data and to reduce participant burden (Veloze et al., 2024). Here we use a large, multimodal student dataset to answer the question to which degree self-report and sensor data derived from wearable devices assessing similar constructs such as stress are correlated. Such work is important because other fields have recently demonstrated that the assumption that different methods reliably assess the same construct is often not met. For example, self-report and task-based measures of different constructs such as self-control have shown to be very weakly correlated (Dang et al., 2020), and there are considerable discrepancies between self-reported and logged measures of digital media use (Parry et al., 2021). In the remainder of the Introductory part, we introduce passive sensor data, outline its relevance for psychopathology research, provide a brief overview of the literature on the validity of passive sensing research, and then conclude with a summary of the present study.

### Passive Sensing

Digital phenotyping, that is, data collected from passive sensors, is seen as very promising in mental health science (Adams et al., 2017; Ebner-Priemer & Santangelo, 2020; Huckvale et al., 2019; Insel, 2018). This broad term includes physiological data collected from wearable devices such as smartwatches and other passive sensing data such as location or communication data collected via smartphones (see Mohr et al., 2017, for an overview). In this article, we focus on data from wearable smartwatches while occasionally referring to the broader digital phenotyping literature.

Due to the high temporal resolution of such data, large amounts of data can easily be collected per individual (Onnela, 2021), enabling more scalable forms of assessment (Veloze et al., 2024). Another hope is that digital measures will provide more “objective” assessments of constructs for which self-report measures may be biased or flawed (Insel, 2018). The continuous monitoring of individuals may open the door to personalized early interventions and thereby decrease existing barriers to treatments (Huckvale et al., 2019). Besides being a useful resource for researchers, people with mental health problems can also use passive sensor data for self-monitoring (Nadal et al., 2021). Commercial-grade devices that are practical to wear in everyday life play a crucial role in this regard. It has also been argued that digital phenotyping may be particularly useful for transdiagnostic research (Ringwald et al., 2025), as passive markers might relate to various symptoms across disorders. The extent to which digital phenotyping can deliver on these promises depends critically on many factors such as the quality of the data and the extent to which psychological and physiological measures are associated with one another.

In this article, we will, therefore, investigate three constructs that are highly relevant to mental health. We assess the overlap between self-report and wearable measures of stress, tiredness, and sleep, which can arguably all be measured with both sensors and self-report. We now describe the transdiagnostic relevance of these variables and summarize their use in the digital phenotyping literature.

### Stress, Sleep, and Tiredness as Transdiagnostic Markers

There is a long line of research into stressful life events and chronic stress and their relationship with disorders such as major

depression (see, e.g., Hammen, 2005) or substance use disorders (Sinha, 2001). Daily stressors and their relations to maladaptive traits have also been investigated in research on personality pathology (Vize et al., 2024). In the digital phenotyping literature, several studies have focused on predicting stress from passive sensing data in specific populations such as college students (see, e.g., Adler et al., 2022), whose academic stressors may increase their risk to develop mental illness (Auerbach et al., 2016).

Sleep problems are among the most common symptoms in the *Diagnostic and Statistical Manual of Mental Disorders*, fifth edition, (*DSM-5*), and also a robust contributor to psychopathology (Forbes et al., 2024; Freeman et al., 2020). Therefore, disturbances of sleep are widely considered a useful transdiagnostic marker for many psychiatric diagnoses (Baglioni et al., 2016; Harvey et al., 2011), and it is no surprise that sleep was the most commonly investigated passive variable in a review of the depression literature (De Angel et al., 2022). In a transdiagnostic study on digital markers of psychopathology, the sleep-related variable “later bedtime” was robustly associated with a summary of psychopathology domains (Ringwald et al., 2025).

Tiredness or fatigue is conceptually related to sleep problems and can be an indicator thereof (Coulombe et al., 2010; Shochat et al., 2014). Fatigue has long been discussed as a common thread in the intertwined nature of psychopathology and medical diagnoses (Berrios, 1990). Due to its potential intraday fluctuations, tiredness is useful for experience sampling studies and has, for example, been used to investigate the interplay of sleep and affect (Mill et al., 2016; Triantafillou et al., 2019). Conceptually, low-arousal affect items such as feeling tired have a clear connection to depressive symptoms (Price et al., 2023). In digital phenotyping studies, the prediction of tiredness and fatigue with passive sensing has been studied in specific clinical populations (Rao et al., 2023).

Overall, we believe that these three constructs present a good starting point to investigate relations between self-report and passive sensing measurement modalities, because they are common in mental health problems, have shown to be transdiagnostically useful, and are relevant to studying various forms of psychopathology in daily life.

### Challenges in the Digital Phenotyping Literature

To date, much of the literature has focused on using digital phenotyping data for predictive purposes across a wide range of psychological and psychiatric constructs (see Bufano et al., 2023; Melcher et al., 2020; Mohr et al., 2017, for some example reviews in mental health). However, the literature is still at too early a stage to obtain a comprehensive overview of measurement issues and draw an overall conclusion about the utility of digital phenotyping for predicting constructs related to psychopathology. While there are some promising and interesting early results (see, e.g., Mohr et al., 2017; Smets et al., 2018), some disappointing results have dampened some of the initial enthusiasm (Ebner-Priemer & Santangelo, 2020) and the literature is generally characterized by many small sample studies (see, e.g., De Angel et al., 2022, for depression). We also suspect that, similar to other areas of research, larger samples and preregistered studies will lead to nonreplication of some promising early results (see Tackett et al., 2017, for an overview and discussion of this topic for clinical psychology). The emphasis on predictive models, often in the form of complex machine-learning pipelines with limited interpretability, further complicates understanding the relationship between specific

passive and self-report measures (De Angel et al., 2022). Conversely, the interpretability of predictive models will also benefit from a better understanding of the underlying passive sensing variables. In either case, to make progress in this field, a better understanding of how passive sensor measures relate to traditional self-report measures or clinical ratings—in other words, construct validation in the broadest sense—is crucial (Davidson, 2022; Langener, Siepe, et al., 2024; Strauss et al., 2022). This is also our core goal in the present article.

There are several approaches to providing evidence of construct validity for passive sensors. One approach is to validate measures captured by sensors against existing gold-standard laboratory devices (see Nelson et al., 2020; Velozo et al., 2024, for some examples). For example, wearables measuring stress-related arousal have been able to capture similar information to gold-standard laboratory-based psychophysiology measures, with some limitations (van Lier et al., 2020). Another approach is to validate measures against self-report data, such as social media use (Mahalingham et al., 2023), drug use (Bertz et al., 2018), sleep duration (Lauderdale et al., 2008), and physical activity (Prince et al., 2008).<sup>1</sup> These latter studies tend to conclude that self-report measures of some behavioral or physiological constructs are often imperfect and biased. We believe this is a reasonable conclusion for well-defined and narrow constructs for which valid and reliable laboratory devices exist.

However, there are reasons to believe that the discrepancy between more objective and self-report data is not always solely due to measurement issues on the side of self-report (Das Swain et al., 2022). For example, the lack of alignment between digital media use and logging data is in part due to poor self-report, but may also arise due to sources of bias on the side of logging (Jürgens et al., 2020). Reviews on digital phenotyping in social media use (Chancellor & De Choudhury, 2020) and the social environment (Langener et al., 2023) have found a lack of explicit validation and theoretic rationale for the use of passive sensor measures for certain psychological constructs, such as using global positioning system (GPS) and accelerometer data as indicators of social interactions (Langener et al., 2023). Conceptually, broad constructs such as stress can subsume various physiological and psychological facets, for which self-report and sensor data might be differentially useful (Das Swain et al., 2022). Generally, the more complex the construct to be interrogated, the more difficult validation efforts become, especially when attempting to measure constructs encompassing various psychological and physiological facets, such as many constructs in mental health research (Tutunji, Kogias, et al., 2023).

A complicating feature in obtaining validity evidence for digital phenotyping methods is directly related to one of their greatest strengths: they produce large amounts of data, which poses several challenges in data analysis (Langener, Stulp, et al., 2024; Onnela, 2021). Researchers must decide on a variety of potential preprocessing and aggregation steps needed to make the temporal resolution of sensor data compatible with self-reported data. In the absence of strong theoretical guidance on the best decisions, this range of possibilities leads to many degrees of freedom for the researcher, which can have a large potential impact on the conclusions of a study. For example, Niemeijer et al. (2022) have shown the importance of different preprocessing and modeling choices on the performance of sensor measures for predicting subjective sleep quality. Other work shows that different temporal resolutions for aggregating passive sensor measures collected via smartphones led to different substantive conclusions and predictive model performance (Langener, Stulp, et al.,

2024). These results suggest that studies investigating the overlap between self-report and passive sensor data should consider plausible alternative pathways in the analysis pipeline.

## The Present Study

In this study, we investigated whether and to what extent wearable sensor metrics are concurrently associated with self-report measures of the same construct. Specifically, we assessed the overlap between self-report and wearable measures of stress, tiredness, and sleep in a large sample of students over several months. This serves to improve our understanding of the extent to which these transdiagnostically relevant constructs can be assessed using passive sensors. Methodologically, we adopted a multilevel modeling strategy of time-series data to investigate the interindividual heterogeneity of associations between self-report and sensor measures. To account for the wide range of plausible analytic choices, we complement our main analyses with a wide range of additional analyses, which are available—together with code and additional information—in our additional online materials (<https://osf.io/txgnq/>; Siepe, Tutunji, et al., 2025).

## Method

### Transparency and Openness

We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study, and we follow the Journal Article Reporting Standards (JARS, Appelbaum et al., 2018). This study was not preregistered. Data collection for the project to build a warning system for depression in students (WARN-D) is ongoing and we want to avoid having different small parts of the data shared across many projects. We will therefore make data available (excluding potentially identifiable data) on the WARN-D project hub (<https://osf.io/frqdv/>) when all data are collected, cleaned, and checked. We share the participant identifiers we used for this article in the additional online materials to make the article reproducible in the future. Data collection was approved by the Leiden University Research Ethics Committee (2021-09-06-E.I.FriedV2-3406). The project is funded by the European Research Council in the Horizon 2020 research and innovation program (Grant 949059).

Data were analyzed using the R programming language (Version 4.4.0, R Core Team, 2024). Our main analyses were conducted using nlme (Version 3.1-164, Pinheiro et al., 2023). For visualization, we mostly relied on ggplot2 (Wickham, 2016) and ggdist (Kay, 2024). We created a reproducible environment for the project via renv (Version 1.0.7, Ushey & Wickham, 2024). A list of all other packages used is provided in the additional online materials.

### Participants and Procedure

Data were collected as part of the WARN-D study which aims to build a personalized warning system for depression in higher-education students. To achieve this goal, 2,000 students were followed over multiple months. To be enrolled in data collection, participants

<sup>1</sup> These references do not all focus on wearable devices per se but provide an overview of self-report versus passively collected data more broadly.

had to be at least 18 years of age and be enrolled as students at a Dutch institution of higher education. All inclusion and exclusion criteria for the study can be found in [Fried et al. \(2023\)](#). Notably, participants were excluded when they displayed moderate to high levels of depression, mania, thought disorders, substance-use problems; were undergoing treatment for these mental health issues; experienced moderate to severe suicidal ideation; or indicated that they would find it stressful to see estimates of calories burned as displayed by the smartwatch. Data collection was divided into four cohorts with roughly 500 participants each. We used data from all cohorts but excluded participants who will be used as part of training/validation sets for future prediction projects (see the preregistration by [Tutunji, Proppert, et al., 2023](#)), leaving us with a final sample size of 1,193 participants. We used the maximum number of participants available to obtain the highest level of precision in our estimates.

The data collection procedure in WARN-D consists of a baseline, an 85-day daily monitoring/ecological momentary assessment (EMA) phase, and eight follow-up surveys. It is described in more detail by [Fried et al. \(2023\)](#). At the time of writing, all regular EMA data collection has finished while follow-up surveys are still ongoing. During EMA data collection, participants received four surveys per day over 85 days as well as a weekly survey on Sundays. Surveys were sent via the Ethica application on their personal smartphones. Additionally, they were provided with a Garmin VivoSmart 4 smartwatch which they were supposed to wear during the full 85-day period. Of the 1,193 participants, 80.64% identified as women, 16.08% identified as male, and 3.28% indicated another gender identity. The average age was 22.5 ( $SD = 3.94$ , range = [18, 61]). The majority (51.30%) of participants were of Dutch/Belgian/German nationality, 40.07% were of another European nationality, and the other 8.63% had a non-European nationality.<sup>2</sup>

## Measures

Our interest was to compare self-report and wearable data of stress, sleep, and tiredness.

### Self-Report Variables

Self-reported momentary stress was assessed four times daily with a single item: “I feel stressed right now.” Momentary tiredness was assessed with the question “I feel tired right now.” In the morning survey, participants could indicate their sleep quality with the item “Last night, I slept well.” All three variables were assessed on a 1–7 Likert scale, ranging from “not at all” to “very much.”

### Digital Phenotyping Data

Digital phenotyping data were collected using Garmin VivoSmart 4 smartwatches. The watch automatically tracked parameters such as heart rate, sleep, movement, activity, and stress. In the following, we will use the terms self-report stress, self-report sleep, and self-report tiredness as well as sensor stress, sensor sleep, and sensor body battery to distinguish between self-report and wearable measures, respectively. The smartwatch provided various sleep indices per night. We chose total sleep duration (in seconds) as the sensor sleep measure, and converted it to sleep in hours. Note that a previous systematic review indicates that the Garmin watch may overestimate total sleep duration compared to polysomnography, often considered a gold standard for sleep assessment ([Schyvens et al., 2024](#)).

The smartwatch computed a stress score based on heart rate, heart rate variability, and activity measures. It was quantified continuously from 0 to 100, but Garmin provides four brackets of values for interpretation ([Garmin, 2024b](#)): 0–25 as resting state, 26–50 as low stress, 51–75 as medium stress, and 76–100 as high stress. The stress score was calculated based on a proprietary algorithm by Firstbeat Technologies ([Garmin, 2024b](#)). A white paper by the company ([Firstbeat Technologies Ltd., 2014](#)) provides some information on the underlying analysis, but the algorithm is not explained in sufficient detail to be reproducible. The smartwatch returned a stress score every 3 min and has been used before, for example, to predict change processes in psychotherapy ([Hehlmann et al., 2021](#)).

The smartwatch measure that may best capture the construct of being tired is the so-called “body battery” score calculated by the watch. It is supposed to measure the body’s energy resources which are influenced by factors such as sleep, exercise, and stress ([Garmin, 2024a](#)). Similar to the stress score, it was calculated by taking into account heart rate (variability) and activity data and sampled every 3 min. It was also continuously scored from 0 to 100, but there are two main brackets for interpretation ([Garmin, 2024a](#)): 0–25 as a charging/parasympathetic state, and 26–100 as a tiring/sympathetic state. Underlying these brackets is the assumption that stressful experiences or a sympathetic state drain bodily resources. To the best of our knowledge, this sensor feature has not yet been validated in peer-reviewed research, although it is used as a feature by [Daniels et al. \(2024\)](#). We chose to include this feature as a potential passive marker of tiredness.

## Data Analysis

### Preprocessing

As wearable data were assessed at a higher sampling frequency than the EMA data, the sensor data needed to be aggregated before analysis to match it to the concurrent EMA prompts. For this aggregation, we needed to choose (a) the summary measure for aggregation (e.g., the mean or the  $SD$  of passive sensor data such as stress), (b) the aggregation window (e.g., the time frame over which we aggregate, such as 15 or 240 min), and (c) the lag size relative to the EMA prompt (i.e., summarizing sensor data before, around, or after the EMA prompt).

In the absence of strong theoretical expectations, we chose the following procedure. As a summary measure, we selected the mean, although we also estimated models using the  $SD$  and the maximum (specifically, the 0.95 quantile of the data to avoid potential outlier effects) as a summary for stress and tiredness scores. As aggregation windows, we chose 15, 30, 60, 120, and 240 min. To align the sensor data with the EMA prompt, we aggregated the data in three distinct time frames (which we call “lag”): before the EMA prompt, around the time of the EMA prompt, and after the EMA prompt. For example, a combination of 30-min aggregation and a lag “around” the EMA data means that we calculated the mean of the sensor in the 15 min before and the 15 min after the EMA beep. All possible combinations of aggregation window and lag specification resulted in  $5 \times 3 = 15$  data sets for analysis. We conducted all analyses on each of the individual data sets. We always chose the data set that is aggregated around the EMA beep with a 30-min aggregation window as the main analysis. We estimated the same model on all

<sup>2</sup> One participant’s demographic and baseline data were missing due to a data collection error. We excluded this individual.

15 data sets with different aggregation and lag specifications as sensitivity analyses for different temporal aggregation choices. We additionally provide further secondary analyses that are described in more detail below.

All negative sensor stress values, representing either the lack of enough data or that a participant was performing a physical activity, were removed and coded as missing data. Further handling of missing data is described below as part of our analysis pipeline.

### Main Analysis

We decided on a single main analysis for each of the outcomes based on recommendations in the methodological literature (such as Hamaker & Muthén, 2020; Myin-Germeys & Kuppens, 2022; Wang & Maxwell, 2015), our theoretical expectations, and the properties of the data. For all main analyses, we estimated linear mixed effects models with random intercepts and random slopes with the `nlme` package in R. We deleted missing data points via listwise deletion during model fitting, within-person centered all nonfactorial predictor variables to separate between-person from within-person effects, and used restricted maximum likelihood estimation.

We specified relatively simple models without any covariates because we were interested in the overlap of different ways of assessment, and not in any predictive capacities when controlling for covariates. We excluded participants with <25% of valid data points for all main analyses. While this cutoff is arbitrary, we chose it to retain as many participants as possible while still obtaining reasonably precise estimates of within-person associations. This left us with 730, 709, and 781 included participants for the stress, sleep, and tiredness analyses, respectively. For all variables, most excluded individuals lacked both sufficient EMA and sensor data, followed by insufficient sensor data. Only a few participants were excluded because of missing EMA data only. Additional information is available in the additional online materials.

In addition, we provide several secondary analyses in the additional online materials. We did not perform a full multiverse analysis in which we would cross each analysis decision in a fully factorial manner (see, e.g., Weermeijer et al., 2022). Rather, we decided on the most appropriate alternative analyses for each target variable to keep computational effort reasonable and the results interpretable. A list of all secondary analyses is provided in Table 1.

All multilevel models for the main analyses consisted of the following structure:

$$y_{it} = \beta_0 + \beta_1 \cdot (x_{it} - \bar{x}_i) + \beta_2 \cdot \bar{x}_i + \beta_3 \cdot d_{it} + u_{0i} + u_{1i} \cdot (x_{it} - \bar{x}_i) + e_{it}, \quad (1)$$

**Table 1**  
Overview of Secondary Analyses

Analysis	Stress	Tiredness	Sleep
Interaction with age, gender, depression, cohort (S1)	✓	✓	✓
Change of residual correlation structure (S2)	✓	✓	✓
Informed “binning” of outcome (S3)	✓	✓	✗
Alternative outcome operationalization (S4)	✓	✓	✗

*Note.* This table contains an overview of all secondary analyses available in the additional online materials along with a numerical index (e.g., S1) for easy retrieval. S1–S4 = secondary analyses 1–4.

where  $y_{it}$  represents the sensor outcome  $y$  of individual  $i$  at time  $t$ . The value of the EMA predictor variable of individual  $i$  at time  $t$  is denoted as  $x_{it}$ , where  $\bar{x}_i$  represents its within-person mean. We additionally grand-mean centered  $\bar{x}_i$  for ease of interpretation.  $d_{it}$  denotes the value of an additional categorical predictor for the time of day, with the first prompt of the day serving as the reference category. We added this predictor because we expected potential (small) intrain-day trends for stress and tiredness based on the results of a previous study using a subset of the present data (Siepe, Rieble, et al., 2025).

The fixed intercept is denoted as  $\beta_0$ . The fixed slope of the within-person centered response is denoted as  $\beta_1$  and can be interpreted as the average within-person association.  $\beta_2$  is the fixed slope of the person-specific mean and can be interpreted as the between-person association.  $\beta_3$  contains the effects of the dummy-coded categorical predictor and can be interpreted as differences in the predicted outcome variable at different times of the day. All estimates represent conditional effects, that is, effects holding all other predictors constant. For brevity, we do not repeat this point in the Results section. The random effects are represented by  $u_{0i}$  and  $u_{1i}$ , which are the random intercepts and slopes of the within-person effect for individual  $i$ , respectively. Both random effects follow a normal distribution. The residuals  $e_{ij}$  are assumed to follow an autocorrelated structure with lag-1, for which we ignored potentially different time intervals between prompts.

For the main analysis of stress, we regressed mean sensor stress during the aggregation window on self-reported stress. For the tiredness main analysis, we regressed mean body battery during the aggregation window on self-reported tiredness. After initial problems with estimating the random effects covariance matrix, we divided the sensor outcome variable by 100 for both analyses to avoid large differences in the variance of the outcome and predictors. We report all estimates on the original scale below for ease of interpretation. For the sleep main analysis, we regressed sensor total sleep duration on self-reported sleep quality in the morning after. As we expected no persisting temporal trends across the course of the study, we performed no detrending and did not include a time-of-day variable, as sleep variables were only assessed at the first prompt of the day.

To better understand the size of an effect, we interpreted the size of the regression coefficient  $\beta_1$  with respect to the scale of the outcome variable. As another form of interpretation, we calculated the standardized effects obtained by standardizing outcomes and predictors with the person-specific variance. Additionally, we estimated the marginal  $R^2$  (only considering the variance of fixed effects) and the conditional  $R^2$  (additionally considering the variance of random effects) based on the method by Nakagawa et al. (2017) as well as the root-mean-square error (RMSE) as implemented in the performance R package (Lüdtke et al., 2021) as estimates of model performance. We used the best linear unbiased predictor estimates for individual random effects in combination with the fixed effects estimates to calculate and visualize person-specific associations between sensor and self-report data. Unless declared otherwise, we used the conventional error level of .05.

### Secondary Analyses

To save space, we report more detailed results of our secondary analyses and robustness checks for our main analyses in the additional online materials. For all outcomes, we also included interactions of

the main effect of interest ( $\beta_1$ ) with age, gender, depression severity at baseline, and study cohort (from one to four). Due to the small sample size of individuals who did not identify as women or men, we did not include them in the estimation of the interaction effect and coded gender as binary.<sup>3</sup> These analyses help to understand whether the within-person association between self-report and sensor variables differs based on demographic characteristics, depressive symptomatology, or the study cohort. We added the latter to investigate potential seasonal influences and changes in watch quality over time. Depression severity was assessed using an adapted version of the Patient Health Questionnaire-9 (PHQ-9; Kroenke et al., 2001), see Fried et al. (2023) for more information. We also estimated all models without an autocorrelated residual structure. For stress and sleep, we estimated additional models where we “binned” the sensor variables based on the categorization by Garmin provided above. This analysis helps to understand if the categorical understanding of stress and body battery values as put forth by Garmin leads to clearer associations with self-report variables. For stress and tiredness, we also used the *SD* and the maximum value during the aggregation window as alternative operationalizations. Using, for example, the maximum of sensor stress values as an outcome could be meaningful if self-reported stress reflects the highest stress level around the EMA beep instead of an average.

**Results**

**Descriptives**

As explained in the Method section, our data consists of 15 data sets with different lags and aggregation windows of wearable data, whereas the EMA data is identical across these data sets. We calculated person-specific summary statistics for each variable of interest in each data set. To provide a general overview of all data, we aggregated summary statistics across data sets for Table 2. To summarize: we first calculate person-specific summaries, such as the mean or the standard deviation, for each variable in a data set. We then calculate the mean across individuals for each of these summaries, obtaining the average of person-specific summaries per variable per data set. We can then again average these across data sets.

To give insight into interindividual variability in summary statistics, Figure 1 contains distributions of person-specific means for the items used in the main analysis. Many individuals had a low mean of self-report stress and a relatively high self-report sleep quality. The person-specific averages of sensor stress and body battery are

concentrated rather closely to the scale average of 50. Orienting ourselves at Siepe, Rieble, et al. (2025), we present additional descriptive statistics and item properties in the additional online materials.

**Analysis Results**

Table 2 summarizes the main results, that is, regression estimates on the data set that aggregated sensor data in the 30-min window around the EMA prompt. Additionally, the person-specific estimates of the association between EMA and sensor variables are visualized in Figure 2. These were obtained by adding the estimate of an individual random effect to the fixed effect. In the following, we round results to two decimals, unless values are small enough that the third decimal contains relevant information.

**Stress**

As shown in the leftmost column of Table 3, the overall within-person association of self-report and sensor stress was positive, but rather small ( $\beta_1 = .49, 95\% \text{ CI} = [0.35, 0.63]$ ). The standardized estimate for this relationship was 0.026 ( $95\% \text{ CI} = [0.019, 0.033]$ ), which means that a one *SD* change in self-report stress corresponds to 0.026 *SD* change in sensor stress. The random slopes of the within-person effect had a *SD* of 1.29, implying considerable variability across people. This also means that the association was estimated to be negative for a substantial amount of individuals (see Figure 2).

In the sensitivity analyses of different temporal aggregation choices, we found the same general pattern of results, where  $\beta_1$  ranged from .17 to .67 with a mean of .46 ( $SD = 0.13$ ). See Figure 3 for the results of the sensitivity analyses.

The between-person component of self-report stress had a small positive association with sensor stress ( $\beta_2 = .56, 95\% \text{ CI} = [-0.02, 1.14]$ ) that was not significant at an  $\alpha$  level of .05. If the effect were robust, this would mean that individuals with a higher average self-report stress also report higher levels of sensor stress. We additionally found a strong effect of the categorical predictor of time-of-day, showing that sensor stress was lowest for the morning prompt and considerably higher at later prompts of the day.

While the conditional and marginal  $R^2$  point estimates of .22 and .13 indicate some amount of explained variance of the overall model, this can largely be attributed to the effects of the time-of-day predictor. When omitting this predictor, conditional and marginal  $R^2$  drop to .085 and .001, respectively. The RMSE was 21.66.

The same general result holds across our different secondary analyses. In secondary analysis S1, we found no significant associations between age, gender, or depression at baseline with the association between sensor and self-report stress. We found a significant effect of the study cohort, where participants in cohorts two and four showed slightly weaker associations between the sensor and self-report measure. For the second cohort, the average effect was 0. Changing the autocorrelation structure in S2 had no meaningful impact on results. When binning the outcome into four stress levels as used by Garmin (no, low, medium, and high stress) for secondary analysis S3 within an ordinal regression model, we observed a weak relationship (but this is not straightforward to compare to our linear mixed effects models). In S4, using the standard deviation of the

**Table 2**  
*Descriptive Statistics*

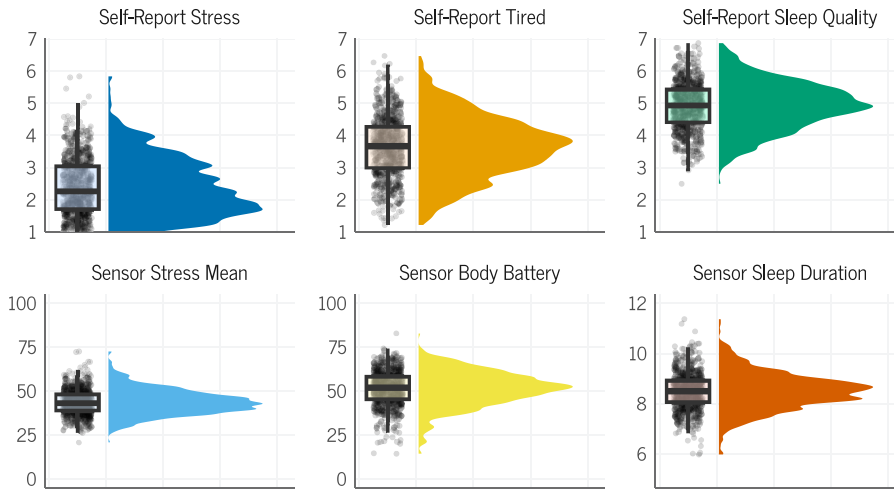
Variable	Range	iM	iSD	Missingness
Self-report stress	1–7	2.43	1.18	0.33
Self-report tiredness	1–7	3.64	1.37	0.33
Self-report sleep quality	1–7	4.88	1.21	0.34
Sensor stress	0–100	42.78	21.66	0.42
Sensor body battery	0–100	51.38	25.00	0.39
Sensor sleep duration	0–100	8.47	1.47	0.46

*Note.* Descriptive statistics were calculated for individuals included in the main analyses. The starting letter “i” refers to the individual summary statistics. Missingness was calculated as a median proportion. Sensor sleep duration is calculated in hours. Range refers to the possible range of the respective scales. iM = mean of person-specific means; iSD = mean of person-specific standard deviations.

<sup>3</sup> We additionally report the analysis with a separate category for nonbinary gender in the additional online materials.

This document is copyrighted by the American Psychological Association or one of its allied publishers. This article is intended solely for the personal use of the individual user and is not to be disseminated broadly. All rights, including for text and data mining, AI training, and similar technologies, are reserved.

**Figure 1**  
*Raincloud Plots for Person-Specific Means in Main Analysis*



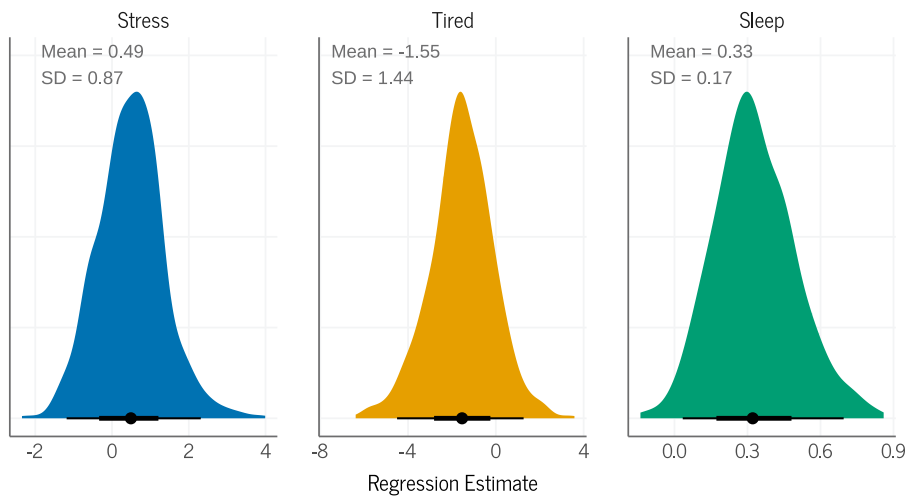
*Note.* Individuals excluded in the main analysis because of the amount of missing data are also excluded here. The boxplot displays the first and third quartile of the distribution as upper and lower hinges, and the median as a horizontal bar. The whiskers extend to 1.5 times the interquartile range. Dots represent individual means. See the online article for the color version of this figure.

sensor stress led to even weaker associations with self-reported stress. Using the maximum sensor stress during the aggregation period as an outcome led to slightly stronger associations (mean within-person relationship across all 15 data sets of 0.59 compared to 0.46 in the main analysis). Full results are available in the additional online materials.

**Tiredness**

We found a negative association between self-reported tiredness and sensor-based body battery in our main analysis ( $\beta_1 = -1.55$ , 95% CI = [-1.68, -1.42], see the middle column of Table 3). The relationship is negative as a higher sensor body battery is

**Figure 2**  
*Individual Estimates for Main Analyses*



*Note.* This figure uses data from the data set aggregated in a 30-min window around the EMA beep. The densities represent the estimated distribution of person-specific estimates for the respective analysis. The person-specific estimates were obtained by adding the empirical Bayes estimate to the fixed effect. The black bars represent the mean (black dot), 66% (thick bar), and 95% (thin bar) of the distribution. These do not represent the intervals of statistical uncertainty, but rather a dispersion of person-specific point estimates. Note that the standard deviation of individual estimates is smaller than the estimated random effect standard deviation due to shrinkage. EMA = ecological momentary assessment. See the online article for the color version of this figure.

This document is copyrighted by the American Psychological Association or one of its allied publishers. This article is intended solely for the personal use of the individual user and is not to be disseminated broadly. All rights, including for text and data mining, AI training, and similar technologies, are reserved.

**Table 3**  
*Results for Main Analyses*

Parameter	Dependent sensor variable		
	Stress (1)	Body battery (2)	Sleep duration (3)
Intercept	28.27** (0.28)	71.68** (0.03)	8.43** (0.35)
Self-report stress WP	0.49** (0.07)		
Self-report stress BP	0.56 (0.30)		
Self-report tiredness WP		-1.55** (0.07)	
Self-report tiredness BP		-0.99** (0.36)	
Self-report sleep quality WP			0.33** (0.01)
Self-report sleep quality BP			0.12** (0.04)
Midday prompt	18.68** (0.16)	-14.58** (0.09)	
Afternoon prompt	20.40** (0.16)	-28.64** (0.10)	
Evening prompt	22.64** (0.16)	-38.23** (0.09)	
Random intercept <i>SD</i>	6.60	9.21	0.70
Random slope <i>SD</i>	1.29	1.62	0.21
<i>R</i> <sup>2</sup> (conditional)	0.22	0.42	0.28
<i>R</i> <sup>2</sup> (marginal)	0.13	0.30	0.07
RMSE	21.66	20.46	1.33

*Note.* Random intercept *SD* and random slope *SD* refer to the standard deviation of the random effects. Values in brackets below a point estimate denote the standard error. We do not report standard errors of variance components here, but their confidence intervals are available in the additional online materials. WP = within-person effect; BP = between-person effect; RMSE = root-mean-square error.  
\*\* *p* < .05.

associated with lower tiredness. As the sensor measure was assessed on the same scale as the sensor stress measure, this indicates a stronger relationship between self-report and sensor measures of tiredness compared to stress. The standardized effect of this relationship was

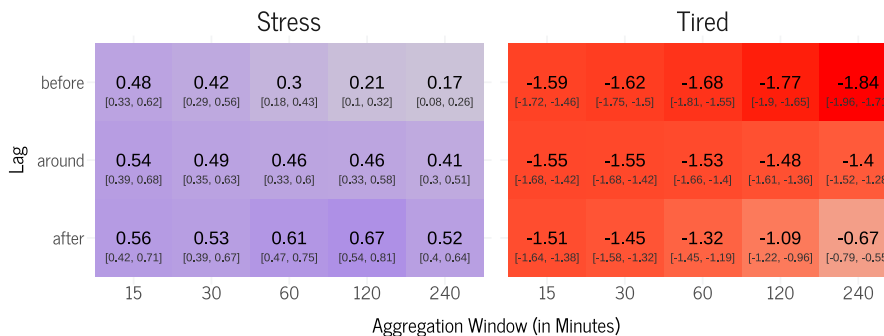
-0.082 (95% CI = [-0.087, -0.076]). Across the 15 different aggregation and lag types in our sensitivity analyses, we estimated a mean point estimate of  $\beta_1$  of -1.47 (*SD* = 0.28). The specific point estimates with their 95% CIs for each aggregation window and lag size are provided in Figure 3. There seems to be a tendency for associations to be the strongest when the sensor data are aggregated in a relatively long timeframe before the EMA prompt, and the weakest when sensor data are aggregated after the EMA prompt. We found a strong effect of the categorical predictor of time-of-day: body battery values were considerably lower at later times during the day. As indices of model fit, conditional and marginal *R*<sup>2</sup> values were estimated as .42 and .30 and—similar to the stress analyses—dropped considerably, to .20 and .05, without the inclusion of the time-of-day variable. The RMSE was 20.46.

The association remained mostly stable across secondary analyses. In secondary analysis S1, we found a significant interaction effect of the within-person effect with age (point estimate 0.04, 95% CI = [0.007, 0.074]). The interaction with age, albeit significant, is too small to interpret. Participants in cohorts two and four had stronger average within-person associations of the self-report and sensor measure, with average effects being ~40% larger compared to the first cohort. The interactions with gender and depression were not significant. Changing the residual correlation structure (S2) resulted in slightly higher within-person associations, but did not change the overall gist of the results. The binning of the outcome into the two categories provided by Garmin (“parasympathetic” and “sympathetic,” secondary analysis S3) resulted in a similarly modest association as our main analysis. Using alternative summary statistics of body battery during the aggregation window as the outcome (S4) led to considerably weaker (standard deviation as outcome) or roughly similar (maximum as outcome) results compared to the main analysis. Full results are again available in the additional online materials.

**Sleep**

We found a small, positive association between self-reported sleep quality and sensor-based sleep duration in our main analysis ( $\beta_1 = .33$ , 95% CI = [0.31, 0.35], see the rightmost column of

**Figure 3**  
*Fixed Effects of Stress and Tiredness Across Aggregation Windows and Lags*



*Note.* This figure contains the point estimates of the within-person association of self-report with sensor data (in other words,  $\beta_1$ ) across different lags and aggregation windows. For example, an aggregation of 120 min “before” means that the outcome of the regression was the mean of sensor data within the 2 hr before the EMA prompt. The brackets below the point estimates indicate the 95% CIs. Tiles are color-coded based on the (absolute) strength of the estimated relationship. EMA = ecological momentary assessment; CIs = confidence intervals. See the online article for the color version of this figure.

This document is copyrighted by the American Psychological Association or one of its allied publishers. This article is intended solely for the personal use of the individual user and is not to be disseminated broadly. All rights, including for text and data mining, AI training, and similar technologies, are reserved.

Table 3). This means that a difference of one point on the self-report response scale corresponds to a difference of around 20 min in sleep duration, meaning that if an individual reports higher sleep quality compared to their average, they also tend to have a longer sensor sleep duration. The standardized effect for this relationship was 0.26 (95% CI = [0.25, 0.28]). The standard deviation of the random slopes for sleep duration 0.21 (95% CI = [0.20, 0.23]) indicates a relatively large variability in within-person associations. The point estimate for the between-person association between self-report sleep quality and sensor sleep duration was 0.12 (95% CI = [0.05, 0.19]), indicating that individuals with higher average self-report sleep quality had a higher average sensor sleep duration. As sleep was assessed retrospectively in the morning and there is only one sensor value per night, the aggregation window and lag size do not play a role here. Conditional and marginal  $R^2$  were estimated as .28 and .065, while the RMSE was 1.33.

In our secondary analysis S1, we found no significant interaction of the within-person association with age, gender, or depression. Participants in cohorts two and four showed higher within-person associations between the sensor and self-report measure compared to the first cohort. Specifically, the average association was ~30% stronger in cohort two compared to cohort one. Point estimates of the association were virtually identical to the main analysis when specifying a different autocorrelation structure (S2). Full results are available in the additional online materials.

## Discussion

We analyzed the concurrent associations between self-report and wearable measures of stress, sleep, and tiredness in a large student sample. Overall, largely independent of preprocessing and modeling choices, we found very small associations for the stress measures, suggesting a lack of overlap between the sensor and self-report measures. For sleep and tiredness, we found stronger relationships (standardized point estimates of 0.26 and  $-0.082$ , respectively) with strong interindividual heterogeneity, suggesting that the sensor and self-report versions may tap into a somewhat similar construct.

Our secondary analyses indicate that age, gender, and depression were not meaningful moderators of the association between self-report and sensor measures. However, average within-person associations for stress were weaker in cohorts two and four whereas they were stronger for tiredness and sleep in the same cohorts. This is a surprising result, and we can only speculate about its causes. One potential factor may be that cohorts one and three were assessed in the winter (December through February), and cohorts two and four in the summer (June through August), implying substantial temperature differences in the Netherlands. While various plausible explanations come to mind—for example, differences in lifestyle or sweat production—we are unaware of any research into the topic. We do not find that sensor data quality systematically goes down with each cohort, implying that the smartwatches likely do not deteriorate meaningfully over the data collection time frame of around 2 years.

We found no evidence for depression as a moderator of the association between passive sensing and self-report data. Looking at the literature, Collins et al. (2025) have found that the discrepancy between both data types was larger for individuals with higher depressive symptoms, which they link to potential recall biases or memory deficits. Price et al. (2025) used actigraphy-based

prediction models for depressive symptoms in both a general population sample and a clinical sample (based on PHQ-9). Their results were mixed: they found slightly higher score accuracy of the prediction models in the clinical population, but not all of these predictions were significantly better than chance. Contrary to these previous studies, we investigated the moment-to-moment associations of passive sensor and self-report data. Our findings suggest no evidence for better or worse validity of passive sensor data in people with higher depressive symptoms. This finding may be partly explained by the weak overlap between passive sensor and self-report data and the large interindividual heterogeneity in associations.

Overall, our findings suggest that some features collected by commercial smartwatches may be useful for obtaining a multimodal insight into transdiagnostic constructs related to mental health. At the same time, the finding regarding stress suggests a lack of conceptual overlap between different data sources. We discuss these results in the following section for sleep, tiredness, and stress, and then highlight the implications of our finding for psychopathology research more broadly.

## Sleep

We observed a positive relationship between sensor sleep duration and self-report sleep quality in all analyses and for most individuals. This association was estimated to be positive for most individuals. Combined with the fact that the fixed effect estimate implies that a 1-point increase in self-reported sleep quality is associated with around 20 min of more sensor sleep, this indicates the potential utility of the wearable data for unobtrusive sleep tracking. This makes such data relevant for a wide range of mental health research, as sleep problems are part of many different diagnoses in the *DSM-5* (Forbes et al., 2024). Previous studies on the association between self-report sleep quality and sensor sleep duration found no evidence for an overlap between self-reported sleep quality and sensor sleep duration (Teo et al., 2019) or limited predictive power of sensor sleep duration for future sleep quality (Staples et al., 2017). However, these studies used different ways of assessing self-report and sensor sleep duration and had a considerably lower sample size. While we can expect sensor sleep duration and self-reported sleep quality to be related, they do not measure exactly the same construct, which may partially explain a lack of stronger associations in our sample.

## Tiredness

The results for the body battery sensor measure, which has not yet been evaluated against self-report data, indicate that it is weakly negatively related to momentary experiences of being tired. This result gives some preliminary validity evidence for researchers using pre-processed Garmin data. At the same time, person-specific estimates showed considerable heterogeneity and a lack of association between self-report and sensor measures for a substantial part of the sample. Also, the RMSE of 20.46 indicates, roughly speaking, that a typical difference between the observed and predicted value in our model is around 20. This corresponds to one-fifth of the full scale range, indicating large differences between the predictions of our model and observed values. For psychopathology research, this means that the body battery may be a somewhat useful proxy for fatigue, which in turn can be used to tap into the effects of

sleep problems. Future studies could try to show if body battery is more strongly related to other self-report measures and if it shows predictive validity for future mental health outcomes.

## Stress

The results for stress show an astonishing lack of overlap between the sensor and self-report measures. The association between self-report and sensor stress seems to be descriptively slightly lower when aggregating sensor stress before the prompt and in a larger aggregation window, but given the size of confidence intervals, it is unclear how robust this effect is. Overall, the lack of overlap was robust to preprocessing choices and other analysis strategies. There are multiple plausible explanations for this result.

First, sensor-based measures of stress rely on physiological signals of arousal. These signals are often valence nonspecific, meaning that excitement or positive arousal may also elicit an increase in heart rate and a decrease in heart rate variability which sensors may interpret as stress signals. Indeed, previous work has shown that such sensor-based measures relate to both negative and positive affect in daily life (Tutunji, Kogias, et al., 2023; van Halem et al., 2020).

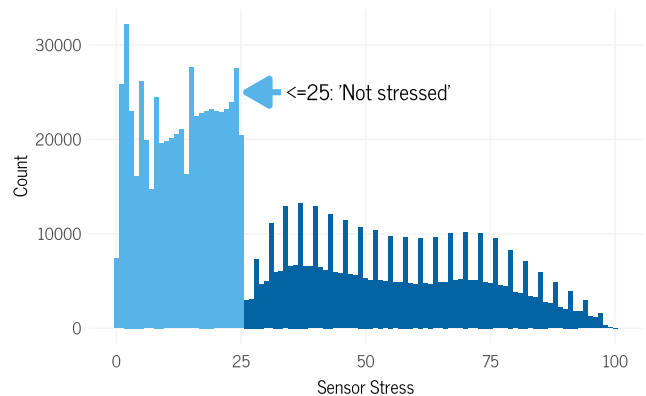
Second, there may be a so-called “semantic gap” (Das Swain et al., 2022) between the physiological and psychological concepts of stress. This means that physiological representations of what one might call stress and individual psychological definitions of stress may diverge, and they may not be able to measure the same underlying construct. This has also been the case for laboratory-based stress research using psychophysiology measures. In such studies, there is at times little overlap between the subjective component of stress following stress induction and the physiological measures acquired (Campbell & Ehlert, 2012). An investigation of postvaccination reactions by Guan et al. (2022) provides some preliminary evidence of such a semantic gap with the Garmin VivoSmart 4. In their study, wearable heart rate and stress measures were elevated for some time after vaccination, even in individuals who did not report side effects of the vaccine.

Third, another plausible explanation is measurement issues, potentially both in wearable and self-reported stress. The algorithm used to estimate individual stress scores is proprietary, and little information is available about its specifics. Given that the smartwatch used in this study is a commercial product, the stress score calculation may be optimized not only for high accuracy but also for a good user experience.

Analyses of the raw stress data indicate that the distribution of stress scores is not continuous, but rather thresholded at 25, which is the highest value in the “no stress” category. We illustrate the issue in Figure 4. Specifically, across all participants, around 56.7% of all stress scores fall into the interval between 0 and 25. We can only speculate that this may serve the purpose of avoiding indicating that individuals are “stressed” too often, but it is problematic from the perspective of accurately capturing the underlying phenomenon of stress. We have rarely seen similar figures on the raw distributions of passive sensing data in mental health research. Recent calls for more descriptive work in experience sampling studies (Siepe, Rieble, et al., 2025), therefore, also apply to passive sensing research.

At the same time, self-reported stress may also contain measurement issues, including potential measurement error (we only

**Figure 4**  
Distribution of all Sensor Stress Values



*Note.* This figure contains the absolute counts of specific sensor stress values. We randomly sampled one million stress values from all available sensor stress data across all four cohorts. Bars are colored based on Garmin’s classification into no stress ( $\leq 25$ ) and stress ( $>25$ ). See the online article for the color version of this figure.

measured stress with a single item) and recall biases. Furthermore, stress may be conceived of as an umbrella term for multiple distinct constructs (see, e.g., Crosswell & Lockwood, 2020; Goodday & Friend, 2019; Vaessen et al., 2021) whose heterogeneity cannot be fully captured by a single item. Interestingly, there is evidence that this applies to physiological markers of stress as well (Goodday & Friend, 2019). Finally, the limited variability and potential floor effects for the self-report stress item could have additionally impeded the possibility of finding associations with sensor data. Irrespective of the reasons for our results, they serve as cautionary reminders not to simply replace self-report with passive measures without proper validation.

These results do not imply that the sensor stress measure is not useful or in any way incorrect in the data set under consideration. If we follow the idea that there may be a semantic gap between self-report and wearable measures, wearable measures may still be useful for prediction because they may capture relevant information that is somewhat orthogonal to the self-report measures (Das Swain et al., 2022). Combining passive and self-report measures for the same or similar constructs, as many studies already do (Velozo et al., 2024), can provide both potential benefits for predictive modeling as well as avenues for research into construct validation. For example, studying associations between both sources of data across a range of aggregation windows and time lags can provide insight into how individuals aggregate their experiences when responding to self-report prompts (see Leertouwer et al., 2021, for similar comparisons of different types of self-report data). We believe that such work, in combination with theoretical work on the concepts one intends to measure (Bringmann et al., 2022), will advance our ability to understand and predict mental health. If the field of digital phenotyping wants to fulfill its promises, it should, therefore, not neglect to build a strong foundation of construct validation and measurement work. Until we have reached a better understanding of the relationship between passive sensor measures and self-reported variables, researchers should be cautious not to assume that these different data sources necessarily measure similar constructs.

## Limitations

For the constructs assessed by wearables, we used data that were internally preprocessed by the Garmin watch, which is sometimes called “proxy” data (Velozo et al., 2024). Using data processed via openly available algorithms could improve the transparency and validity of results (Velozo et al., 2024) and enhance our understanding of the role which specific physiological markers play in mental health science (Weber et al., 2022). At the same time, we need to face the reality that commercial-grade sensors in mental health research are likely here to stay. They come with simple data management, are relatively inexpensive, and are easy to use and wear in everyday life. Beyond that, they are already worn by millions of customers in their everyday lives. Therefore, we believe that future research into the statistical properties and validity of these commercial sensors is and will continue to be worthwhile.

While the large amount of data is a distinct strength of our study, there was a substantial proportion of individuals with a relatively low compliance. Although we aimed to follow state-of-the-art recommendations in the methodological literature and explored various modeling approaches, we could have modified and extended our modeling strategy in various ways. For example, we assumed heterogeneity in residual variances between individuals (see Nestler, 2024, for an alternative), used observed instead of latent person-specific means for centering (see Hamaker & Muthén, 2020, for an alternative), and did not explore more flexible functional forms of the association between predictors and outcomes, such as polynomials or machine learning approaches. We did not preregister our exploratory analyses but tried to account for plausible alternative analysis pipelines via sensitivity and secondary analyses. Future studies could use preregistration protocols (Langener, Siepe, et al., 2024) to test more precise and specific hypotheses about digital phenotyping data.

## Conclusions

In this study, we investigated the overlap between wearable and self-report measures purportedly measuring the same or similar constructs. We found a robust association between self-reported sleep quality and sensor sleep duration, and a smaller association between self-report tiredness and sensor “body battery.” Wearable and self-report measures of stress showed a lack of overlap for most individuals, likely highlighting both differences in the construct that is supposed to be assessed as well as potential measurement issues. Age, gender, and depressive symptomatology were no meaningful moderators of these associations. We are convinced that further work on measurement in the digital phenotyping literature is crucial for clinical utility and will move the field forward.

## References

Adams, Z. W., McClure, E. A., Gray, K. M., Danielson, C. K., Treiber, F. A., & Ruggiero, K. J. (2017). Mobile devices for the remote acquisition of physiological and behavioral biomarkers in psychiatric clinical research. *Journal of Psychiatric Research, 85*, 1–14. <https://doi.org/10.1016/j.jpsychires.2016.10.019>

Adler, D. A., Wang, F., Mohr, D. C., & Choudhury, T. (2022). Machine learning for passive mental health symptom prediction: Generalization across different longitudinal mobile sensing studies. *PLoS ONE, 17*(4), Article e0266516. <https://doi.org/10.1371/journal.pone.0266516>

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative

research in psychology: The APA publications and communications board task force report. *American Psychologist, 73*(1), 3–25. <https://doi.org/10.1037/amp0000191>

Auerbach, R. P., Alonso, J., Axinn, W. G., Cuijpers, P., Ebert, D. D., Green, J. G., Hwang, I., Kessler, R. C., Liu, H., Mortier, P., Nock, M. K., Pinder-Amaker, S., Sampson, N. A., Aguilar-Gaxiola, S., Al-Hamzawi, A., Andrade, L. H., Benjet, C., Caldas-de-Almeida, J. M., Demyttenaere, K., ... Bruffaerts, R. (2016). Mental disorders among college students in the World Health Organization world mental health surveys. *Psychological Medicine, 46*(14), 2955–2970. <https://doi.org/10.1017/S0033291716001665>

Baglioni, C., Nanovska, S., Regen, W., Spiegelhalter, K., Feige, B., Nissen, C., Reynolds, C. F., & Riemann, D. (2016). Sleep and mental disorders: A meta-analysis of polysomnographic research. *Psychological Bulletin, 142*(9), 969–990. <https://doi.org/10.1037/bul0000053>

Berrios, G. E. (1990). Feelings of fatigue and psychopathology: A conceptual history. *Comprehensive Psychiatry, 31*(2), 140–151. [https://doi.org/10.1016/0010-440X\(90\)90018-N](https://doi.org/10.1016/0010-440X(90)90018-N)

Bertz, J. W., Epstein, D. H., & Preston, K. L. (2018). Combining ecological momentary assessment with objective, ambulatory measures of behavior and physiology in substance-use research. *Addictive Behaviors, 83*, 5–17. <https://doi.org/10.1016/j.addbeh.2017.11.027>

Bringmann, L. F., Elmer, T., & Eronen, M. I. (2022). Back to basics: The importance of conceptual clarification in psychological science. *Current Directions in Psychological Science, 31*(4), 340–346. <https://doi.org/10.1177/09637214221096485>

Bufano, P., Laurino, M., Said, S., Tognetti, A., & Menicucci, D. (2023). Digital phenotyping for monitoring mental disorders: Systematic review. *Journal of Medical Internet Research, 25*(1), Article e46778. <https://doi.org/10.2196/46778>

Campbell, J., & Ehlert, U. (2012). Acute psychosocial stress: Does the -emotional stress response correspond with physiological responses? *Psychoneuroendocrinology, 37*(8), 1111–1134. <https://doi.org/10.1016/j.psyneuen.2011.12.010>

Chancellor, S., & De Choudhury, M. (2020). Methods in predictive techniques for mental health status on social media: A critical review. *NPJ Digital Medicine, 3*(1), Article 43. <https://doi.org/10.1038/s41746-020-0233-7>

Collins, A. C., Price, G. D., Moreno, V. A., Mackin, D. M., Oh, J. Y., Heinz, M. V., & Jacobson, N. C. (2025). A comparison of objective and subjective measures of physical activity, sedentary and sleep behaviors between persons with and without depressive symptoms. *Journal of Affective Disorders, 375*, 22–26. <https://doi.org/10.1016/j.jad.2025.01.092>

Coulombe, J. A., Reid, G. J., Boyle, M. H., & Racine, Y. (2010). Sleep problems, tiredness, and psychological symptoms among healthy adolescents. *Journal of Pediatric Psychology, 36*(1), 25–35. <https://doi.org/10.1093/jpepsy/jsq028>

Crosswell, A. D., & Lockwood, K. G. (2020). Best practices for stress measurement: How to measure psychological stress in health research. *Health Psychology Open, 7*(2), Article 2055102920933072. <https://doi.org/10.1177/2055102920933072>

Dang, J., King, K. M., & Inzlicht, M. (2020). Why are self-report and behavioral measures weakly correlated? *Trends in Cognitive Sciences, 24*(4), 267–269. <https://doi.org/10.1016/j.tics.2020.01.007>

Daniels, K., Vonck, S., Robijns, J., Spooren, A., Hansen, D., & Bonnechère, B. (2024). *Unveiling the digital phenotype: A protocol for a prospective study on physical activity behavior in community-dwelling older adults*. <https://doi.org/10.21203/rs.3.rs-3896647/v1>

Das Swain, V., Chen, V., Mishra, S., Mattingly, S. M., Abowd, G. D., & De Choudhury, M. (2022). *Semantic gap in predicting mental wellbeing through passive sensing*. CHI Conference on Human Factors in Computing Systems (pp. 1–16). <https://doi.org/10.1145/3491102.3502037>

- Davidson, B. I. (2022). The crossroads of digital phenotyping. *General Hospital Psychiatry*, 74, 126–132. <https://doi.org/10.1016/j.genhosppsych.2020.11.009>
- De Angel, V., Lewis, S., White, K., Oetzmann, C., Leightley, D., Oprea, E., Lavelle, G., Matcham, F., Pace, A., Mohr, D. C., Dobson, R., & Hotopf, M. (2022). Digital health tools for the passive monitoring of depression: A systematic review of methods. *NPJ Digital Medicine*, 5(1), Article 3. <https://doi.org/10.1038/s41746-021-00548-8>
- Ebner-Priemer, U., & Santangelo, P. (2020). Digital phenotyping: Hype or hope? *The Lancet Psychiatry*, 7(4), 297–299. [https://doi.org/10.1016/S2215-0366\(19\)30380-3](https://doi.org/10.1016/S2215-0366(19)30380-3)
- Firstbeat Technologies Ltd. (2014, November 4). *Stress and recovery analysis method based on 24-hour heart rate variability* [White paper]. Retrieved May 26, 2024, from [https://assets.firstbeat.com/firstbeat/uploads/2015/11/Stress-and-recovery\\_whitepaper\\_20145.pdf](https://assets.firstbeat.com/firstbeat/uploads/2015/11/Stress-and-recovery_whitepaper_20145.pdf)
- Forbes, M. K., Neo, B., Nezami, O. M., Fried, E. I., Faure, K., Michelsen, B., Twose, M., & Dras, M. (2024). Elemental psychopathology: Distilling constituent symptoms and patterns of repetition in the diagnostic criteria of the DSM-5. *Psychological Medicine*, 54(5), 886–894. <https://doi.org/10.1017/S0033291723002544>
- Freeman, D., Sheaves, B., Waite, F., Harvey, A. G., & Harrison, P. J. (2020). Sleep disturbance and psychiatric disorders. *The Lancet Psychiatry*, 7(7), 628–637. [https://doi.org/10.1016/S2215-0366\(20\)30136-X](https://doi.org/10.1016/S2215-0366(20)30136-X)
- 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>
- Garmin. (2024a). *Body battery*. Garmin Homepage. Retrieved May 26, 2024, from <https://www.garmin.com/en-US/garmin-technology/health-science/body-battery/>
- Garmin. (2024b). *Stress tracking*. Garmin Homepage. Retrieved May 26, 2024, from <https://www.garmin.com/en-US/garmin-technology/health-science/stress-tracking/>
- Goodyay, S. M., & Friend, S. (2019). Unlocking stress and forecasting its consequences with digital technology. *npj Digital Medicine*, 2(1), Article 75. <https://doi.org/10.1038/s41746-019-0151-8>
- Guan, G., Mofaz, M., Qian, G., Patalon, T., Shmueli, E., Yamin, D., & Brandeau, M. L. (2022). Higher sensitivity monitoring of reactions to COVID-19 vaccination using smartwatches. *npj Digital Medicine*, 5(1), Article 140. <https://doi.org/10.1038/s41746-022-00683-w>
- Hamaker, E. L., & Muthén, B. (2020). The fixed versus random effects debate and how it relates to centering in multilevel modeling. *Psychological Methods*, 25(3), 365–379. <https://doi.org/10.1037/met0000239>
- 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>
- Hammen, C. (2005). Stress and depression. *Annual Review of Clinical Psychology*, 1(1), 293–319. <https://doi.org/10.1146/annurev.clinpsy.1.102803.143938>
- Harvey, A. G., Murray, G., Chandler, R. A., & Soehner, A. (2011). Sleep disturbance as transdiagnostic: Consideration of neurobiological mechanisms. *Clinical Psychology Review*, 31(2), 225–235. <https://doi.org/10.1016/j.cpr.2010.04.003>
- Hehlmann, M. I., Schwartz, B., Lutz, T., Gómez Penedo, J. M., Rubel, J. A., & Lutz, W. (2021). The use of digitally assessed stress levels to model change processes in CBT—A feasibility study on seven case examples. *Frontiers in Psychiatry*, 12, Article 613085. <https://doi.org/10.3389/fpsy.2021.613085>
- Huckvale, K., Venkatesh, S., & Christensen, H. (2019). Toward clinical digital phenotyping: A timely opportunity to consider purpose, quality, and safety. *NPJ Digital Medicine*, 2(1), Article 88. <https://doi.org/10.1038/s41746-019-0166-1>
- Insel, T. R. (2018). Digital phenotyping: A global tool for psychiatry. *World Psychiatry*, 17(3), 276–277. <https://doi.org/10.1002/wps.20550>
- Jürgens, P., Stark, B., & Magin, M. (2020). Two half-truths make a whole? On bias in self-reports and tracking data. *Social Science Computer Review*, 38(5), 600–615. <https://doi.org/10.1177/0894439319831643>
- Kay, M. (2024). *ggdist: Visualizations of distributions and uncertainty* (R Package Version 3.3.2). <https://doi.org/10.5281/zenodo.3879620>
- 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>
- Langener, A. M., Siepe, B. S., Elsherif, M., Niemeijer, K., Andresen, P. K., Akre, S., Bringmann, L., Cohen, Z. D., Choukas, N. R., Drexler, K., Fassi, L., Green, J., Hoffmann, T., Kas, M., Kurten, S., Schoedel, R., Stulp, G., Turner, G., & Jacobson, N. C. (2024, March 19). *A template and tutorial for preregistering studies using passive smartphone measures*. <https://doi.org/10.31234/osf.io/p4xf8>
- Langener, A. M., Stulp, G., Jacobson, N. C., Costanzo, A., Jagesar, R. R., Kas, M. J., & Bringmann, L. F. (2024). It's all about timing: Exploring different temporal resolutions for analyzing digital-phenotyping data. *Advances in Methods and Practices in Psychological Science*, 7(1), Article 25152459231202677. <https://doi.org/10.1177/25152459231202677>
- 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 egocentric networks: Scoping review. *JMIR Mental Health*, 10(1), Article e42646. <https://doi.org/10.2196/42646>
- Lauderdale, D. S., Knutson, K. L., Yan, L. L., Liu, K., & Rathouz, P. J. (2008). Self-reported and measured sleep duration: How similar are they? *Epidemiology (Cambridge, MA)*, 19(6), 838–845. <https://doi.org/10.1097/EDE.0b013e318187a7b0>
- Leertouwer, I. J., Cramer, A. O. J., Vermunt, J. K., & Schuurman, N. K. (2021). A review of explicit and implicit assumptions when providing personalized feedback based on self-report EMA data. *Frontiers in Psychology*, 12, Article 764526. <https://doi.org/10.3389/fpsyg.2021.764526>
- Lüdecke, D., Ben-Shachar, M. S., Patil, I., Waggoner, P., & Makowski, D. (2021). Performance: An R package for assessment, comparison and testing of statistical models. *Journal of Open Source Software*, 6(60), Article 3139. <https://doi.org/10.21105/joss.03139>
- Mahalingham, T., McEvoy, P. M., & Clarke, P. J. (2023). Assessing the validity of self-report social media use: Evidence of no relationship with objective smartphone use. *Computers in Human Behavior*, 140, Article 107567. <https://doi.org/10.1016/j.chb.2022.107567>
- Melcher, J., Hays, R., & Torous, J. (2020). Digital phenotyping for mental health of college students: A clinical review. *Evidence Based Mental Health*, 23(4), 161–166. <https://doi.org/10.1136/ebmental-2020-300180>
- 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>
- Mill, A., Realo, A., & Allik, J. (2016). Emotional variability predicts tiredness in daily life. *Journal of Individual Differences*, 37(3), 181–193. <https://doi.org/10.1027/1614-0001/a000206>
- Mohr, D. C., Zhang, M., & Schueller, S. M. (2017). Personal sensing: Understanding mental health using ubiquitous sensors and machine learning. *Annual Review of Clinical Psychology*, 13(1), 23–47. <https://doi.org/10.1146/annurev-clinpsy-032816-044949>
- Myin-Germeys, I., & Kuppens, P. (Eds.). (2022). *The open handbook of experience sampling methodology: A step-by-step guide to designing, conducting, and analyzing ESM studies* (2nd ed.). Center for Research on Experience Sampling and Ambulatory Methods.
- Nadal, C., Earley, C., Enrique, A., Viganò, N., Sas, C., Richards, D., & Doherty, G. (2021). Integration of a smartwatch within an internet-delivered intervention for depression: Protocol for a feasibility randomized

- controlled trial on acceptance. *Contemporary Clinical Trials*, 103, Article 106323. <https://doi.org/10.1016/j.cct.2021.106323>
- Nakagawa, S., Johnson, P. C., & Schielzeth, H. (2017). The coefficient of determination  $R^2$  and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of the Royal Society Interface*, 14(134), Article 20170213. <https://doi.org/10.1098/rsif.2017.0213>
- Nelson, B. W., Low, C. A., Jacobson, N., Areán, P., Torous, J., & Allen, N. B. (2020). Guidelines for wrist-worn consumer wearable assessment of heart rate in biobehavioral research. *NPJ Digital Medicine*, 3(1), Article 90. <https://doi.org/10.1038/s41746-020-0297-4>
- Nestler, S. (2024). A mixed-effects model in which the parameters of the auto-correlated error structure can differ between individuals. *Multivariate Behavioral Research*, 59(1), 98–109. <https://doi.org/10.1080/00273171.2023.2217418>
- Niemeijer, K., Mestdagh, M., & Kuppens, P. (2022). Tracking subjective sleep quality and mood with mobile sensing: Multiverse study. *Journal of Medical Internet Research*, 24(3), Article e25643. <https://doi.org/10.2196/25643>
- Onnela, J.-P. (2021). Opportunities and challenges in the collection and analysis of digital phenotyping data. *Neuropsychopharmacology*, 46(1), 45–54. <https://doi.org/10.1038/s41386-020-0771-3>
- Parry, D. A., Davidson, B. I., Sewall, C. J., Fisher, J. T., Mieczkowski, H., & Quintana, D. S. (2021). A systematic review and meta-analysis of discrepancies between logged and self-reported digital media use. *Nature Human Behaviour*, 5(11), 1535–1547. <https://doi.org/10.1038/s41562-021-01117-5>
- Pinheiro, J., & Bates, D., & R Core Team. (2023). *nlme: Linear and nonlinear mixed effects models* (R Package Version 3.1-164). <https://CRAN.R-project.org/package=nlme>
- Price, G. D., Collins, A. C., Mackin, D. M., Heinz, M. V., & Jacobson, N. C. (2025). Use of passively collected actigraphy data to detect individual depressive symptoms in a clinical subpopulation and a general population. *Journal of Psychopathology and Clinical Science*, 134(1), 31–40. <https://doi.org/10.1037/abn0000933>
- Price, G. D., Heinz, M. V., Song, S. H., Nemesure, M. D., & Jacobson, N. C. (2023). Using digital phenotyping to capture depression symptom variability: Detecting naturalistic variability in depression symptoms across one year using passively collected wearable movement and sleep data. *Translational Psychiatry*, 13(1), Article 381. <https://doi.org/10.1038/s41398-023-02669-y>
- Prince, S. A., Adamo, K. B., Hamel, M. E., Hardt, J., Gorber, S. C., & Tremblay, M. (2008). A comparison of direct versus self-report measures for assessing physical activity in adults: A systematic review. *International Journal of Behavioral Nutrition and Physical Activity*, 5(1), Article 56. <https://doi.org/10.1186/1479-5868-5-56>
- R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Rao, C., DiLascio, E., Demanse, D., Marshall, N., Sopala, M., & De Luca, V. (2023). Association of digital measures and self-reported fatigue: A remote observational study in healthy participants and participants with chronic inflammatory rheumatic disease. *Frontiers in Digital Health*, 5, Article 1099456. <https://doi.org/10.3389/fdgh.2023.1099456>
- Ringwald, W. R., King, G., Vize, C., & Wright, A. G. C. (2025, March 6). *Advancing psychiatry with transdiagnostic targets for digital phenotyping*. [https://doi.org/10.31234/osf.io/3tyjp\\_v2](https://doi.org/10.31234/osf.io/3tyjp_v2)
- Schyvens, A.-M., Oost, N. C. V., Aerts, J.-M., Masci, F., Peters, B., Neven, A., Dirix, H., Wets, G., Ross, V., & Verbraecken, J. (2024). Accuracy of fitbit charge 4, Garmin VivoSmart 4, and WHOOP versus polysomnography: Systematic review. *JMIR mHealth and uHealth*, 12(1), Article e52192. <https://doi.org/10.2196/52192>
- Shochat, T., Cohen-Zion, M., & Tzischinsky, O. (2014). Functional consequences of inadequate sleep in adolescents: A systematic review. *Sleep Medicine Reviews*, 18(1), 75–87. <https://doi.org/10.1016/j.smrv.2013.03.005>
- Siepe, B. S., Rieble, C. L., Tutunji, R., Rimpler, A., März, J., Proppert, R. K. K., & Fried, E. I. (2025). Understanding ecological-momentary-assessment data: A tutorial on exploring item performance in ecological-momentary-assessment data. *Advances in Methods and Practices in Psychological Science*, 8(1), Article 25152459241286877. <https://doi.org/10.1177/25152459241286877>
- Siepe, B. S., Tutunji, R., Rieble, C., Proppert, R. K. K., & Fried, E. I. (2025, January). *Online supplement for: Associations between ecological momentary assessment and passive sensor data in a large student sample*. <https://osf.io/txgnq/>
- Sinha, R. (2001). How does stress increase risk of drug abuse and relapse? *Psychopharmacology*, 158(4), 343–359. <https://doi.org/10.1007/s002130100917>
- Smets, E., Rios Velazquez, E., Schiavone, G., Chakroun, I., D'Hondt, E., De Raedt, W., Cornelis, J., Janssens, O., Van Hoecke, S., Claes, S., Van Diest, I., & Hoof, C. V. (2018). Large-scale wearable data reveal digital phenotypes for daily-life stress detection. *NPJ Digital Medicine*, 1(1), Article 67. <https://doi.org/10.1038/s41746-018-0074-9>
- Staples, P., Torous, J., Barnett, I., Carlson, K., Sandoval, L., Keshavan, M., & Onnela, J.-P. (2017). A comparison of passive and active estimates of sleep in a cohort with schizophrenia. *NPJ Schizophrenia*, 3(1), Article 37. <https://doi.org/10.1038/s41537-017-0038-0>
- Strauss, G. P., Raugh, I. M., Zhang, L., Luther, L., Chapman, H. C., Allen, D. N., Kirkpatrick, B., & Cohen, A. S. (2022). Validation of accelerometry as a digital phenotyping measure of negative symptoms in schizophrenia. *Schizophrenia*, 8(1), Article 37. <https://doi.org/10.1038/s41537-022-00241-z>
- Tackett, J. L., Lilienfeld, S. O., Patrick, C. J., Johnson, S. L., Krueger, R. F., Miller, J. D., Oltmanns, T. F., & Shrout, P. E. (2017). It's time to broaden the replicability conversation: Thoughts for and from clinical psychological science. *Perspectives on Psychological Science*, 12(5), 742–756. <https://doi.org/10.1177/1745691617690042>
- Teo, J. X., Davila, S., Yang, C., Hii, A. A., Pua, C. J., Yap, J., Tan, S. Y., Sahlén, A., Chin, C. W.-L., Teh, B. T., Cook, S. A., Yeo, K. K., Tan, P., Lim, W. K., & Rozen, S. G. (2019). Digital phenotyping by consumer wearables identifies sleep-associated markers of cardiovascular disease risk and biological aging. *Communications Biology*, 2(1), Article 361. <https://doi.org/10.1038/s42003-019-0605-1>
- Triantafyllou, S., Saeb, S., Lattie, E. G., Mohr, D. C., & Kording, K. P. (2019). Relationship between sleep quality and mood: Ecological momentary assessment study. *JMIR Mental Health*, 6(3), Article e12613. <https://doi.org/10.2196/12613>
- Tutunji, R., Kogias, N., Kapteijns, B., Krentz, M., Krause, F., Vassena, E., & Hermans, E. J. (2023). Detecting prolonged stress in real life using wearable biosensors and ecological momentary assessments: Naturalistic experimental study. *Journal of Medical Internet Research*, 25, Article e39995. <https://doi.org/10.2196/39995>
- Tutunji, R., Proppert, R. K. K., Rieble, C. L., & Fried, E. I. (2023, December 22). *Defining a generic holdout sample for combined exploratory and predictive analyses in the WARN-D dataset*. <https://doi.org/10.17605/OSF.IO/W9NXY>
- Ushey, K., & Wickham, H. (2024). *Renv: Project environments* (R Package Version 1.0.7). <https://CRAN.R-project.org/package=renv>
- Vaessen, T., Rintala, A., Otsabryk, N., Viechtbauer, W., Wampers, M., Claes, S., & Myin-Germeys, I. (2021). The association between self-reported stress and cardiovascular measures in daily life: A systematic review. *PLoS ONE*, 16(11), Article e0259557. <https://doi.org/10.1371/journal.pone.0259557>
- van Halem, S., Van Roekel, E., Kroencke, L., Kuper, N., & Denissen, J. (2020). Moments that matter? On the complexity of using triggers based on skin conductance to sample arousing events within an experience sampling framework. *European Journal of Personality*, 34(5), 794–807. <https://doi.org/10.1002/per.2252>

- van Lier, H. G., Pieterse, M. E., Garde, A., Postel, M. G., de Haan, H. A., Vollenbroek-Hutten, M. M., Schraagen, J. M., & Noordzij, M. L. (2020). A standardized validity assessment protocol for physiological signals from wearable technology: Methodological underpinnings and an application to the E4 biosensor. *Behavior Research Methods*, *52*(2), 607–629. <https://doi.org/10.3758/s13428-019-01263-9>
- Veloze, J. D. C., Habets, J., George, S. V., Niemeijer, K., Minaeva, O., Hagemann, N., Herff, C., Kuppens, P., Rintala, A., Vaessen, T., Riese, H., & Delespaul, P. (2024). Designing daily-life research combining experience sampling method with parallel data. *Psychological Medicine*, *54*(1), 98–107. <https://doi.org/10.1017/S0033291722002367>
- Vize, C. E., Kaurin, A., & Wright, A. G. (2024). Personality pathology and momentary stress processes. *Clinical Psychological Science*, *12*(4), 686–705. <https://doi.org/10.1177/21677026231192483>
- Wang, L. P., & Maxwell, S. E. (2015). On disaggregating between-person and within-person effects with longitudinal data using multilevel models. *Psychological Methods*, *20*(1), 63–83. <https://doi.org/10.1037/met0000030>
- Weber, J., Angerer, P., & Apolinário-Hagen, J. (2022). Physiological reactions to acute stressors and subjective stress during daily life: A systematic review on ecological momentary assessment (EMA) studies. *PLoS ONE*, *17*(7), Article e0271996. <https://doi.org/10.1371/journal.pone.0271996>
- Weermeijer, J., Lafit, G., Kiekens, G., Wampers, M., Eisele, G., Kasanova, Z., Vaessen, T., Kuppens, P., & Myin-Germeys, I. (2022). Applying multiverse analysis to experience sampling data: Investigating whether preprocessing choices affect robustness of conclusions. *Behavior Research Methods*, *54*(6), 2981–2992. <https://doi.org/10.3758/s13428-021-01777-1>
- Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag.

Received July 1, 2024

Revision received March 12, 2025

Accepted April 15, 2025 ■