bioRxiv preprint doi: https://doi.org/10.1101/2021.06.29.450360; this version posted July 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Using wearable biosensors and ecological momentary assessments for the detection of prolonged stress in real life

Short Title: Detecting Stress with Wearable Biosensors

Authors

Rayyan Tutunji[1*], Nikos Kogias[1], Bob Kapteijns[1], Martin Krentz[1], Florian Krause[1], Eliana Vassena[1,2], Erno Hermans[1]

Affiliations

1. Donders Institute for Brain, Cognition and Behaviour, Radboud University Medical Center, Nijmegen, The Netherlands.
2. Behavioural Science Institute, Radboud University, Nijmegen, The Netherlands

*Corresponding Author

Rayyan Tutunji, rayyan.tutunji@donders.ru.nl

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.29.450360; this version posted July 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Abstract

Emerging efforts toward prevention of stress-related mental disorders have created a need for unobtrusive real-life monitoring of stress-related symptoms. We used ecological momentary assessments (EMA) combined with wearable biosensors to investigate whether these can be used to detect periods of prolonged stress. During stressful high-stake exam (versus control) weeks, participants reported increased negative affect and decreased positive affect. Intriguingly, physiological arousal was decreased on average during the exam week. Time-resolved analyses revealed peaks in physiological arousal associated with both self-reported stress and self-reported positive affect, while the overall decrease in physiological arousal was mediated by lower positive affect during the stress period. We then used machine learning to show that a combination of EMA and physiology yields optimal classification of week types. Our findings highlight the potential of wearable biosensors in stress-related mental-health monitoring, but critically show that psychological context is essential for interpreting physiological arousal detected using these devices.

Teaser

Smartwatches combined with daily diaries of mood can detect stress periods using individualized machine learning models.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.29.450360; this version posted July 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Introduction

Stress-related mental disorders such as major depression and anxiety disorders have gained increased recognition in the public eye. While a vast body of research exists regarding these disorders, studies have mostly focused on retrospective assessments of individuals who are already afflicted with these conditions. More recently, an increased interest has emerged in determining what makes some individuals more resilient to developing these disorders than others (1–4). Investigating resilience requires investigation of individual variation in stress reactivity prior to the development of psychological illness(1). A driving force behind this approach is the need to establish early warning signs of subsequent onset of stress-related disorders. Early interventions are known to improve psychological outcomes in patients(5), and reduce the economic burden of psychiatric illness on society(6). The ability to unobtrusively detect states of stress in daily life would enable early ecological interventions in those at risk, by either flagging risk states to health-care providers, or by delivering in-the-moment personalized interventions during these periods(7, 8).

Previous attempts looking at daily-life stress have used Experience Sample Methods (ESM(9), also known as Ecological Momentary Assessments or EMA)(10) to derive ecologically valid experiences of stress in daily life. These paradigms use repeated questionnaire assessments ("beeps") in the daily life of individuals to gain a better understanding of various psychological processes such as addiction(11), interpersonal relationships(12), and stress reactivity(13–15). Studies using these methods in stress and stress-related disorders have identified specific behavioral patterns in everyday life that may explain, or in some instances predict onset of psychiatric illness(16, 17). Such studies have also given insight into associations between the impact of stress exposure on affect, showing the effects of stress on positive and negative mood and its links to depression(18, 19). While these studies have provided insight into the dynamics of disease and behavior in daily life, they are often intrusive and require active participation of clients or patients. Additionally, extensive longitudinal sampling may not be feasible for all psychiatric populations(20). Relying on subjective measures may inadvertently result in unreliable data due to

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.29.450360; this version posted July 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

careless responses, or a participants' lower insight into symptoms and states associated(21). Furthermore, the sparse sampling of subjective states using EMA may miss the time windows in which stress responses occur. This has led to a growing interest in establishing adequate passive and ambulatory mental-health monitoring that may be more reliably used in a general population.

The emergence of widely accessible wearable biosensors has raised the question whether these devices can be used for ecological *physiological* assessments (EPA), either as an add-on or as an alternative to EMA, in mental health monitoring. While measuring stress-hormone reactivity in daily life remains very difficult(22–24), wearable biosensors offer continuous recording of autonomic physiological markers such as skin conductance (SC) and heart rate (HR). These measures have been extensively validated in laboratory-based studies using controlled stress-induction protocols(25), showing increased HR and SC and decreased HR variability in response to stressors(26, 27). Changes in SC and HR have also been associated with increased psychological stress(28). Notably, however, these autonomic physiological parameters are associated with general arousal(29), and this includes high-arousal states of *positive* affect as well(27). This indicates that the use of EPA may be more complicated in daily life than in the lab: While acute stress may trigger arousal, arousal itself may not necessarily signal the presence of acute stress.

Although autonomic physiological responses have been extensively studied in the lab and linked to stress and arousal, their links to stressors in real life are not well understood. Few attempts have been made at investigating the physiology of stress in daily life, mostly using fixed scenarios such as driving(30), or using burdensome equipment, such as ECG belts, that may be difficult to apply in the general population(31). An important study using large-scale wearable data using ECG and wrist-worn devices by Smets and colleagues (2018) showed that these findings could be replicated when classifying individual stress levels in real life as high, moderate, and low(32). However, one limitation of this study is the lack of environmental stressors, and the underlying assumption that subjective stress measures can be taken as the "ground truth". Indeed, the overall reports of stressed states in this study were relatively low when

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.29.450360; this version posted July 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

compared to the non-stress states. Additionally, while understanding the change of physiological arousal related to single time experiences of stressors is important, it is the change in the measures related to the accumulation of stress over a prolonged period that may be more important in the context of mental health and resilience.

In the current study, we therefore aimed to investigate the validity of passive EPA monitoring of physiological arousal and active EMA measures to detect prolonged stress exposure. We investigated a population of first-year medical and biomedical students, who have been shown to experience increased psychological distress relative to peers in other programs(*33*). Participants underwent two weeks with combined EMA and EPA assessments, one prior to a high-stake examination (i.e., stress week) and the other during a regular period (i.e., control week). Participants answered questions regarding subjective stress in the EMA questionnaires, including event-related stress (i.e., most prominent event between surveys), activity-related (i.e., current activity), social-related stress (i.e., social context), and physical stress (i.e., physical discomfort). Additional outcome measures were recorded from both the EMA and EPA including mood (positive and negative), and autonomic arousal measures of HR and SC (Figure 1).

We first validated our protocol by examining differences in subjective stress measures from EMA between the weeks. We then assessed the impact of prolonged stress exposure on mood and physiology outcomes. Finally, we used individualized machine-learning models to classify per time point (beep) whether participants were in a stress or control week, using either mood measures, physiological measures, or a combination of both. In addition to confirming increased subjective stress in the exam week, we predicted that there would be a shift in both physiological (increased autonomic physiological responses) and mood (more negative affect) outcomes as a function of stressful experiences. Based on previous findings, we expected that both EPA and EMA measures would be successful in classifying prolonged stress states. We also expected that models combining both sources of data (EPA+EMA) would outperform either of the previous models.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.29.450360; this version posted July 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

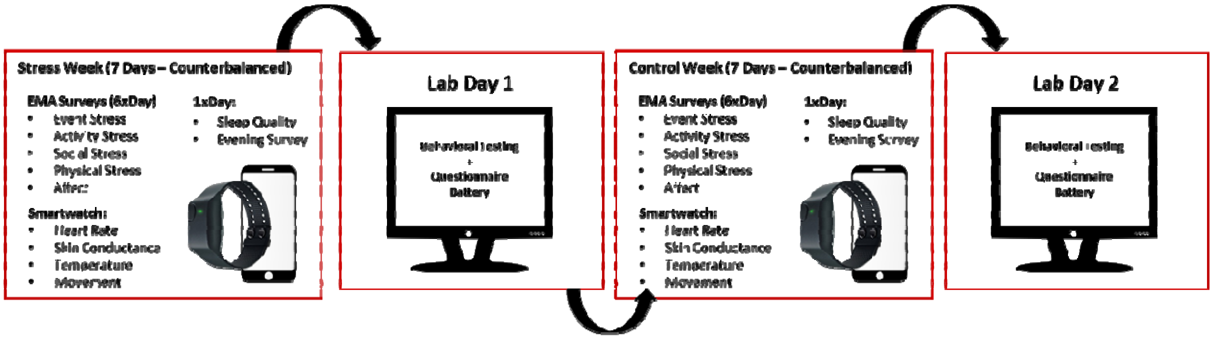


Fig 1. Study Timeline. Diagram portraying sequence of participation in the study with counterbalanced weeks

Results

Examination periods are associated with increased self-reported stress.

We first validated our stress manipulation by examining if there were overall differences in the subjective stress measures between the control and stress weeks using generalized linear mixed effects models with the different types of EMA subjective stress measures as dependent variables (i.e., event stress, activity stress, social stress, and physical stress), week type as a fixed effect (stress or control), and subject as random effect in a single model. We additionally controlled for exercise, movement, sex, study program, and modeled the survey instance (i.e., time of day of beeps), day, and week order (stress or control week first) to control for potential confounds in the data. We found a significant increase in prominent stressful events (i.e., event-related stress in periods between beeps, β=0.31, 95%CI [0.18, 0.43], std. error=0.06, t-stat=4.92, p<0.001) and current reports of stress (i.e., activity-related stress at the time of beeps, β=0.49, 95%CI [0.29, 0.69], std. error=0.1, t-stat=4.81, p<0.001). Social stress was not significantly different between the two weeks (β=0.02, 95%CI [-0.01, 0.05], std. error=0.02, t-stat=1.19, p=0.236). The control items measuring physical stress also did not differ significantly between the weeks (β=-0.06, 95%CI [-0.24, 0.12], std. error=0.9, t-stat=-0.65, p=0.519), showing that increases in our subjective stress measures were likely due to our experimental manipulation, as opposed to other environmental or physical changes (Figure 2A). There were no effects of the other covariates in the model aside from males reporting slightly

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.29.450360; this version posted July 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

higher social stress than females, and some effects of movement and exercise of physical stress (Supplementary material Table S1 for full model results).

We next investigated the effects of stress exposure (i.e., week type) on mood and physiology as our main outcome measures. The same covariates used in modeling subjective stress were used here. Models for the physiology features also included mean skin temperature and change in temperature as measured by the slope during the selected time windows as fixed effects with random slopes. This was done as we expected these variables to have a significant effect on heart rate and skin conductance. For example, warmer body temperatures might result in increased sweat production, and thus increased skin conductance levels. Relative movement during a period may result in sensor displacement that was potentially unaccounted for in the processing pipeline and was thus also included as a covariate in the physiology models(*34*). Adjusted p-values were calculated using FDR correction and are reported below. In accordance with our expectations, we saw an increase in negative affect (β=0.11, 95%CI [0.07, 0.16], std. error=0.02, t-stat=4.8, $p<0.001$), and decrease in positive affect (β=-0.08, 95%CI [-0.11, -0.05], std. error=0.01, t-stat=-5.23, $p<0.001$) during the stress week (Figure 2B, full model details in supplementary material Table S2). Contrary to our expectations from laboratory stress studies, we found a decrease in arousal-related measures derived from skin conductance and heart rate during the exam week (Figure 2B). There was a decrease in the number of skin conductance responses (log-Mean=-0.23, 95%CI [-0.39, -0.08], std. error=0.08, t-stat=-3.00, $p<0.05$). Maximum heart rate was also lower during the stress week (β=-0.11, 95%CI [-0.17, -0.04], std. error=0.03, t-stat=-3.30, $p<0.05$). Additionally, we confirmed expected effects of movement and skin temperature confounds on skin conductance and heart rate measures (supplementary material Table S3).

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.29.450360; this version posted July 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

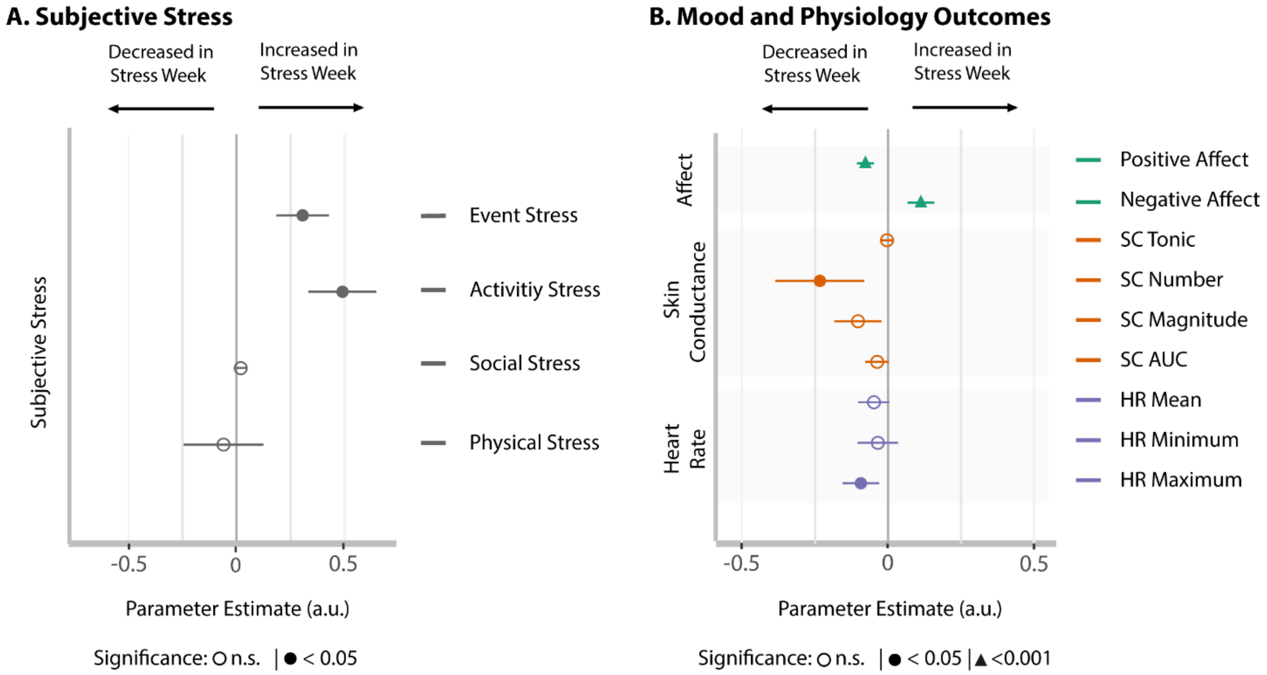


Fig 2. Fixed effects estimates of between-week difference. (**A**) Event-related stress (pertaining to the most prominent event since the last survey), and activity-related stress (relating to the current activity participants are engaged in) are significantly higher in the stress week compared to the control week. **(B)** This is accompanied by increased negative affect, decreased positive affect, and decreases in averages of multiple arousal-related physiological measures. Error bars represent confidence intervals.

Momentary subjective stress is associated more strongly with mood than physiology.

To explore the dynamics underlying the unexpected average decrease in measures of physiological arousal during the stress week, we investigated the link between moment-to-moment fluctuations in subjective stress and outcome measures (mood and physiological arousal). Separate models were constructed for each of the mood and physiology outcomes, with subjective stress variables as fixed effects and subject as a random effect. Due to high correlations between activity stress and both event (r=0.42, p<0.001) and social stress (r=0.34, p<0.001), interaction terms were also modeled for these two variables. Additional confounds were modelled as fixed effects with random slopes including sex, beep, temperature, mean displacement, and physical activity. Subjective stress measures were examined for multicollinearity using

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.29.450360; this version posted July 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

the variable inflation score (VIF) in each of the models without the interaction terms using the R package "performance"(*35*). VIF's for all results were below five, indicating low multicollinearity. Results (see Figure 3) were corrected for multiple comparisons using FDR correction.

In models investigating the moment-to-moment relationship between subjective stress and mood, we found a positive association between negative affect and activity related stress (β=0.08, 95%CI [0.02, 0.14], std. error=0.03, t-stat=2.75, p<0.05), social stress (β=0.23, 95%CI [0.19, 0.28], std. error=0.02, t-stat=10.15, p<0.001), and physical stress (β=0.13, 95%CI [0.10, 0.16], std. error=0.02, t-stat=8.21, p<0.001). The opposite was true for positive affect, with negative associations for event-related stress (β=-0.12, 95%CI [-0.19, -0.05], std. error=0.03, t-stat=-3.45, p<0.001), activity-related stress(β=-0.20, 95%CI [-0.28, -0.12], std. error=0.04, t-stat=-4.95, p<0.001), social stress (β=-0.30, 95%CI [-0.s36, -0.24], std. error=0.03, t-stat=-9.75, p<0.001), and physical stress(β=-0.15, 95%CI [-0.19, -0.11], std. error=0.02, t-stat=-7.63, p<0.001; see supplementary material Table S4 for full results). Regarding moment-to-moment associations with physiology measures, the number of skin conductance responses was positively associated with increased activity-related stress (β=0.08, 95%CI [0.01, 0.16], std. error=0.04, t-stat=2.2, p=0.045). The magnitude of skin conductance responses was associated with both activity (β=0.07, 95%CI [0.01, 0.13], std. error=0.02, t-stat=2.16, p=0.046) and event stress (β=0.07, 95%CI [0.02, 0.12], std. error=0.03, t-stat=2.16, p=0.016). For heart-rate measures, only minimum heart rate was negatively associated with social stress (β=-0.02, 95%CI [-0.03, -0.00], std. error=0.01, t-stat= -2.69, p=0.016). Full details are reported in supplementary material Table S5. Thus, moment-to-moment fluctuations in subjective stress are associated with expected mood changes and increases in physiological arousal, and therefore, the observed average decrease of physiological arousal measures during stress weeks is not explained by physiological changes associated with subjective stress.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.29.450360; this version posted July 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

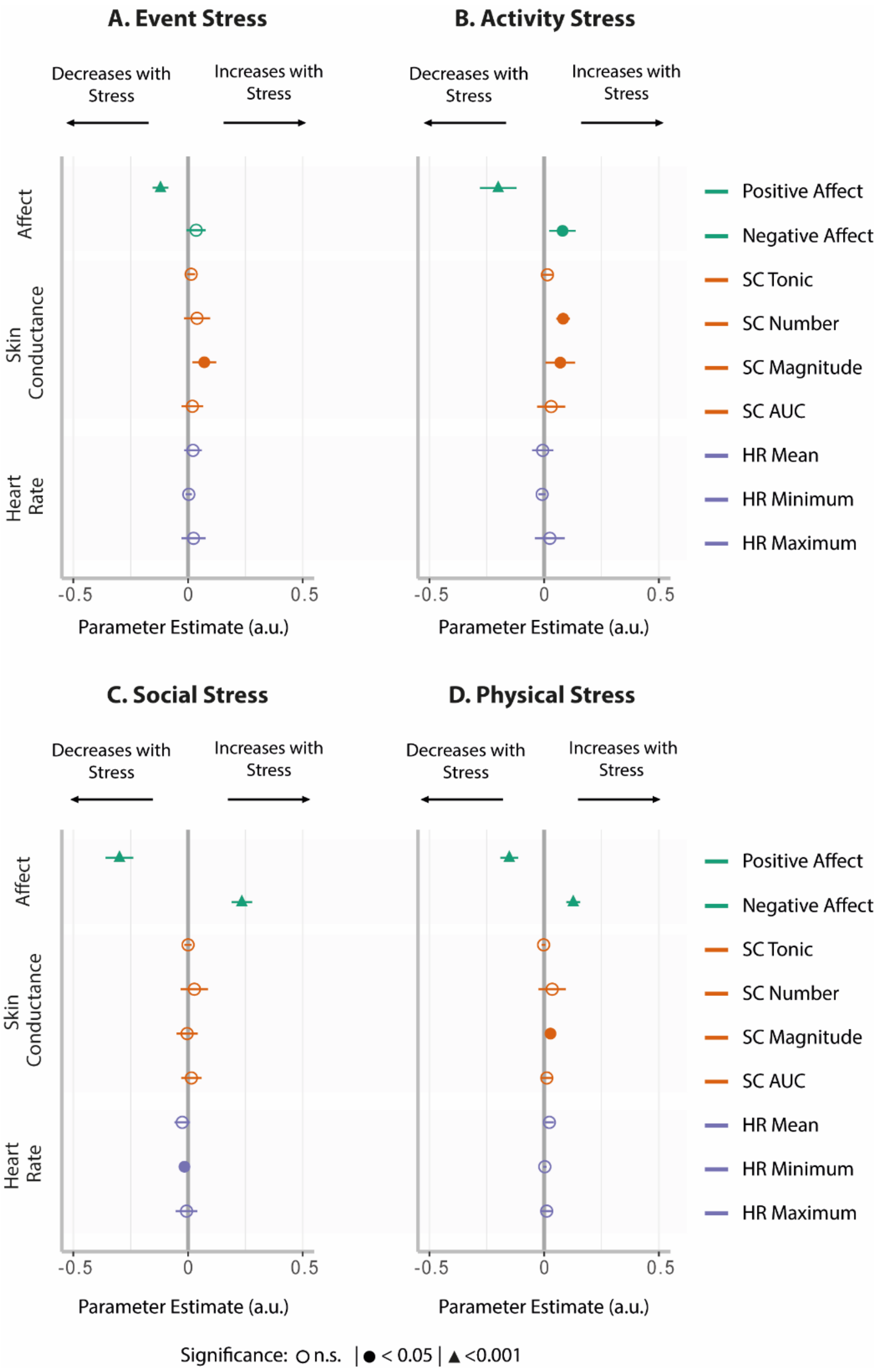


Fig 3. Effect estimates for the associations between moment-to-moment fluctuations in subjective stress and measures of mood and physiology. Subjective stress measures are generally associated with a decrease in positive affect, an increase in negative affect, and increases in some of the measures of physiological arousal. P-values corrected for multiple comparisons using FDR. Error bars represent confidence intervals.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.29.450360; this version posted July 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Positive mood is related to increased arousal and mediates week changes.

To investigate whether the observed average decrease of physiological arousal measures during stress weeks could instead be linked to reduced positive affect, we investigated the moment-to-moment association between affect and physiological arousal using the same covariates as the previous models. Due to a strong negative correlation between positive and negative affect (r=-0.57, P<0.001), interaction effects between positive and negative affect were added to the models. Multicollinearity was checked on the base models without the interaction terms. All VIF's were below five. Increased positive affect was related to the number of skin conductance responses (β=0.03, 95%CI [0.01, 0.05], std. error=0.01, t-stat=2.55, p=0.018), and increased mean heart rate (β=0.01, 95%CI [0.00, 0.02], std. error=0.01, t-stat=2.83, p<0.01), minimum heart rate (β=0.01, 95%CI [0.00, 0.03], std. error=0.01, t-stat=2.55, p=0.015), and maximum heart rate (β=0.02, 95%CI [0.00, 0.03], std. error=0.01, t-stat=2.56, p=0.021, Figure 4, full model details in supplementary material Table S6). Thus, in addition to subjective stress, also positive affect is positively associated with momentary physiological arousal.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.29.450360; this version posted July 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

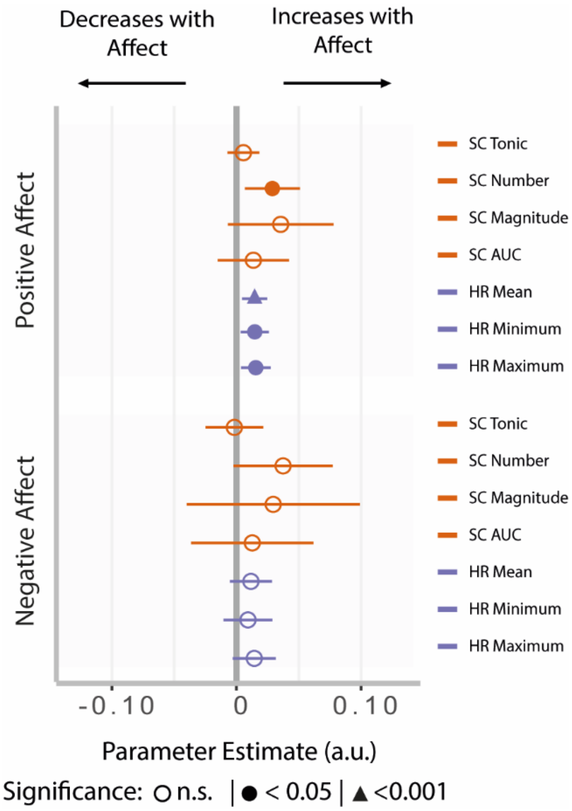


Fig 4. Relationship of momentary affect and physiology. Arousal-related physiological measures (magnitude of skin conductance responses and mean and minimum heart rate) were linked to positive affect, but not to negative affect. P-values are corrected for multiple comparisons using FDR. Error bars represent confidence intervals.

Next, to confirm that the observed average decrease in physiological arousal observed during the stress weeks is due to the decrease in positive affect, we assessed whether positive affect statistically mediated the effects of week type on physiological arousal. For this analysis, we specifically focused on the arousal measures that were also linked to subjective stress: The number of skin conductance responses and their magnitudes. Effect estimates were computed via Monte Carlo simulation (n=5,000). Results indicated that positive affect mediated a significant proportion of the relationship between skin conductance magnitude and week type (Approximately 7.3%, Mediating Estimate= -0.014, 95%CI [-0.03, 0.00], p= 0.028) but not the whole relationship (Direct Estimate=-0.166, 95%CI [-0.23, -0.10], p<0.001), indicating additional mechanisms that may also underly this relationship. The effect of week type on number of skin conductance responses was not mediated by positive affect.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.29.450360; this version posted July 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Machine-learning classification of week types using mood and physiology.

We next examined to what extent prolonged stress (i.e., week type; stress versus control) can be classified using machine learning based on measures of affect, physiological arousal, or a combination of both. We made use of random-forest machine-learning models and a leave-one-beep-out approach (LOBO), where models were tested on a single-subject level by training them on a single subject's data omitting one beep. Models were then tested on the left-out beep, with this process repeated until all beeps had been removed once. Model 1 attempting to classify week type from positive and negative affect resulted in a mean subject-level error rate of 33.45% (SD=±2.21). Model 2 tested if week type could be classified from the physiology data alone, resulting in a mean subject-level error rate of 36.11% (SD=±2.72). We finally determined the classification error for the combination of mood and physiology, resulting in the lowest error rate (M=29.87%, SD=±3.45).

Each model was tested against its bootstrapped error rate (n=10,000) to determine whether the model performed above chance level. P-values were first calculated from the distribution by examining the proportion of bootstrapped error rates that exceed the actual model. All models performed significantly above chance on an individual level for all but one subject (Figure 6, individual p-values for each subject and error rates are reported in associated online notebook indicated at the end of the supplementary material). Group-level effects were further tested with paired-samples t-tests comparing the LOBO models to the mean bootstrapped error for each model. Model 1 (Affect, M_{diff}=-16.29, $t_{(80)}$=-64.06, p<0.001), Model 2 (Physiology, M_{diff}=-13.87, $t_{(78)}$=-50.38, p<0.001), and Model 3 (Combination, M_{diff}=-19.45, $t_{(78)}$=-48.94, p<0.001) all performed above chance at the group level.

Paired-samples t-tests comparing the within-subject error rates between the LOBO models showed that the model using only EPA data (M2) had the highest error rates, and performed significantly worse than the model using affect (M1) items (M_{diff}=2.60, $t_{(78)}$=14.65, p<0.001). The model using affect measures alone came in second relative to the model using a combination of measures (M_{diff}=3.64, $t_{(78)}$=19.20, p<0.001).

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.29.450360; this version posted July 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

However, worth noting is that the overall difference between models 1 and 2 was at 2.60% on average, which shows that Model 2 using EPA wristwatch data performed almost on par with the EMA affect model. While overall the EMA mood models performed better, in some subjects the models had almost equivalent performance.

Within-subject models offer better predictions than between-subject

We next sought to investigate the generalizability of these models from a within-subject approach to a population-level one (using between-subject classification training). This was done through the commonly used leave-one-subject-out (LOSO) cross-validation instead. The same measures were used for these models again: model 1 (M1) attempted to classify week from affect, model 2 classified week from physiology (M2), and model 3 used the combination of mood and physiology for classification (M3). M1 using affect (45.85%, SD±9.50), M2 using physiology 48.42%, SD=±8.05), and M3 using the combination (42.44%, SD=±9.00) were tested against their bootstrapped counterpart similar to the LOBO models.

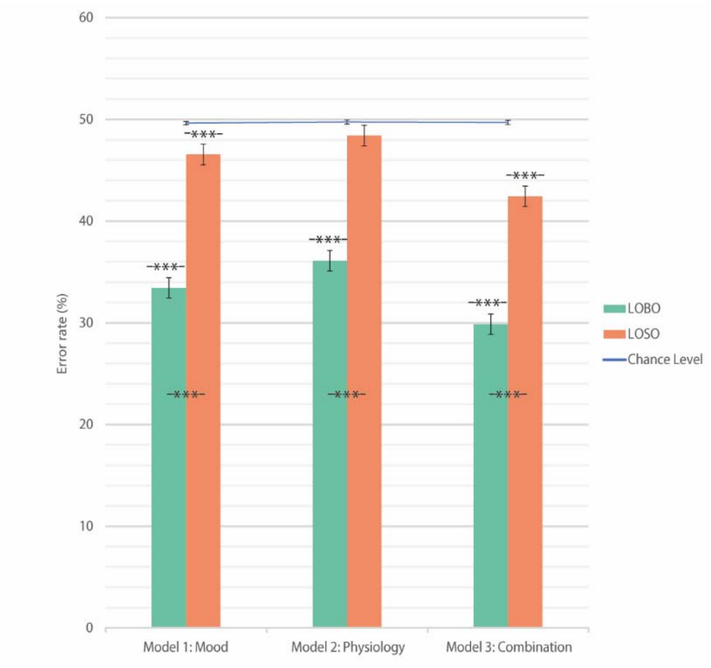


Fig 5. Random-forest classification error estimates. Average error estimates and error bars (representing standard errors of the mean) for each of the random-forest

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.29.450360; this version posted July 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

models. Combinations of mood and physiology yield superior classification, and individually trained and tested models (LOBO, Leave-One-Beep-Out) perform better than models trained on group-level data (LOSO, Leave-One-Subject-Out). Chance levels estimated from permutation test and confidence interval are shown in blue. Significance levels between bars indicate between model comparisons, and above bar indicate model comparison to chance levels. ***P<0.001. -

For some individual subjects LOSO models performed significantly above chance level (Model 1-Affect n=45(54.1%), Model 2-Physiology n=30(37.9%), and Model 3-Combination n= 55(69.6%)) in classifying week type (subject level p-values reported in supplementary material section 7). Group level analysis using a paired sample t-test showed that only model 1 (Affect, M_{diff}=-3.63, $t_{(80)}$=-3.59, p<0.001), and model 3 (Combinations, M_{diff}=-6.55, $t_{(78)}$=-6.81, p<0.001) performed better than chance. Model 2 did not perform above chance at the group level (Physiology, M_{diff}=-1.34, $t_{(78)}$=-1.54, p=0.128).

We additionally directly compared the classification errors between the LOSO and LOBO models for each subject using a paired-sample t-test. For all three models, the LOBO within subject approach performed better than the LOSO. LOSO model 1 using mood as a classifier performed significantly worse than the equivalent LOBO model (M_{diff}=11.43, $t_{(80)}$=12.17, p<0.001). The LOSO error estimates for model 2 using physiology as a classifier were also significantly worse than the LOBO counterpart (M_{diff}=8.11, $t_{(78)}$=8.61, p<0.001). Model 3 LOSO error rates using the combination were the lowest among the LOSO family of models, but similarly performed worse than the LOBO model counterpart (M_{diff}=12.61, $t_{(78)}$=11.74, p<0.001). This demonstrates that training classifiers on individual data results in vastly superior classification compared to training models on population-level models.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.29.450360; this version posted July 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Discussion

In the current study, we investigated physiological and psychological responses to an ecological prolonged stressor in daily life (i.e., an exam week in students), with the goal of determining the usability of passive monitoring technologies for detection of prolonged stress. We employed EMA and EPA to track subjective stress, mood, and arousal-related physiology. Our findings confirmed an overall increase in subjective stress levels during the exam week. As hypothesized, during the stress week, negative mood increased, and positive mood decreased. Contrary to what was expected, lower skin conductance and heart rate arousal measures were observed during the stress week. At a beep-to-beep time scale, increased subjective stress was associated with increased negative mood, decreased positive mood, and increased skin conductance responses. Interestingly, positive affect was also associated with skin conductance responses, and mediated the changes seen between the two weeks. This indicates that the observed decreases in physiological arousal measures were (at least partially) due to a reduction in positive mood. Finally, using a machine-learning approach, we showed that the combination of individual mood and physiology best dissociated stress from control weeks. These results highlight the potential of using passive monitoring using wearable sensors, but they also caution that mood measures are still important in distinguishing positive and negatively valent arousal.

In our results, we observed an increase in self-reported stress during the exam week in line with previous findings on examination stress, thus providing a validation for our paradigm(36, 37). Interestingly, the effects of week type also included expected changes in mood items, with increased negative and decreased positive affect. However, while we initially expected arousal measures to be increased during the stress week, the opposite was true in our results. The observed overall decrease in physiological arousal during stress weeks appears at odds with the positive association between subjective stress and increased arousal shown in our moment-to-moment analysis and in previous works(26, 30, 32, 38). What becomes evident here is that there is a distinct mechanism that differentiates prolonged stress from acute stress. While prolonged stress leads to increased moment-to-moment peaks in self-reported acute stress, it also results

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.29.450360; this version posted July 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

more generally in decreased positive affect and decreased overall average arousal. Given that positive affect and arousal measures are both reduced during the stress week, it appears that the dominant effect on arousal is tied to (reduced) positive affect rather than to peaks in subjective stress. The mechanistic link supporting this claim is shown in our mediation analysis, which confirms that positive affect (partially) mediates the effect of week type on reduced arousal. While this may seem counterintuitive, physiological arousal through skin conductance and heart rate measures has shown to be responsive to both positive and negative events, showing that physiological arousal is not valence specific(27, 38–41). Thus, the net effect of prolonged stress exposure (as operationalized in this study) stems in part from a reduction in overall arousal driven by reductions in positive mood that persist outside of peak moments of acute stress.

We subsequently tested the ability of machine-learning models in classifying stress and control weeks based on either physiology, mood, or both combined. Physiology models could classify prolonged stress exposure almost as well as models using mood alone (3.85% difference on average). However, and more importantly, the combination of both resulted in the best predictive accuracy (approximately 29.9% error). Thus, it is apparent that the addition of mood questions to physiological arousal provides valuable information to classification models. This builds on the initial findings of our mixed models and mediation analysis, showing that accounting for valence through mood is important in trying to separate stress-induced arousal from positive arousal. This problem is highlighted in studies using skin conductance trigger-based EMA to detect stress, which resulted in capturing positive arousal instead(39) . Our findings highlight the need for continued development of passive, non-invasive measurements for stress detection. This would allow us to bypass some limitations of EMA, such as response biases, (lack-of) insight into mood states, the age and sex of participants, and the instructions given by experimenters(42–44). With the recent growth of interest in the application of physiological assessments in clinical populations, our findings indicate that combining a wrist-worn device with a minimally invasive mood assessment might offer a feasible approach to detecting stress in clinical populations that surpasses a full battery of EMA measures.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.29.450360; this version posted July 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

In addition to showing the utility of physiological monitoring, our results also show the importance of individualized approaches in stress detection. We assessed the generalizability of our machine-learning approach by comparing individualized models (i.e., Leave-One-Beep-Out, LOBO) to group-level models (i.e., Leave-One-Subject-Out, LOSO). Classification models trained and tested on an individual's own data performed significantly better than those trained at the group level. This shows that our individualized approach offers drastic improvements to classification of stress states. Both our own and previous studies using LOSO models in this pursuit had mixed results for individual participants, with only around 18% of participants in other studies achieving high classification performance (i.e., low error rates)(32). With the movement towards more personalized approaches in psychiatry, the outcomes of the individualized LOBO method further emphasize the need for a personalized approach. The reason for this is likely that the same experience can generate different physiological and psychological responses in different people based on a multitude of factors such as sex, appraisal, or clinical traits and features(45). For example, it may be reasonable to assume that patients with anxiety may display a very different physiological response to stress than those with depression (hyper vs hypo activation) or than those with impulsive aggression(40). Individualized models would allow for greater prediction than a one-size-fits all approach, with improved generalizability of the methods. One foreseeable implementation in general practice may include a period of passive data collection so that adequate data for individual patients can be acquired to train and test models with a continuous data stream.

Some limitations that warrant discussion include the implementation of physiology assessments. While marketable devices are available, the underlying processing and feature extraction is not easily implementable in clinical practice. Recent tools(45) and recommendations have made the process simpler(46), but decisions on what processing steps to implement, what software and platform to use, and what features to select are beyond the scope of many clinicians. Through highlighting important features in our own results, we hope to elucidate what physiology measures might be relevant in practice to continue the development of accessible platforms in clinical practice. Furthermore, processing of

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.29.450360; this version posted July 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

physiology data requires that poor quality signals be removed from the full data set. However, despite discarding 10% of surveys due to poor quality physiology, we were still able to maintain 77% completion rates, which are comparable to many other studies using EMA alone(*43*). The need for computational power is also apparent in the model estimations in our study. While the individualized LOBO and group-level LOSO models were relatively easy to estimate, bootstrap estimations are much more computationally expensive and can require days to compute. These issues on a whole are not easy to solve but increasing technological advancements will make future research using these methods even more accessible.

In conclusion, our study shows that EPA can potentially be used for monitoring stress-related mental health, but highlights the fact that psychological context remains critical in interpreting changes in physiological arousal in terms of acute stress versus positive affect. We show that a combination of EMA and EPA is optimal for detecting prolonged stress, and we furthermore highlight the need for an individualized approach in this effort. Personalized approaches in psychiatry have been gaining momentum in recent years, and our findings further support this development. If successfully implemented at wider scale, our findings may have implications on disease prevention that may help reduce the overall disease burden of stress-related disorders through personalized early-warning systems and treatment strategies, though more work is needed to explore differences in these mechanisms in various clinical populations.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.29.450360; this version posted July 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Materials and Methods

Experimental Design

We recruited 84 right-handed, first year bachelor's students in the medical or biomedical science majors from Radboud Health Academy spanning three academic years (2017, 2018, and 2019). One participant withdrew during testing, resulting in a total sample size of 83 participants used in the analysis. The programs were selected due to their structured examination weeks that occur every 5th and 10th week of a semester, allowing us to examine a period with higher stress levels during examination weeks as an ecological prolonged stressor. While course work and examinations for both programs were identical, participants' major was recorded to be used in statistical models as a confound. Only participants with no history of psychiatric illness were included in the study. Recruitment was stopped following the COVID-19 outbreak (in March 2020). All procedures carried out were approved by the regional medical ethical review board (CMO Arnhem-Nijmegen).

Participants completed two weeks of ecological momentary assessments (EMA), one occurring during an examination period (i.e., stress week) and the other occurring on average 16 days (min=10, max=33) outside of these periods (i.e., control week, demographics in Table 1). Compliance rates were overall high with 84% of surveys being completed within the allocated one-hour window during both stress and control weeks. When accounting for missing and poor-quality physiology (EPA) data, completion rates dropped to between 76% and 77% which was within the median ranges for EMA studies reported the metanalysis by Vachon et al (2019, Table 1)(*43*). Gender distribution was similar to that of students enrolled at the university (57% female, according to Radboud University website). We were unable to fully counterbalance the order of weeks due to termination of recruitment upon the onset of the COVID-19 crisis. Week order was therefore controlled for in all statistical models. At the end of each of these weeks, participants completed a series of behavioral tasks, and a psychological test battery. Participants also underwent two counterbalanced fMRI sessions which are outside the scope of this paper and will be reported elsewhere.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.29.450360; this version posted July 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 1. Descriptive statistics

Sex	**Female**	51 (61.4%)	**Male**	32 (38.6%)		
Course Program	**Medicine**	61	**Bio-Medical Sciences**	22		
	Exam Week			**Control Week**		
First Week	27 (32.5%)			56 (67.5%)		
	1st Qu.*	**Mean**	**3rd Qu.***	**1st Qu.***	**Mean**	**3rd Qu.***
Survey Completion	34 (81%)	35.51 (85%)	39 (93%)	34 (81%)	35.89 (85%)	40 (95%)
Completion with usable Physiology	29 (69%)	32.15 (76.55%)	37 (88%)	29 (69%)	32.36 (77.05%)	37 (88%)

** 1st and 3rd Quantiles indicating 50% of participants had completion rates in given range*

Assessing daily-life stress through EMA and EPA

To assess stress reactivity in daily life, we employed EMA and EPA. Participants completed two different testing weeks, with one of these periods culminating in a high-stakes exam (i.e., stress week) and the other with no examinations (i.e., control week). This allowed us to objectively examine a period with higher stress levels and subsequently determine individualized patterns of stress reactivity. Participants initially had an intake meeting during which the study procedures were explained before the start of the testing weeks. During these weeks, participants received six surveys a day at fixed intervals. Surveys were hosted on CastorEDC(*47*), and participants received links to the surveys via SMS texts. Participants were given a one-hour window to fill in the surveys to adjust for differences in class schedules within the population, as phones were not permitted during these times. While this time window is longer than that used in some studies(*14, 48*), it is still within range of time windows used in other studies(*49*). Surveys assessed different psychological aspects related to stress, using questionnaires and constructs based on other studies in EMA literature (see Vaessen et al. 2017 for overview of question types)(*50*). Additionally, the first questionnaire of the day contained an assessment for sleep quality, while the last contained items regarding self-reflection. The full questionnaires are provided in the associated online repository listed in the supplementary materials section 7.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.29.450360; this version posted July 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

The EMA surveys themselves consisted of questions regarding subjective stress used in the validation of our experimental paradigm, and mood questions (positive and negative affect) relating to our subjective outcome measures filled in on a 7-point Likert scale. Questions in the validation set probed four types of stress as follows: i) Event-related stress assessed the most prominent event that occurred in between EMA beeps ii) Activity-related stress questions probed the activity participants were engaged in upon receiving the beep iii) Social-related stress addressed stress that may arise from the social context participants were present in (either being alone, or with someone) iv) Physical-related stress was used as a control measure to account for environmental and physical demands. Mood outcome questions consisted of four items assessing positive mood, and five items assessing negative mood.

In addition to filling in EMA surveys, participants were also instructed to wear an Empatica E4 wristband (Empatica, Milano, Italy) that collected ambulatory EPA data throughout the stress and control weeks. Ambulatory data is collected continuously and in the background without the need of participants to actively engage in the collection of this type of data. Participants were instructed to charge the watch and simultaneously synchronize the data to anonymized researcher-specific accounts once a day for one hour preferably when showering to minimize data loss. A detailed explanation was given to participants on their operation with a practice session recorded during the intake interview. The E4 devices collected blood pulse volume, electrodermal activity, three-axis movement, and body temperature.

EPA Data Processing

EPA Data cleaning was performed using Python (V3.6.1)(*51*). Additional packages used for preprocessing included NumPy (V1.18.1)(*52*) and Pandas (V1.0.3)(*53*). Time stamps for each survey instance were used to classify surveys as belonging to a stress or control week. Ten-minute time windows prior to each survey were selected for the extraction of physiology features acquired from the E4. Pre-processed IBI data were deemed too sparse to offer meaningful temporal domain analysis, with an average of 27% of IBIs successfully detected in our selected time window. This is within the margins of the manufacturer's signal loss estimates in daily use. We instead selected average heart rate features from the

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.29.450360; this version posted July 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

resulting processed files from Empatica. The devices use a strict proprietary detection algorithm in the detection of IBIs, so these files can be used with minimal processing to derive global heart rate features. These features included the mean, minimum, and maximum heart rate. Raw skin conductance was processed for offline use with the PyPhysio package (V2.1)(*45*). A minimum threshold of 0.01 µsiemens was set for the skin conductance levels deemed of acceptable quality based on previous recommendations of a threshold between 0.01-0.05 µsiemens(*46*). Data was first despiked to remove artifacts due to sudden hand motions using standard settings in the library. Data was then denoised to remove remaining artifacts through windowed filtering of changes in the signal greater than 0.02 µsiemens between subsequent samples. Additionally, an Elliptic filter with cut-off frequency set between 0.8 and 1.1 was applied to the data. Skin conductance data were subsequently de-convolved using a Bateman impulse response function into phasic and tonic components from which specific features were extracted (mean tonic activity, and magnitude, area under the curve, and the number of phasic responses). The raw temperature measures were used to calculate the mean skin temperature, as well as the slope as a function of change in skin temperature within the acquired time window. Two participants had a watch with faulty temperature sensors. These measures were substituted from the population mean and standard deviation to avoid loss of participants' data due to missing data points in statistical models. The other sensors on this device were tested and no errors were detected in other recordings. Finally, the root mean squared displacement in each time window was calculated from the accelerometer data. The extracted features were collected into a single data frame used for statistical analysis.

Statistical Analysis

All statistical analyses were conducted in R (R, version 3.6.1). EMA surveys contained several questions relating to four stress scales: 1) Event stress 2) Activity Stress, 3) Social Stress and 4) Physical stress (Box 1 for detailed information). Items on a reversed scale were inverted. Items for each scale were summed to create a single score for each of the scales. The same was done for items relating to positive and negative

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.29.450360; this version posted July 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

affect. Total item scores were then rescaled, and a subject centered measure was derived. Surveys that were not filled in within the assigned time window were excluded from further analyses.

Initial analysis examined overall differences in the population between the two weeks to establish the validity of the experimental manipulation through generalized linear mixed effects models using the "lmer" package(54). A maximal fitting approach was used in constructing our models to reduce Type-I errors in which random slopes were estimated for all fixed effects(55). EMA and EPA measures were modeled as dependent variables, with week type as the primary predictor of interest. Sex, program, order of the weeks, as well as day relative to start and beep number were modeled as fixed effects. We additionally modeled nuisance regressors using temperature and movement for the EPA models. We used subject as a random effect and modeled random slopes for each of the predictors. A random intercept was set for the random effect modeling our predictor of interest (i.e., week type). Except for week type, which was the predictor of interest, we did not model random slopes for factors with seven or fewer levels. Model fits were checked, and model families were adjusted to achieve optimal fit based on Akaike Information Criterion (AIC) and residual normality. We additionally sought to replicate previous findings associating momentary stress with physiology signals, and to further explore findings from our week assessments. To this end, we examined the relationships underlying the continuous physiology and psychological outcomes in relation to the combined subjective stress measures. Models were fit with the same approach used in the first analysis. Mediation analysis was then used to explain the apparent differences in the relationships between the week type and momentary analyses.

Machine-Learning Models

One of our primary goals was assessing the ability to use ambulatory, non-intrusive measures to determine whether someone is currently in a stressed state. To this end, random forest models were used to determine the ability to classify whether subjects were in the stress or control week using the collected EMA and/or ambulatory EPA data using the randomForest package in R(56). Random forest models were selected due to the demonstrated high accuracy(57), previous use in similar studies(32, 58), and

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.29.450360; this version posted July 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

simplicity of implementation that may make them more usable in a broader setting. To determine the variables used in training the models, we conceptualized mood and changes in physiology as outcomes of stressed states based on previous findings(18, 19). That is, we expected that feelings of stress would directly impact mood and be related to changes in physiology. Therefore, we selected only the mood items from the EMA measures in these analyses.

Model predictions were estimated using a Leave-One-Beep-Out (LOBO) cross-validation method similar to a Leave-One-Trial-Out (LOTO) approach used in other fields using custom functions(59, 60). Models estimating the classification errors were constructed at a single-subject level. For each participant, a model was estimated on their n-1 beep data set, with a prediction being tested on the removed beep. This was repeated until all beeps had been removed once for the subject and tested against their remaining data set. Three models were tested as follows: Model 1 tested the ability to classify week type from positive and negative affect. Model 2 tested the ability to predict week type from ambulatory data collected via the E4 wristband. Model 3 tested the combination of physiology and mood data in predicting week type. A bootstrapped error was estimated through permutation tests by randomly resampling the week type and testing each of the models again. Resampling was carried out using 10,000 iterations per model to achieve the true error distributions. Each model was tested against the distribution resulting from the permutation tests to determine if predictions were indeed above chance. Models were then tested against the average error rates of the permutation tests to determine group level effects using paired sample t-tests. Models were then tested against each other to determine which model had the lowest error rates. Finally, we tested the generalizability of the random forest models to a population level using a Leave-One-Subject-Out (LOSO) analysis(60). In a LOSO analysis, models were trained on N-1 participants dataset, where an entire participant's data was removed from the dataset and a model trained on the remaining participants. Classification errors for the removed participant were then calculated, and the process repeated until each participant had been removed once from the dataset. Model predictions using the LOBO were then

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.29.450360; this version posted July 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

compared to that of the LOSO method to estimate the generalizability of machine-learning models on the data.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.29.450360; this version posted July 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

References

1. R. Kalisch, D. G. Baker, U. Basten, M. P. Boks, G. A. Bonanno, E. Brummelman, A. Chmitorz, G. Fernàndez, C. J. Fiebach, I. Galatzer-Levy, E. Geuze, S. Groppa, I. Helmreich, T. Hendler, E. J. Hermans, T. Jovanovic, T. Kubiak, K. Lieb, B. Lutz, M. B. Müller, R. J. Murray, C. M. Nievergelt, A. Reif, K. Roelofs, B. P. F. Rutten, D. Sander, A. Schick, O. Tüscher, I. van Diest, A. L. van Harmelen, I. M. Veer, E. Vermetten, C. H. Vinkers, T. D. Wager, H. Walter, M. Wessa, M. Wibral, B. Kleim, The resilience framework as a strategy to combat stress-related disorders. *Nature Human Behaviour*. **1**, 784–790 (2017).

2. B. S. McEwen, In pursuit of resilience: stress, epigenetics, and brain plasticity. *Annals of the New York Academy of Sciences*. **1373**, 56–64 (2016).

3. E. J. Hermans, G. Fernández, Heterogeneity of cognitive-neurobiological determinants of resilience. *Behavioral and Brain Sciences*. **38**, e103 (2015).

4. C. Osório, T. Probert, E. Jones, A. H. Young, I. Robbins, Adapting to Stress: Understanding the Neurobiology of Resilience. *Behavioral Medicine*. **43**, 307–322 (2017).

5. M. J. Giummarra, A. Lennox, G. Dali, B. Costa, B. J. Gabbe, Early psychological interventions for posttraumatic stress, depression and anxiety after traumatic injury: A systematic review and meta-analysis. *Clinical Psychology Review*. **62**, 11–36 (2018).

6. C. F. Reynolds 3rd, P. Cuijpers, V. Patel, A. Cohen, A. Dias, N. Chowdhary, O. I. Okereke, M. A. Dew, S. J. Anderson, S. Mazumdar, F. Lotrich, S. M. Albert, Early intervention to reduce the global health and economic burden of major depression in older adults. *Annual review of public health*. **33**, 123–135 (2012).

7. S. M. Schueller, A. Aguilera, D. C. Mohr, Ecological momentary interventions for depression and anxiety. *Depression and Anxiety*. **34**, 540–545 (2017).

8. M. E. McDevitt-Murphy, M. T. Luciano, R. J. Zakarian, Use of Ecological Momentary Assessment and Intervention in Treatment With Adults. *Focus (American Psychiatric Publishing)*. **16**, 370–375 (2018).

9. M. Csikszentmihalyi, R. Larson, Validity and reliability of the experience-sampling method. *Journal of Nervous and Mental Disease*. **175** (1987), pp. 526–536.

10. S. Shiffman, A. A. Stone, M. R. Hufford, Ecological Momentary Assessment. *Annual Review of Clinical Psychology*. **4**, 478–481 (2008).

11. J. Swendsen, Contributions of mobile technologies to addiction research. *Dialogues in clinical neuroscience*. **18**, 213–221 (2016).

12. E. Bar-Kalifa, H. Sened, Using Network Analysis for Examining Interpersonal Emotion Dynamics. *Multivariate Behavioral Research*. **55**, 211–230 (2019).

13. L. F. Bringmann, N. Vissers, M. Wichers, N. Geschwind, P. Kuppens, F. Peeters, D. Borsboom, F. Tuerlinckx, A Network Approach to Psychopathology: New Insights into Clinical Longitudinal Data. *PLoS ONE*. **8** (2013), doi:10.1371/journal.pone.0060188.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.29.450360; this version posted July 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

14. D. Collip, J. T. W. Wigman, I. Myin-Germeys, N. Jacobs, C. Derom, E. Thiery, M. Wichers, J. van Os, From Epidemiology to Daily Life: Linking Daily Life Stress Reactivity to Persistence of Psychotic Experiences in a Longitudinal General Population Study. *PLoS ONE*. **8**, 1–6 (2013).

15. B. Lenaert, M. Colombi, C. van Heugten, S. Rasquin, Z. Kasanova, R. Ponds, Exploring the feasibility and usability of the experience sampling method to examine the daily lives of patients with acquired brain injury. *Neuropsychological Rehabilitation*. **0**, 1–13 (2017).

16. K. Hoorelbeke, N. van den Bergh, M. Wichers, E. H. W. Koster, Between vulnerability and resilience: A network analysis of fluctuations in cognitive risk and protective factors following remission from depression. *Behaviour Research and Therapy*. **116**, 1–9 (2019).

17. M. Wichers, M. J. Schreuder, R. Goekoop, R. N. Groen, Can we predict the direction of sudden shifts in symptoms? Transdiagnostic implications from a complex systems perspective on psychopathology. *Psychological Medicine*. **49**, 380–387 (2019).

18. E. C. D. van der Stouwe, N. A. Groenewold, E. H. Bos, P. de Jonge, M. Wichers, S. H. Booij, How to assess negative affective reactivity to daily life stress in depressed and nondepressed individuals? *Psychiatry Research* (2019), , doi:10.1016/j.psychres.2019.03.040.

19. D. M. Dunkley, M. Lewkowski, I. A. Lee, K. J. Preacher, D. C. Zuroff, J. L. Berg, J. E. Foley, G. Myhr, R. Westreich, Daily Stress, Coping, and Negative and Positive Affect in Depression: Complex Trigger and Maintenance Patterns. *Behavior Therapy*. **48**, 349–365 (2017).

20. M. Wichers, A. C. Smit, E. Snippe, Early warning signals based on momentary affect dynamics can expose nearby transitions in depression: A confirmatory single-subject time-series study. *Journal for Person-Oriented Research*. **6**, 1–15 (2020).

21. P. J. Quee, L. van der Meer, R. Bruggeman, L. de Haan, L. Krabbendam, W. Cahn, N. C. L. Mulder, D. Wiersma, A. Aleman, Insight in Psychosis: Relationship With Neurocognition, Social Cognition and Clinical Symptoms Depends on Phase of Illness. *Schizophrenia Bulletin*. **37**, 29–37 (2011).

22. N. Takai, M. Yamaguchi, T. Aragaki, K. Eto, K. Uchihashi, Y. Nishikawa, Effect of psychological stress on the salivary cortisol and amylase levels in healthy young adults. *Archives of Oral Biology*. **49**, 963–968 (2004).

23. N. L. Sin, A. D. Ong, R. S. Stawski, D. M. Almeida, Daily positive events and diurnal cortisol rhythms: Examination of between-person differences and within-person variation. *Psychoneuroendocrinology*. **83** (2017), doi:10.1016/j.psyneuen.2017.06.001.

24. N. Smyth, A. Clow, L. Thorn, F. Hucklebridge, P. Evans, Delays of 5—15 min between awakening and the start of saliva sampling matter in assessment of the cortisol awakening response. *Psychoneuroendocrinology*. **38**, 1476–1483 (2013).

25. L. Schwabe, L. Haddad, H. Schachinger, HPA axis activation by a socially evaluated cold-pressor test. *Psychoneuroendocrinology*. **33**, 890–895 (2008).

26. T. Pereira, P. R. Almeida, J. P. S. Cunha, A. Aguiar, Heart rate variability metrics for fine-grained stress level assessment. *Computer Methods and Programs in Biomedicine*. **148**, 71–80 (2017).

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.29.450360; this version posted July 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

27. A. Löw, P. J. Lang, J. C. Smith, M. M. Bradley, Both predator and prey: Emotional arousal in threat and reward. *Psychological Science*. **19**, 865–873 (2008).

28. L. Schwabe, H. Schächinger, Ten years of research with the Socially Evaluated Cold Pressor Test: Data from the past and guidelines for the future. *Psychoneuroendocrinology*. **92**, 155–161 (2018).

29. L. P. J, M. K. Greenwald, M. M. Bradley, A. Hamm, Looking at pictures: Affective, facial, visceral, and behavioral reactions. *Psychophysiology*. **30**, 261–273 (1993).

30. J. A. Healey, R. W. Picard, Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on Intelligent Transportation Systems*. **6**, 156–166 (2005).

31. K. Hovsepian, M. Al'Absi, E. Ertin, T. Kamarck, M. Nakajima, S. Kumar, cStress: Towards a Gold Standard for Continuous Stress Assessment in the Mobile Environment. *Proceedings of the ... ACM International Conference on Ubiquitous Computing . UbiComp (Conference)*. **2015**, 493–504 (2015).

32. E. Smets, E. Rios Velazquez, G. Schiavone, I. Chakroun, E. D'Hondt, W. de Raedt, J. Cornelis, O. Janssens, S. van Hoecke, S. Claes, I. van Diest, C. van Hoof, Large-scale wearable data reveal digital phenotypes for daily-life stress detection. *npj Digital Medicine*. **1**, 67 (2018).

33. B. Maser, M. Danilewitz, E. Guérin, L. Findlay, E. Frank, Medical Student Psychological Distress and Mental Illness Relative to the General Population: A Canadian Cross-Sectional Survey. *Academic Medicine*. **94** (2019) (available at https://journals.lww.com/academicmedicine/Fulltext/2019/11000/Medical_Student_Psychological_Distress_and_Mental.42.aspx).

34. S. Doberenz, W. T. Roth, N. I. Maslowski, E. Wollburg, S. Kim, Methodological Considerations in Ambulatory Skin Conductance Monitoring. *International Journal of Psychophysiology*. **39**, 237–245 (2012).

35. D. Lüdecke, D. Makowski, P. Waggoner, I. Patil, Performance: assessment of regression models performance. *R package version 0.4*. **5** (2020).

36. L. Brodersen, R. Lorenz, Perceived stress, physiological stress reactivity, and exit exam performance in a prelicensure Bachelor of Science nursing program. *International Journal of Nursing Education Scholarship*. **17**, 1–12 (2020).

37. E. M. J. Peters, Y. Müller, W. Snaga, H. Fliege, A. Reißhauer, T. Schmidt-Rose, H. Max, D. Schweiger, M. Rose, J. Kruse, Hair and stress: A pilot study of hair and cytokine balance alteration in healthy young women under major exam stress. *PLOS ONE*. **12**, e0175904 (2017).

38. S. D. Kreibig, Autonomic nervous system activity in emotion: A review. *Biological Psychology*. **84**, 394–421 (2010).

39. S. van Halem, E. van Roekel, L. Kroencke, N. Kuper, J. Denissen, 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*. **807**, 794–807 (2020).

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.29.450360; this version posted July 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

40. P. de Looff, M. L. Noordzij, M. Moerbeek, H. Nijman, R. Didden, P. Embregts, Changes in heart rate and skin conductance in the 30 min preceding aggressive behavior. *Psychophysiology*. **56** (2019), doi:10.1111/psyp.13420.

41. P. J. Lang, M. M. Bradley, B. N. Cuthbert, Motivated attention: Affect, activation, and action. *Attention and orienting: Sensory and motivational processes*. **97**, 135 (1997).

42. A. Rintala, M. Wampers, I. Myin-Germeys, W. Viechtbauer, Response compliance and predictors thereof in studies using the experience sampling method. *Psychological Assessment*. **31** (2019), doi:10.1037/pas0000662.

43. H. Vachon, W. Viechtbauer, A. Rintala, I. Myin-Germeys, Compliance and Retention With the Experience Sampling Method Over the Continuum of Severe Mental Disorders: Meta-Analysis and Recommendations. *Journal of medical Internet research*. **21**, e14475–e14475 (2019).

44. A. A. Stone, C. K. F. Wen, S. Schneider, D. U. Junghaenel, Evaluating the Effect of Daily Diary Instructional Phrases on Respondents' Recall Time Frames: Survey Experiment. *Journal of Medical Internet Research*. **22** (2020), doi:10.2196/16105.

45. A. Bizzego, A. Battisti, G. Gabrieli, G. Esposito, C. Furlanello, pyphysio: A physiological signal processing library for data science approaches in physiology. *SoftwareX*. **10**, 100287 (2019).

46. W. Boucsein, D. C. Fowles, S. Grimnes, G. Ben-Shakhar, W. T. Roth, M. E. Dawson, D. L. Filion, Publication recommendations for electrodermal measurements. *Psychophysiology*. **49**, 1017–1034 (2012).

47. Castor Electronic Data Capture, Castor EDC (2019).

48. N. Jacobs, I. Myin-Germeys, C. Derom, P. Delespaul, J. van Os, N. A. Nicolson, A momentary assessment study of the relationship between affective and adrenocortical stress responses in daily life. *Biological Psychology*. **74**, 60–66 (2007).

49. D. Schultchen, J. Reichenberger, T. Mittl, T. R. M. Weh, J. M. Smyth, J. Blechert, O. Pollatos, Bidirectional relationship of stress and affect with physical activity and healthy eating. *British journal of health psychology*. **24**, 315–333 (2019).

50. T. Vaessen, M. van Nierop, J. Decoster, P. Delespaul, C. Derom, M. de Hert, N. Jacobs, C. Menne-Lothmann, B. Rutten, E. Thiery, J. van Os, R. van Winkel, M. Wichers, I. Myin-Germeys, Is sensitivity to daily stress predictive of onset or persistence of psychopathology? *European Psychiatry*. **45** (2017), pp. 167–173.

51. G. van Rossum, F. L. Drake, *The python language reference manual* (Network Theory Ltd., 2011).

52. C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, T. E. Oliphant, Array programming with NumPy. *Nature*. **585** (2020), doi:10.1038/s41586-020-2649-2.

53. W. McKinney, in *Proceedings of the 9th Python in Science Conference* (Austin, TX, 2010), vol. 445, pp. 51–56.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.29.450360; this version posted July 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

54. A. Kuznetsova, P. B. Brockhoff, R. H. B. Christensen, lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software; Vol 1, Issue 13 (2017)* (2017) (available at https://www.jstatsoft.org/v082/i13).

55. D. J. Barr, R. Levy, C. Scheepers, H. J. Tily, Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*. **68**, 255–278 (2013).

56. A. Liaw, M. Wiener, Classification and regression by randomForest. *R news*. **2**, 18–22 (2002).

57. E. Smets, P. Casale, U. Großekathöfer, B. Lamichhane, W. de Raedt, K. Bogaerts, I. van Diest, C. van Hoof, S. Serino, A. Matic, D. Giakoumis, G. Lopez, P. Cipresso, Eds. (Springer International Publishing, Cham, 2016), pp. 13–22.

58. C. H. Cho, T. Lee, M. G. Kim, H. P. In, L. Kim, H. J. Lee, Mood prediction of patients with mood disorders by machine learning using passive digital phenotypes based on the circadian rhythm: Prospective observational cohort study. *Journal of Medical Internet Research*. **21** (2019), doi:10.2196/11029.

59. S. Gluth, N. Meiran, Leave-One-Trial-Out, LOTO, a general approach to link single-trial parameters of cognitive models to neural data. *eLife*. **8**, e42607 (2019).

60. A. Koul, C. Becchio, A. Cavallo, Cross-Validation Approaches for Replicability in Psychology. *Frontiers in Psychology*. **9** (2018), p. 1117.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.29.450360; this version posted July 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Acknowledgments

The authors would like to acknowledge Mike van Engelenburg from the technical group for his assistance with setting up the EMA platform, and Annemieke Smeets for her assistance with recruitment.

Funding:

This work was supported by the European Research Council (ERC2015-CoG 682591).

Author Contributions:

Conceptualization and set-up: RT, EH, MK, FK

Data Collection: RT, NK, BK

Data analysis: RT, EV, EH

Writing RT, EV, EH

Editing: All authors

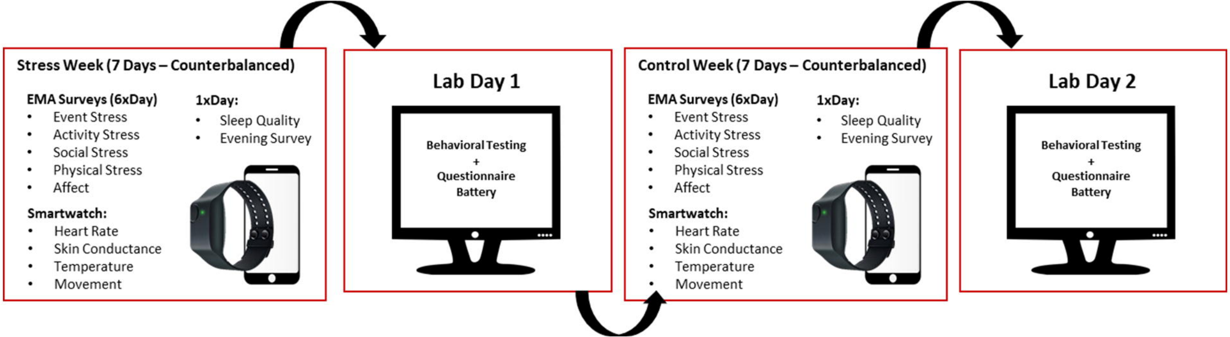
Funding: EH

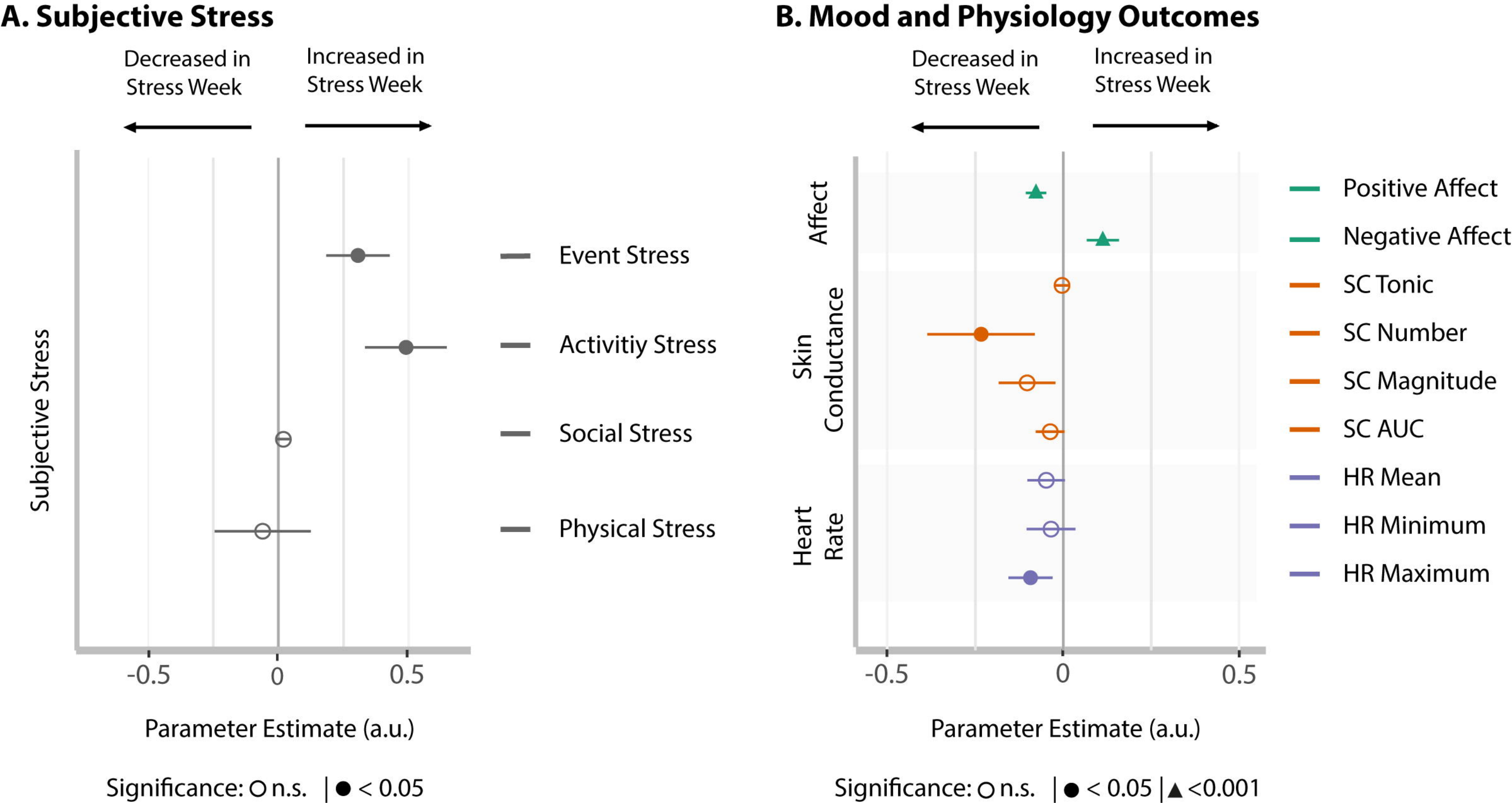
Competing interests:

The authors declare that they have no competing interests.

Data and code availability:

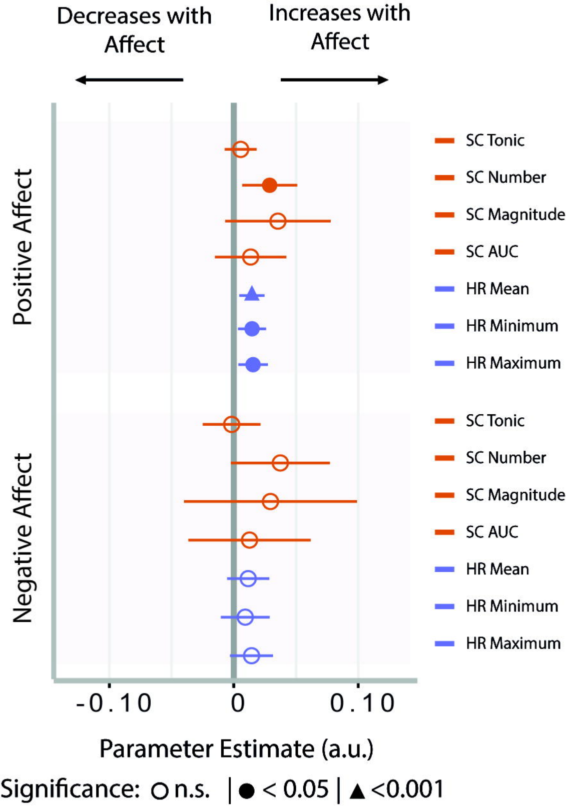
Due to the sensitive nature of the data, data is a made available through our institutional repository (data.donders.ru.nl) upon reasonable request. The code and the analysis notebook are made publicly available on GitHub and can be found in the supplementary text.

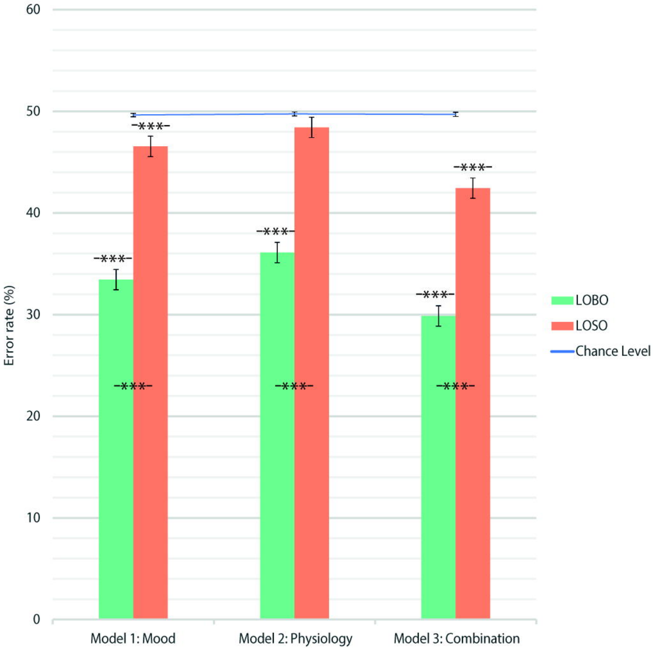

Stress Week (7 Days – Counterbalanced)
EMA Surveys (6xDay)
• Event Stress
• Activity Stress
• Social Stress
• Physical Stress
• Affect
Smartwatch:
• Heart Rate
• Skin Conductance
• Temperature
• Movement
1xDay:
• Sleep Quality
• Evening Survey
Lab Day 1
Behavioral Testing
+
Questionnaire Battery
Control Week (7 Days – Counterbalanced)
EMA Surveys (6xDay)
• Event Stress
• Activity Stress
• Social Stress
• Physical Stress
• Affect
Smartwatch:
• Heart Rate
• Skin Conductance
• Temperature
• Movement
1xDay:
• Sleep Quality
• Evening Survey
Lab Day 2
Behavioral Testing
+
Questionnaire Battery


A. Subjective Stress
Decreased in Stress Week
Increased in Stress Week
Subjective Stress
Event Stress
Activitiy Stress
Social Stress
Physical Stress
-0.5
0
0.5
Parameter Estimate (a.u.)
Significance: ○ n.s. | ● < 0.05
B. Mood and Physiology Outcomes
Decreased in Stress Week
Increased in Stress Week
Affect
Skin Conductance
Heart Rate
Positive Affect
Negative Affect
SC Tonic
SC Number
SC Magnitude
SC AUC
HR Mean
HR Minimum
HR Maximum
-0.5
0
0.5
Parameter Estimate (a.u.)
Significance: ○ n.s. | ● < 0.05 | ▲ <0.001

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.29.450360; this version posted July 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

A. Event Stress
Decreases with Stress
Increases with Stress
Affect
Skin Conductance
Heart Rate
-0.5
0
0.5
Parameter Estimate (a.u.)
B. Activity Stress
Decreases with Stress
Increases with Stress
-0.5
0
0.5
Parameter Estimate (a.u.)
Positive Affect
Negative Affect
SC Tonic
SC Number
SC Magnitude
SC AUC
HR Mean
HR Minimum
HR Maximum
C. Social Stress
Decreases with Stress
Increases with Stress
Affect
Skin Conductance
Heart Rate
-0.5
0
0.5
Parameter Estimate (a.u.)
D. Physical Stress
Decreases with Stress
Increases with Stress
-0.5
0
0.5
Parameter Estimate (a.u.)
Positive Affect
Negative Affect
SC Tonic
SC Number
SC Magnitude
SC AUC
HR Mean
HR Minimum
HR Maximum
Significance: ○ n.s. | ● < 0.05 | ▲ <0.001


Decreases with Affect
Increases with Affect
Positive Affect
Negative Affect
SC Tonic
SC Number
SC Magnitude
SC AUC
HR Mean
HR Minimum
HR Maximum
SC Tonic
SC Number
SC Magnitude
SC AUC
HR Mean
HR Minimum
HR Maximum
-0.10
0
0.10
Parameter Estimate (a.u.)
Significance: ○ n.s. | ● < 0.05 | ▲ <0.001


60
50
40
30
20
10
0
Error rate (%)
-***-
-***-
-***-
-***-
-***-
-***-
-***-
-***-
LOBO
LOSO
Chance Level
Model 1: Mood
Model 2: Physiology
Model 3: Combination