Non-compliance in an Ecological Momentary Assessment Study on Students` Mental Health.

Essen, Joël

Universiteit Leiden

Psychologie
Faculteit der Sociale Wetenschappen

Non-compliance in an Ecological Momentary Assessment Study on Students` Mental Health.

Joël Essen

Research Master Thesis *Clinical & Health Psychology*
Date: 28.09.2022
Supervisor: Ricarda Proppert
Second reader: Dr. Evin Aktar
Word Count: 8856

Abstract

Ecological momentary assessment (EMA) is a data collection method in which participants' current behaviors and experiences are sampled repeatedly in their natural environment. EMA has advantages over retrospective research methods, in that it reduces retrospective bias, increases ecological validity, and offers the possibility to observe dynamical changes of variables. However, EMA protocols are burdensome for participants and may interfere with their daily activities. This can lead to non-compliance over the course of a study. Missing data can subsequently decrease statistical power, and even induce bias. This paper explored whether missing data can be predicted by various variables related to students' primary motivation to participate, mental health, stress levels, and demographics. We analyzed data of the first cohort (N = 418) of the ongoing WARN-D project on student mental health. Participants completed a comprehensive baseline survey and took part in an 85-day long EMA study. We predicted overall rates of non-compliance by participant characteristics at baseline (Analysis 1) and weekly rates of non-compliance by time-varying factors during the EMA stage (Analysis 2). Analysis 1 showed that overall non-compliance can be predicted by baseline measures such as age, depression, substance use, and primary motivation to participate. Analysis 2 showed that weekly assessed time-varying measures like time into study, enjoyment of the study, weekly stress, anxiety, and depression may predict weekly rates of non-compliance. Participant's sex and smartphone operating system were not related to overall non-compliance. Summarizing, non-compliance rates of participants can be predicted by participant characteristics at baseline as well as by time-varying predictors. Our findings may inform future research on potential mechanisms behind noncompliance in EMA designs that should be considered to maximize participation rates while avoiding biased conclusions.

Laymen's Abstract

Ecological momentary assessment (EMA) is a data collection method, where participants are observed repeatedly in their natural environment for an extended period. Typical collected data are current behaviors and experiences. Doing so brings various advantages compared to other research designs that ask about past experiences and the memory thereof. One advantage is the short time between experience and assessment because people tend to report past experiences differently, when these are longer ago. However, EMA protocols place a considerable burden on participants, and may interrupt their daily activities. This can have the effect that participants increasingly miss surveys over the course of a study. If this happens only with participants having a special characteristic, the data may not be a good representation of reality. This in turn is problematic because it can lead to wrong or exaggerated conclusions. The current work explored whether non-compliance is associated with students' motivation to participate in our study, to their mental health, stress levels, and demographics. Included in this study was the first cohort (N = 418) of the ongoing WARN-D project on student mental health. We found that overall non-compliance is related to age, depression, substance use, and primary motivation to participate. Participants' sex and smartphone operating system had no effect on the rate of non-compliance. Further, we explored the effects of time-varying predictors like time into study, enjoyment of the study, weekly stress, anxiety, and depression on weekly non-compliance. To conclude, non-compliance of participants can be predicted by participant characteristics as well as time-varying predictors. Future studies may use this information to maximize participation and prevent biased conclusions.

Non-Compliance in an Ecological Momentary Assessment Study on Students` Mental Health

The ongoing process of digitization as well as changes in society continuously provide us with new opportunities to study and treat mental health problems. One possibility is to use electronic smart devices for real time assessment which enables researchers to zoom into participants' daily lives. Standard research methods assessing mental health problems retroactively often disregard the complex dynamical nature of mental health constructs. These methods are vulnerable to recall bias and low generalizability (Fortea et al., 2021). Daily diary protocols, also known as Ecological Momentary Assessment (EMA) methods, may counter these limitations (Fortea et al., 2021; Gillan & Rutledge, 2021). In contrast to previous pen and paper assessment methods, EMA nowadays often implements the usage of smartphones and smartwatches. EMA are data collection methods repeatedly sampling participants' current behaviors and experiences in real time and real-world settings (Shiffman et al., 2008). EMA focuses on current feelings and behaviors rather than capturing participants' autobiographical memories. Hence, EMA is less affected by recall and other biases which can, for example, cause overestimation of symptoms in clinical patients (Stone & Shiffman, 2002). Furthermore, through multiple assessments over time, with EMA data, researchers can develop a dynamic profile for participants' behavior and mood. These profiles may allow us to better characterize, understand, and work on mental health problems (Stavrakakis et al., 2015). Moreover, because the assessment is taking place in participants' everyday environments, EMA is considered to have better ecological validity and higher sensitivity to detect slight changes (Bolger et al., 2003). Despite the strengths of EMA, some challenges remain. One critical issue comes with an increased burden for participants. Completing multiple surveys per day, over a period of weeks or months, can be time consuming and interruptive of participants' everyday life. This leads to non-compliance which decreases statistical power, potentially introduces biases, and hence leads to false conclusions (Messiah et al., 2011; Shiffman et al., 2008; Sun et al., 2021).

Rubin, (1976), describes three types of missing data: data missing completely at random (MCAR), data missing at random (MAR), and data not missing at random

(NMAR). Data is considered MCAR if this is due to a random development (e.g., technical error that affects participants equally). In other words, the cause of non-compliance is not related to any other variable of interest and therefore estimated parameters of the population are unbiased. Considering data MAR, we assume that non-compliance is only associated with other observed variables (Rubin, 1976). Imagine a study on mental health involving repeated measures, which consists of a representative random sample of law students and psychology students. More psychology students completed the full study, meaning that there are more missing observations in the group of law students, but nonetheless the groups appear similar on all other attributes. In this example, whether a participant responded to the survey on mental health seems to be related to the type of study program they follow. This means that non-compliance is MAR as it can be explained by other observed variables. If missing data is not MCAR or MAR, it must be considered NMAR (Rubin, 1976). Participants' non-compliance for a variable is dependent on unobserved data. Following the above example, imagine that not only law students were less inclined to complete this study, but particularly those law students with little mental health problems. As participants' mental health issues remain unobserved until they answer the survey, this mechanism of non-compliance cannot be accounted for, and data is NMAR. NMAR data is more problematic than the other types, as we face higher obstacles analyzing the data and thus receiving valid estimates of the studied effects.

It is important to look at predictors of non-compliance in EMA studies since this type of research design might create additional obstacles to compliance. Furthermore, these obstacles are potentially even higher for certain populations such as socially disadvantaged participants (Acorda et al., 2021). This is a relevant topic since NMAR data can distort the estimation of within-person effects between factors (Rubin, 1976). For example, if a participant is more likely to miss surveys when feeling depressed, the estimate of the relationship between depression and other observed variables may be biased.

Potentially Relevant Predictors of Non-Compliance in EMA

Previous research on non-compliance in EMA has provided some initial information about potential factors that could be related to participants' compliance (Courvoisier et al., 2012; Gershon et al., 2019; Messiah et al., 2011; Murray, Brown, et al., 2022; Murray, Ushakova, et al., 2022; Rintala et al., 2019; Sokolovsky et al., 2014; Turner et al., 2017; Wen et al., 2017) These predictors can be roughly categorized in aspects of the used study designs, baseline predictors, and momentary predictors.

A meta-analysis found that studies that offer financial incentives seem to have better compliance compared to those that do not (Wrzus & Neubauer, 2022). However, other design characteristics such as duration and number of surveys presented per day only showed minimal effects. Furthermore, other studies showed that study duration had an increasing effect on non-compliance (Ono et al., 2019). Baseline predictors that were investigated are characteristics such as sex, age, mental health, substance use and other trait characteristics. Older age might have a positive effect on compliance (Ono et al., 2019). Additionally, being male or a polysubstance user can be related to lower participation (Messiah et al., 2011). So far, there is no unambiguous evidence whether mental health is related to non-compliance in EMA. Some studies found increased non-compliance among participants with a mental health diagnosis or stronger mental health problems (Gershon et al., 2019; Rintala et al., 2019) However, others could not replicate these findings (van Genugten et al., 2020). Other predictors that were observed are enthusiasm, being outside, higher negative affect (Murray, Brown, et al., 2022), and higher- (Sokolovsky et al., 2014) and lower levels of positive affect (Williams-Kerver et al., 2021). All these predictors were associated with higher non-compliance.

Summarizing, some predictors for non-compliance have been recommended by previous research. However, results cannot always be confirmed across studies or are based on only a small number of specific samples. Thus, there remains uncertainty about which, if any, of the observed variables are true predictors of non-compliance in EMA.

The Present Study

The goal of this study is to examine the extent to which overall non-compliance in an EMA stage can be predicted based on participant characteristics at baseline in a larger sample of students. Furthermore, we investigated effects of variables collected weekly during the EMA stage, to predict preceding participation on EMA surveys. Using weekly assessed retrospective- as well as baseline variables to predict non-compliance in an EMA is a novel strategy. This design can be valuable as it allows us to combine insights from static and dynamic predictors of non-compliance which can improve our understanding of participants non-compliance.

Given the value of detecting participants' characteristics associated with non-compliance in EMA studies and due to the limited amount of prior information in this new line of research, we explored a wide range of potential predictor. As recommended by prior research we investigated the effects of educational level, substance use, and sex. ((Messiah et al., 2011; Tsiampalis & Panagiotakos, 2020). Furthermore, we included SES as it seems to be a relevant predictor for depression and general life stressors (Freeman et al., 2016). Life stressors in turn are relevant for non-compliance insofar that participation might be a greater burden in some situations (e.g., work or a more serious health issue). Furthermore, people with lower financial security might have weaker smartphones (e.g., weaker battery, more technical problems) and less time to respond to surveys. We included primary motivation to participate since diverse types of motives for participation (e.g., financial reimbursement, interest in mental health, interest in trying out a smartwatch) might influence participants' compliance. Since students participating in English are most probably international students and might have a different mother language, we included language as a predictor.

. Participants completed daily as well as weekly surveys on Sundays. We tried to predict weekly non-compliance based on variables stemming from the weekly, as well as the daily EMA data. Weekly physical activity was included since prior research suggests that being a sport science student might be associated with higher noncompliance (Messiah et al., 2011). Potentially, this difference stems from time constraints as higher involvement in physical activity could lead to less involvement with smartphones. Weekly stress levels might be relevant since higher stress could prevent people from continuous participation.

On the other hand, participants with high stress levels could also show higher compliance due to higher interest in monitoring mental health.

Other variables were included based on elaborations of the WARN-D research team. These included predictors that are commonly studied in EMA protocols. We included weekly depression and anxiety since acute episodes of depression and anxiety might hinder participation. However, higher scores on these variables could also predict higher compliance since participants suffering from depression and anxiety might be more invested. Finally, the experienced enjoyment of the study will be included as it is likely a good predictor since people's joy in participating might directly indicate their compliance. Participants' joy was never directly assessed in prior EMA research about non-compliance.

For the current study, we investigated the following three research questions:

1. Is overall non-compliance in EMA unrelated to observed participant characteristics at baseline?
2. Which specific participant characteristics at baseline can be related to participants' non-compliance in our sample?
3. Can the rate of weekly non-compliance be predicted by time-varying mental health variables, self-reported reasons for non-compliance or time into the study?

Methods

Procedure

The present study is part of the WARN-D research project led by Dr. Eiko Fried at Leiden University. WARN-D is a prospective longitudinal study following students enrolled in the Netherlands over a period of two years to gain a better understanding of the stressors and experiences that may contribute to mental health problems. For the current paper we will make use of the data of this first cohort of the WARN-D project to explore potential predictors of non-compliance.

A WARN-D cohort goes through three research stages: a baseline survey (Stage 1), a 3-month EMA period (Stage 2), and a 2-year long follow-up period (Stage 3). At the

beginning of the study, after receiving consent, we asked participants to complete a questionnaire screening whether they qualify for participation.

Throughout the study we used two different data collection tools. First, we asked participants to respond to a self-report baseline questionnaire via the Qualtrics survey platform (*Qualtrics*, 2022). Second, during the EMA-stage, participants received four surveys per day at a semi-random prompt schedule from December 6th, 2021, until February 28th, 2022 (85 consecutive days). These EMA surveys were presented via the Android- or the iOS-Ethica data app (Ethica, 2021). Participants received prompts in the morning (9:49 a.m. to 10:19), around lunch time (1:39 p.m. to 2:19 p.m.), in the afternoon (5:39 p.m. to 6:19 p.m.) and in the evening (9:04 p.m. to 9:49 p.m.). After each notification participants had 20 minutes to respond to the survey. On Sundays, between 11:45 a.m. and 12:15 p.m., we presented an additional retrospective survey regarding participants' mood and behavior during the preceding week. We reimbursed participants depending on the number of completed surveys. In total one participant was able to receive 7.50 € for completing the baseline survey and up to 45 € for completing the EMA stage. Additionally, participants that completed stage 2 were eligible for receiving a personalized report about their development over the course of the EMA stage. These reimbursements were supposed to be motivating to continue participation since continuously delivering data leads to higher financial rewards and a higher quality report (Rimpler, 2022). The reimbursement strategy is described in more detail in the WARN-D protocol paper (Fried et al., 2022). Hypothesis and analyses for the present study were not pre-registered.

Participants

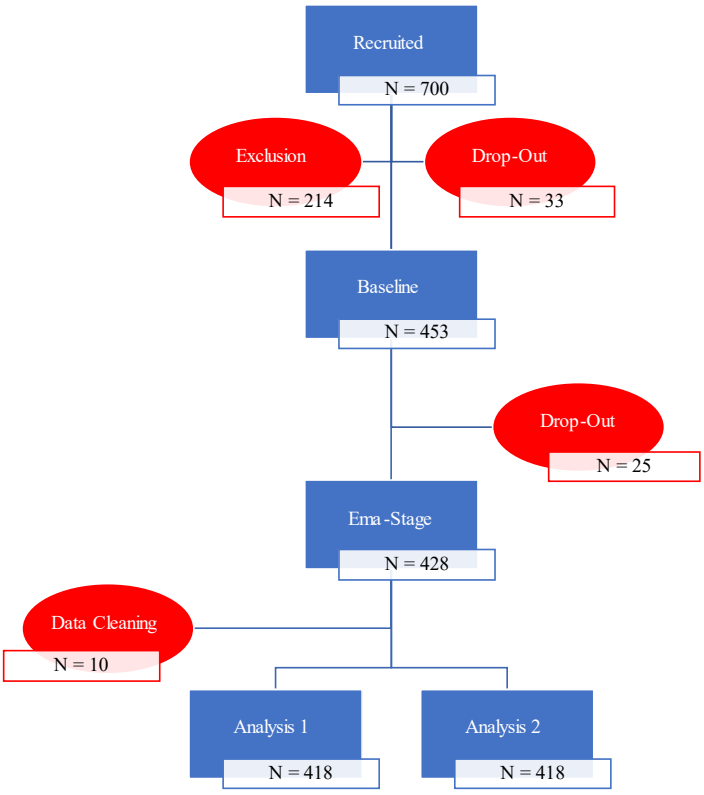
In total we recruited 700 participants via posters, e-mail, social media (Facebook, Instagram, Twitter), and word-of-mouth. After screening for exclusion, 453 participants took part in the baseline survey and 428 in the EMA stage. We excluded participants if they matched self-reported exclusion criteria for schizophrenia, psychosis or thought disorder; major depressive disorder, mania or bipolar disorder, severe substance use disorder, suffered from moderate to severe suicidal ideation, or stated that receiving information about burned calories would stress them.

After data cleaning we considered 418 participants for the analyses. An overview of the sample characteristics can be found in Table 1. A visualization of the participant flow for the current study can be found in Figure 1.

All the remaining participants were at least 18 years old, fluent in reading English or Dutch and students at a Dutch (applied) university or vocational school. Furthermore, all the participants were required to be in possession of a functional Android or iOS smartphone and have a European bank account. The WARN-D study was approved by the research ethics committees of the European Research Council and the Leiden University Research Ethics Committee (No. 2021-09-06-E.I.Fried-V2-3406).

Figure 1

Participant Flow-Chart



*Note. T*his study concerns participants of cohort 1 of the WARN-D study. The WARN-D protocol paper describes screening and exclusion criteria in more detail (Fried et al., 2022).

Measures

The present study took an exploratory approach and therefore evaluated a wide range of variables that may be associated with non-compliance. First, from baseline measures we included age, language, sex, educational level pursued, physical activity, depression scores, anxiety scores, substance use, primary motivation to participate, subjective socioeconomic status and OS. Second, from weekly assessed retrospective surveys we included reasons for missed prompts, experienced enjoyment of the current study, weekly anxiety, depression, and stress levels. All measures are described in detail in Appendix A.

Overall Non-Compliance. Participants' non-compliance was assessed by looking at the summed-up number of missed surveys over the course of the study. Not receiving a prompt due to an empty phone battery was considered as noncompliant. Participants could have missed a minimum of 0 and a maximum of 352 EMA surveys.

Weekly Non-Compliance. We received the outcome variable for Analysis 2, by calculating participants missed daily EMA surveys per week. Every participant had 12 observations. For each observation one participant can have a minimum of 0 up to 28 missing surveys. The distribution of weekly non-compliance scores is visualized in Figure 2.

Age, Language and Educational Level Pursued. Demographics were assessed using items designed by the core WARN-D team. We asked participants to provide their age on a scale from 18-90. We asked participants for their preferred language responding when responding to surveys. Participants were able to choose between English and Dutch. We asked participants which educational degree they are pursuing. We merged the priorly six categories of education pursued into 3 ("applied university degree pursued", "university degree pursued" and "other education pursued").

Sex. We assessed participants' sex using an item adopted from Caring Universities ((Vrije Universiteit Amsterdam, 2022), following the guidelines from "Williams Institute Best Practices for Asking Questions about Sexual Orientation on Surveys" (Almazan et al., 2009). We coded female sex as 1 and male sex as 0.

Smartphone Operating System (OS). We assessed participants OS using smartphone metadata. Participants either participated with iOS (coded as 1) or Android (coded as 2). 5 participants changed OS during the study. They were coded in a third category referred to as “Switched OS” (coded as 3).

Subjective Socioeconomic Status (SES). We assessed SES using the MacArthur Scale of Subjective Social Status (Adler et al., 2000). A challenge here was that many measures of students’ SES are based on parental information that students do not always have access to. We showed participants a ladder and asked them to select the rank that best represents where they think they stand compared to others in society. There were 10 steps on the ladder coded as 1 (bottom) to 10 (top).

Substance Use. We derived substance use scores using the Alcohol, Smoking and Substance Involvement Screening Test (ASSIST) ((Humeniuk & World Health Organization., 2010). For this study we used participants’ summed scores of all subscales. Therefore, participants could score a maximum of 42 and a minimum of 0.

Physical Activity. We assessed physical activity using items adapted from the short International Physical Activity Questionnaire (IPAQ-SF) (Lee et al., 2011). Participants were asked about how much time per week they would normally spend on (1) vigorous physical activities, (2) more moderate physical activities and (3) walking. All scores were transformed into minutes and summed together. Hence, participants could score a maximum of 30.240 minutes and minimum of 0 minutes on physical activity per week.

Primary Motivation to Participate. We looked at participants’ primary motivations by asking participants to rank specific motivations by importance from top to bottom. The categories were: “Payment”, “Try out a smartwatch”, “Improve my fitness by tracking my activities”, “Improve my wellbeing by tracking my mood”, “I hope it helps with a severe mental health problem I suffer from”, “Out of curiosity (to understand myself better)”, “Receiving the personalized feedback report on my mood development from the study”, “Supporting science”, “General interest in mental health”, “I hope it helps with a severe mental health problem that a friend or family member is suffering from” and “Something else". For this study we decided to only use the most important reason for each participant.

This means that categories were coded as mutually exclusive while participants priorly were allowed to provide multiple reasons.

Attention Deficit (Hyperactivity) Disorder (AD(H)D). AD(H)D was assessed by asking participants whether they ever in their life had AD(H)D. They were able to respond with either No (0) or Yes (1).

Depression & Weekly Depression. We derived depression scores by using an adaptation of the PHQ-9 (Kroenke et al., 2001). Compound symptoms (e.g., hypersomnia and insomnia) were pulled apart and items on hopelessness, and decreased sex drive were added. Further an item regarding irritability was added to the weekly retrospective survey. For the analyses of the current work, we left out the item about major depressive impairment since it is not part of the official scoring system. This leaves us with a total of 14 items at baseline and 15 items in the weekly retrospective surveys. For both baseline- and weekly depression, on a single item a participant could score from 0 to 3. Therefore, the maximum participants were able to score was 42 at baseline and 45 at weekly surveys. The minimum a participant was able to reach on both measurements was 0.

Anxiety & Weekly Anxiety. We recorded anxiety scores via the GAD-7 (Williams, 2014) at baseline and weekly with the retrospective surveys. Due to overlap with the adapted PHQ-9, we left out items 5 (restlessness) and 6 (irritability) from the GAD-7. Therefore, for both, baseline- and weekly anxiety, the maximum score a participant was able to reach was 15. The minimum a participant was able to reach was 0.

Weekly Stress. We assessed stress with the weekly retrospective surveys. We asked participants to which degree they would identify with the statement “This week was stressful to me.”. Participants responded on a 7-point Likert scale from 1 (“Not at all”) to 7 (“Very much”). The WARN-D research team developed this item.

Enjoyment of the WARN-D Study. We recorded enjoyment with the weekly retrospective surveys. We asked participants to respond to an item asking about their experience of participating in the WARN-D study this week. [2] Participants responded on a Likert scale from -3 (very negative) to 3 (very positive).

Reasons for Missed Prompts. We assessed reasons for missed prompts by asking participants about their two most important reasons for missing surveys this week. They were able to pick two items out of the following answer possibilities: “Missed no or very few surveys”, “Didn't see notification in time” “Saw notification but too busy/could not answer”, “Not motivated/interested”, “Not feeling well”, “Was asleep”, “Forgot”, “Technical problems”. A binary dummy variable was created for each category. However, categories were not mutually exclusive. Participants were scored with 1 for the two categories selected. Participants were scored with 0 for all categories they did not select.

Time into Study. We defined time into study as the number of the week in which we recorded an observation. Hence, the score on “time into study” can be minimally 1 (the first week) or maximally 12 (the last week).

Due to the daily schedule of EMA surveys and weekly schedule of the retrospective surveys, the WARN-D team adapted the phrasing asking participants either about their momentary experience or about their experiences during the previous week. Given that clear guidelines and validated measures still need to be developed for EMA, the measures used, and adjustments made for this study are based on existing literature, evaluations, prior work and currently ongoing projects of the WARN-D team and other EMA experts. A more detailed description of all measures used, and adjustments made for this study can be found Appendix A or in the WARN-D protocol paper (Fried et al., 2022).

Statistical procedure

All analyses were conducted in the free statistical environment R (R Core Team, 2022). For the R-code used for the analysis see supplementary materials (https://osf.io/5qh8m/). To investigate the first two research questions, we estimated a multiple linear regression (MLR) regarding the relationship of participants’ overall non-compliance with age, sex, user language, educational institute, physical activity, substance use, primary reason to participate, subjective socioeconomic status, anxiety, AD(H)D, depression, and participants’ OS.

To investigate the third research question, we conducted a stepwise multilevel regression analysis. This means we step- by-step added predictors to subsequent models

while keeping the predictors of prior models to find the best fit. Model 1 is the unconditional means model which only predicts from the mean of each subject individually. This model provides us with information about how much of the total variance in the outcome variable varies within and between persons. It predicts no change in non-compliance for the 12 consecutive weeks. Additionally, we included random intercepts to imply that participants differ on their likelihood of being non-compliant at the start. Since we expected changes in non-compliance over time, we included the variable time into study as a predictor for model 2. Model 2 is the unconditional growth model, which predicts a linear change of non-compliance per subject, while all subjects show the same slope. However, as we expected differences in slopes between participants, we created the conditional growth model 3, which allows each subject to have a random slope.

To see whether there is a main effect, we included the type of OS as a predictor in model 4. For model 5, we added the dummy variables regarding people's reasons to participate. For model 6, we added the psychological factors weekly-depression, -anxiety and -stress. For model 7 and 8 we subsequently added weekly depression and then weekly anxiety as random effects.

To estimate the best fitting model, we used the FML method, since we wanted to include fixed as well as random effects. Additionally, to compare successive models, we used the likelihood ratio test (LRT) since the tested models are nested.

Results

Analysis 1

Pre-Processing

Descriptive statistics of the sample included in analysis 1 are provided in Table 1 and Table 2. In the current study, 62.724 surveys were missed, and 87.932 surveys were completed, giving an overall completion rate of 58.36%. There were no compliance thresholds imposed, meaning that participants did not need to respond to specific number of surveys to be included. All scores on continuous variables were standardized. For the categorical variables, binary dummy variables were constructed. In analysis each dummy

variable was compared to its corresponding reference group. The effect of indicating female sex must be interpreted as opposing to indicating male sex. The effect of choosing Dutch language must be interpreted as opposed to choosing the English language. The effect of the different dummy categories for education level must be interpreted as opposed to the effect of "Other degree pursued". The effect of the different dummy categories for "Primary motivation to participate" must be interpreted as opposed to the effect of choosing "financial reimbursement" as primary motivation. The effect of having had AD(H)D must be interpreted as opposed to the effect of not having had AD(H)D. Finally, the effect of the different dummy categories for OS must be interpreted as opposed to the effect of iOS as an OS.

Since we found missing values for participants' OS and primary motivation to participate, we used listwise deletion for 10 cases. Deleting these cases seems to not be harmful to our results since missing data on these items is unlikely to be MNAR. 418 participants were analyzed for the final model.

Table 1

Descriptive Statistics of Continuous Variables Included in MLR

Variables	Min	Max	M	Mdn	SD
Non-compliance	3	350	144.17	112	95.01
Age	18	53	22.65	22	4.02
SES	1	10	6.90	7	1.49
Depression	0	35	9.29	8	6.04
Substance use	0	111	20.51	15	18.35
Anxiety	0	20	6.18	5	4.60
Physical activity	3	8640	776.70	520	883.13

Note. N = 418. Non-compliance is calculated as the overall number of missed surveys. A description of scales for all constructs can be found in Appendix A.

Table 2

Descriptive Statistics of Categorical Variables in MLR

Variables	N	Percentage
Sex		
Female	354	15.31%
Male	64	84.69%
User language		
English	200	47.84%
Dutch	218	52.15%
Education level pursued		
Applied university degree	39	9.33%
University degree	360	86.12%
Other degree	19	1.2%
OS		
iOS	182	43.54%
Android	231	55.26%
Switched OS	5	1.20%
Primary motivation to participate		
Payment	48	11.48%
Try out smartwatch	30	7.18%
Improve my fitness	14	3.34%
Improve wellbeing	48	11.48%
Hope it helps with own mental health problem	8	1.91%
Understand myself better	108	25.83%

Receive personalized feedback report	41	9.80%
Supporting science	68	16.27%
General interest in mental health	36	8.61%
Hope it helps with mental health problem of other	5	1.20%
Something else	12	2.87%
AD(H)D		
Yes	52	12.44%
No	366	87.56%

Note. A description of scales for all constructs can be found in Appendix A.

Multiple Linear Regression

We used MLR to test (1) whether overall non-compliance in EMA can be considered MCAR and (2) whether overall non-compliance in EMA can be considered MAR in this sample. Therefore, we predicted participants overall non-compliance by age, sex, user language, education pursued, physical activity, substance use, primary reason to participate, subjective socioeconomic status, anxiety, AD(H)D, depression, participants' OS.

The overall regression model was statistically significant, $R^2 = 0.17$, $F(23, 394) = 3.40$, $p < .001$. This indicates that 17% percent of the variance in overall non-compliance can be explained by this model. Thus, we cannot assume data is MCAR.

Regarding research question 2, looking at individual predictors we found a significant positive relationship of depression, $\beta = 0.14$, $t = 2.09$, $p = .038$. This effect indicates that participants with an increase of one standard deviation in depression scores on average miss around 18 surveys more. Similarly, substance use shows a positive effect on weekly non-compliance, $\beta = 0.22$, $t = 4.36$, $p < .001$. This effect indicates that participants with an increase in one standard deviation on this variable tend to miss 21 one surveys more. Further we found a significant negative effect of age, $\beta = -0.13$, $t = -$

2.67, p = .008. This effect indicates that being 1 standard deviation above the average age can be associated with completing 3 surveys more during our study. Finally, we found significant effects of the dummy categories regarding participants' primary motivation. All following effects need to be interpreted as opposed to the reference group of having selected financial reimbursement as primary motivation. "Try out a smartwatch" has a negative effect on non-compliance, β = -0.12, t = -2.10, p = .036. "Understand myself better" has a negative effect on non-compliance, β = -0.21, t = -2.86, p = .004. Receiving a "Personalized feedback report" shows a negative effect on non-compliance, β = -0.12, t = -2.01, p = .045. "General interest in mental health" shows a negative effect on non-compliance, β = -0.22, t = -3.65, p < .001. "Something else" as primary motivation has a negative effect on non-compliance, β = -0.11, t = -2.16, p = .031. Since all these negative effects are calculated in comparison to the reference group "financial reimbursement", it can be associated with an increase in overall non-compliance. For a detailed overview of all parameter estimates see Table 3.

Table 3

Summary of Standardized Model Parameter Fits

Variables	β	SE	t	p	95% CI
Intercept		46.33	6.15	< .001	
Age	-0.13	1.15	-2.67	.008	[-2.39, 2.13]
Sex*Female	-0.08	12.75	-1.62	.10	[-25.15, 24.99]
Language*Dutch	-0.06	9.68	-1.14	.25	[-19.08, 18.97]
SSS	-0.04	3.25	-0.77	.44	[-9.62, 9.54]
Pursued*Applied University	-0.03	27.00	-0.36	.72	[-53.11, 53.06]
Pursued*University	-0.13	23.38	-1.58	.11	[-46.10, 45.83]
Depression	0.14	1.04	2.09	.03	[-12.41, 12.58]
Substance use	0.22	0.26	4.36	< .001	[-9.07, 9.50]
Anxiety	-0.07	1.35	0.31	.30	[-12.41, 12.27]

Physical activity	0.04	0.005	0.73	.46	[-9.46, 9.53]
AD(H)D*Yes	-0.04	13.99	-0.82	.41	[-27.55, 27.47]
Try out smartwatch	-0.12	21.59	-2.10	.03	[-42.56, 42.31]
Improve my fitness	-0.10	27.79	-1.96	.05	[-54.73, 54.53]
Improve wellbeing	-0.09	18.72	-1.46	.14	[-36.91, 36.72]
Hope it helps with own mental health problem	-0.04	35.12	-0.82	.41	[-69.10, 69.02]
Understand myself better	-0.21	16.04	-2.86	.004	[-31.75, 31.32]
Personalized feedback report	-0.12	19.49	-2.01	.04	[-38.44, 38.19]
Supporting science	-0.12	17.39	-1,84	.06	[-34.32, 34.07]
General interest in mental health	-0.22	20.36	-3.65	< .001	[-40.25, 39.81]
Hope it helps with mental health of other	-0.06	43.12	-1.13	.19	[-84.84, 84.71]
Something else	-0.11	29.49	-2.16	.03	[-58.08, 57.86]
OS*Android	0.04	9.40	0.86	.39	[-18.43, 18.52]
OS*Switched OS	-0.03	41.12	-0.61	.54	[-80.87, 80.81]

Note. Effects of categorical variables must be interpreted regarding their reference categories (see pre-processing of analysis 1).

Model Assumption Checks

There seemed to be no multicollinearity as all VIF scores were close to 1 and below 5 and tolerance levels were above 0.2. The assumptions of homogeneity of variance, linearity, normality of residuals, and homoscedasticity of the predicted outcome were visually inspected via scatterplots. All these assumptions are met. There were no signs of funneling, suggesting homoscedasticity, however a small up- and down-swings in residuals indicate that the distribution of residuals is heavier-tailed than the theoretical distribution.

The Durbin-Watson test indicated that the assumption of independence of residuals was not met, Durbin-Watson value = 1.78, p = 0.03. Values for Cook's Distance showed

31 individual cases having a strong influence on the model. Deleting these cases results in a better fit, $R^2 = 0.24$, $F(23, 363) = 5.03$, $p < .001$. This indicates that the variance in non-compliance for these deleted cases cannot be well explained by our models.

Analysis 2

Pre-Processing

For the second analysis we used participants' weekly non-compliance as the outcome variable. Weekly non-compliance was aggregated by summing the number of daily missed surveys of one week. The variables assessed with the weekly retrospective surveys of the corresponding week were included as predictors. This leaves us with a minimum of 1 and a maximum of 12 observations per participant included in the study which are 3887 observations. Descriptive statistics of the data set used for analysis 2 are provided in Table 4 and Table 5. All scores on continuous variables were standardized. We created binary dummies for categorical variables. Here, the effect of each dummy category was calculated independent of a reference category. Meaning that the effect of scoring on each category must be interpreted as opposed to the effect of not scoring on this category. For the variable "reasons for missed prompts" participants were asked to select the two most important reasons for missed prompts. Therefore, the dummy variables for "reasons for missed prompts" are not mutually exclusive.

Looking at the data we found some missing observations among predictor variables. Participants sometimes were not able to finish the survey within the given 20-minute time span. We used imputation by participant means for missing scores in the weekly retrospective surveys. We imputed values for the continuous variables weekly stress (5 missing), weekly depression (25 missing), weekly anxiety (26 missing) and enjoyment of the study (32 missing). For missing observations on the categorical dummies for participants' reasons to miss prompts (35 missing) we used imputation by participant mode. Further, for 4 participants OS was unknown and another participant only provided one incomplete observation. Here, we did not have any reference to compute missing values from and hence these participants were excluded from the analysis. This left us with 3887 observations (min. 1 and max. 12 measurement points per

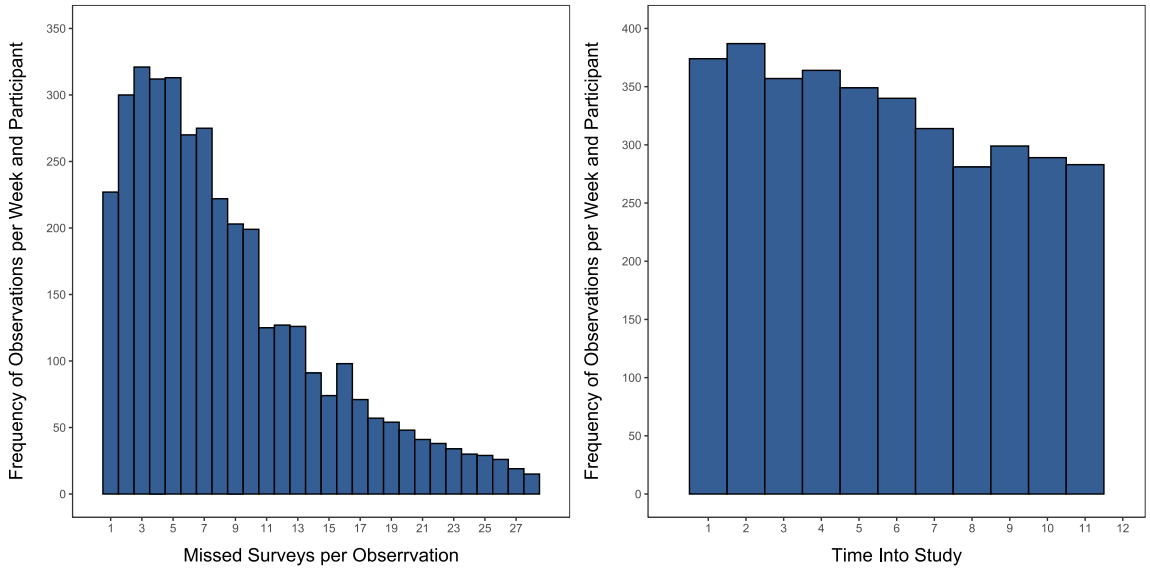
participant) of 418 participants. Since the number of imputed and deleted cases is low compared to the total number of observations, we can assume imputation and exclusion does not highly affect estimated model parameters.

Data Exploration

We created two histograms (Figure 2) showing the distribution of all observations of participants' weekly noncompliance (left) and observations of participants' weekly non-compliance across weeks (right). On the right histogram one can see that not all participants have observations in each single week since the number of observations varies and generally decreases per week. Furthermore, the distribution of weekly non-compliance scores shows that no implausible values occur, and that scores on weekly non-compliance are not normally distributed. We calculated the correlation between weekly non-compliance and non-compliance with the weekly retrospective surveys to check whether a prediction across weeks and across people is legitimated. We conducted a Shapiro-Wilk test for both weekly non-compliance- and missed weekly surveys to check whether our variables are normally distributed. The Shapiro-Wilk test on weekly non-compliance showed a significant departure from normality, $W = 0.92$, $p < .001$. Similarly, the Shapiro Wilk test on missed weekly surveys showed a significant departure from normality, $W = 0.80$, $p < .001$. Thus, we used Spearman's ρ for correlation analysis since it works for non-normally distributed data and is more robust with larger sample sizes. We detected a significant correlation between weekly non-compliance and missed weekly surveys, $\rho = 0.84$, $p < .001$. Furthermore, we investigated how many observations (weekly scores of non-compliance) in the data set have a corresponding weekly survey to predict from. Figure 3 shows a violin plot of completed weekly surveys against weekly non-compliance. One can observe that participants with higher non-compliance in a week are also likely to miss the corresponding weekly survey.

Figure 2

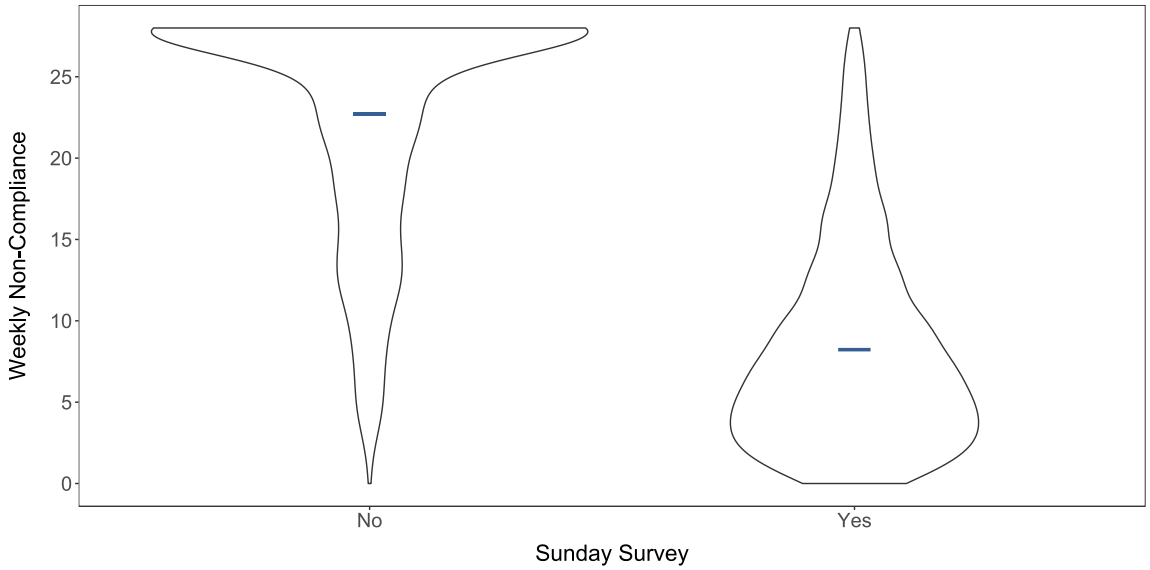
Distribution of Observations of Weekly Non-Compliance



Note. A histogram showing the distribution *of participants' weekly non-compliance scores* (left) and a histogram showing all observations of weekly non-compliance scores across weeks (right). One observation refers to on weekly assessment of one participant.

Figure 3

Violin Plot of Completed weekly Sunday Surveys Against Weekly Non-Compliance



Note. The presence of a corresponding weekly survey of the same week (x-axis) plotted against the distribution of participants' scores on weekly non-compliance (y-axis). Participants' weekly non-compliance was measured looking at weekly the number of missed surveys per week. The bar represents the mean non-compliance in each group. The width represents the frequency of observations of each individual value for participants' weekly non-compliance.

Table 4

Descriptive Statistics of Categorical Variables for the Multilevel Regression Analysis

	N (Observations)	*Min*	*Max*	*M*	*Mdn*	*SD*
Weekly non-compliance	3901	0	28	8.25	7	6.28
Weekly stress	3896	1	7	4.25	4	1.71
Weekly depression	3876	0	38	8.78	8	6.520
Weekly anxiety	3875	5	20	9.12	9	3.46
Enjoyment of study	3869	1	7	4.76	5	1.33

Note. A description of scales for all constructs can be found in Appendix A.

Table 5

Descriptive Statistics of Categorical Variables for the Multilevel Regression Analysis

	N (Observations)	Yes	No	Percentage
OS				
iOS	3888	1687	2200	43.39%
Android	3888	2144	1744	55.14%
Switched OS	3888	57	3830	1.47%
Reasons for missed prompts				
“Missed no or very few surveys”	3866	453	3413	11.72%
“Didn’t see notification in time”	3866	2509	1357	64.90%
“Saw notification but could not answer”	3866	1833	2033	47.41%
“Not motivated/interested”	3866	309	3557	7.99%
“Not feeling well”	3866	141	3725	3.65%
“Was asleep”	3866	1009	2857	26.10%

"Forgot"	3866	183	3683	4.73%
"Technical Problems"	3887	437	3429	11.30%

Note. One observation refers to on weekly assessment of one participant. A description of scales for all constructs can be found in Appendix A.

Multilevel regression analysis

We conducted a stepwise multilevel regression analysis to answer the third research question. Model 1 is the unconditional means model which has the mean of each subject individually as only a predictor. Furthermore, we included random intercepts, implying that participants differ on their likelihood of being non-compliant at the start. We calculated the intraclass correlation (ICC) from this model, according to the following formula:

$$ICC = \frac{\sigma^2 o}{\sigma^2 o + \sigma^2 e} = \frac{29.24}{29.24 + 16.20} = 0.643$$

This resulted in an ICC value of about .64, indicating that about 64% of the total variance is attributable to between-person variation whereas about 36% is attributable to within-person variation. This is a high ICC and multilevel modeling is needed to take this inter-dependency of observations into account. However, this also means there is high within-person variance to model using the time-varying predictor time into study.

We included the variable time into study as a predictor for model 2. Model 2 is the unconditional growth model, which predicts a linear change of non-compliance per subject, with all subjects showing the same slope. The conditional growth model 3, allows each subject to have a random slope. Looking at the models 2 and 3 we can only observe a small effect of time into study.

To see whether there is a main effect, we included the type of OS as a predictor in model 4. For model 5, we added the dummy variables regarding participants' reasons to participate. For model 6, we added the psychological factors weekly-depression, -anxiety and -stress. For model 7 and 8 we subsequently added weekly depression and then weekly anxiety as random effects. As it showed the best fit, we continued with model 7 for the final interpretation. A summary of the model parameters of the final model can be

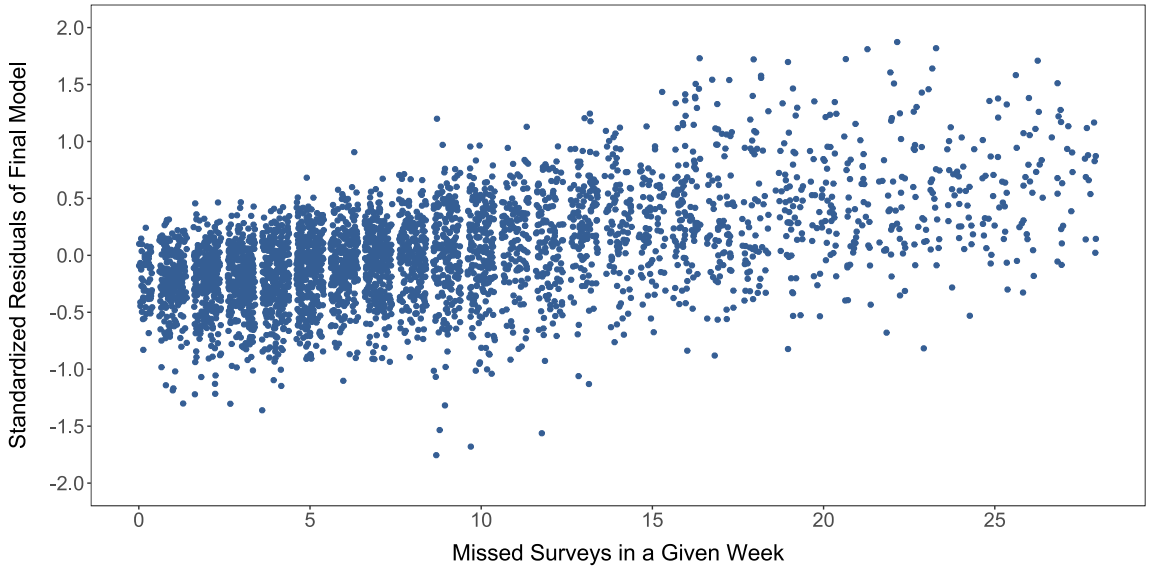
found in Table 6. A summary of model fit statistics can be found in Table 7. A summary of all models and the comparison of models can be found in Appendix B.

Assumption Checks

Looking at the predicted values compared to the observed scores (Figure 4), data points seem distributed randomly therefore the assumption of linearity does not seem violated. Furthermore, we created QQ-plots to check the normality of level-1 and level-2 residuals (Figure 5). Since the level-1 residuals plotted against the theoretical distribution follow an approximately straight line we can assume normality. For the level 2-residuals we created two QQ-plots, one concerning the intercept and one concerning the slope of the effect. Similarly, in each individual plot, all data points form an approximately straight line, meaning that we can assume normality. To check the assumption of residual homoscedasticity, we conducted an ANOVA of between subjects' residuals, $F(1, 3885) = 5.05$, $p = .02$. We conclude that the assumption of homoscedasticity is not met.

Figure 4

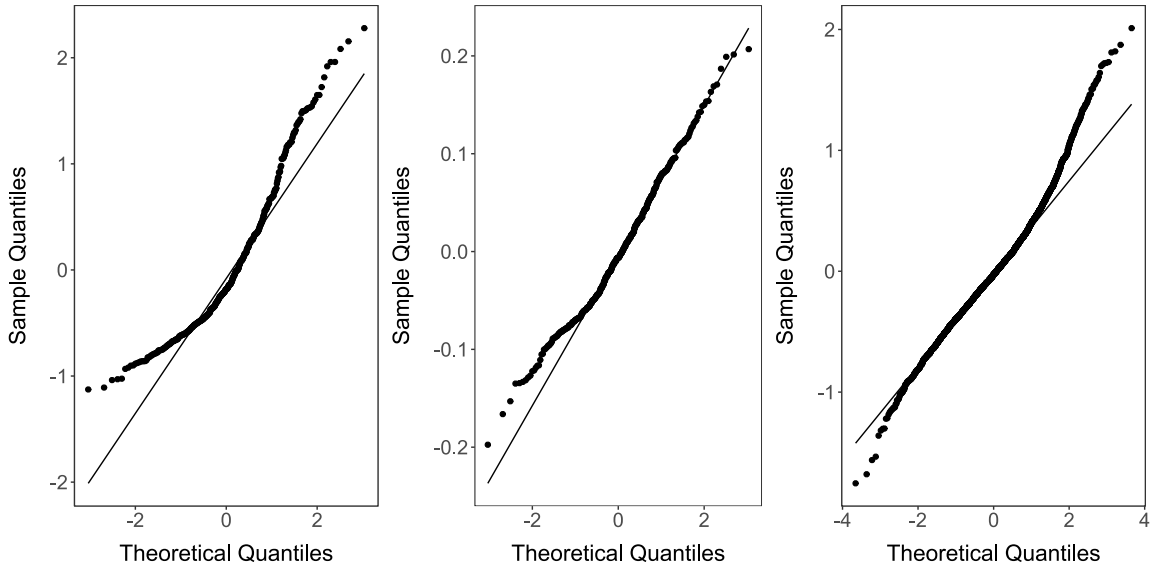
Scatterplot to Investigate Linearity Assumption



Note. Plotting observed weekly non-compliance scores against the final model residuals (model 7) to check the linearity assumption. We applied a slight jitter to increase differentiation between observations.

Figure 5

QQ-plot to Investigate Normality Assumption of Level-1 and Level-2 Residuals



Note. QQ plots to check the assumption of normality for intercepts (left), slopes (middle) and residuals (right).

Final Model

The best fitting model we found is model 7, predicting from time into study, phone OS, weekly enjoyment of study reasons for missed prompts, weekly depression, stress, and anxiety. The specific parameter estimates are listed in Table 6. The unexplained variance for subject specific intercepts is 23.24%. The explained within subject variance for the effect of time into study is 0.7%. The explained within subject variance for weekly depression is 1.7%. Furthermore, we found a small negative correlation between the regression coefficients for time into study and weekly depression r = - .12, SD = 0.13. The negative correlation between the effects of time into study and weekly depression indicates that participants who deviate stronger from the average estimate of time into study, show a slight tendency to deviate less from the average estimate of weekly depression. Summarizing, this shows that participants' compliance

that is more affected by weekly depression is less affected by time into study and vice versa.

Table 6

Model parameter estimates and fit for final model (model 7)

Parameter	β	SD/SE	p
Fixed effects			
Intercept	-0.21	0.13	.12
Time into study	0.07	0.01	< .001
Phone OS	0.00	0.08	.99
Enjoy Study	-0.16	0.01	< .001
“Missed no or very few surveys”	-0.32	0.04	< 0.001
“Didn’t see notification in time”	-0.02	0.03	.46
“Saw notification but could not answer”	0.04	0.03	.19
“Not motivated/interested”	0.29	0.05	< 0.001
“Not feeling well”	0.05	0.06	.40
“Was asleep”	0.01	0.03	.75
“Forgot”	0.02	0.05	.71
“Technical Problems”	0.39	0.04	< 0.001
Weekly depression	-0.04	0.02	0.09
Weekly stress	0.00	0.01	0.92
Weekly anxiety	-0.01	0.02	0.48
Random Effects			
Intercept	0.55	0.74	
Time into study	0.01	0.08	
Weekly depression	0.02	0.14	
Residuals	0.23	0.48	

log likelihood	-10653.0
deviance	21306.0
df Residuals	3826
AIC	21350.0
BIC	21487.6

Note. Summary of final model parameters and fit. For fixed effects we reported SDs. For random effects we reported SEs.

Table 7

Comparison of model fit statistics for models 1-8

Model	LL	deviance	df residuals	AIC	BIC	Δ χ^2	Δ df	p
1	-1154.7	23109.4	3898	23115.4	23134.2			
2	-11311.3	22622.6	3897	22630.6	22655.7	488.19	1	.000***
3	-11084.0	22168.0	3895	22180.0	22217.6	457.05	2	.000***
4	-11039.0	22078.0	3881	22092.0	22135.8	0.32	1	.57
5	-10678.2	21356.5	3838	21388.5	21488.6	499.1	10	.000***
6	-10657.9	21315.8	3829	21353.8	21472.7	9.45	3	.024*
7	-10653.0	21306.0	3826	21350.0	21487.6	9.08	3	.028*
8	-10649.3	21298.6	3822	21350.6	21513.2	7.06	4	.13

Note. LL= log likelihood. Δ χ^2 refers to the differences in χ^2 between models. Alle models were tested against the subsequent model. Model 4 did not result in a better fit. Hence, we continued testing model 3 against model 5.

Discussion

In the current study we examined the extent to which non-compliance can be predicted based on specific participant characteristics. Therefore, we looked at three different research questions. The first one was whether 1) overall non-compliance is

unrelated to predictor variables. The second one investigated the 2) relationship between specific participant characteristics and overall non-compliance. The third research question considered whether the 3) rate of weekly non-compliance can be predicted by weekly assessed time varying variables.

We found that 1) overall non-compliance is related to predictor variables. This allowed us to further investigate these relationships. Furthermore, we found that 2) some participant characteristics are related to overall non-compliance. Lastly, we found that also 3) weekly assessed retrospective variables can predict weekly non-compliance throughout an EMA stage.

Looking at our data set, 1), we could observe that non-compliance in our sample is not MCAR. This means that at least some variance between participants' tendency to miss surveys can be explained by the baseline variables included in the regression model. As mentioned earlier, ignoring the dependence between predictor variables and non-compliance can lead to systematic bias. Further, 2), we found that age, depression, substance use, and participants' primary motivation are relevant predictors for non-compliance in our sample. Similarly, to the study of Ono et al. (2019), in our sample, higher age indicated slightly higher compliance. Higher depression and substance use were associated with slightly lower compliance. The latter is in line with the findings of Messiah et al. (2011), who state that being a polysubstance user seems to be related to lower participation. Further, we compared different motivations for participation using "financial reimbursement" as a reference category. We found that participants with all other motives than financial reimbursement as their primary motive had higher compliance rates. Thus, in our study, financial reimbursement was associated with higher non-compliance. This is interesting as Wrzus & Neubauer (2022), found that studies offering financial incentives seem to have better compliance compared to those that do not. Pro-social motivations such as "Supporting Science" and "Hope that participating can help with a mental health problem of another person" are not testing significantly against financial motivation. Further, more intrinsic motivations (e.g., participating to understand oneself better or due to general interest in science) show the largest negative effects on overall non-compliance compared to financial reimbursement. Additionally, unlike in the study of Messiah et al. (2011), we could not find an association between sex

and non-compliance. That could mean that the differences in compliance for sex found in prior studies might be explained better by other variables.

Regarding research question 3, we found that an average participant in our study misses 6-7 (25%) surveys per week with an increase of roughly one survey every two weeks. Furthermore, higher scores of weekly assessed enjoyments of the study are related to lower weekly non-compliance. Participants' weekly depression and anxiety scores have a small decreasing effect on weekly non-compliance. However, there seems to be a larger difference in the effect of weekly depression between individual participants. This could mean that differences in weekly depression between different participants might be more relevant to participants' weekly non-compliance than different observations of depression between weeks.

Additionally, participants' self- reported reasons for missed prompts had indeed an effect on weekly non-compliance. We found that people indicating no or very few missed surveys have on average 2-3 more surveys in the respective week. Not seeing the survey in time or being too busy only showed minor negative effects. Again, having low interest or motivation had a larger effect on weekly non-compliance with an increase of almost 2 missed surveys for respective weeks. We found that when technical issues were the reason for missed surveys participants missed 2-3 surveys more on average. Finally, in our sample, participants' weekly stress and type of OS did not predict weekly non-compliance.

Strengths and Limitations

The design of this study comes with some reasonable strengths. Participants were asked for their reasons to participate, about their enjoyment during the study and for the reasons why they missed prompts during the study. Enjoyment during a study was never used in this setting but is likely to be good predictors of compliance. Receiving this information can help us to improve confidence in our findings and increase compliance for future EMA designs. Further, evaluating the relationship between primary motivation for participation and compliance might help us to find a well-balanced reward for extensive longitudinal studies. A well-balanced reward is crucial for compliance since too high

financial rewards might alter behavior and too low financial rewards might lead to lower compliance.

The estimated effects of baseline depression and substance use on overall non-compliance are small and there seems to be no effect of anxiety and AD(H). Designing a study that is inclusive of these populations was important since the WARN-D research project aims to observe people that are at risk of developing mental health problems. Similarly, participants' age and OS did not predict participants' not compliance.

Finally, one novel aspect about the current study comes with the setup of the study since we used weekly assessed retrospective- as well as baseline variables to predict non-compliance in an EMA. We included state and trait scores for depression and anxiety to capture dynamics within these constructs. This is a valuable procedure as it allows us to combine insights from static and dynamic predictors of non-compliance. Receiving this information can improve our understanding of how and why participants miss surveys in EMA.

However, some challenges to this approach remain. Looking at our results, we observed that that participants with weekly non-compliance are also likely to miss the corresponding weekly retrospective surveys. Further, participants that do not miss a lot of surveys per week are more likely to respond to the corresponding weekly retrospective survey. This is problematic since we aimed to predict weekly non-compliance from variables assessed in weekly surveys, however, for participants with the highest non-compliance, we tend to have the least information to predict from. Therefore, we must consider missing data MNAR, and hence our model parameters are likely to be biased. For future analysis one could think about predicting non-compliance in EMA from additional measures that are more likely to be complete (e.g., smartwatch data). Another recommendation for future research would be to differentiate between participants missing surveys and participants dropping out. When participants drop out, they obviously do not respond to either daily EMA or weekly retrospective surveys. A differentiation between diverse types of non-compliance could be beneficial to explore the different mechanisms behind participants missing surveys. For the current study we used an item asking for reasons for missed prompts, but we cannot trace back which

exact prompt was missed because of which specific reason. For different types of non-compliance there might be differential predictors.

Further, for our sample the assumption of independence of observations was not met. This might be because some participants were in touch and might have influenced each other. Another reason for this could be that the variable reasons for missed prompts participants were recorded in a way that participants were asked to select their two most important reasons for missed prompts which automatically leads to not selecting the other options. Therefore, observations between the different dummy variables were not independent from each other. To avoid bias, future studies should improve the assessment of reasons for missed prompts.

It must be mentioned that data for the first WARN-D cohort was collected during a Covid lockdown. This could potentially lead to less generalizability to future studies due to confounding factors (e.g., higher, or lower completion rates, different mood in general, …). Further, working with a student sample leaves us with a mean age of 22.6 years. Future research should consider a more diverse sample including older people as well. Communication with participants via Instagram could have influenced their engagement. Meaning that if a post reminding people to participate was posted on Instagram, participants that usually would have missed a prompt could have been motivated or reminded to engage. This should be considered in future research about non-compliance when using social media as a communication tool.

Moreover, people that selected financial reimbursement as primary motivation did show higher rates of non-compliance. This could have several reasons. One likely possibility is that the scoring we used for participants' motivation did not fully reflect people's true multifaceted motivations. At baseline, we asked participants to rank their motivations in an order. However, for simplicity, we decided to only use participants' primary motivations. Another possibility could be that people participating for financial reimbursement did not think that reimbursement in our study was high enough to compensate for their efforts. Potentially people found an activity that is more financially beneficial. Another reason could be that financial compensation generally is not a good motivation to participate. This would be contrary to the findings of Smyth et al. (2021), who claim that higher financial compensation increased willingness to participate. These

different findings could be due to the amount and timing of compensation. In our study, participants received their compensations at three points: after 3 months, after 1 year, and after 2 years. On these 3 occasions, we calculate how many questionnaires participants did respond to and they are reimbursed accordingly. Maybe, if participants are reimbursed immediately after responding to a survey, financial compensation has a more beneficial effect.

Conclusion

Summarizing, we found that the compliance rate of participants can be predicted by participant characteristics as well as weekly recorded time-varying predictors. Therefore, we can say that some of the data is MAR. For future studies, these variables could be considered and hence maybe some bias prevented. Moreover, learning about predictors of participation in EMA studies could help us to enhance recruitment and measurement strategies (Murray, Ushakova, et al., 2022). Specific participants who are less likely to participate in an EMA study could be oversampled and targeted with higher resources. Similarly, if the sample does not need to be representative, one could target participants that are likely to participate to save valuable resources. Further, a study's load could be adjusted regarding participants' needs, placing a lower burden if needed (Murray, Ushakova, et al., 2022).

For future research we would recommend a differentiation between different types of non-compliance. This seems especially important since participants reported technical problems as one of the most important reasons for missing prompts. Further, potentially a study design should rather use intrinsic rewards or at least well-balanced financial rewards. It would be highly interesting to see which other motivations work better than financial reimbursement. Additionally, since participants show more non-compliance the longer, they are taking part in a study, it might be beneficial to conduct a time series analysis when working with this kind of research design.

While there are still many unanswered questions about the prediction of participant non-compliance, with this study we were able to add further insights to the

topic. Discovering reasons for non-compliance and preventively addressing them, future research can further work towards incorporating relevant variables for sampling plans, reaching maximal compliance and hence produce better results and save valuable resources. Furthermore, considering these variables for future analyses could help prevent bias.

References

Acorda, D., Businelle, M., & Maria, D. S. (2021). Perceived impacts, acceptability, and recommendations for ecological momentary assessment among youth experiencing homelessness: Qualitative study. *JMIR Formative Research*, *5*(4). https://doi.org/10.2196/21638

Adler, N. E., Epel, E. S., Castellazzo, G., & Ickovics, J. R. (2000). Relationship of subjective and objective social status with psychological and physiological functioning: Preliminary data in healthy white women. *Health Psychology*, *19*(6), 586–592. https://doi.org/10.1037/0278-6133.19.6.586

Almazan, E., Ayala, G., Lee Badgett, M. v, Bye, L., Chae, D. H., Cochran, S., Díaz, R., Klawitter, M., Landers, S., Saewyc, E., & Sell, R. (2009). *Best Practices for Asking Questions about Sexual Orientation on Surveys*.

Bolger, N., Davis, A., & Rafaeli, E. (2003). Diary Methods: Capturing Life as it is Lived. *Annual Review of Psychology*, *54*, 579–616. https://doi.org/10.1146/annurev.psych.54.101601.145030

Courvoisier, D. S., Eid, M., & Lischetzke, T. (2012). Compliance to a Cell Phone-Based Ecological Momentary Assessment Study: The Effect of Time and Personality Characteristics. *Psychological Assessment*, *24*(3), 713–720. https://doi.org/10.1037/aO026733

Ethica (No. 550). (2021). Ethica Data Services Inc. .

Fortea, L., Tortella-Feliu, M., Juaneda-Seguí, A., de la Peña-Arteaga, V., Chavarría-Elizondo, P., Prat-Torres, L., Soriano-Mas, C., Lane, S. P., Radua, J., & Fullana, M. A. (2021). Development and Validation of a Smartphone-Based App for the

Longitudinal Assessment of Anxiety in Daily Life. *Assessment*. https://doi.org/10.1177/10731911211065166

Freeman, A., Tyrovolas, S., Koyanagi, A., Chatterji, S., Leonardi, M., Ayuso-Mateos, J. L., Tobiasz-Adamczyk, B., Koskinen, S., Rummel-Kluge, C., & Haro, J. M. (2016). The role of socio-economic status in depression: Results from the COURAGE (aging survey in Europe). *BMC Public Health*, *16*(1). https://doi.org/10.1186/s12889-016-3638-0

Fried, E. I., Proppert, R. K. K., & Rieble, C. (2022). *WARN-D Protocol Paper* (No. 12761885). 17. https://doi.org/none

Gershon, A., Kaufmann, C. N., Torous, J., Depp, C., & Ketter, T. A. (2019). Electronic Ecological Momentary Assessment (EMA) in youth with bipolar disorder: Demographic and clinical predictors of electronic EMA adherence. *Journal of Psychiatric Research*, *116*, 14–18. https://doi.org/10.1016/J.JPSYCHIRES.2019.05.026

Gillan, C. M., & Rutledge, R. B. (2021). *Annual Review of Neuroscience Smartphones and the Neuroscience of Mental Health*. https://doi.org/10.1146/annurev-neuro-101220

Humeniuk, Rachel., & World Health Organization. (2010). *The Alcohol, smoking and substance involvement screening test (ASSIST) : manual for use in primary care*. World Health Organization.

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). *The PHQ-9 Validity of a Brief Depression Severity Measure*.

Lee, P. H., Macfarlane, D. J., Lam, T. H., & Stewart, S. M. (2011). Validity of the international physical activity questionnaire short form (IPAQ-SF): A systematic review. In *International Journal of Behavioral Nutrition and Physical Activity* (Vol. 8). https://doi.org/10.1186/1479-5868-8-115

Messiah, A., Grondin, O., & Encrenaz, G. (2011). Factors associated with missing data in an experience sampling investigation of substance use determinants. *Drug and Alcohol Dependence*, *114*(2–3), 153–158. https://doi.org/10.1016/j.drugalcdep.2010.09.016

Murray, A. L., Brown, R., Zhu, X., Speyer, L. G., Yang, Y., Xiao, Z., Ribeaud, D., & Eisner, M. (2022). *Momentary predictors of compliance in an ecological momentary assessment study of young adults' mental health*. https://doi.org/10.31234/OSF.IO/5R9JS

Murray, A. L., Ushakova, A., Zhu, X., Yang, Y., Xiao, Z., Brown, R., Speyer, L., Ribeaud, D., & Eisner, M. (2022). *Who participates in ecological momentary assessment (EMA) studies? Predicting participation in a general population EMA study*.

Ono, M., Schneider, S., Junghaenel, D. U., & Stone, A. A. (2019). What Affects the Completion of Ecological Momentary Assessments in Chronic Pain Research? An Individual Patient Data Meta-Analysis. *J Med Internet Res 2019;21(2):E11398 Https://Www.Jmir.Org/2019/2/E11398*, *21*(2), e11398. https://doi.org/10.2196/11398

Qualtrics. (2022). Qualtrics LLC.

R Core Team. (2022). *R: The R Project for Statistical Computing*. https://www.r-project.org/

Rimpler, A. (2022). *Generating Feedback Reports for Ecological Momentary Assessment Data*. https://hdl.handle.net/1887/3453447

Rintala, A., Wampers, M., Myin-Germeys, I., & Viechtbauer, W. (2019). Response compliance and predictors thereof in studies using the experience sampling method. *Psychological Assessment*, *31*(2), 226–235. https://doi.org/10.1037/PAS0000662

Rubin, D. B. (1976). *Inference and Missing Data*. *63*(3), 581–592. https://about.jstor.org/terms

Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological momentary assessment. *Annual Review of Clinical Psychology*, *4*, 1–32. https://doi.org/10.1146/annurev.clinpsy.3.022806.091415

Smyth, J. M., Jones, D. R., Wen, C. K. F., Materia, F. T., Schneider, S., & Stone, A. (2021). Influence of ecological momentary assessment study design features on reported willingness to participate and perceptions of potential research studies: an experimental study. *BMJ Open*, *11*(7), e049154. https://doi.org/10.1136/BMJOPEN-2021-049154

Sokolovsky, A. W., Mermelstein, R. J., & Hedeker, D. (2014). Factors Predicting Compliance to Ecological Momentary Assessment Among Adolescent Smokers. *Nicotine & Tobacco Research*, *16*(3), 351–358. https://doi.org/10.1093/NTR/NTT154

Stavrakakis, N., Booij, S. H., Roest, A. M., de Jonge, P., Oldehinkel, A. J., & Bos, E. H. (2015). Supplemental Material for Temporal Dynamics of Physical Activity and

Affect in Depressed and Nondepressed Individuals. *Health Psychology*. https://doi.org/10.1037/hea0000303.supp

Stone, A. A., & Shiffman, S. (2002). *Capturing Momentary, Self-Report Data: A Proposal for Reporting Guidelines*. https://academic.oup.com/abm/article/24/3/236/4633694

Sun, J., Rhemtulla, M., & Vazire, S. (2021). Eavesdropping on Missing Data: What Are University Students Doing When They Miss Experience Sampling Reports? *Personality and Social Psychology Bulletin*, *47*(11), 1535–1549. https://doi.org/10.1177/0146167220964639

Tsiampalis, T., & Panagiotakos, D. B. (2020). Missing-data analysis: Socio-demographic, clinical and lifestyle determinants of low response rate on self-reported psychological and nutrition related multi- item instruments in the context of the ATTICA epidemiological study. *BMC Medical Research Methodology*, *20*(1). https://doi.org/10.1186/s12874-020-01038-3

Turner, C. M., Coffin, P., Santos, D., Huffaker, S., Matheson, T., Euren, J., DeMartini, A., Rowe, C., Batki, S., & Santos, G. M. (2017). Race/ethnicity, education, and age are associated with engagement in ecological momentary assessment text messaging among substance-using MSM in San Francisco. *Journal of Substance Abuse Treatment*, *75*, 43–48. https://doi.org/10.1016/J.JSAT.2017.01.007

van Genugten, C. R., Schuurmans, J., Lamers, F., Riese, H., Penninx, B. W. J. H., Schoevers, R. A., Riper, H. M., & Smit, J. H. (2020). Experienced Burden of and Adherence to Smartphone-Based Ecological Momentary Assessment in Persons

with Affective Disorders. *Journal of Clinical Medicine 2020, Vol. 9, Page 322*, *9*(2), 322. https://doi.org/10.3390/JCM9020322

Vrije Universiteit Amsterdam. (2022). *Caring Universities*.

Wen, C. K. F., Schneider, S., Stone, A. A., & Spruijt-Metz, D. (2017). Compliance With Mobile Ecological Momentary Assessment Protocols in Children and Adolescents: A Systematic Review and Meta-Analysis. *J Med Internet Res 2017;19(4):E132 Https://Www.Jmir.Org/2017/4/E132*, *19*(4), e6641. https://doi.org/10.2196/JMIR.6641

Williams, N. (2014). The GAD-7 questionnaire. *Occupational Medicine*, *64*(3), 224. https://doi.org/10.1093/occmed/kqt161

Williams-Kerver, G. A., Schaefer, L. M., Hazzard, V. M., Cao, L., Engel, S. G., Peterson, C. B., Wonderlich, S. A., & Crosby, R. D. (2021). Baseline and momentary predictors of ecological momentary assessment adherence in a sample of adults with binge-eating disorder. *Eating Behaviors*, *41*. https://doi.org/10.1016/J.EATBEH.2021.101509

Wrzus, C., & Neubauer, A. B. (2022). Ecological Momentary Assessment: A Meta-Analysis on Designs, Samples, and Compliance Across Research Fields. *Assessment*. https://doi.org/10.1177/10731911211067538/ASSET/IMAGES/LARGE/10.1177_10731911211067538-FIG6.JPEG

Appendix A

Constructs and Corresponding Scales

Construct	Item	Scale	Reference
Total & weekly non-compliance		Missed surveys overall (0 - 352) Missed weekly surveys (0 - 28)	(Fried et al., 2022)
Age	*"How old are you?"*	Numerical (Years), 18-90	(Fried et al., 2022)
Language	*"What is your preferred language?"*	0 = EN, 1 = NL	(Fried et al., 2022)
Education level	*"What kind of degree are you currently pursuing?"*	4 = Vocational school degree (MBO or equivalent); 5 = Applied University / HBO Bachelor's degree or equivalent; 6 = Applied University / HBO Master's degree or equivalent; 7 = University / WO Bachelor's degree or equivalent, 8 = University / WO Master's degree or equivalent; 10 = Other	
Sex	*"What was your sex at birth, as it appears on your birth certificate?"*	0 = Male, 1 = Female "	Adopted from Caring Universities (Vrije Universiteit Amsterdam, 2022), following the guidelines from "Williams

			Institute Best Practices for Asking Questions about Sexual Orientation on Surveys" (Almazan et al., 2009).
Subjective socioeconomic status	*""At the top of the ladder (rank 10) are the people who are the best off, those who have the most money, best education, and the best jobs. At the bottom of the ladder (rank 1) are the people who are the worst off, those who have the least money, least education, worst jobs, or no job. Please select the number that best represents where you think you stand on the ladder."*	1-10 (no labels)	MacArthur Scale of Subjective Social Status (Adler et al., 2000)
Substance use			
	"In your life, which of the following substances have you ever used? We are only interested in recreational use—if you used the substances for medical reasons, click "no".	0 = No; 3 = Yes	Alcohol, Smoking and Substance Involvement Screening Test (ASSIST) (Humeniuk & World Health Organization., 2010)
	"In the past 3 months, have you used the following substance(s)? We are only interested in recreational use—if you used the substances for medical reasons, click "never"".	0 = Never; 2 = Once or twice; 3 = Monthly; 4 = Weekly; 6 = Daily or almost daily; 7 =Multiple times a day	

	"During the past 3 months, how often have you had a strong desire or urge to use these substances?"	0 = Never; 3 = Once or twice; 4 = Monthly; 5 = Weekly; 6 = Daily or almost daily	
	"During the past 3 months, how often has your use of these substance led to health, social, legal, or financial problems?"	0 = Never; 4 = Once or twice; 5 = Monthly; 6 = Weekly; 7 = Daily or almost daily	
	"During the past 3 months, how often have you failed to do what was normally expected of you because of your use of these substances?"	0 = Never; 5 = Once or twice; 6 = Monthly, 7 = Weekly; 8 = Daily or almost daily	
	"Has a friend or relative or anyone else ever expressed concern about your use of these substances?"	0 = No, Never; 6 = Yes, in the past 3 months; 3 = Yes, but not in the past 3 months	
	"Have you ever tried and failed to control, cut down or stop using these substances?"	0 = No, Never; 6 = Yes, in the past 3 months; 3 = Yes, but not in the past 3 months	
Physical activity	*"During the last 7 days, on how many days did you do 10 minutes or more vigorous physical*	1-7 (Days)	Items designed by the WARN-D team adapted from the short

activities like cardio (e.g., running, or fast bicycling) or exercising at the gym? Vigorous physical activities take hard physical effort and make you breathe much harder than normal."		International Physical Activity Questionnaire (IPAQ-SF) (Lee et al., 2011)
"On days where you performed vigorous physical activities, how much time did you usually spend on it? *For example, if you spent 3.5 hours on them, please fill the question in like this: "*	0-24 (hours), 0-60 (minutes)	
"During the last 7 days, on how many days did you do 10 minutes or more moderate physical activities like cycling or carrying a large bag home from the supermarket? Moderate physical activities that take moderate physical effort and make you breathe somewhat harder than normal."	1-7 (Days)	
"On days on which you performed moderate physical activities, how much time did you usually spend on them?"	0-24 (hours), 0-60 (minutes)	
"During the last 7 days, on how many days did you walk 10 minutes or more? This includes at work, at school, and at home, walking to travel from place to place, and any other walking that you have done for recreation, sport, exercise, or leisure."	1-7 (Days)	

	"On the days that you walked, how much time did you usually spend walking?"	0-24 (hours), 0-60 (minutes)
Depression & weekly depression	*"Over the past 2 weeks, how often have you been bothered by the following problems?"* *"Little interest or pleasure in doing things",* *"Feeling down or depressed",* *Feeling hopeless",* *"Trouble falling asleep or staying asleep",* *"Sleeping too much",* *"Feeling tired or having little energy",* *"Poor appetite",* *"Overeating",* *"Feeling bad about yourself – or that you're a failure or have let yourself or your family down",* *"Trouble concentrating on things, such as reading or watching television",* *"Moving or speaking so slowly that other people could have noticed"* *"Being so fidgety or restless that you have been moving around a lot more than usual"* *"Thoughts that you would be better off dead or of hurting yourself in some way"* *"Little interest in sex"*	0 = Not at all; 1 = Several Days, 2 = More Than Half the Days; 3 = Nearly Every Day	An adaptation of the PHQ-9 with 14 items for baseline and 15 items for the weekly Sunday survey. Compound symptoms (e.g., hypersomnia and insomnia) were pulled apart and items on hopelessness, and decreased sex drive were added, leading to a total of 15 items. An additional item on irritability was included to assess weekly depression within the Sunday surveys. For further information about this see the WARN-D protocol paper (Fried et al., 2022).
Anxiety & weekly anxiety	*"Over the last 2 weeks, how often have you been bothered by the following problems?"* *"Feeling nervous, anxious, or on edge ",* *"Not being able to stop or control worrying",*	0 = Not at all; 1 = Several days; 2 = More than half the days; 3 = Nearly every day	GAD-7 (Williams, 2014)

	"Worrying too much about different things" *"Trouble relaxing",* *"Being so restless that it is hard to sit still"* *"Becoming easily annoyed or irritable",* *"Feeling afraid, as if something awful might happen"*		
Weekly Stress	*"This week was stressful for me."*	1-7 (Not at all - Very much)	(Fried et al., 2022)
Primary motivation to participate	*"Why are you participating in our study? Please select the motivations that apply to you by dragging them into the box on the right, and rank by importance (that is, pull the most important one to the top)."*	1 = Payment; 2 = Try out a smartwatch; 3 = Improve my fitness by tracking my activities; 4 = Improve my wellbeing by tracking my mood; 5 = I hope it helps with a mental health problem I suffer from; 6 = Out of curiosity, to understand myself better; 7 = Receiving the personalized feedback report on my mood development from the study; 8 = Supporting science; 9 = General interest in mental health; 11 = I hope it helps with a severe mental health problem that a friend or family member is suffering from;	(Fried et al., 2022)

		10 = Something else: (open textfield)	
AD(H)D	*"Have you ever in your life had any of the following emotional or mental health problems?" "Attention deficit (hyperactivity) disorder (AD(H)D)"*	0 = No; 1 = Yes	(Fried et al., 2022)
Enjoyment of study	*"The experience of participating in WARN-D this week was"*	-3 ±0 +3 (very negative – very positive)	(Fried et al., 2022)
Reasons for missed prompts	*"What are the two most important reasons for missing surveys this week?"*	1 = Missed no or few surveys 2 = Didn't see notification in time 3 =Saw notification but could not answer 4 = Not motivated or interested 5 = Not feeling well 6 = Was asleep 7 = Forgot 8 = Technical problems	(Fried et al., 2022)
Time into study		1-12 (weeks)	Assessed through metadata
OS		1 = iOS, 2 = Android, 3 = Switched OS	Assessed through metadata

Note. Substance Use items were asked individually for tobacco products, alcoholic beverages, cannabis, cocaine, amphetamine type stimulants, inhalants, sedatives or sleeping pills and opioids. Depression and Anxiety were assessed at baseline as well as in weekly

Sunday surveys and correspondingly adjusted. Due to overlap with the adapted PHQ-9, we left out items 5 (restlessness) and 6 (irritability) from the GAD-7. Participants who switched their phones to a different OS were coded as 3 (Switched OS). A precise description about all items and how they were created and eventually adjusted can be found in the WARN-D protocol paper (Fried et al., 2022).

Appendix B

Multi-level model parameters

Parameter	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
Fixed effects								
Intercept	0.20 (0.04)	-0.18 (0.05)	-0.24 (0.04)	-0.31 (0.14)	-0.20 (0.13)	-0.19 (0.14)	-0.21 (0.13)	-0.21 (0.13)
Week		0.07 (0.00)	0.09 (0.01)	0.09 (0.01)	0.07 (0.01)	0.07 (0.01)	0.07 (0.01)	0.07 (0.01)
Phone OS				0.05 (0.08)	-0.01 (0.08)	-0.01 (0.08)	0.00 (0.08)	0.00 (0.08)
Enjoy Study					-0.16 (0.01)	-0.16 (0.01)	-0.16 (0.01)	-0.16 (0.01)
Reason1					-0.32 (0.04)	-0.32 (0.04)	-0.32 (0.04)	-0.32 (0.04)
Reason2					-0.02 (0.03)	-0.03 (0.03)	-0.02 (0.03)	-0.02 (0.03)
Reason3					0.04 (0.03)	0.04 (0.03)	0.04 (0.03)	0.04 (0.03)
Reason4					0.29 (0.05)	0.29 (0.05)	0.29 (0.05)	0.29 (0.05)
Reason5					0.03 (0.06)	0.04 (0.06)	0.05 (0.06)	0.05 (0.06)
Reason6					0.00 (0.03)	0.01 (0.03)	0.01 (0.03)	0.01 (0.03)
Reason7					0.02 (0.05)	0.02 (0.05)	0.02 (0.05)	0.02 (0.05)
Reason8					0.39 (0.04)	0.39 (0.04)	0.39 (0.04)	0.39 (0.04)
Weekly depression						-0.04 (0.02)	-0.04 (0.02)	-0.03 (0.02)

Parameter	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
Weekly stress						0.00 (0.01)	0.00 (0.01)	0.00 (0.01)
Weekly anxiety						-0.01 (0.02)	-0.01 (0.02)	-0.02 (0.02)
Random effects								
Intercept	0.74 (0.86)	0.84 (0.92)	0.66 (0.81)	0.66 (0.81)	0.56 (0.75)	0.57 (0.75)	0.55 (0.74)	0.55 (0.74)
Week			0.01 (0.09)	0.01 (0.09)	0.01 (0.09)	0.01 (0.09)	0.01 (0.08)	0.01 (0.08)
Weekly depression							0.02 (0.13)	0.02 (0.16)
Weekly anxiety								0.00 (0.07)
Residuals	0.41 (0.64)	0.35 (0.59)	0.27 (0.52)	0.27 (0.52)	0.24 (0.49)	0.24 (0.49)	0.23 (0.48)	0.23 (0.48)

Note. Summary of model parameters for models 1-8. SD/SE in brackets. Fixed Effects = SD/S; Random Effects = SE.