"**©American Psychological Association, [2018]. This paper is not the copy of record and may not exactly replicate the authoritative document published in the APA journal. Please do not copy or cite without author's permission. The final article is available, upon publication, at: 10.1037/pas0000662. [Epub ahead of print]"**

Response compliance and predictors thereof in studies using the experience sampling method

Aki Rintala
Department of Neurosciences, Center for Contextual Psychiatry, KU Leuven, Leuven, Belgium

Martien Wampers
UPC KU Leuven, Leuven, Belgium / Department of Neuroscience, Center for Contextual Psychiatry, KU Leuven, Belgium

Inez Myin-Germeys
Department of Neurosciences, Center for Contextual Psychiatry, KU Leuven, Leuven, Belgium

Wolfgang Viechtbauer
Department of Psychiatry and Neuropsychology, School for Mental Health and Neuroscience, Maastricht University, the Netherlands / Department of Neurosciences, Center for Contextual Psychiatry, KU Leuven, Leuven, Belgium

Corresponding author: Aki Rintala, Department of Neurosciences, KU Leuven/ Research Group Psychiatry/ Center for Contextual Psychiatry, Kapucijnenvoer 33 bus 7001 (blok h), 3000 Leuven, Belgium. E-mail: aki.rintala@kuleuven.be, Tel: +32474132388.

Conflicts of Interest: None declared.

Source of Funding: This study was supported by the Remote Assessment of Disease and Relapse – Central Nervous System (RADAR-CNS) research programme from the Innovative Medicines Initiative 2 Joint Undertaking under a grant agreement number 115902. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA. Inez Myin-Germeys was funded by an ERC consolidator grant (ERC-2012-StG, project 309767 – INTERACT) and by an FWO Odysseus grant (GOF8416N).

Abstract

Intensive repeated measurement techniques, such as the experience sampling method (ESM), put high demands on participants and may lead to low response compliance, which in turn may affect data quality. Therefore, the objective of this study was to investigate ESM compliance and predictors thereof based on a pooled dataset of 10 ESM studies with a total of 92,394 momentary assessments from 1,717 individuals with different mental health conditions. All included studies used an ESM paper-and-pencil diary protocol of 4 to 6 study days with 10 random time assessments per day. Analyses were conducted using multilevel mixed-effects logistic regression models. Results indicated overall acceptable compliance with an average

response rate of 78% (95%CI 0.74 to 0.82). However, compliance declined across days (p<.001), reaching a low on the 5th day with 73% (95%CI: 0.68 to 0.77). Compliance also varied significantly across assessments depending on the time within a day (p<.001), with highest compliance between 12 p.m. and 1.30 p.m. (83%; 95%CI: 0.80 to 0.86) and lowest compliance between 7.30 a.m. and 9 a.m. (56%; 95%CI: 0.50 to 0.62). Persons with psychosis were less compliant than healthy participants (70% vs. 83%, respectively; p<.001). Also females (p=.002) and older participants (p<.001) were slightly more compliant. The findings suggest acceptable compliance in an ESM protocol of 4 to 6 study days with a high frequency of 10 assessments per day despite fluctuations across and within study days. Further evidence on compliance and its predictors in different ESM protocols is needed, especially in clinical populations.

Keywords: momentary assessment, compliance, experience sampling method, data quality

Public Significance Statements: This study suggest acceptable compliance in experience sampling method (ESM) protocols of 4 to 6 study days with high frequency of 10 assessments per day. This type of ESM protocol can be considered as an option when choosing a protocol for EMA/ESM research.

Momentary assessment techniques, such as ecological momentary assessment (EMA) and the experience sampling method (ESM), are structured paper-and-pencil or electronic diary techniques to frequently assess experiences and behavior in the realm of daily life. Use of such methods in mental health research has been rapidly increasing in the last decades (e.g., Aan het Rot, Hogenelst, & Schoevers, 2012; Fahrenberg, Myrtek, Pawlik, & Perrez, 2007; Morren, Dulmen, Ouwerkerk, & Bensing, 2009; Myin-Germeys, Oorschot, Collip, Lataster, Delespaul, & van Os 2009; Shiffman, Stone, & Hufford, 2008). ESM has been used to capture the intensity and variability in momentary experiences, such as mood, thoughts, symptoms, and behaviors in everyday life (Ebner-Priemer & Trull, 2009; Trull & Ebner-Priemer, 2009).

Momentary assessment techniques have several advantages over traditional retrospective assessments. The latter may be distorted by memory biases, as individuals have to rely on their

memory when answering questions, whereas the momentary assessments of ESM inherently minimize this recall bias (Solhan, Trull, Jahng, & Wood, 2009). In addition, ESM makes it possible to capture an individual's emotional, behavioral, and cognitive experiences in an ecologically valid way, i.e., while they occur in an individual's natural environment (Trull & Ebner-Priemer, 2009).

Despite the advantages of data collection techniques such as EMA/ESM, the high frequency of the daily momentary assessments evidently makes it a demanding assessment tool that can be a serious burden for participants (Delespaul, 1995; Palmier-Claus, Myin-Germeys, Barkus, Bentley, Udachina, Delespaul, Lewis, & Dunn 2011). Consequently, several methodological issues arise when designing an EMA/ESM study, e.g., for how many study days should data be collected, how many assessments per day would be feasible for participants to answer, or how many questions can be asked at each moment without compromising compliance and the quality of the information that is planned to be captured with the method.

If a particular EMA/ESM protocol results in a low level of compliance, the collected data are unlikely to be an adequate and valid reflection of the intensity and variability of the participants' momentary experiences in daily life, which thereby would undermine the core purpose of using the method in the first place. Strategies such as financial compensation, study briefing, and communication during the sampling procedure are commonly used in EMA/ESM studies to ensure acceptable compliance (Morren et al., 2009; Palmier-Claus et al., 2011). Previous methodological studies investigating compliance in EMA/ESM using electronic diaries have found compliance rates ranging from 66% to 86% using 4 to 7 study days with 5 to 7 random time assessments per day (e.g., Courvoisier, Eid, & Lischetzke, 2012; Green et al., 2006;

Messiah, Grondin, & Encrenaz, 2011; Schüz, Walters, Frandsen, Bower, & Ferguson, 2014; Sokolovsky, Mermelstein, & Hedeker, 2014). In paper-and-pencil diaries, self-reported compliance measured by the number of answered moments was similar and ranged from 66% to 93% (Ben-Zeev & Young, 2010; Broderick, Schwartz, Shiffman, Hufford, & Stone, 2003; Geschwind, Peeters, Drukker, van Os, & Wichers, 2011; Havermans, Nicolson, & deVries, 2007; Stone, Shiffman, Schwartz, Broderick, & Hufford, 2003; Swendsen, 1998).

While overall compliance rates are usually reported in EMA/ESM studies, only few studies have investigated and reported the variability in compliance between study days or between assessment times within a day. Fuller-Tyszkiewicz et al. (2013) investigated compliance in a general population-based female sample (n = 105) using a protocol of 7 study days with 7 random time assessments per day, and observed a decline in mean compliance from the first (89%) to the last study day (76%). Similarly, Courvoisier et al. (2012) observed that compliance remained stable for the first 4 study days, ranging from 75% to 76%, but then dropped to 67% on the last study day in a general population-based sample (n = 305) when using a protocol of 7 study days with 6 random time assessments per day. Compliance also seemed to vary within a day as was illustrated in two studies using a protocol of 7 study days with 5 to 7 random time assessments per day (Messiah et al., 2011; Courvoisier et al., 2012). In both studies, compliance was especially low in the morning (between 8 or 9 a.m. and 11 a.m.) compared to other time intervals during the day. However, a protocol of 7 days with 8 random time assessments per day indicated that participants were more compliant at the first beeps when ESM sampling started at noon (Silvia, Kwapil, Eddington, & Brown, 2013). Knowledge about differences in compliance between study days and assessment times within a day is highly relevant when one must choose an ESM protocol to set up a study. For instance, a high assessment frequency

within a day may cause participants to miss questions, rush through questions, or even intentionally skip one (Morren et al., 2009). These issues might result in lowered compliance, and hence decrease the data quality of ESM. Despite these methodological challenges, many ESM studies have used high sampling frequency, even up to 10 assessments per day, because it gives the opportunity to assess highly variable daily life experiences (e.g., mood) throughout the day (Myin-Germeys, Kasanova, Vaessen, Vachon, Kirtley, Viechtbauer, & Reininghaus 2018).

Other protocol issues, such as the use of additional data collection methods alongside ESM and the number of questions in the diary, might influence compliance. For example, some previous studies asked participants to collect a saliva sample (to measure cortisol levels) at each assessment moment (Collip, Habets, Marcelis, Gronenschild, Lataster, Lardinois, Nicolson, & Myin-Germeys, 2013; Habets, Collip, Myin-Germeys, Gronenschild, van Bronswijk, Hofman, Lataster, Lardinois, Nicolson, van Os, & Marcelis, 2012; Jacobs et al., 2005). Two of these studies (Collip, Nicolson, Lardinois, Lataster, van Os, & Myin-Germeys, 2011; Jacobs et al., 2005), using an ESM protocol of 6 study days and 10 random time assessments per day, reported compliance rates similar to those of studies that did not use such additional sampling. Furthermore, some authors have suggested to limit the length of the ESM questionnaire to 20-30 questions in order to induce better compliance (Burton, Weller, & Sharpe, 2007; Morren et al., 2009). However, the impact of these protocol issues on compliance has not yet been formally investigated.

Compliance is not only influenced by study or protocol characteristics, but also by personal characteristics. Only few studies to date have investigated personal characteristics in relation to

compliance in EMA/ESM research. Messiah et al. (2011) investigated substance use among university students (n = 224) using ESM and found that male participants tended to be less compliant. However, previous EMA/ESM studies have not found associations between compliance and personal characteristics (i.e., age and gender) among general (Courvoisier et al., 2012) or psychotic populations (Hartley, Varese, Vasconcelos e Sa, Udachina, Barrowclough, Bentall, Lewis, Dunn, Haddock, & Palmier-Claus, 2014). The latter study raises another interesting issue related to possible differences in compliance in EMA/ESM research between clinical and general populations. Despite the enormous rise in EMA/ESM studies in mental health research and new developments extending the methodology to daily life clinical interventions (i.e., ecological momentary interventions; Geschwind et al., 2011; Myin-Germeys, Klippel, Steinhart, & Reininghaus, 2016), very little is known about the possible influence of clinical status on EMA/ESM compliance.

In sum, only few studies have investigated compliance in a high frequency EMA/ESM protocol and little is known about relevant predictors of compliance in such studies. However, knowledge about these predictors is crucial as it may guide the development of EMA/ESM protocols. Hence, further methodological studies around this topic are highly needed. The objective of the present study was to examine compliance and predictors thereof in intensive high frequency ESM protocols (4-6 study days with 10 semi-randomized assessments per day) in a large sample of participants with different mental health conditions.

Method

Participants

Analyses were conducted using a pooled dataset of 10 studies comprising a total of 1,717 participants. From the 1,717 participants, sufficient data for inclusion in the analysis were available from 1,647 (96%) participants. Sixty-five participants were excluded due to missing information on mental health status, one participant was excluded due to missing data on gender, and four participants were excluded due to missing age values[1]. The final sample comprised 1,186 (72%) female and 461 (28%) male subjects with a mean (SD, range) age of 34 (11.8, 16–65) years. Among the participants, 895 (54%) were classified as healthy subjects, 291 (18%) were persons with psychosis, 244 (15%) with depression, 176 (11%) with a familial risk for psychosis (i.e., having a first-degree relative with a psychotic disorder), and 41 (2%) with a psychometric risk for psychosis (i.e., persons scoring high on a subclinical psychosis scale). An overview of the included studies in the pooled dataset is presented in Appendix 1.

ESM protocol

All 10 studies in the pooled dataset used an identical ESM protocol where self-reported data were collected using a paper-and-pencil diary and a digital wristwatch for either 4, 5, or 6 consecutive study days (e.g., Collip et al., 2011; Collip et al., 2013; Collip, Wigman, Myin-Germeys, Jacobs, Derom, Thiery, Wichers, & van Os 2013; Geschwind et al., 2011; Wigman, van Os, Borsboom, Wardenaar, Epskamp, Klippel, Viechtbauer, Myin-Germeys, & Wichers, 2015). Participants received 10 randomized signals (hereafter called 'beeps') per day within 90-

[1] Part of the 65 excluded subjects with no mental health status were bipolar subjects, which we cannot reconstruct from the data. A comparison between excluded vs. included subjects is therefore problematic, since we know that part of those excluded come from a different study population, which our analyses does not encompass. We ran additional analyses with excluded participants included in the models and the obtained results were very similar to those presented in the results (e.g., overall compliance 78% a range of 72% to 83% for compliance across study days, a range of 55% to 83% for compliance within a day).

minute intervals between 7.30 a.m. and 10.30 p.m. After every beep, participants were asked to fill in a diary assessing current thoughts, mood, context of activity, location, social situations, and appraisals of the current situation. A typical diary in the pooled dataset used a 7-point Likert scale format (e.g., "*I feel cheerful*" with 1 = 'not at all' to 7 = 'very much'). A few questions were open-ended (e.g., "*What am I doing?*") or used bipolar (e.g., event-related question "*This event was*" with -3 = 'very unpleasant' to +3 = 'pleasant') or binary scales (e.g., "*I am alone*" with answer options of 'Yes' or 'No'). A diary example is presented in Appendix 2. All studies included in the pooled dataset were approved by the local ethics committee.

Randomized beeps were programmed in a digital wristwatch by a researcher and these times were masked from the participants. Participants had to report the time when they responded to the beep. After the study, a researcher matched the diary entries based on the self-reported response times reported in the diary with the randomized beeps triggered by the digital wristwatch for every study day.

Definition of compliance

Compliance to a given beep was defined as having a recorded response time that fell within a time window of 5 minutes before and 15 minutes after the beep. Based on Delespaul (1995), this time window may be considered acceptable when using a paper-and-pencil diary and a digital wristwatch. In particular, a participant might need some time to interrupt his or her current activities and might report a response time from a different time reference than the digital wristwatch itself (e.g., a kitchen or cell phone clock) that is not synchronized with the wristwatch. Hence, the outcome of interest was dichotomous (0 = not answered within the time

window, 1 = answered within the time window) and was measured for each subject between 40 to 60 times depending on the length of the study.

Predictors

Predictor variables were divided into three categories: time, study, and personal characteristics. A list of predictors and corresponding hypotheses are shown in Table 1.

Personal characteristics. Three variables were extracted from the dataset: age, gender, and study population. Age was considered as a continuous variable (coded (age – 20) / 10 to avoid an overly small coefficient), and gender was coded as "0" for males and "1" for females. Study population was examined as a five-level factor according to the classification described earlier, with healthy subjects used as the reference category.

Time characteristics. Time characteristics consisted of three different variables: chronological study day (i.e., 1 through 4, 5, or 6), calendar day, and time within a day (i.e., the beep number within a given day from 1 to 10). For the analyses, chronological study day was examined as a six-level factor using the first study day as the reference category. For calendar day, Sunday was considered the first day and was used as the reference category for this seven-level factor. Finally, time of the day was coded as a 10-level factor using nine dummy variables taking the first beep of the day (i.e., between 7.30 a.m. and 9 a.m.) as the reference category.

Study characteristics. We identified two predictors related to study characteristics: whether studies used saliva sampling at every assessment and the number of questions asked in the ESM

diary. Four out of the 10 studies in our pooled dataset measured cortisol levels from saliva samples collected alongside ESM at each beep. A dummy variable was coded as "0" for studies that did not use saliva sampling and "1" for studies that used saliva sampling. We examined the number of questions as a continuous variable (coded as (number of questions – 42) / 10). The number of questions in the diaries varied slightly between studies and ranged from 42 to 52 questions, counting only questions that were asked of every participants (i.e., we did not count questions that were presented as a result of branching logic).

Data Analysis

Compliance and its association with the various characteristics was analyzed using multilevel mixed-effect logistic regression models. For overall compliance, we fitted an empty model with just a model intercept. In the other models, we first added one predictor variable at a time (univariate models) and then fitted a model with all predictors included simultaneously (multivariable model). All models included random effects for study, subjects within study, study day within subjects, and beep number within subjects, with the last two random effects entered as crossed random effects. This model formulation implies four different degrees of correlation for outcomes corresponding to 1) two different subjects within the same study, 2) for different beeps within the same study day for a given subject (e.g., beep 1 and 2 on study day 1), 3) for the same beep number on different study days for a given subject (e.g., beep 1 on study day 1 and study day 2), and 4) for different beep numbers on different days for a given person (e.g., beep 1 on study day 1 and beep 2 on study day 2). We expected the magnitude of these four types of correlations to reflect the similarity of the circumstances under which the outcomes (i.e., compliance) were observed. In other words, the correlation is expected to be

highest for outcomes coming from the same day for a given subject, somewhat lower for outcomes with the same beep number across different days, even lower for different beep numbers across days, and lowest for different subjects within the same study.

We report the estimated intercept (i.e., log odds) and slope(s) (i.e., log odds ratio(s)) of each model with corresponding Wald-type tests. Factors as a whole were tested with Wald-type chi-square tests. Based on the intercept-only model and the univariate models with categorical predictors, we computed the predicted average compliance rate with corresponding 95% confidence intervals (95%CI) for each level of the factor variable. For models with continuous predictors (i.e., age and number of questions), we report some illustrative predicted average compliance rates as a function of the predictor (with 95%CI). Finally, based on the intercept-only model (i.e., for overall compliance) and the multivariable model, we computed and report the estimated values for the four types of correlations described above. Analyses were conducted using R 3.3.3 (R Development Core Team, 2016) with packages *lme4* (Bates, Mächler, Bolker, & Walker, 2015), *car* (Fox & Weisberg, 2011), and *multcomp* (Hothorn, Bretz, & Westfall, 2008).

Results

Overall response compliance as estimated based on the intercept-only model was 78% (95%CI 0.74 to 0.82). The results from the univariate and multivariable models are presented in Table 2.

Univariate analyses

Personal characteristics. Higher age was related to better compliance ($p < .001$). For example, persons 30 years of age had a compliance of 76% (95%CI 0.72 to 0.79) compared to 60 year olds with a compliance of 86% (95%CI 0.84 to 0.89). For gender, female participants were slightly more compliant than male participants (81% vs. 75% respectively, $p < .001$). Compliance also varied significantly across study population (χ^2 (4) = 57.1, $p < .001$). Persons with psychosis were less compliant than healthy participants (70% vs. 83% respectively, $p < .001$). In addition, persons with a familial risk for psychosis were slightly less compliant compared to healthy participants (79% vs. 83% respectively, $p = .044$). On the other hand, no significant differences in compliance were found between psychometric risk for psychosis ($p = .230$) or persons with depression ($p = .306$) compared to healthy participants.

Time characteristics. Compliance gradually declined across chronological study days (χ^2 (5) = 407.9, $p < .001$), starting at a high of 83% on the first and reaching a low on the 5th study day (73%). On the 6th study day, compliance across days seemed to stabilize (74%). With respect to the calendar day (χ^2 (6) = 101.2, $p < .001$), overall compliance was higher during the weekdays, with the highest compliance rate observed on Wednesdays (81%) and on Thursdays (81%). The lowest compliance across calendar days was observed during the weekends on Saturdays (75%) and Sundays (76%). Compliance also varied significantly across the time within a day (χ^2 (9) = 1839.6, $p < .001$). The highest compliance was measured between 12 p.m. and 1.30 p.m. (83%), while the lowest compliance was observed between 7.30 a.m. and 9 a.m. (56%).

Study characteristics. ESM protocols using saliva sampling at each beep did not have

significantly different compliance compared to ESM protocols without saliva sampling (p = .850). Furthermore, no significant relationship was found between the number of questions and the compliance rate (p = .763).

Multivariable analysis

In the multivariable model, all of the findings obtained from the univariate analyses remained significant (Table 2) with the exception of the difference between the group with a familial risk for psychosis and healthy participants (p = .107) and the difference between Fridays and Sundays (p = .211). In addition, compliance on Saturdays was now found to be slightly lower compared to Sundays (p = .037).[2]

We also examined all models with two-way interactions (8 × 7 / 2 = 28 models) as exploratory analyses. One model (with the interaction between the time within a day and the calendar day) failed to converge. After a Bonferroni correction, 9 interactions were significant. Results for the interaction models are provided as part of the supplementary materials. For the most part, the interactions were subtle and did not alter any of the main conclusions. However, the interaction between the 'saliva sampling' and 'study population' variables indicated that saliva sampling was associated with an unexpected increase in compliance in the psychosis group, opposite to what we see in the other groups where there was only an immaterial drop in compliance with the use of saliva sampling. Also, the interaction between the 'saliva sampling'

[2] Predicted compliance percentages can be computed based on the multivariable model using the coefficients reported in Table 2. For example, using the intercept-coefficient of the multivariable model with the estimated log odds for 25 year old males from the healthy participant group on day 1 on a Monday between 10.30 a.m. and 12 p.m. without saliva sampling and a 52-questionnaire are: 0.19 + (25–20)/10 × 0.21 + 0.24 + 1.17 + (52–42)/10 × 0.16 = 1.865, which translates into a compliance percentage of 100 × exp(1.865) / (1 + exp(1.865)) = 86.6%.

and the 'number of questions' variables indicated that compliance only dropped with the use of saliva sampling when the number of questions in the diary were lower. However, the latter interaction is difficult to interpret, since studies involving saliva sampling tended to use a higher number of questions in the diary overall to begin with, so these two variables are heavily confounded in the first place (which also partly explains why the coefficients for these two variables switch signs in the univariate versus multivariable analyses).

Correlations

As expected, a very high correlation was observed for compliance recorded at different beeps on the same study day within a given subject (r = 0.80 based on the intercept-only model, r = 0.85 based on the multivariable model). Also, compliance for the same beep on different study days for a given subject were correlated quite strongly (r = 0.75 and r = 0.74 for the intercept-only and multivariable model, respectively). A more moderate correlation was found for compliance corresponding to different beeps on different study days for a given subject (r = 0.55 and r = 0.59). On the other hand, there was almost no correlation between compliance of two different subjects within the same study (r = 0.05 and r = 0.02).

Discussion

The objective of this study was to examine compliance and its predictors in a pooled ESM dataset using a high-frequent ESM sampling scheme with a large study sample including general population subjects and persons with depression, familial or psychometric risk for psychosis, and persons with psychosis. The main findings indicate that compliance varied

across study days and within study days. Overall compliance was 78%, which is in line with previous studies that reported compliance rates ranging from 66% to 86% (Broderick et al., 2003; Green et al., 2006; Messiah et al., 2011; Schüz et al., 2014; Sokolovsky et al., 2014). The overall compliance rate found in the present case could be considered acceptable; in fact, compliance rates closer to 100% might indicate reactivity to the method, meaning that participants start adapting their behavior or even their environment to ensure that they do not miss any beeps. Given the naturalistic setting under which ESM data are collected, it is expected that participants will inevitably miss some beeps (e.g., due to a noisy environment, driving a car, or being at work and unable to respond) (Palmier-Claus et al., 2011).

We found that compliance drops across study days, reaching the highest level on the first study day (83%) and the lowest on the 5th day (73%). On the 6th study day, compliance was 74% which seems to indicate a stabilization of compliance within ESM protocols using 6 study days. Despite the drop across study days, overall compliance remained above 70%. However, we cannot be certain if the compliance would drop or stabilize in similar high-frequent assessment ESM studies using more than 4 to 6 study days. Hence, methodological studies investigating compliance in ESM protocols using more than 6 study days with a high frequency of assessments per day are needed to clarify if the trend of declining compliance continues across further study days.

With respect to the compliance rates within a day, our study also suggests that some beeps are more likely to be missed than others. The first beep of the day (i.e., the morning beep) is missed significantly more often than the other beeps of the day, which is in agreement with previous studies (Courvoisier et al., 2012; Messiah et al., 2011; Sokolovsky et al., 2014). However, a

study by Silvia et al. (2013) showed that participants were more compliant at the first beeps when ESM sampling started at noon. Our findings with low compliance in the morning between 7.30 a.m. to 9 a.m. (56%) might be due to the fact that participants are still asleep or are focused on their morning routine. When setting an ESM protocol, one should carefully consider the timing of the first beep to ensure adequate compliance in the morning. The present results suggest that it might be better to avoid starting the sampling immediately in the early morning. However, it might still be important to capture daily life experiences such as feelings, activity, and stress that occur during awakening times. In the future, the starting time of the sampling could be individualized by just asking participants about their preferred start time of the diary, by means of sensor tracking to register when a person is awake, or via a smartphone application to register the signal from the built-in or an external alarm clock. These approaches might increase compliance to the first assessment of the day.

Our findings also suggest that various personal characteristics influence compliance. Females and older individuals tended to be more compliant compared to males and younger participants. For example, participants aged 30 were estimated to have an overall compliance of 76% compared to participants aged 60 whose overall compliance rate was estimated at 86%. Sokolovsky et al. (2014) observed a compliance rate of 68% in an EMA study among adolescent smokers, which is in line with our finding. However, a study focused on psychotic patients using a similar ESM protocol as was used in the studies included in the present dataset did not observe any associations between compliance and demographic characteristics such as age and gender (Hartley et al., 2014). These differences between our findings and those by Hartley et al. (2014) might be due to power issues in sample sizes (n = 291 vs. n = 120, respectively). Further methodological EMA/ESM studies are needed to further clarify how such personal

characteristics are related to compliance.

All clinical populations reached a level of compliance that was comparable to the general population subjects with the exception of persons with psychosis who were significantly less compliant (70%) than the healthy participants (83%). Interaction analyses also revealed that lower compliance in the morning beep (i.e., 7.30 a.m. to 9 a.m.) was driven by participants with psychosis. This might be due to the fact that the participants with psychosis might not have daily obligations to attend to in the morning (e.g., going to work) compared to the healthy participants. One previous study that investigated compliance in a psychotic population reported a very similar overall compliance rate of 73% (Hartley et al., 2014). These results indicate that more considerations are needed to enhance compliance among specific clinical populations, and possible illness-specific predictors (e.g., disease severity, medication, or illness-specific symptoms) of non-response should be investigated to better understand possible reasons for lower compliance and hence how ESM protocols can be tailored for certain clinical populations. Future studies targeting this population might consider employing approaches that could increase compliance when conducting an ESM study, such as tying the amount of monetary compensation for study participation to the number of answered beeps, study briefing (e.g., researcher emphasizing to the participants to fill in as many beeps as possible), or increasing the amount of communication during the study procedure (Palmier-Claus et al., 2011). However, one must be careful not to interfere with the participants' daily life by giving too many reminders or providing too much feedback during the study period. Future research on the influence of reward approaches to enhance compliance and especially compliance in different clinical populations could provide useful information to optimize ESM protocols.

In our pooled dataset, 4 out of 10 studies used saliva sampling alongside the ESM protocol (Collip et al., 2011; Collip et al., 2013; De Wild-Hartmann, Wichers, van Bemmel, Derom, Thiery, Jacobs, van Os, & Simons, 2013; Peeters, Nicholson, & Berkhof, 2003). Our analyses did not reveal any significant differences in terms of compliance between studies with and without saliva sampling, which might indicate that the addition of a further data collection method does not automatically lead to lower compliance. Previous studies using saliva sampling as part of the ESM protocol have reported compliance rates ranging from 74% to 96%, which is in line with our results (Jacobs et al., 2005; Kudielka, Broderick, & Kirschbaum, 2003; Moeller, Lieb, Meyer, Loetscher, Krastel, & Meinlschmidt, 2014). However, caution must be exercised in generalizing these findings, as only a limited number of studies in our pooled dataset actually used saliva sampling.

Additionally, our findings did not indicate an association between the number of questions in the ESM diary and compliance. However, in our pooled dataset, the number of questions was quite similar across studies, ranging from 42 to 52 questions. It is possible that this lack of variability explains our null finding with respect to this variable. In this context, it is also worth noting that some authors have suggested to limit the number of questions to 20-30 in order to induce better compliance (Burton et al., 2007; Morren et al., 2009), which is actually much lower than the number of questions included in the studies in our pooled dataset. Hence, our study shows that adequate compliance can be obtained when using a relatively high number of questions, even when using a high frequency ESM protocol. However, to gain more evidence on the influence of the length of the diary, more methodological EMA/ESM studies are needed to clarify how the number and even the content of the questions affects compliance. Furthermore, methodological studies that investigate other forms of missing data such as

skipping questions in a diary and how this depends on the question format (e.g., Likert scale versus open-ended questions) might give us further insight on how to improve EMA/ESM diaries to enhance compliance.

Finally, models that were used in the analyses implied four different degrees of correlation for the observed compliance within and between subjects. In essence, these correlations indicate at which moments compliance tends to be more similar. As expected, we found the highest correlation for different beeps on the same day within subjects (e.g., on certain days it might generally be easier or more difficult to fill in the 10 diary entries; also, subjects may forget to put on the wristwatch or take the diary with them when leaving their home on certain days, leading to very similar – i.e., very low – compliance across all 10 beeps within that day). The next highest correlation was the one for the same beep number across different days within subjects. This is likely to reflect typical behavior patterns of subjects across different days (e.g., 'late risers' / 'night owls' will often miss the early beep but fill in the evening beep, leading to increased similarity and hence correlation for the compliance at the same beep number across days). The third highest correlation was the one for different beeps on different days. We can interpret this as reflecting differences in how willing subjects generally are to fill in the diary. Finally, the model allowed for a correlation among different subjects within the same study. If variability in overall compliance across studies was high (especially relative to the amount of variability in compliance across subjects), then this would be reflected in a high value for this last correlation component. However, we found this correlation to be very close to zero, indicating that variability across subjects was much higher than across studies.

Strengths and limitations

One strength of this study is its large sample size of 1,647 participants from a pooled dataset that gives new insight in compliance using an ESM protocol with a high-frequent sampling scheme. To our knowledge, this is the first methodological study to use multiple datasets to examine compliance in ESM studies with high-frequent daily assessments. In addition, our study provides information on compliance and its predictors not only in a general population sample, but also in individuals at risk for psychosis and with different mental health disorders, namely psychosis and depression.

At the same time, this study has some limitations. Our study focused only on the paper-and-pencil diary and wristwatch approach to collect ESM data, a method that is likely to fade out of practice given the easy and widespread availability of smartphones that can be used for data collection. Therefore, the relevance of the present results might be questioned. However, comparisons between paper-and-pencil versus electronic data collections methods have not revealed any noteworthy differences (Green et al., 2006).

Another potential limitation is the use of a self-reported response time variable to assess compliance. Two previous paper-and-pencil studies verified compliance by recording the opening and closing of the diary binder, which resulted in much lower compliance rates of 11% and 39% (Broderick et al., 2003; Stone et al., 2003). In these two studies, self-reported compliance was much higher (85% and 90%), suggesting that participants may have filled in paper-and-pencil diaries retrospectively (e.g., at the end of the day) (Broderick et al., 2003; Stone et al., 2003). However, these studies used EMA protocols of 21 and 24 study days with only three momentary assessments per day that occurred at fixed time points, in contrast to our

4 to 6 study days with random time sampling scheme and 10 momentary assessments per day. With fixed time points, it might be easier for participants to cheat since a seemingly appropriate response time can be filled in retrospectively. On the other hand, it would be more difficult for participants to do so in a paper-and-pencil diary study with high-frequent random time assessments, because this would require that participants keep track of the signaling times (Jacobs et al, 2005). Participants were encouraged not to change their daily life routines (i.e., participants were explained that it was acceptable to miss beeps if they were in a difficult situation such as driving a car) and fill in the diary entries in a correct manner. Still, we cannot be certain that participants in our present study actually filled out the diaries at the reported response times and hence compliance estimates might be biased upwards to some extent. However, if participants would have wanted to retrospectively complete the diaries, they would have had to actively keep track of the beep times, which might have happened in a few cases, but is unlikely to have been a common practice. Also, compensation (e.g., monetary incentives) for study participation was not tied to the number of completed assessments in any of the studies included in our dataset, further reducing the motivation to engage in such behavior. Nevertheless, we recommend that future studies conduct similar compliance analyses on ESM protocols using an electronic diary device where back-filling of the questionnaires retrospectively is impossible.

Additionally, there was no information in the pooled dataset about other factors that might influence compliance (such as the level of education and the marital or work status of the participants) and conclusions about the influence of the various predictors is based on purely observational evidence (e.g., use of saliva sampling was not randomized within studies either as a within- or between-subjects factor). Therefore, our analyses were restricted by the data

available in the pooled dataset and the way the data were collected.

It should be noted that these findings are not generalizable to all ESM protocols, as this study investigated compliance and predictors thereof in a rather homogeneous pooled dataset of studies using an ESM protocol of 4 to 6 study days with 10 beeps per day. As a further step, a meta-analysis is recommended to investigate compliance and its predictors in ESM protocols with more variability in study days and frequencies of daily assessment times to better understand if certain types of protocols are preferable in terms of achieving high compliance. Nevertheless, this present study provides unique information on compliance and its associations from one specific ESM protocol with high frequency assessments per day using a paper-and-pencil diary and a wristwatch approach.

Conclusions

Results show an overall acceptable compliance of 78% in ESM protocols of 4 to 6 study days with a high assessment frequency of 10 beeps per day using a paper-and-pencil diary despite fluctuations across and within study days. Persons with psychosis tended to be less compliant than healthy participants, but still reached a compliance rate of 70%. Hence, protocols of this type can be considered a possible option for experience sampling studies in mental health research. However, further evidence on the effects of different ESM protocols on compliance is needed, especially in clinical populations.

Acknowledgements

The authors wish to thank the research group of ESM MERGE for proving the data to this study.

References

Aan het Rot, M., Hogenelst, K., & Schoevers, R. A. (2012). Mood disorders in everyday life: A systematic review of experience sampling and ecological momentary assessment studies. *Clinical Psychology Review*, *32*(6), 510–523. https://doi.org/10.1016/j.cpr.2012.05.007

Bak, M., Delespaul, P., Krabbendam, L., Huistra, K., Walraven, W., & Van Os, J. (2009). Capturing coping with symptoms in people with a diagnosis of schizophrenia: Introducing the MACS-24. International Journal of Methods in Psychiatric Research, 18(1), 4–12. http://dx.doi.org/10.1002/mpr.272

Barge-Schaapveld, D. Q., & Nicolson, N. A. (2002). Effects of Antidepressant Treatment on the Quality of Daily Life: An Experience Sampling Study. The Journal of Clinical Psychiatry, 63(6), 477–485.

Barge-Schaapveld, D. Q., Nicolson, N. A., Berkhof, J., & Devries, M. W. (1999). Quality of life in depression: Daily life determinants and variability. Psychiatry Research, 88(3), 173–189. http://dx.doi.org/10.1016/S0165-1781(99)00081-5

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 51. https://doi.org/10.18637/jss.v067.i01

Ben-Zeev, D., & Young, M. A. (2010). Accuracy of Hospitalized Depressed Patients' and Healthy Controls' Retrospective Symptom Reports. *The Journal of Nervous and Mental Disease*, *198*(4), 280–285. https://doi.org/10.1097/NMD.0b013e3181d6141f

595 Broderick, J. E., Schwartz, J. E., Shiffman, S., Hufford, M. R., & Stone, A. A. (2003).
596 Signaling does not adequately improve diary compliance. *Annals of Behavioral*
597 *Medicine : A Publication of the Society of Behavioral Medicine*, *26*(2), 139–48.
598 https://doi.org/10.1207/S15324796ABM2602_06
599 Burton, C., Weller, D., & Sharpe, M. (2007). Are electronic diaries useful for symptoms
600 research? A systematic review. *Journal of Psychosomatic Research*, *62*(5), 553–561.
601 https://doi.org/10.1016/j.jpsychores.2006.12.022
602 Collip, D., Habets, P., Marcelis, M., Gronenschild, E., Lataster, T., Lardinois, M., … Myin-
603 Germeys, I. (2013). Hippocampal volume as marker of daily life stress sensitivity in
604 psychosis. *Psychological Medicine*, *43*(7), 1377–1387.
605 https://doi.org/10.1017/S003329171200219X
606 Collip, D., Nicolson, N. a, Lardinois, M., Lataster, T., van Os, J., & Myin-Germeys, I. (2011).
607 Daily cortisol, stress reactivity and psychotic experiences in individuals at above average
608 genetic risk for psychosis. *Psychological Medicine*, *41*(11), 2305–15.
609 https://doi.org/10.1017/S0033291711000602
610 Collip, D., Wigman, J. T. W., Myin-Germeys, I., Jacobs, N., Derom, C., Thiery, E., … van
611 Os, J. (2013). From Epidemiology to Daily Life: Linking Daily Life Stress Reactivity to
612 Persistence of Psychotic Experiences in a Longitudinal General Population Study. *PLoS*
613 *ONE*, *8*(4), e62688. https://doi.org/10.1371/journal.pone.0062688
614 Courvoisier, D. S., Eid, M., & Lischetzke, T. (2012). Compliance to a cell phone-based
615 ecological momentary assessment study: the effect of time and personality
616 characteristics. *Psychol Assess*, *24*(3), 713–720. https://doi.org/10.1037/a0026733
617 De Wild-Hartmann, J. A., Wichers, M., Van Bemmel, A. L., Derom, C., Thiery, E., Jacobs,
618 N., … Simons, C. J. P. (2013). Day-to-day associations between subjective sleep and

affect in regard to future depression in a female population-based sample. *British Journal of Psychiatry*, *202*(6), 407–412. https://doi.org/10.1192/bjp.bp.112.123794

Delespaul, P. (1995). *Assessing schizophrenia in daily life: the Experience Sampling Method*. Maastricht : Universitaire Pers Maastricht.

Ebner-Priemer, U. W., & Trull, T. J. (2009). Ecological Momentary Assessment of Mood Disorders and Mood Dysregulation. *Psychological Assessment*, *21*(4), 463–475. https://doi.org/10.1037/a0017075

Fahrenberg, J., Myrtek, M., Pawlik, K., & Perrez, M. (2007). Ambulatory assessment - monitoring behavior in daily life settings: A behavioral-scientific challenge for psychology. *European Journal of Psychological Assessment*, *23*(4), 206–213. https://doi.org/10.1027/1015-5759.23.4.206

Fox, J., & Weisberg, S. (2011). *An R Companion to Applied Regression*. Sage Publications. Retrieved from http://socserv.socsci.mcmaster.ca/jfox/Books/Companion

Fuller-Tyszkiewicz, M., Skouteris, H., Richardson, B., Blore, J., Holmes, M., & Mills, J. (2013). Does the burden of the experience sampling method undermine data quality in state body image research? Body Image, 10(4), 607–613. http://dx.doi.org/10.1016/j.bodyim.2013.06.003

Geschwind, N., Peeters, F., Drukker, M., van Os, J., & Wichers, M. (2011). Mindfulness training increases momentary positive emotions and reward experience in adults vulnerable to depression: A randomized controlled trial. *Journal of Consulting and Clinical Psychology*, *79*(5), 618–628. https://doi.org/10.1037/a0024595

Green, A. S., Rafaeli, E., Bolger, N., Shrout, P. E., & Reis, H. T. (2006). Paper or plastic? Data equivalence in paper and electronic diaries. *Psychological Methods*, *11*(1), 87–105. https://doi.org/10.1037/1082-989X.11.1.87

643 Habets, P., Collip, D., Myin-Germeys, I., Gronenschild, E., van Bronswijk, S., Hofman, P., …
644 Marcelis, M. (2012). Pituitary volume, stress reactivity and genetic risk for psychotic
645 disorder. *Psychological Medicine*, *42*(7), 1523–1533.
646 https://doi.org/10.1017/S0033291711002728
647 Hartley, S., Varese, F., Vasconcelos e Sa, D., Udachina, A., Barrowclough, C., Bentall, R. P.,
648 … Palmier-Claus, J. (2014). Compliance in experience sampling methodology: the role
649 of demographic and clinical characteristics. *Psychosis*, *6*(1), 70–73.
650 https://doi.org/10.1080/17522439.2012.752520
651 Havermans, R., Nicolson, N. A., & deVries, M. W. (2007). Daily Hassles, Uplifts, and Time
652 Use in Individuals With Bipolar Disorder in Remission. *The Journal of Nervous and*
653 *Mental Disease*, *195*(9), 745–751. https://doi.org/10.1097/NMD.0b013e318142cbf0
654 Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous inference in general parametric
655 models. *Biometrical Journal*, *50*(3), 346–363. https://doi.org/10.1002/bimj.200810425
656 Jacobs, N., Nicolson, N. A., Derom, C., Delespaul, P., Van Os, J., & Myin-Germeys, I.
657 (2005). Electronic monitoring of salivary cortisol sampling compliance in daily life. *Life*
658 *Sciences*, *76*(21), 2431–2443. https://doi.org/10.1016/j.lfs.2004.10.045
659 Kudielka, B. M., Broderick, J. E., & Kirschbaum, C. (2003). Compliance with saliva
660 sampling protocols: electronic monitoring reveals invalid cortisol daytime profiles in
661 noncompliant subjects. *Psychosomatic Medicine*, *65*(2), 313–319.
662 https://doi.org/10.1097/01.PSY.0000058374.50240.BF
663 Lardinois, M., Myin-Germeys, I., Bak, M., Mengelers, R., Van Os, J., & Delespaul, P. A. E.
664 G. (2007). The dynamics of symptomatic and non-symptomatic coping with psychotic
665 symptoms in the flow of daily life. Acta Psychiatrica Scandinavica, 116(1), 71–75.
666 http://dx.doi.org/10.1111/j.1600-0447.2007.01022.x

667 Lataster, J., Myin-Germeys, I., Wichers, M., Delespaul, P., van Os, J., & Bak, M. (2011).
668 Psychotic exacerbation and emotional dampening in the daily life of patients with
669 schizophrenia switched to aripiprazole therapy: a collection of standardized case reports.
670 Therapeutic Advances in Psychopharmacology, 1(5), 145–151.
671 http://dx.doi.org/10.1177/2045125311419552
672 Lataster, T., Valmaggia, L., Lardinois, M., van Os, J., & Myin-Germeys, I. (2013). Increased
673 stress reactivity: a mechanism specifically associated with the positive symptoms of
674 psychotic disorder. Psychological Medicine, 43(7), 1389–400.
675 http://dx.doi.org/10.1017/S0033291712002279
676 Messiah, A., Grondin, O., & Encrenaz, G. (2011). Factors associated with missing data in an
677 experience sampling investigation of substance use determinants. *Drug and Alcohol*
678 *Dependence*, *114*(2–3), 153–158. https://doi.org/10.1016/j.drugalcdep.2010.09.016
679 Moeller, J., Lieb, R., Meyer, A. H., Quack Loetscher, K., Krastel, B., & Meinlschmidt, G.
680 (2014). Improving ambulatory saliva-sampling compliance in pregnant women: A
681 randomized controlled study. *PLoS ONE*, *9*(1), e86204.
682 https://doi.org/10.1371/journal.pone.0086204
683 Morren, M., van Dulmen, S., Ouwerkerk, J., & Bensing, J. (2009). Compliance with
684 momentary pain measurement using electronic diaries: A systematic review. *European*
685 *Journal of Pain*, *13*(4), 354–365. https://doi.org/10.1016/j.ejpain.2008.05.010
686 Myin-Germeys, I., Kasanova, Z., Vaessen, T., Vachon, H., Kirtley, O., Viechtbauer, W., &
687 Reininghaus, U. (2018). Experience sampling methodology in mental health research:
688 new insights and technical developments. *World Psychiatry*, *17*(2), 123–132.
689 https://doi.org/10.1002/wps.20513
690 Myin-Germeys, I., Klippel, A., Steinhart, H., & Reininghaus, U. (2016). Ecological

momentary interventions in psychiatry. *Current Opinion in Psychiatry*, *29*(4), 258–263. https://doi.org/10.1097/YCO.0000000000000255

Myin-Germeys, I., Oorschot, M., Collip, D., Lataster, J., Delespaul, P., & van Os, J. (2009). Experience sampling research in psychopathology: opening the black box of daily life. *Psychological Medicine*, *39*(9), 1533–1547. https://doi.org/10.1017/S0033291708004947

Myin-Germeys, I., Van Os, J., Schwartz, J. E., Stone, A. A., & Delespaul, P. A. (2001). Emotional reactivity to daily life stress in psychosis. Archives of General Psychiatry, 58(12), 1137–1144. https://doi.org/10.1001/archpsyc.58.12.1137

Palmier-Claus, J. E., Myin-Germeys, I., Barkus, E., Bentley, L., Udachina, A., Delespaul, P., … Dunn, G. (2011). Experience sampling research in individuals with mental illness: Reflections and guidance. *Acta Psychiatrica Scandinavica*, *123*(1), 12–20. https://doi.org/10.1111/j.1600-0447.2010.01596.x

Peeters, F., Berkhof, J., Delespaul, P., Rottenberg, J., & Nicolson, N. A. (2006). Diurnal mood variation in major depressive disorder. Emotion, 6(3), 383–391. http://dx.doi.org/10.1037/1528-3542.6.3.383

Peeters, F., Nicholson, N. A., & Berkhof, J. (2003). Cortisol responses to daily events in major depressive disorder. *Psychosomatic Medicine*, *65*(5), 836–841. https://doi.org/10.1097/01.PSY.0000088594.17747.2E

R Development Core Team. (2016). *R: A Language and Environment for Statistical Computing*. *R Foundation for Statistical Computing Vienna Austria*. https://doi.org/10.1038/sj.hdy.6800737

Schüz, N., Walters, J. A. E., Frandsen, M., Bower, J., & Ferguson, S. G. (2014). Compliance with an EMA monitoring protocol and its relationship with participant and smoking

715 characteristics. *Nicotine and Tobacco Research*, *16*(Suppl 2), S88-92.
716 https://doi.org/10.1093/ntr/ntt142
717 Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological momentary assessment.
718 *Annual Review of Clinical Psychology*, *4*, 1–32.
719 https://doi.org/10.1002/9781118384404.ch20
720 Silvia, P. J., Kwapil, T. R., Eddington, K. M., & Brown, L. H. (2013). Missed Beeps and
721 Missing Data: Dispositional and Situational Predictors of Nonresponse in Experience
722 Sampling Research. *Social Science Computer Review*, *31*(4), 471–481.
723 https://doi.org/10.1177/0894439313479902
724 Sokolovsky, A. W., Mermelstein, R. J., & Hedeker, D. (2014). Factors predicting compliance
725 to ecological momentary assessment among adolescent smokers. *Nicotine and Tobacco*
726 *Research*, *16*(3), 351–358. https://doi.org/10.1093/ntr/ntt154
727 Solhan, M. B., Trull, T. J., Jahng, S., & Wood, P. K. (2009). Clinical Assessment of Affective
728 Instability: Compating EMA indices, questionnaire reports, and retrospective recall.
729 *Psychol Assessment*, *21*(3), 425–436. https://doi.org/10.1037/a0016869.Clinical
730 Stone, A. A., Shiffman, S., Schwartz, J. E., Broderick, J. E., & Hufford, M. R. (2003). Patient
731 compliance with paper and electronic diaries. *Controlled Clinical Trials*, *24*(2), 182–199.
732 https://doi.org/10.1016/S0197-2456(02)00320-3
733 Swendsen, J. D. (1998). The helplessness-hopelessness theory and daily mood experience: An
734 idiographic and cross-situational perspective. *Journal of Personality and Social*
735 *Psychology*, *74*(5), 1398–1408. https://doi.org/10.1037/0022-3514.74.5.1398
736 Thewissen, V., Bentall, R. P., Lecomte, T., van Os, J., & Myin-Germeys, I. (2008).
737 Fluctuations in self-esteem and paranoia in the context of daily life. Journal of Abnormal
738 Psychology, 117(1), 143–153. http://dx.doi.org/10.1037/0021-843X.117.1.143

739 Trull, T. J., & Ebner-Priemer, U. W. (2009). Using Experience Sampling Methods/Ecological
740 Momentary Assessment (ESM/EMA) in Clinical Assessment and Clinical Research:
741 Introduction to the Special Section. *Psychological Assessment, 21*(4), 457–462.
742 https://doi.org/10.1037/a0017653
743 Vaessen, T., Kasanova, Z., Hernaus, D., Lataster, J., Collip, D., van Nierop M., & Myin-
744 Germeys, I. Cortisol reactivity to daily-life stressors in psychosis (in preparation).
745 Wigman, J. T. W., van Os, J., Borsboom, D., Wardenaar, K. J., Epskamp, S., Klippel, A., …
746 Wichers, M. (2015). Exploring the underlying structure of mental disorders: cross-
747 diagnostic differences and similarities from a network perspective using both a top-down
748 and a bottom-up approach. *Psychological Medicine, 45*(11), 2375–2387.
749 https://doi.org/10.1017/S0033291715000331
750
751

752 Table 1

753 *List of predictors and hypotheses*

Predictors	Level	Type	Hypotheses
Personal characteristics			
Age	Subject	Continuous	Younger participants have lower compliance
Gender	Subject	Categorical	Female participants have higher compliance
Study population	Subject	Categorical	Clinical populations have lower compliance
Time characteristics			
Chronological study day	Day	Categorical	Compliance decreases in the following study days
Calendar day	Day	Categorical	Compliance is lower during the weekends
Time within a day	Beep	Categorical	The first beep in the morning has the lowest compliance
Study characteristics			
Saliva sampling	Study	Categorical	Studies using saliva sampling have lower compliance
Number of questions	Study	Continuous	Studies with a higher number of questions have lower compliance

754

Table 2

Results of personal, time, and study characteristics on response compliance

Predictor	Level	Number of observations	Univariate model					Multivariable model		
			β	Z	*p*	Compliance (%)	95%CI	β	Z	*p*
								0.19*	0.76*	
Personal characteristics										
Age†	Intercept	n.a.	0.90	9.10						
	Age	92,200	0.24	8.07	< **.001**			0.21	7.23	< **.001**
Gender	Male**	27,340	1.09	9.27		75	0.70 to 0.79			
	Female	64,860	0.36	4.63	< **.001**	81	0.77 to 0.84	0.25	3.15	**.002**
Study population	Healthy participants**	47,620	1.57	14.70		83	0.80 to 0.86			
	Psychosis	16,920	-0.70	-6.73	< **.001**	70	0.66 to 0.75	-0.57	-5.20	< **.001**

	Familial risk for psychosis	10,560	-0.24	-2.02	**.044**	79	0.74 to 0.83	-0.19	-1.61	.107
	Psychometric risk for psychosis	2,460	0.27	1.20	.230	86	0.80 to 0.91	0.08	0.34	.730
	Depression	14,640	-0.18	-1.02	.306	80	0.74 to 0.85	-0.19	-1.13	.259
Time characteristics										
Chronological study day	1**	16,470	1.62	12.91		83	0.80 to 0.87			
	2	16,470	-0.12	-2.99	**.003**	82	0.78 to 0.85	-0.10	-2.46	**.014**
	3	16,470	-0.28	-7.28	**< .001**	79	0.75 to 0.83	-0.26	-6.43	**< .001**
	4	16,470	-0.48	-12.51	**< .001**	76	0.71 to 0.80	-0.45	-11.38	**< .001**
	5	16,200	-0.64	-16.52	**< .001**	73	0.68 to 0.77	-0.63	-15.78	**< .001**
	6	10,120	-0.56	-12.38	**< .001**	74	0.69 to 0.79	-0.60	-12.83	**< .001**
Calendar day	Sunday**	14,600	1.13	9.36		76	0.71 to 0.80			
	Monday	12,040	0.19	4.40	**< .001**	79	0.75 to 0.83	0.24	5.40	**< .001**
	Tuesday	12,160	0.18	4.12	**< .001**	79	0.75 to 0.82	0.15	3.42	**< .001**

	Wednesday	12,380	0.30	6.86	< **.001**	81	0.77 to 0.84	0.19	4.21	< **.001**
	Thursday	12,930	0.29	6.70	< **.001**	81	0.77 to 0.84	0.12	2.81	**.005**
	Friday	13,160	0.21	4.92	< **.001**	79	0.75 to 0.83	0.05	1.25	.211
	Saturday	14,930	-0.00	-0.11	.912	75	0.71 to 0.80	-0.09	-2.09	**.037**
Time within a day	7.30 a.m. – 9 a.m.**	9220	0.24	2.01		56	0.50 to 0.62			
	9 a.m. – 10.30 a.m.	9220	0.85	21.02	< **.001**	75	0.70 to 0.79	0.85	20.97	< **.001**
	10.30 a.m. – 12 p.m.	9220	1.17	28.19	< **.001**	80	0.76 to 0.84	1.17	28.11	< **.001**
	12 p.m. – 1.30 p.m.	9220	1.37	32.50	< **.001**	83	0.80 to 0.86	1.37	32.41	< **.001**
	1.30 p.m. – 3 p.m.	9220	1.30	31.15	< **.001**	82	0.79 to 0.86	1.30	31.06	< **.001**
	3 p.m. – 4.30 p.m.	9220	1.23	29.51	< **.001**	81	0.77 to 0.85	1.23	29.42	< **.001**
	4.30 p.m. – 6 p.m.	9220	1.27	30.40	< **.001**	82	0.78 to 0.85	1.27	30.32	< **.001**
	6 p.m. – 7.30 p.m.	9220	1.31	31.15	< **.001**	82	0.79 to 0.86	1.31	31.06	< **.001**
	7.30 p.m. – 9 p.m.	9220	1.16	28.10	< **.001**	80	0.76 to 0.84	1.16	28.02	< **.001**
	9 p.m. – 10.30 p.m.	9220	0.78	19.26	< **.001**	74	0.69 to 0.78	0.78	19.20	< **.001**

Study characteristics

Saliva sampling	No**	33,780	1.27	8.34		78	0.73 to 0.83			
	Yes	58,420	0.04	0.19	.850	79	0.73 to 0.84	-0.18	-0.68	.499
Number of	Intercept	n.a.	1.37	5.19						
questions††	Number of questions	92,200	-0.12	-0.30	.763			0.16	0.35	.726

β = estimate value of coefficient; Z = Z-value; 95%CI = 95 % confidence interval; p = p-value; † = age was coded (age – 20) / 10 and treated as a continuous variable in the model; †† = number of items was coded (number of items – 42) / 10 and treated as a continuous variable in the model; n.a. = not applicable; * = intercept coefficient and Z-value for the multivariable model; ** = reference category

Appendix 1

Included ESM MERGE studies

Status	Study	N	References
Psychosis	Aripiprazole	27	(Lataster et al., 2011)
	Genetic Risk and Outcome of Psychosis (GROUP)	72	(Collip et al., 2011; Lataster, Valmaggia, Lardinois, van Os, & Myin-Germeys, 2013)
	Maastricht Coping Study (MACS)	18	(Bak et al., 2009; Lardinois et al., 2007)
	Maastricht Psychosis Study (MAPS)	48	(Myin-Germeys, Van Os, Schwartz, Stone, & Delespaul, 2001)
	Stress-reactivity in Psychosis (STRIP)	47	(Vaessen, Kasanova, Hernaus, Lataster, Collip, van Nierop, & Myin-Germeys, in preparation)
	ZAPP	79	(Thewissen, Bentall, Lecomte, van Os, & Myin-Germeys, 2008)
Familial risk for psychosis	GROUP	81	(Collip et al., 2011; Lataster et al., 2013)
	MAPS	48	(Myin-Germeys et al., 2001)
	STRIP	49	(Vaessen et al., in preparation)
Psychometric risk for psychosis	ZAPP	41	(Thewissen et al., 2008)

Depression	Antidepressants RCT	70	(Barge-Schaapveld & Nicolson, 2002; Barge-Schaapveld, Nicolson, Berkhof, & Devries, 1999)
	MindMaastricht	129	(Geschwind, Peeters, Drukker, van Os, & Wichers, 2011)
	Mood and cortisol reactivity to daily stress	45	(Peeters, Berkhof, Delespaul, Rottenberg, & Nicolson, 2006; Peeters, Nicholson, & Berkhof, 2003)
Healthy participants	Antidepressants RCT	25	(Barge-Schaapveld & Nicolson, 2002; Barge-Schaapveld et al., 1999)
	GROUP	85	(Collip et al., 2011; Lataster et al., 2013)
	MAPS	50	(Myin-Germeys et al., 2001)
	Mood and cortisol reactivity to daily stress	39	(Peeters et al., 2006, 2003)
	STRIP	51	(Vaessen et al., in preparation)
	Twins	610	(Collip et al., 2013; De Wild-Hartmann et al., 2013)
	ZAPP	38	(Thewissen et al., 2008)

References

Bak, M., Delespaul, P., Krabbendam, L., Huistra, K., Walraven, W., & Van Os, J. (2009). Capturing coping with symptoms in people with a diagnosis of schizophrenia: Introducing the MACS-24. *International Journal of Methods in Psychiatric Research*, *18*(1), 4–12. http://dx.doi.org/10.1002/mpr.272

Barge-Schaapveld, D. Q., & Nicolson, N. A. (2002). Effects of Antidepressant Treatment on the Quality of Daily Life: An Experience Sampling Study. *The Journal of Clinical Psychiatry*, *63*(6), 477–485.

Barge-Schaapveld, D. Q., Nicolson, N. A., Berkhof, J., & Devries, M. W. (1999). Quality of life in depression: Daily life determinants and variability. *Psychiatry Research*, *88*(3), 173–189. http://dx.doi.org/10.1016/S0165-1781(99)00081-5

Collip, D., Nicolson, N. a, Lardinois, M., Lataster, T., van Os, J., & Myin-Germeys, I. (2011). Daily cortisol, stress reactivity and psychotic experiences in individuals at above average genetic risk for psychosis. *Psychological Medicine*, *41*(11), 2305–15. http://dx.doi.org/10.1017/S0033291711000602

Collip, D., Wigman, J. T. W., Myin-Germeys, I., Jacobs, N., Derom, C., Thiery, E., … van Os, J. (2013). From Epidemiology to Daily Life: Linking Daily Life Stress Reactivity to Persistence of Psychotic Experiences in a Longitudinal General Population Study. *PLoS ONE*, *8*(4), e62688. http://dx.doi.org/10.1371/journal.pone.0062688

De Wild-Hartmann, J. A., Wichers, M., Van Bemmel, A. L., Derom, C., Thiery, E., Jacobs, N., … Simons, C. J. P. (2013). Day-to-day associations between subjective sleep and affect in regard to future depression in a female population-based sample. *British Journal of Psychiatry*, *202*(6), 407–412. http://dx.doi.org/10.1192/bjp.bp.112.123794

Geschwind, N., Peeters, F., Drukker, M., van Os, J., & Wichers, M. (2011). Mindfulness training increases momentary positive emotions and reward experience in adults vulnerable to depression: A randomized controlled trial. *Journal of Consulting and Clinical Psychology*, *79*(5), 618–628. http://dx.doi.org/10.1037/a0024595

Lardinois, M., Myin-Germeys, I., Bak, M., Mengelers, R., Van Os, J., & Delespaul, P. A. E. G. (2007). The dynamics of symptomatic and non-symptomatic coping with psychotic symptoms in the flow of daily life. *Acta Psychiatrica Scandinavica*, *116*(1), 71–75. http://dx.doi.org/10.1111/j.1600-0447.2007.01022.x

Lataster, J., Myin-Germeys, I., Wichers, M., Delespaul, P., van Os, J., & Bak, M. (2011). Psychotic exacerbation and emotional dampening in the daily life of patients with schizophrenia switched to aripiprazole therapy: a collection of standardized case reports. *Therapeutic Advances in Psychopharmacology*, *1*(5), 145–151. http://dx.doi.org/10.1177/2045125311419552

Lataster, T., Valmaggia, L., Lardinois, M., van Os, J., & Myin-Germeys, I. (2013). Increased stress reactivity: a mechanism specifically associated with the positive symptoms of psychotic disorder. *Psychological Medicine*, *43*(7), 1389–400. http://dx.doi.org/10.1017/S0033291712002279

Myin-Germeys, I., Van Os, J., Schwartz, J. E., Stone, A. A., & Delespaul, P. A. (2001). Emotional reactivity to daily life stress in psychosis. *Archives of General Psychiatry*, *58*(12), 1137–1144.

Peeters, F., Berkhof, J., Delespaul, P., Rottenberg, J., & Nicolson, N. A. (2006). Diurnal mood variation in major depressive disorder. *Emotion*, *6*(3), 383–391. http://dx.doi.org/10.1037/1528-3542.6.3.383

Peeters, F., Nicholson, N. A., & Berkhof, J. (2003). Cortisol responses to daily events in major depressive disorder. *Psychosomatic Medicine*, *65*(5), 836–841. http://dx.doi.org/10.1097/01.PSY.0000088594.17747.2E

Thewissen, V., Bentall, R. P., Lecomte, T., van Os, J., & Myin-Germeys, I. (2008). Fluctuations in self-esteem and paranoia in the context of daily life. *Journal of Abnormal*

Psychology, *117*(1), 143–153. http://dx.doi.org/10.1037/0021-843X.117.1.143

Vaessen, T., Kasanova, Z., Hernaus, D., Lataster, J., Collip, D., van Nierop M., & Myin-Germeys, I. Cortisol reactivity to daily-life stressors in psychosis (in preparation)

Appendix 2

ESM questionnaire example

What was I thinking (just before the beep went off?)	Open-ended question
This thought was…	
Pleasant	Likert scale from 1 (Not) to 7 (Very)
Clear	Likert scale from 1 (Not) to 7 (Very)
Common	Likert scale from 1 (Not) to 7 (Very)
I have trouble concentrating	Likert scale from 1 (Not) to 7 (Very)
I feel…	
Cheerful	Likert scale from 1 (Not) to 7 (Very)
Uncertain	Likert scale from 1 (Not) to 7 (Very)
Lonely	Likert scale from 1 (Not) to 7 (Very)
Relaxed	Likert scale from 1 (Not) to 7 (Very)
Anxious	Likert scale from 1 (Not) to 7 (Very)
Satisfied	Likert scale from 1 (Not) to 7 (Very)
Irritated	Likert scale from 1 (Not) to 7 (Very)
Sad	Likert scale from 1 (Not) to 7 (Very)
Guilty	Likert scale from 1 (Not) to 7 (Very)
Overall, I am feeling happy	Likert scale from 1 (Not) to 7 (Very)
Right now…	
I like myself	Likert scale from 1 (Not) to 7 (Very)
I am ashamed of myself	Likert scale from 1 (Not) to 7 (Very)
I am a failure	Likert scale from 1 (Not) to 7 (Very)

I am a good person	Likert scale from 1 (Not) to 7 (Very)
Right now, I feel that others…	
Dislike me	Likert scale from 1 (Not) to 7 (Very)
Might hurt me	Likert scale from 1 (Not) to 7 (Very)
I…	
Feel suspicious	Likert scale from 1 (Not) to 7 (Very)
Feel safe	Likert scale from 1 (Not) to 7 (Very)
Feel I can't get rid of my thoughts	Likert scale from 1 (Not) to 7 (Very)
Feel unreal	Likert scale from 1 (Not) to 7 (Very)
Hear voices	Likert scale from 1 (Not) to 7 (Very)
See 'things'	Likert scale from 1 (Not) to 7 (Very)
Feel fear of losing control	Likert scale from 1 (Not) to 7 (Very)
Where am I?	Open-ended question
I am alone?	Yes/No
[If not]	
With whom?	Open-ended question
How many men?	Open-ended question
How many women?	Open-ended question
How many children?	Open-ended question
In the company of these people, I feel…	
Comfortable	Likert scale from 1 (Not) to 7 (Very)
Threatened	Likert scale from 1 (Not) to 7 (Very)
Accepted	Likert scale from 1 (Not) to 7 (Very)
Frightened	Likert scale from 1 (Not) to 7 (Very)

What was I doing (just before the beep went off?)	Open-ended question
I would prefer doing something else	Likert scale from 1 (Not) to 7 (Very)
I am active	Likert scale from 1 (Not) to 7 (Very)
This requires a lot of effort	Likert scale from 1 (Not) to 7 (Very)
I am skilled at doing this	Likert scale from 1 (Not) to 7 (Very)
I am challenged by it	Likert scale from 1 (Not) to 7 (Very)
Since the last beep, the most important event that happened to me was:	Open-ended question
This was: very unpleasant / pleasant	Bipolar scale from -3 (very unpleasant) to +3 (pleasant)
Why did this event happen?	Open-ended question
It had to do with…	
Myself	Likert scale from 1 (Not) to 7 (Very)
Other people	Likert scale from 1 (Not) to 7 (Very)
Circumstances	Likert scale from 1 (Not) to 7 (Very)
I was responsible for this event	Likert scale from 1 (Not) to 7 (Very)
This beep disturbed me	Likert scale from 1 (Not) to 7 (Very)
It is now exactly	Reporting the response time by hour and minutes