AMERICAN PSYCHOLOGICAL ASSOCIATION

Psychological Methods

© 2021 American Psychological Association
ISSN: 1082-989X

2023, Vol. 28, No. 6, 1335–1357
https://doi.org/10.1037/met0000447

# Detecting Mean Changes in Experience Sampling Data in Real Time: A Comparison of Univariate and Multivariate Statistical Process Control Methods

Evelien Schat[1], Francis Tuerlinckx[1], Arnout C. Smit[2, 3], Bart De Ketelaere[4], and Eva Ceulemans[1]

[1] Quantitative Psychology and Individual Differences, Department of Psychology and Education Sciences, KU Leuven

[2] Interdisciplinary Center Psychopathology and Emotion Regulation (ICPE), University Center for Psychiatry, University Medical Center Groningen, University of Groningen

[3] Clinical Psychology, Faculty of Behavioral and Movement Sciences, Vrije Universiteit Amsterdam

[4] Division of Mechatronics, Biostatistics and Sensors, Department of Biosystems, KU Leuven

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

***Abstract***

Detecting early warning signals of developing mood disorders in continuously collected affective experience sampling (ESM) data would pave the way for timely intervention and prevention of a mood disorder from occurring or to mitigate its severity. However, there is an urgent need for online statistical methods tailored to the specifics of ESM data. Statistical process control (SPC) procedures, originally developed for monitoring industrial processes, seem promising tools. However, affective ESM data violate major assumptions of the SPC procedures: The observations are not independent across time, often skewed distributed, and characterized by missingness. Therefore, evaluating SPC performance on simulated data with typical ESM features is a crucial step. In this article, we didactically introduce six univariate and multivariate SPC procedures: Shewhart, Hotelling's $T^2$, EWMA, MEWMA, CUSUM and MCUSUM. Their behavior is illustrated on publicly available affective ESM data of a patient that relapsed into depression. To deal with the missingness, autocorrelation, and skewness in these data, we compute and monitor the day averages rather than the individual measurement occasions. Moreover, we apply all procedures on simulated data with typical affective ESM features, and evaluate their performance at detecting small to moderate mean changes. The simulation results indicate that the (M)EWMA and (M)CUSUM procedures clearly outperform the Shewhart and Hotelling's $T^2$ procedures and support using day averages rather than the original data. Based on these results, we provide some recommendations for optimizing SPC performance when monitoring ESM data as well as a wide range of directions for future research.

***Translational Abstract***

Detecting early warning signals of developing mood disorders in continuously collected data, would pave the way for timely intervention and prevention of a mood disorder from occurring or to mitigate its severity. We focus on data collected in experience sampling (ESM) studies, where individuals report on their momentary affect at a number of occasions throughout the day, for multiple days. To detect such early warning signals, there is an urgent need for online statistical methods tailored to the specifics of ESM data. Statistical process control (SPC) procedures, originally developed for monitoring industrial processes, seem promising tools. However, affective ESM data violate major assumptions of the SPC procedures. Therefore, evaluating SPC performance on simulated data with typical ESM features is a crucial step. In this paper, we didactically introduce three univariate and three multivariate SPC procedures. Their behavior is illustrated on publicly available affective ESM data of a patient that relapsed into depression. To deal with the assumption violations, we compute and monitor the day averages rather than the individual measurement occasions. Moreover, we apply all procedures on simulated data with typical affective ESM features, and evaluate their

This article was published Online First December 16, 2021.

Evelien Schat https://orcid.org/0000-0003-1169-3984
Francis Tuerlinckx https://orcid.org/0000-0002-1775-7654
Arnout C. Smit https://orcid.org/0000-0001-9465-8687
Bart De Ketelaere https://orcid.org/0000-0002-5140-1643
Eva Ceulemans https://orcid.org/0000-0002-7611-4683

Evelien Schat, Eva Ceulemans, and Francis Tuerlinckx were supported by a research grant from the Research Council of KU Leuven (C14/19/054). Arnout C. Smit was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovative program (ERC-CoG-2015; 681466 to M. Wichers).

The computational resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Hercules Foundation and the Flemish Government department EWI. We thank Kristof Meers for his help in using this supercomputer. The data set analyzed in this article are publicly available at https://osf.io/j4fg8/. The results appearing in the article were presented at the Society for Ambulatory Assessment Conference 2021.

Correspondence concerning this article should be addressed to Evelien Schat, Quantitative Psychology and Individual Differences, Department of Psychology and Education Sciences, KU Leuven, Tiensestraat 102 Box 3713, B-3000 Leuven, Belgium. Email: evelien.schat@kuleuven.be

performance at detecting small to moderate mean changes. The simulation results indicate that certain SPC procedures clearly outperform others, and support using day averages rather than the original data. Based on these results we provide some recommendations for optimizing SPC performance when monitoring ESM data as well as a wide range of directions for future research.

*Keywords:* statistical process control, online monitoring, experience sampling method, detection of mean changes

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Mood disorders, including major depression, are highly prevalent and come with a large cost for individuals, their social environment and society in general (Steel et al., 2014; Vigo et al., 2016; Wittchen, 2012). Early detection of developing mood disorders is therefore of great importance, as this would allow to intervene and to prevent an episode from occurring or to mitigate its severity. Given that mood disorders are characterized by altered emotional and affective experiences, monitoring these affective experiences across time may be a promising solution.

To capture affective fluctuations over time, many researchers use experience sampling (ESM) approaches (Myin-Germeys et al., 2009, 2018). In ESM studies, participants are instructed to report on their momentary affect at a number of occasions throughout the day, for multiple days. The type of data that are generated by ESM studies are called intensive longitudinal data (see e.g., Hamaker & Wichers, 2017; Lafit et al., 2021). Between-person comparisons of ESM data show that healthy persons generally experience higher levels of positive affect and lower levels of negative affect than depressed persons and demonstrate a certain level of resilience, in that intense emotions do not linger long (Dejonckheere, Mestdagh, et al., 2019; Dejonckheere et al., 2018; Hollenstein et al., 2013; Houben et al., 2015). Moreover, retrospective analyses of longer-term within-person ESM studies yield first indications that in case of an imminent depressive episode, a person's affective system may become less resilient (i.e., higher auto correlation) and more variable and may show increased levels of negative affect and decreased levels of positive affect (Cabrieto, Adolf, et al., 2018; Cabrieto et al., 2019; Nelson et al., 2017; Olthof et al., 2020; Smit et al., 2019; Wichers et al., 2020; Wichers & Groot, 2016). Such changes may thus be potential early warning signals of an imminent depression. Hence, online scanning continuously harvested ESM data for the presence of these early warning signals may be fruitful in the prevention and timely treatment of severe depression.

The goal of this article is to didactically introduce existing online methods from other scientific disciplines and to evaluate how useful they are for detecting (small) changes in the level of positive and negative affect across time. We focus on the family of statistical process control (SPC) procedures (Montgomery, 2009). The origin of SPC lies in industry, where it was developed to monitor production processes over time (Shewhart, 1931). Nowadays, SPC techniques are widely used in numerous domains, such as climate change (Hackney et al., 2013), agriculture (Mertens et al., 2008), and pharmaceutics (Silva et al., 2017). Smit et al. (2019) recently reported a first application of a univariate SPC technique in depression research. However, up to now, it remains unclear how well SPC handles the specific characteristics of ESM data.

## Statistical Process Control

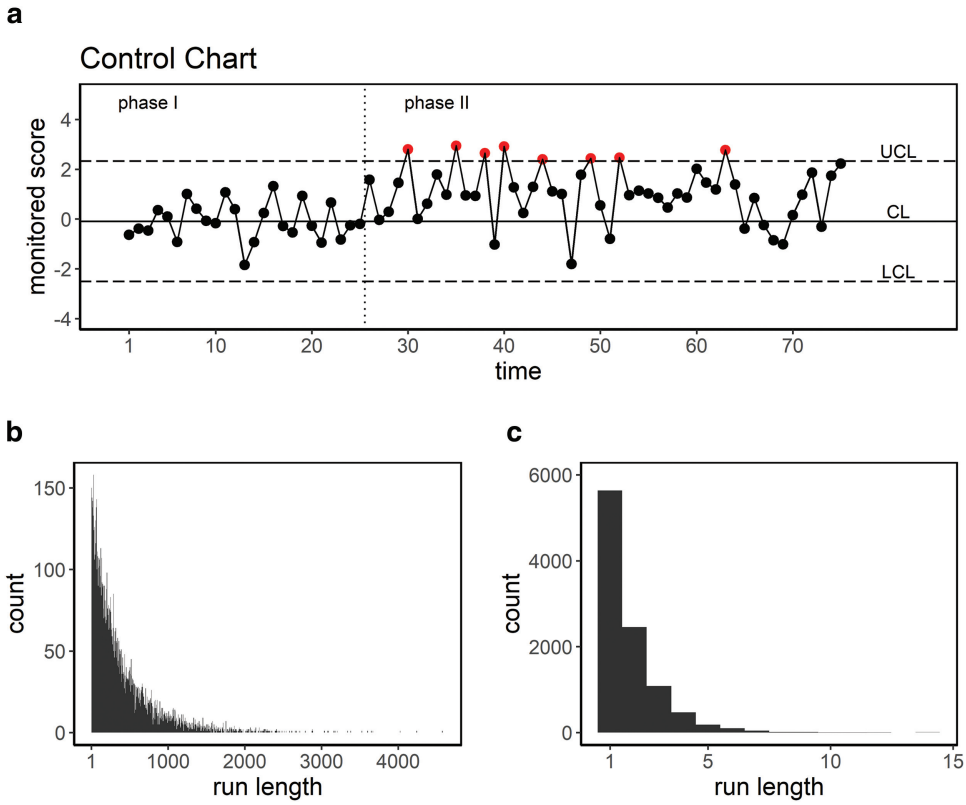
SPC was originally devised to monitor industrial production processes (Shewhart, 1931). The central idea is that quality scores of products will always show some natural variability. The production process is said to be in statistical control if it remains within control limits derived from this variability. However, when a production process is perturbed, the distribution of the quality measures changes, and some of the quality scores may exceed the control limits, indicating that the process is out-of-control.

To apply SPC procedures in practice, two distinct phases are required (for a detailed introduction, see Montgomery, 2009). During Phase I, the natural variability of a set of in-control data is evaluated in order to establish an in-control baseline distribution. The actual online monitoring takes place in Phase II. In this phase, the incoming continuously harvested data are compared to the in-control distribution, to detect and test whether and when the process generating the data goes out-of-control. To assess and visualize the behavior of the process, a control chart is usually drawn. In such a chart, the monitored scores are plotted against time. An example of a Shewhart control chart (see Shewhart Procedure section for more details) is given in Figure 1a, where Phase I consists of 25 measurement occasions; in this phase, the monitored scores were simulated from a normal distribution with $\mu_1 = 0$ and $\sigma_1 = 1$. The remaining 50 measurement occasions belong to Phase II. The scores in this second phase were randomly sampled from a normal distribution with $\mu_2 = 1$ and $\sigma_2 = 1$, implying a mean change of one. The control chart contains a center line (CL), an upper control limit (UCL), and a lower control limit (LCL), which are computed based on the in-control data in Phase I. As long as the Phase II scores fall within the control limits, the process is considered to be in-control. As soon as a Phase II score goes beyond the control limits, the process is flagged as out-of-control. In the control chart in Figure 1a, this happens at measurement occasion 30, as indicated by the first red dot.

The performance of SPC procedures and associated control charts is usually assessed by inspecting the run length (denoted as RL). This run length indicates at which Phase II occasion the process goes out-of-control for the first time. In case the process did not change, this run length should ideally be high because an out-of-control signal would be a false positive, whereas it should be short in case the process does change to allow for fast interventions. However, the run length can hugely vary across different samples from the same Phase I and Phase II distributions, as demonstrated in Figure 1b. This figure shows the typical positively skewed run length distribution of an in-control process. One usually reports the average of this RL distribution (average run length or *ARL*), where one distinguishes between the in-control $ARL_0$ and the out-of-control $ARL_1$. The $ARL_0$ is the average run length, given that the process remains in-control throughout Phase II and is ideally as large as possible. For instance, the $ARL_0$ associated with Figure 1b amounts to 390. The $ARL_1$ is the average run length given that the process changed at the start of Phase II. $ARL_1$ values quantify the power to detect a specific change and should therefore be as small as possible. An example is shown in Figure 1c, where a change of $1\sigma$

**Figure 1**

*Examples of a Control Chart and Run Length Distributions*




***Note***. (a) Example of a control chart. Phase I consists of the first 25 measurement occasions, the remaining 50 measurement occasions constitute Phase II. A mean change was introduced at the start of Phase II. The dashed horizontal lines indicate the UCL and LCL. The solid horizontal line denotes the CL. The red dots indicate out-of-control scores (i.e., scores that are beyond the control limits). (b) Example of a $RL_0$ (run length under the assumption of no change) distribution based on 10,000 simulated data sets, in which both the Phase I and Phase II scores were independently sampled from the same normal distribution. (c) Example of a $RL_1$ (run length when change happens) distribution based on 10,000 simulated data sets, where the Phase I scores were independently sampled from a normal distribution with $\mu_1 = 0$ and the Phase II scores from a normal distribution with $\mu_2 = 1$; both distributions had equal variances. CL = center line; UCL = upper control limit; LCL = lower control limit.

occurred at the start of Phase II, resulting in an $ARL_1$ of 1.8. The $ARL_0$ and $ARL_1$ are in trade off relation with each other (as is usually the case with false alarms and power). Thus, different methods will resolve the trade-off differently.

A wide variety of univariate and multivariate SPC procedures have been proposed, where the univariate ones have been researched most extensively. In this article, we will study the performance of six SPC procedures. First, we will evaluate three standard univariate procedures: the Shewhart procedure (Shewhart, 1931), the exponentially weighted moving average (EWMA; Roberts, 1959) procedure, and the cumulative sum (CUSUM; Page, 1954) procedure. As we will discuss in the next section, these three SPC procedures differ with respect to which score they actually monitor: the original data or a derived score (e.g., a cumulative sum, an exponentially weighted moving average). This difference obviously affects the computation of the control limits and consequently the associated $ARL_0$ and $ARL_1$. However, all three procedures build on the assumption that the original data are independently sampled from a normal distribution. Moreover, a sufficient amount of Phase I data is needed to reliably compute the control limits. Second, we will consider the multivariate extensions of these three procedures, which are the Hotelling's $T^2$ procedure (Hotelling, 1947), the multivariate exponentially weighted moving average (MEWMA; Lowry et al., 1992) procedure, and the multivariate cumulative sum (MCUSUM; Crosier, 1988) procedure, respectively.

## The Current Study

The current study investigates how well the six above mentioned SPC procedures perform when applied to typical affective ESM data. ESM data can indeed be expected to violate one or more of the assumptions (i.e., normality, independence, sufficient amount of Phase I data) underlying these SPC procedures. While positive affect items are typically rather normally distributed, negative affect items tend to be strongly positively skewed in healthy controls (Heininga et al., 2019).

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Moreover, the obtained ESM scores usually are serially dependent rather than independent (Houben et al., 2015; Kuppens et al., 2010), reflecting the tendency of intense emotions to linger for a while. An additional complication here is that the measurement occasions in ESM are usually not equidistant, because participants for instance do not report on their experiences during the night. Finally, whereas for an industrial production process it may be easy to obtain a high number of Phase I observations, when monitoring a single individual, it is often unfeasible to collect a large amount of data under in-control conditions. Therefore, it is important to shed light on how robust SPC procedures are against violations of these assumptions. To this end, we will simulate data with typical ESM features and inspect the resulting $ARL_0$ and $ARL_1$ values.

In this study, we focus on the detection of (small) changes in the level of positive and negative affect, and thus on mean changes across time. It is important to note here that other early warning signs, such as autocorrelation and variance changes often also show up in the mean, as these statistical measures are to some extent interrelated (see e.g., Mestdagh et al., 2018). Moreover, a reanalysis of the unique information in these measures in multiple ESM studies revealed that mean levels of positive and negative affect are often sufficient to indicate that a person is experiencing depressive symptoms (Dejonckheere, Mestdagh et al., 2019).

The remainder of this article is structured as follows. First, six well-known univariate and multivariate SPC procedures are explained through an illustrative example, using publicly available ESM data. Next, we report on a simulation study where we apply these six SPC procedures to simulated data based on empirical ESM data. Lastly, a discussion of the results and directions for future research is presented.

## An Overview of Six Standard SPC Procedures

We first describe the ESM data that we will use throughout this section for illustrative purposes. Next, we introduce the three univariate SPC procedures: Shewhart, EWMA, and CUSUM, followed by the three multivariate extensions: Hotelling's $T^2$, MEWMA, and MCUSUM. We used the R implementation available in the qcc (Scrucca, 2004) and MSQC packages (Santos-Fernandez, 2016), respectively. The R code for the illustrative example and numerical examples of the SPC procedures applied to the ESM data are available on OSF at https://osf.io/kv7hg/.

### ESM Data

The ESM data were provided by a mental health care user with a history of major depressive disorder (Groot, 2010; Wichers & Groot, 2016). The participant was a 57-year-old male and had been using antidepressants for the previous 8.5 years. During the experiment, the participant underwent a dose reduction of the antidepressant venlafaxine. The experiment consisted of three periods: a baseline period (4 weeks), a double-blind period containing the dose reduction (14 weeks), and a follow-up period (16 weeks). During the double-blind period, the participant's antidepressant dose was gradually reduced from 150 mg to 0 mg over a period of 8 weeks. The dose reduction scheme started on Day 42 and ended on Day 98, to which both the participant and researchers were blind. Around Day 127 of the experiment (i.e., the start of the follow-up), a change in depressive symptoms was observed and the participant relapsed into depression. The ESM protocol consisted of 10 measurements per day during which the participant reported on a wide range of momentary states, including affective ones.

Given this setup of the experiment, we used the ESM data of the first 41 days as the Phase I data and the remaining data as the Phase II data. We investigated whether we could confirm the earlier findings for these data, that pointed toward distributional changes up to two months before the relapse (Cabrieto et al., 2019; Smit et al., 2019). A crucial benefit of such an early detection, is that it allows for timely intervention. Thus, we checked whether and when the ESM data seem to go out-of-control, indicating a change in distribution, using six different SPC procedures.

In our analysis, we focused on two affective states: a negative one, "restless" (see Smit et al., 2019); and a positive one, "cheerful." Both affective states were measured on a scale from 1 (*not*) to 7 (*very*). Figure 2a and 2b show the resulting data and associated boxplots for Phase I and II. In both phases, "restless" and "cheerful' are right skewed (Figure 2b). Computing the lag one autocorrelation of each affective state, we see that the data are serially dependent. For instance, in Phase I, the autocorrelation amounts to .24 for "cheerful."[1] Moreover, the participant failed to provide data at many measurement occasions (i.e., 38%, to be precise). All three data characteristics are challenging, because SPC assumes that the monitored scores are independently sampled from a normal distribution. To handle the skewness, the serial dependence and the missing data, we decided to monitor the day averages of these affective states rather than the scores on the individual measurement occasions. First, Figure 2c and 2d show that computing day averages indeed renders the skewed distributions less skewed. Second, the autocorrelation in Phase I reduced to .17 for "cheerful." Third, it allows to easily handle missing data, as for all but one day, the participant responded to at least one measurement occasion.[2]

The day averages of "restless" and "cheerful" were centered around the mean of Phase I (so that the average score in Phase I is 0). This centering operation was needed as for some of the SPC procedures (i.e., CUSUM, MCUSUM, and MEWMA), the control limits are calculated using the spc package in R (Knoth, 2020) or are based on simulations, which assume that the Phase I average equals 0. Figure 2 shows that the day averages seem to fluctuate more after the dose-reduction, which ended on Day 98 (as indicated by the second black line in Figure 2c).

### Univariate Statistical Process Control Methods

#### *Shewhart Procedure*

**Monitored Score.** The Shewhart procedure[3] (Shewhart, 1931) directly monitors the observed scores $x_i$, where $i$ ranges from 1 to $t$. $t$ denotes the total number of measurement occasions and consists of $t_1$ occasions in Phase I and $t_2$ occasions in Phase II (that is, $t = t_1 + t_2$). Other variants of the Shewhart procedure exist (e.g., rational subgroup approach), but we only focus on the Shewhart procedure for individual measurements in this article, given our ESM application. For

[1] The autocorrelation was computed on the subset of equidistant measurement occasions.

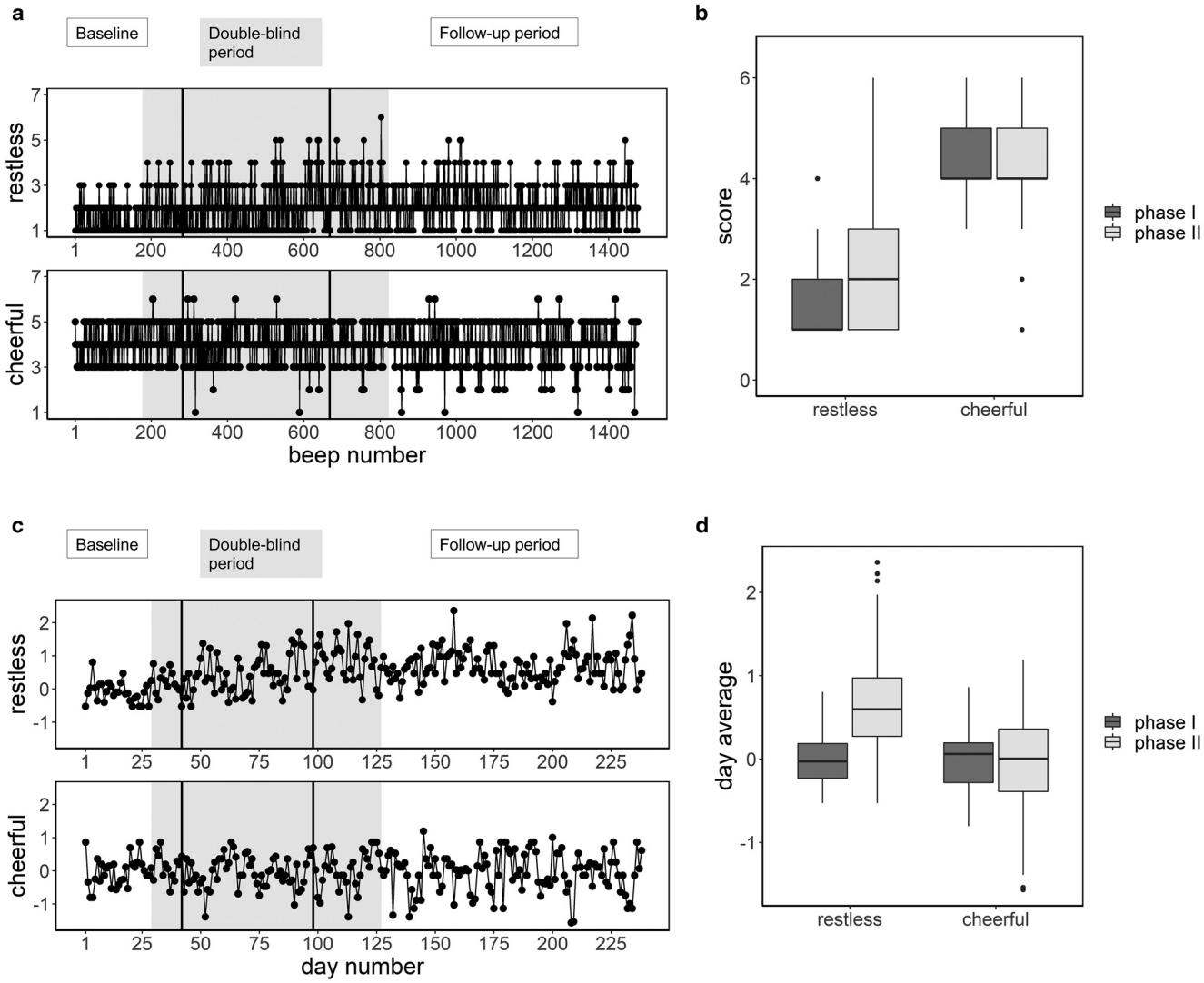
[2] Due to missing values, no day average could be obtained for Day 125. Given that the data are merely used for illustration, this day was left out, resulting in 238 days.

[3] Additionally, a set of decision rules called the Western Electric rules (Western Electric, 1956) can be used to flag an observation as out-of-control: (a) one observation falls beyond the $3\hat{\sigma}_1$ control limits; (b) two out of three consecutive observations fall beyond the $2\hat{\sigma}_1$ limits; (c) four out of five consecutive observations fall beyond the $1\hat{\sigma}_1$ limits; and (d) eight consecutive observations are all located above (resp. below) the center line. Simulation results indicated that the performance of the Shewhart method was not better when using these rules.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

**Figure 2**

*Raw Scores and Day Averages of the Affective States "Restless" and "Cheerful" During a 239-day Antidepressent Reduction ESM Study*




*Note*. (a) Raw scores at beep level of the affective states “restless” and “cheerful”. The experimental periods are indicated by the varying background shading. The start (Day 42, Beep 280) and end (Day 98, Beep 666) of the reduction scheme are indicated by the black vertical lines. (b) Boxplots of the scores of "restless" and “cheerful”, for Phase I and Phase II. (c) Day averages of the affective states “restless” and “cheerful”. The start (Day 42) and end (Day 98) of the reduction scheme are indicated by the black vertical lines. (d) Boxplots of the day averages of “restless” and “cheerful”, for Phase I and Phase II. ESM = experience sampling.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

simplicity, we will not use the qualifier “for individual measurements” in the remainder.

**Calculation of the Control Limits.** The calculation of the control limits starts by quantifying the amount of natural variation in Phase I. The default option[4] in the qcc package is to estimate the population[5] standard deviation $\sigma_1$ of the scores in Phase I by first computing the moving range $MR_i$ of each pair of successive observations, starting at $i = 2$:

$$MR_i = |x_i - x_{i-1}|.$$

Next, the average moving range $\overline{MR}$ is calculated:

[4] Though the Shewhart procedure is typically based on the moving range, the sample standard deviation can also be used to estimate $\sigma_1$. The choice influences SPC performance when the data is autocorrelated. Using the moving range of two successive observations is the default option of the qcc R package. Note that our simulation results reveal that the Shewhart procedure performs rather badly, also for independently and normally distributed observations, where using the sample standard deviation or the moving range are expected to yield the same results.

[5] The population parameters $\mu$ and $\sigma$ are used in the literature on statistical process control. However, as these population parameters are unknown, they are replaced by sample estimates. $\widehat{\mu}_1$ thus equals the sample estimate $\bar{x}_1$ of Phase I. For the multivariate charts, $\Sigma_1$ represents the sample estimate of the covariance matrix of Phase I. Additionally, in the literature, the Phase I average $\widehat{\mu}_1$ is sometimes referred to as $\widehat{\mu}_0$.

$$\overline{MR} = \frac{\sum_{i=2}^{t_1} MR_i}{t_1 - 1},$$

where $t_1$ is the number of measurement occasions in Phase I. Under the assumption of normally and independently distributed sores, dividing $\overline{MR}$ by 1.128 yields an unbiased estimate of $\sigma_1$ (Woodall & Montgomery, 2000), that we will denote by $\hat{\sigma}_1$. Using $\hat{\sigma}_1$, the UCL and LCL are determined as follows:

$$UCL = \hat{\mu}_1 + L_{Shewhat}\hat{\sigma}_1,$$

and

$$LCL = \hat{\mu}_1 - L_{Shewhart}\hat{\sigma}_1,$$

where $\hat{\mu}_1$ is the estimate of the Phase I mean. The parameter $L_{Shewhart}$ determines the width of the in-control zone for Phase II. The $L_{Shewhart}$ value is chosen such that the probability of having an out-of-control observation, given that the process remains in-control in Phase II, is very low. $L_{Shewhart}$ is often set to 3, which implies that the Type I error $\alpha = .0027$ (Montgomery, 2009), if the Phase II scores are assumed to be independently drawn from $\mathcal{N}(\mu_1, \sigma_1^2)$. Given these assumptions, it takes an average of 370 draws before a draw falls outside the [LCL, UCL] interval (i.e., 370 is approximately the mean of the geometric distribution with the event probability set to $\alpha = .0027$, such that $\frac{1}{\alpha} \approx 370$). Therefore, the $ARL_0$ equals 370.

### EWMA Procedure

**Monitored Score.** Rather than monitoring the observed scores themselves, the EWMA procedure (Roberts, 1959) combines past information with current information and tracks a weighted sum of the scores up to now, where the weights depend on how long ago a score was observed. Specifically, the EWMA procedure computes the exponentially weighted moving average $z_i$ at each measurement occasion $i$ ($i = 1, \ldots, t$):

$$z_i = \lambda x_i + (1 - \lambda) z_{i-1}.$$

The starting value $z_o$ is set to the Phase I average $\hat{\mu}_1$. The constant $0 < \lambda \leq 1$ is the weight given to the most recent score. In SPC literature, a weight in the interval $.05 \leq \lambda \leq .25$ is usually recommended, where lower values for $\lambda$ are useful for detecting smaller mean changes (Montgomery, 2009). We set it to .1, which is the default value of the qcc package.[6]

**Calculation of the Control Limits.** Estimating $\sigma_1$ by the sample standard deviation, the UCL and LCL in the EWMA chart are defined as follows:

$$UCL = \hat{\mu}_1 + L_{EWMA}\hat{\sigma}_1\sqrt{\frac{\lambda}{(2-\lambda)}[1 - (1-\lambda)^{2i}]}$$

and

$$LCL = \hat{\mu}_1 - L_{EWMA}\hat{\sigma}_1\sqrt{\frac{\lambda}{(2-\lambda)}[1 - (1-\lambda)^{2i}]},$$

where the term $[1 - (1 - \lambda)^{2i}]$ approaches one as $i$ increases, implying that the control limits do not change anymore (Montgomery, 2009). The parameter $L_{EWMA}$ again determines the range of scores that are considered in-control. To obtain an $ARL_0$ of 370 given $\lambda = .1$, we set $L_{EWMA} = 2.7$, based on the output of the R package spc (Knoth, 2020).

### CUSUM Procedure

**Monitored Score.** The CUSUM procedure also combines past information with current information, by monitoring cumulative sums across the measurement occasions (Page, 1954). It separately sums information pointing toward a positive and a negative mean change, yielding two one-sided CUSUMs, an upper one and a lower one, that can be tracked together in one chart. The upper CUSUM value $C_i^+$ at measurement occasion $i$ is defined as:

$$C_i^+ = \begin{cases} 0 & \text{if } Y_i^+ \leq K \\ Y_i^+ - K & \text{if } Y_i^+ > K \end{cases},$$

where $Y_i^+ = (x_i - \hat{\mu}_1)\hat{\sigma}_1^{-1} + C_{i-1}^+$ and $C_0^+ = 0$. The parameter $K$ is a scale-free allowance parameter, which has to be set relative to the expected mean change. Specifically, it is recommended to choose $K = .5\delta$, where $\delta$ is the expected change size in $\sigma_1$ units. The upper CUSUM $C_i^+$ is reset to 0, when standardizing $x_i$ and adding it to the previous upper CUSUM value yields a value that is not larger than $K$. Otherwise, $C_i^+$ is updated by adding the difference between the standardized $x_i$ score and $K$ to $C_{i-1}^+$.

Similarly, the lower CUSUM value $C_i^-$ at measurement occasion $i$ is defined as:

$$C_i^- = \begin{cases} 0 & \text{if } Y_i^- \geq -K \\ Y_i^- + K & \text{if } Y_i^- < -K \end{cases},$$

where $Y_i^- = (x_i - \hat{\mu}_1)\hat{\sigma}_1{}^{-1} + C_{i-1}^-$ and $C_0^- = 0$.

**Calculation of the Control Limits.** As far as we know, no analytical formulas exist for computing the UCL and LCL. We therefore determined them for an $ARL_0$ of 370 using the spc package in R (Knoth, 2020). Setting $K = .5$, yielded an UCL of 4.77 and a LCL of –4.77. Note that the $C_i^+$ values are compared with the UCL and the $C_i^-$ values with the LCL.

### Application of the Three Univariate SPC Procedures to the ESM Data

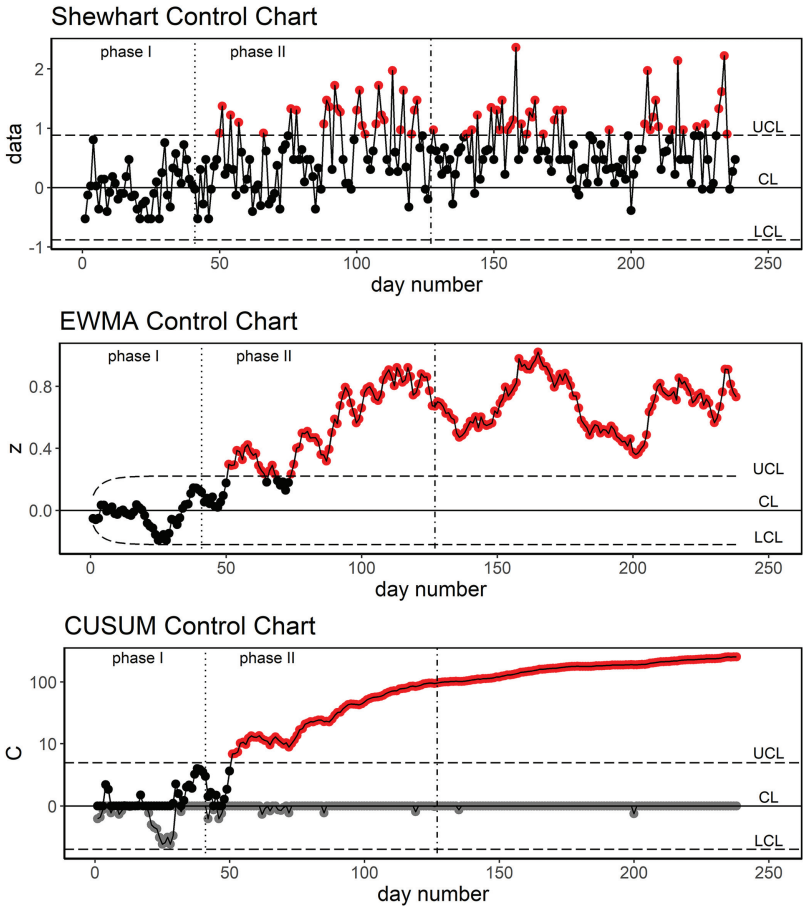
Figure 3 shows the control charts that are obtained when applying the Shewhart, the EWMA, and the CUSUM procedures to the day averages of "restless." The red dots indicate the measurement occasions that are considered to be out-of-control. All three charts show a clear trend indicating that the affective process quickly goes out-of-control, from Day 51 onward. The Shewhart results seem less clear-cut than those of the EWMA and CUSUM chart, because according to the Shewhart chart, many days after Day 51 are still in-control. Indeed, in the CUSUM and EWMA charts, we do not observe out-of-control days in the first 8 days of Phase II (i.e., until Day 51), but afterward almost all days are flagged as out-of-control. This difference in trends is however natural if one considers that CUSUM and EWMA

[6] It should be noted that applications in psychopathology research may be quite different from previous applications and it could be valid to explore other options. Therefore, we also evaluated setting $\lambda = .05$, but obtained very similar simulation results.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

**Figure 3**

*Shewhart Chart, EWMA Chart, and CUSUM Chart of the Day Averages of "Restless"*




*Note.* In the CUSUM chart, the upper CUSUM is shown in black and the lower CUSUM in gray, and the C values on the y-axis are shown on a logarithmic scale. Phase I consists of the first 41 days, the remaining days constitute Phase II, as indicated by the first dashed vertical line. The second dashed vertical line indicates the day of relapse (Day 127). The dashed horizontal lines indicate the UCL and LCL. The solid horizontal line denotes the CL. The red dots indicate the out-of-control days that fall beyond the control limits. EWMA = exponentially weighted moving average; CUSUM = cumulative sum; CL = center line; UCL = upper control limit; LCL = lower control limit.

cumulate information, whereas the Shewhart procedure looks at individual scores, and should therefore not be interpreted too strongly. Actually, given that the practical consequence of a first out-of-control observation would be to check up on the monitored subject immediately, all charts would set off alarm bells rather quickly in Phase II and well in advance of the relapse into depression at Day 127. The CUSUM chart seems to be slightly more sensitive than the EWMA chart, in that the affective process consistently remains out-of-control from Day 51 onward, but this may depend on the parameter values used to construct the control charts (i.e., EWMA parameter $\lambda$, CUSUM parameter $K$). Note that this clear out-of-control trend starts well in advance of the relapse into depression at Day 127. The control charts for the day averages of "cheerful" can be found at https://osf.io/kv7hg/.

## Multivariate Statistical Process Control Methods

We now discuss the multivariate extensions of the three presented univariate SPC procedures. These multivariate extensions share two important features. First, all multivariate procedures transform the multivariate scores into a univariate score, by computing the deviation between the original or a derived score vector at a measurement occasion and the Phase I averages. Herewith, only the size, but not the direction of this deviation (i.e., decrease or increase), is taken into account. This means that multivariate procedures employ a single control limit only and one-sided testing is impossible. Second, all the multivariate control charts take the linear dependencies between the monitored processes into account when transforming the multivariate scores into a univariate one. Intuitively, a simultaneous increase

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

in two independent processes is indeed more indicative of a mean change than a similar increase in two strongly correlated processes.

### *Hotelling's $T^2$ Procedure*

**Monitored Score.** The Hotelling's $T^2$ procedure (Hotelling, 1947) is a multivariate extension of the Shewhart procedure in that the monitored Hotelling $T_i^2$ score only accounts for the observed scores at measurement occasion $i$:

$$\text{Hotelling } T_i^2 = (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_1)' \hat{\boldsymbol{\Sigma}}_1^{-1} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_1),$$

where $\boldsymbol{x}_i$ and $\hat{\boldsymbol{\mu}}_1$ are vectors that respectively contain the scores of the $p$ tracked variables at measurement occasion $i$ and the estimated Phase I averages. $\hat{\boldsymbol{\Sigma}}_1$ is the estimated covariance matrix of Phase I. To illustrate the influence of the covariance matrix, let us consider an example in which two variables are monitored, where $\boldsymbol{x}_i = [1, 1]$ and $\hat{\boldsymbol{\mu}}_1 = [0, 0]$. We first assume strongly dependent variables and use a covariance matrix $\hat{\boldsymbol{\Sigma}}_1 = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$. In this case, Hotelling $T_i^2 = \begin{pmatrix} 1-0 \\ 1-0 \end{pmatrix}' \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}^{-1} \begin{pmatrix} 1-0 \\ 1-0 \end{pmatrix} = 1.05$. Second, we consider independent variables and take $\hat{\boldsymbol{\Sigma}}_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$. Then, Hotelling $T_i^2 = \begin{pmatrix} 1-0 \\ 1-0 \end{pmatrix}' \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}^{-1} \begin{pmatrix} 1-0 \\ 1-0 \end{pmatrix} = 2.0$. This example demonstrates that Hotelling $T_i^2$ is larger when two independent processes deviate in the same direction from the Phase I averages. Finally, we investigate two processes that are strongly negatively correlated and use $\hat{\boldsymbol{\Sigma}}_1 = \begin{bmatrix} 1 & -.9 \\ -.9 & 1 \end{bmatrix}$. Then, Hotelling $T_i^2 = \begin{pmatrix} 1-0 \\ 1-0 \end{pmatrix}' \begin{bmatrix} 1 & -.9 \\ -.9 & 1 \end{bmatrix}^{-1} \begin{pmatrix} 1-0 \\ 1-0 \end{pmatrix} = 20$. This illustrates that Hotelling $T_i^2$ is even larger when two strongly negatively correlated processes deviate in the same direction from the Phase I averages.

**Calculation of the Upper Control Limit.** The upper control limit is defined as (Tracy et al., 1992):

$$UCL = \frac{p(t_1+1)(t_1-1)}{t_1{}^2 - t_1 p} F_\alpha(p, t_1 - p),$$

where $F$ denotes the $F$-distribution and $\alpha$ is the significance level. To obtain an $ARL_0$ of 370, we set $\alpha$ to .0027; like we did in the Shewhart procedure.

### *MEWMA Procedure*

**Monitored Score.** The MEWMA procedure (Lowry et al., 1992) is the multivariate extension of the EWMA procedure. Therefore, MEWMA computes the multivariate exponentially weighted moving averages $\mathbf{z}_i$ at the different measurement occasions $i$:

$$\mathbf{z}_i = \lambda \boldsymbol{x}_i + (1-\lambda)\mathbf{z}_{i-1}.$$

The starting vector $\mathbf{z}_0$ equals $\hat{\boldsymbol{\mu}}_1$. Again, the constant $0 < \lambda \leq 1$, that we set to .1, specifies the weight given to the current observations. Using $\mathbf{z}_i$, we obtain the MEWMA $T_i^2$ values that are monitored:

$$\text{MEWMA } T_i^2 = (\mathbf{z}_i - \hat{\boldsymbol{\mu}}_1)' \boldsymbol{\Sigma}_{\mathbf{z}_i}^{-1} (\mathbf{z}_i - \hat{\boldsymbol{\mu}}_1),$$

where $\boldsymbol{\Sigma}_{\mathbf{z}_i}$ is the MEWMA covariance matrix at measurement occasion $i$, calculated as:

$$\boldsymbol{\Sigma}_{\mathbf{z}_i} = \frac{\lambda}{2-\lambda}[1 - (1-\lambda)^{2t_{2,i}}]\hat{\boldsymbol{\Sigma}}_1.$$

$t_{2,i}$ is the $i$th time point in Phase II.

**Calculation of the Upper Control Limit.** The control limit for an $ARL_0$ of 370 was determined using the spc package (Knoth, 2017; 2020). Given $\lambda = .1$, the UCL is 10.07.

### *MCUSUM Procedure*

**Monitored Score.** The MCUSUM procedure is the multivariate extension of the CUSUM procedure. Although several multivariate extensions have been proposed, we focus on the proposal by Crosier (1988).[7] The proposal replaces the scalars in the univariate CUSUM procedure by the corresponding vectors, and accounts for the covariance of the different monitored variables. Hence, the MCUSUM vectors $\boldsymbol{C}_i^+$ are defined as[8]

$$\boldsymbol{C}_i^+ = \begin{cases} \mathbf{0} & \text{if } Y_i^+ \leq K \\ (\boldsymbol{C}_{i-1}^+ + \boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_1)\left(1 - \dfrac{K}{Y_i^+}\right) & \text{if } Y_i^+ > K \end{cases},$$

where $Y_i^+ = [(\boldsymbol{C}_{i-1}^+ + \boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_1)' \hat{\boldsymbol{\Sigma}}_1^{-1} (\boldsymbol{C}_{i-1}^+ + \boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_1)]^{1/2}$. The starting vector $\boldsymbol{C}_0^+$ is set to $\mathbf{0}$. In line with the CUSUM procedure, the allowance parameter $K$ is set relative to the expected size of the mean change $\delta$, expressed in terms of standard deviations: $K = .5\delta$. Based on the MCUSUM vectors $\boldsymbol{C}_i^+$, we obtain the MCUSUM $T_i$ values that are monitored:

$$\text{MCUSUM } T_i = [\boldsymbol{C}_i^{+\prime} \; \hat{\boldsymbol{\Sigma}}_1^{-1} \; \boldsymbol{C}_i^+]^{1/2}.$$

**Calculation of the Upper Control Limit.** The control limit for an $ARL_0$ of 370 was based on the simulation results of Lee and Khoo (2006). When monitoring two variables and setting the allowance parameter to $K = .5$ (i.e., the default value in the MSQC package), the appropriate UCL amounts to 6.21.

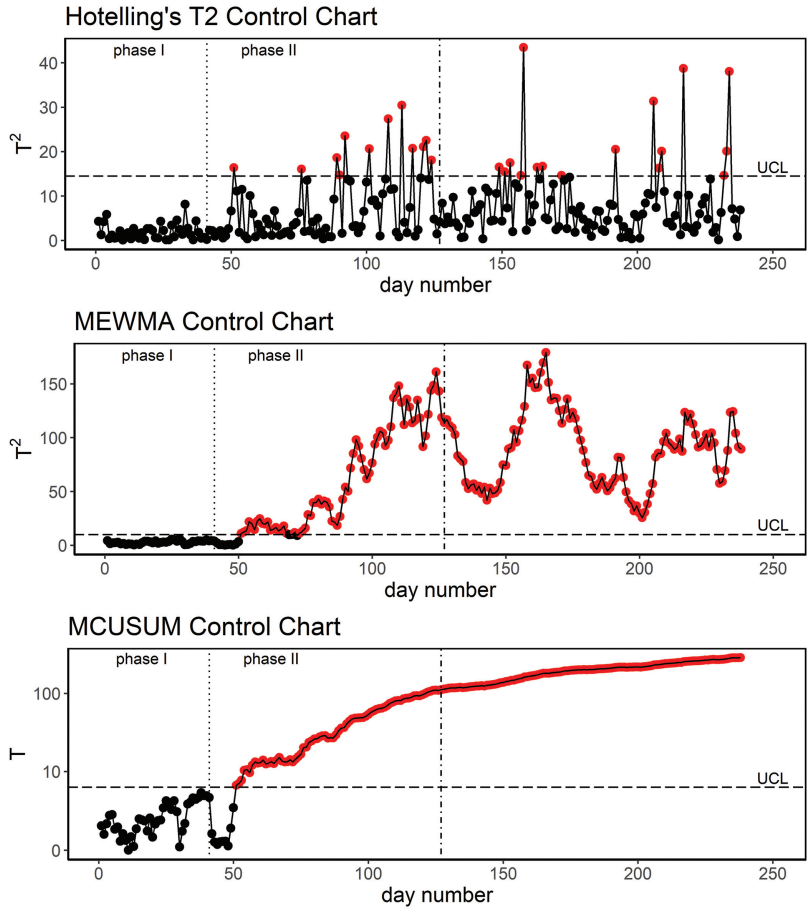
[7] The MC1 procedure proposed by Pignatiello and Runger (1990) is also available in the MSQC package in R (Santos-Fernandez, 2016). The simulation results for the MCUSUM and MC1 procedures do not differ much, aside from the $ARL_0$ values and the $ARL_1$ values for a mean change of .25σ. However, as one procedure does not consistently outperform the other, we only focus on the MCUSUM procedure. Details of the MC1 procedure can be found in Appendix A and the simulation results can be found at https://osf.io/kv7hg/.

[8] The MCUSUM chart applied to univariate data is equal to the CUSUM chart. As the relation between the CUSUM and MCUSUM charts may not be directly apparent, it is demonstrated in Appendix B.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

**Figure 4**

*Hotelling's $T^2$ Chart, MEWMA Chart and MCUSUM Chart of the Day Averages of "Restless" and "Cheerful"*




*Note*. In the MCUSUM chart, the T scores on the y-axis are shown on a logarithmic scale. The Phase I consists of the first 41 days, the remaining days constitute Phase II, as indicated by the first dashed vertical line. The second dashed vertical line indicates the day of relapse (Day 127). The dashed horizontal line indicates the UCL. The red dots indicate the out-of-control days that fall beyond the control limit. MEWMA = exponentially weighted moving average; CUSUM = cumulative sum; CL = center line; UCL = upper control limit; LCL = lower control limit.

### *Application of the Three Multivariate SPC Procedures to the ESM Data*

Figure 4 shows the control charts that result from applying the three multivariate SPC procedures to the day averages of "restless" and "cheerful." Note that the correlation between the day averages of "restless" and "cheerful" amounted to –.48 in Phase I and –.53 in Phase II. The results are largely in line with the conclusions drawn from the corresponding univariate charts. In the Hotelling's $T^2$ chart, some days are flagged as out-of-control, which are interspersed with in-control days. The MEWMA and MCUSUM charts go out-of-control from Day 51 onward, although the MEWMA chart briefly returns in-control around Day 70; this difference in MEWMA and MCUSUM results may again be due to the parameter tuning. Taken together, all charts indicate possible mean changes early on in Phase II.

## Dependence of SPC Results on Data Characteristics

In this section we review the literature on the performance of the six SPC procedures under study. We focus on the influence of four data characteristics: size of the mean change, distribution of the Phase I data, presence of autocorrelation and the amount of data in Phase I.

### *Size of the Mean Change*

The Shewhart and Hotelling's $T^2$ procedures directly monitor the observed scores, making them poor at detecting small mean changes in the underlying process ($\leq 1.5\sigma$) but rather useful for detecting larger changes and sudden spikes in the observed data (Hotelling, 1947; Montgomery, 2009; Shewhart, 1931). The EWMA and CUSUM procedure, as well as their multivariate counterparts, combine past information with current information,

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

making them suitable for detecting small changes in the underlying process (Crosier, 1988; Lowry et al., 1992; Montgomery, 2009; Page, 1954; Roberts, 1959). We therefore expect that the latter two procedures might be better suited for our application.

For EWMA and MEWMA, the choice of the weight $\lambda$ is crucial, in that smaller values of $\lambda$ are to be used for the detection of smaller changes. For some recommendations based on the expected size of the change and the desired $ARL_0$, see Crowder (1987) or Lucas and Saccucci (1990). Note that when $\lambda = 1$, EWMA and MEWMA are equivalent to the Shewhart and Hotelling's $T^2$ procedures, respectively. Indeed, $\lambda = 1$ implies that the monitored scores equal the original scores. Moreover, when calculating the control limits for the EWMA procedure, the term $\sqrt{\frac{\lambda}{(2-\lambda)}[1-(1-\lambda)^{2i}]}$ reduces to one, implying that the limits boil down to the Shewhart ones: $UCL = \hat{\mu}_1 + L_{EWMA}\hat{\sigma}_1$ and $LCL = \hat{\mu}_1 - L_{EWMA}\hat{\sigma}_1$. Consequently, with $L_{EWMA}$ equal to 3, the $ARL_0$ is 370.

For CUSUM and MCUSUM, the allowance parameter $K$ plays an important role as $K$ is set relative to the expected size of the mean change. The procedures are optimal for detecting a change that is equal in size to the expected change, but not for detecting changes of other sizes. Like EWMA, the CUSUM procedure can be turned into the Shewhart procedure, by setting $K$ to the $L_{Shewhart}$ value and the control limits to zero (Woodall & Adams, 1993). In this case, the CUSUM procedure flags a measurement occasion as out-of-control as soon as the cumulative sum differs from 0. In other words, when $|(x_i - \hat{\mu}_1)\hat{\sigma}_1^{-1}| > K$.

### Distribution of the Data

All SPC procedures discussed here assume that the data in both phases are generated from normal distributions. The performance of the Shewhart and Hotelling's $T^2$ procedures are especially influenced by deviations from normality. Even slightly non-normal distributions considerably reduce the $ARL_0$ which in turn increases the number of false alarms (Borror et al., 1999; Stoumbos & Reynolds, 2000; Stoumbos & Sullivan, 2002). Furthermore, the $ARL_1$ for the detection of small mean changes may increase, depending on whether the mean change aligns with the distribution's heavy or thin tail (i.e., alignment with the heavy tail leads to higher $ARL_1$ values). The EWMA and MEWMA procedures appear to be relatively robust to violations of the normality assumption, given that the procedures are properly designed (Borror et al., 1999; Stoumbos & Reynolds, 2000; Stoumbos & Sullivan, 2002; Testik et al., 2003). Specifically, the $\lambda$ parameter should take on values between .05 and .1 in the EWMA procedure and values between .02 and .05 in the MEWMA procedure to remain unaffected by distributional violations. The CUSUM and MCUSUM procedures can also be tuned to be robust to violations of the normality assumption (Chang, 2006; Stoumbos & Reynolds, 2004). Specifically, setting the allowance parameter $K$ between .10 and .30 will yield $ARL_0$ and $ARL_1$ values that are comparable to those for normally distributed data, even for highly skewed or heavy-tailed distributions.

### Autocorrelation

Another critical assumption of SPC procedures is independence of observations over time. In practice this assumption is often violated, leading to incorrect control limits. For instance, for the Shewhart and Hotelling's $T^2$ procedure, sample standard deviation based control limits are too wide, implying longer $ARL_0$ and $ARL_1$ values (Schmid, 1995; Vanhatalo & Kulahci, 2015). On the other hand, moving-range based control limits will be too narrow in case of positive autocorrelation, implying that mean changes are easier to detect, but that the number of false alarms increases as well (Alwan, 1991, 1992). Even low levels of autocorrelation lead to such suboptimal control limits (Montgomery, 2009).

Three approaches to deal with autocorrelated data have been investigated. The first and most simple approach is sampling less frequently from the process under study (Psarakis & Papaleonida, 2007). When more time elapses between sampled observations, the amount of autocorrelation is expected to decrease. However, such subsampling implies that it may take longer to detect a mean change in the process. In the second approach, the standard SPC procedures are used but with adjusted control limits to account for the autocorrelation (see, e.g., Schmid, 1995; Vasilopolous & Stamboulis, 1978; Wardell et al., 1994). The third approach transforms the raw observations such that the transformed observations are independent. Specifically, a time series model is fitted to the data of Phase I. Based on this model, the residuals are obtained for the data in both Phase I and Phase II. The SPC procedures are then applied to these residuals, which are assumed to be independent (see, e.g., Alwan & Roberts, 1988; Harris & Ross, 1991; Lu and Reynolds (1999b), 2001; Mastrangelo & Montgomery, 1995; Montgomery & Mastrangelo, 1991; Noorossana & Vaghefi, 2006). Different time series models have already been implied, including: the AR(1) model (e.g., Bagshaw & Johnson, 1975; Johnson & Bagshaw, 1974), the AR(2) model (Longnecker & Ryan, 1992), the ARMA (1,1) model (Longnecker & Ryan, 1992), the ARIMA(0,1,1) model (Harris & Ross, 1991), and the VAR(1) model (Kalgonda & Kulkarni, 2004). As already indicated in the ESM Data section, we propose a new, fourth approach to deal with autocorrelation, based on computing day averages. However, in the simulation studies we also evaluate the first and third approach, in that we will inspect how (a) including less beeps per day and (b) monitoring AR(1)-residuals rather than the raw day averages affects performance.

### Number of Measurement Occasions in Phase I

SPC procedures heavily rely on estimates of $\mu_1$ and $\sigma_1$, which are obtained using the available data in Phase I. A sufficient amount of Phase I data is therefore required to obtain accurate enough estimates such that the SPC procedures behave as if $\mu_1$ and $\sigma_1$ were known. With little Phase I data, the sampling distributions of these estimates become wider with heavier tails. Due to these heavier tails, the RL distribution has an increased number of short RLs and an increased number of very long RLs (Köksal et al., 2008; Quesenberry, 1993).

So what is considered a sufficient amount of Phase I data? Jensen et al. (2006) noted that more Phase I data is needed than is typically recommended. For example, for the Shewhart procedure less than 50 observations are typically recommended, while research has shown that at least 100 Phase I observations are needed to approach the known-parameter case (Quesenberry, 1993; Rigdon et al., 1994). For the EWMA and CUSUM procedures, values much larger than 100 are required for good parameter estimates (Lu & Reynolds, 1999b, 2001), while about 100 observations in Phase I are typically recommended. As far as we

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

know, no exact recommendations for the number of observations have been given for the multivariate procedures, which is no surprise since differences in dimensionality and covariance structure complicate matters further.

Moreover, what is considered sufficient also interacts with the previously discussed characteristics, especially in the presence of autocorrelation. The effective sample size of autocorrelated data is lower than the actual number of observations, as there are less independent units of information to estimate $\mu_1$ and $\sigma_1$. Furthermore, when filtering out autocorrelation using time series models, the parameters of these time series models (for example, AR(1) model) need to be estimated as well. It has been shown that the accuracy of these estimates has a large influence on the performance of SPC procedures (Kramer & Schmid, 1997; Schmid, 1995). When applying univariate SPC procedures to serially dependent data, 400 Phase I observations are recommended, based on the $ARL_1$ performance (Adams & Tseng, 1998). When applying the EWMA to simple AR(1)-residuals, Lu and Reynolds (1999b) even suggested that 1,000 Phase I observations are needed to obtain an $ARL_0$ close to the prespecified $ARL_0$. Note that the appropriateness of these suggestions may also depend on the amount and type of serial dependency in the data.

## Simulation Studies

Two simulation studies were conducted to evaluate and compare the performance of SPC procedures in detecting mean changes in typical ESM data. The first simulation study focused on univariate procedures and the second simulation study on multivariate procedures. The R code to reproduce the simulation is available at https://osf.io/kv7hg/.

### Design and Rationale of the Two Studies

#### *Data Characteristics*

To make our simulation study relevant, we simulated positive or negative affective time series that mimic typical ESM data. Herewith, we systematically manipulated the four data characteristics that we emphasized in the previous section: size of the mean change, distribution of the data, presence of autocorrelation and number of Phase I data. Additionally, we also varied the number of measurement occasions per day. This allowed us investigate their main and interaction effects on SPC performance. As far as we know, published studies have only focused on a subset of these characteristics.

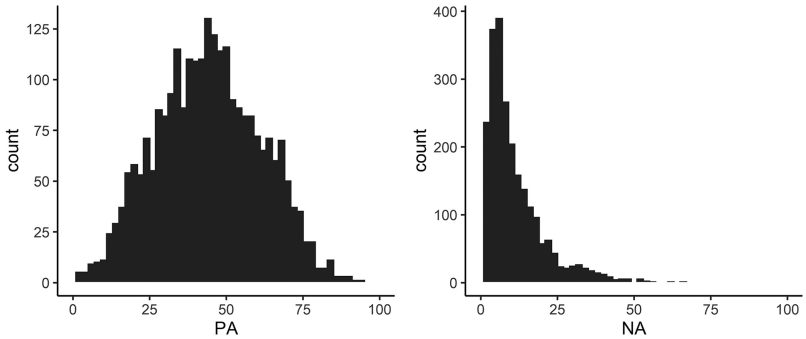
**Size of the Mean Change.** We used a mean change of 0 to study the $ARL_0$ performance and mean changes between .25σ and 1σ to study the $ARL_1$. We focused on changes $\leq 1\sigma$, as these are more likely to occur in affective ESM data. A negative change (i.e., decrease) was introduced for positive affect and a positive change (i.e., increase) was introduced for negative affect, as this corresponds to a person falling into a depressed state. The change was introduced at the first measurement occasion in Phase II. We expected the EWMA and CUSUM procedures, together with their multivariate counterparts, to have lower $ARL_1$ values than the Shewhart and Hotelling's $T^2$ procedures. The $ARL_0$ was expected to be the same for all procedures, as all parameters and corresponding control limits were set such that the $ARL_0$ would equal 370, given that no assumptions were violated.

**Distribution of the Data.** In the univariate study, the distribution of the data was manipulated by simulating either positive or negative affective states. These data were generated based on the ESM data from healthy controls that were collected and reported on by Heininga et al. (2019). Forty healthy controls participated in this study who had no current or previous psychiatric illness. For 7 consecutive days, the participants reported on a range of affective states at 10 semirandom times a day. Positive and negative affect were calculated by averaging the positive affect items (relaxed, happy, euphoric) and negative affect items (depressed, stressed, anxious, anger, restless), respectively. The items were all measured on a scale from 0 (*not at all*) to 100 (*very*). Figure 5 shows the distributions of positive (PA) and negative affect (NA). The distribution of positive affect seems to approximate a normal distribution whereas that of negative affect is clearly right skewed. Therefore, we fitted a gamma distribution to the negative affect data using the fitdistrplus package in R (Delignette-Muller & Dutang, 2015). Based on the resulting parameter estimates, negative affect scores were simulated from a standardized Gamma (1.18, .12) distribution, whereas positive affect scores were simulated from a standard normal distribution. We expected these distributional differences to impact SPC performance, especially for the Shewhart procedure.

In the second study, we focused on bivariate data. Specifically, we manipulated whether two PA, two NA or one PA and one NA

**Figure 5**

*Histograms of Positive Affect (PA) and Negative Affect (NA) Scores in the ESM Data of Heininga et al. (2019)*




*Note*. ESM = experience sampling.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

variables were monitored. To this end, we sampled the affective scores from a multivariate distribution with correlated variables. Specifically, the magnitudes of the correlations were also based on the ESM data of Heininga et al. (2019) and amounted to .40 between PA states, .40 between NA states and –.15 between PA and NA states. To simulate the three types of data, we first sampled two appropriately correlated variables from a standard normal distribution with a given covariance matrix. Next, the probability integral transform was used to obtain bivariate data with the marginal distributions following a Uniform(0, 1) distribution. Lastly, the inverse transform sampling method was used to obtain data from the desired distributions (i.e., normal distribution for positive affect and gamma distribution for negative affect). This transformation changed the correlations slightly (e.g., .36 instead of .40 and –.14 instead of –.15).

**Autocorrelation.** To investigate the impact of autocorrelation on SPC performance, we set the autocorrelation of the variables to either 0 or to .30, by means of a recursive AR(1) filter (Hamilton, 1989). The latter amount of autocorrelation is similar to the values reported in Kuppens et al. (2010).

**Number of Measurement Occasions per Day.** We manipulated the number of measurement occasions (i.e., beeps) per day, by either assuming one, two, five, or 10 beeps per day (for all days) or by assuming the following cyclical pattern (i.e., Day 1: 10 beeps, Day 2: five beeps, Day 3: two beeps, Day 4: one beep, Day 5: 10 beeps, and so on). It is realistic that the number of beeps filled in by a participant differ per day, and thus we investigate an extreme scenario with the cyclical pattern. To mimic an underlying process unfolding continuously throughout time, we started by sampling 20 equidistant beeps per day (i.e., 10 beeps during the day and 10 during the night), which showed an autocorrelation of 0 or of .30. Next, we omitted the night beeps and, where needed (i.e., less than 10 beeps a day), we selected the used beeps from the day beeps as follows: For the one beep settings, we always picked the first day beep; for two beeps, we used the first and sixth day beep; and the second, fourth, sixth, eighth, and 10th day beeps constituted the five beeps settings. For each affective variable, we then computed the day averages across these selected beeps. As discussed earlier, using day averages allows to further investigate and handle the influence of autocorrelation. Using day averages effectively decreased the amount of autocorrelation for the five and 10 beeps per day settings, whereas the autocorrelation for the other settings was already negligible. Specifically, based on the day averages, the average autocorrelation for the data without mean change amounted to 0 for all settings. The lower autocorrelation due to beep averaging is expected to boost SPC performance. Second, averaging scores per day allows to increase to size of the mean change, because the day averages will have lower variance than the original scores. Indeed, computing the size of the mean change (i.e., 0, .25, .50, .75 and 1$\sigma$) in terms of Cohen's *d* yields larger mean changes than those introduced. For example, on the basis of the nonautocorrelated univariate positive affect day averages, the average mean changes in terms of Cohen's *d* are [0, .36, .72, 1.07, 1.43] for two beeps, [0, .57, 1.13, 1.70, 2.26] for five beeps, [0, .80, 1.60, 2.40, 3.21] for 10 beeps, and [0, .38, .77, 1.15, 1.53] for the cyclical pattern.[9] Obviously, increasing the size of the mean change is expected to have a beneficial effect on the *ARL* performance; based on the above changes in effect size, we hypothesize that the cycle and two beeps results will be similar. Finally, using day averages will render the very skewed distribution of the negative affect scores less skewed. We expect this to mainly affect the Shewhart procedure, as the EWMA and CUSUM procedures always average over multiple observations and thereby reduce the skewness by default.

**Number of Days in Phase I.** We varied the amount of Phase I data by setting the number of measurement days to 20, 50, 100, 200, or 500. We expect more Phase I data to lead to better parameter estimates (i.e., estimates of $\mu_1$, $\sigma_1$, time series model parameters) and thus to better SPC performance.

To summarize,[10] the following five data characteristics were varied in both simulation studies and were fully crossed with 10,000 replicates per cell of the design:

1. Size of the mean change: 0, .25, .50, .75, and 1$\sigma$.
2. Distribution of the data: a normal distribution for PA, a gamma distribution for NA in the univariate study; two normals, two gamma's, and one of both in the bivariate study.
3. Autocorrelation: 0 and .30.
4. Number of beeps per day: one, two, five, 10, and cyclical pattern (denoted as cycle).
5. Number of days in Phase I: 20, 50, 100, 200, and 500.

### *Analyses and Performance Measures*

Each simulated dataset was analyzed six times. Specifically, in the univariate simulation study, we applied each of the three univariate SPC procedures twice: once to the raw day averages and once to the corresponding AR(1)-residuals. Similarly, the multivariate simulation study scrutinized the performance of the three multivariate SPC procedures, when applied to either the raw day averages or to the corresponding AR(1)-residuals.[11]

The performance of the SPC procedures was measured in terms of $ARL_0$ and $ARL_1$. Following Qiu and Li (2011), we combined both measures in one *ARL* curve, obtained by plotting the *ARL* as a function of the size of the mean change. The $ARL_0$ performance then corresponds to the mean change of 0 and the $ARL_1$ performance to the remaining mean changes. Ideally, the *ARL* curve starts with an $ARL_0$ of around 370 for a mean change of 0 (as explained before this corresponds to a Type I error probability of .0027 under normality and independence) and shows a steep downward trend as the size of the mean change increases. As mentioned in the

[9] The average mean changes in terms of Cohen's *d* for the cyclical pattern are most similar to those of two beeps per day. Due to unequal sample sizes, the effective sample size for the cyclical pattern, as calculated using the harmonic mean, is equal to $\frac{4}{\left(\frac{1}{1}\right)+\left(\frac{1}{2}\right)+\left(\frac{1}{5}\right)+\left(\frac{1}{10}\right)} = 2.22$.

[10] The Phase I and II data were simulated from the same distribution. The data were then standardized, followed by the introduction of the autocorrelation. Next, the mean change was imposed on the data in phase II. The parameters of the NA distribution may have slightly changed due to these steps, however, the skewness in the data was still clearly present.

[11] When monitoring AR(1)-residuals, an AR(1) model was fit to the day averages of Phase I. Based on this AR(1) model, the residuals of the data in both Phase I and Phase II were obtained.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Number of Measurement Occasions in Phase I section, insufficient Phase I data can lead to an increased number of very high run lengths. Due to computational reasons we cut off the run length of Phase II at 10,000 days. In case no out-of-control day was detected within 10,000 days, the first out-of-control day was set at 10,001 days. The number of replicates per design cell for which we set the run length to 10,001 can be found at https://osf.io/kv7hg/.

## Univariate Results

The results for the AR(1)-residuals did not differ much from the results for the raw day averages. Therefore, we will focus here on the latter results since monitoring raw day averages is obviously simpler than having to compute AR(1)-residuals first. The results for the AR(1)-residuals can however be consulted at https://osf.io/kv7hg/.

Figure 6 shows the *ARL* curves averaged over all "number of beeps per day" and "number of Phase I data" settings of the raw day average results. There is a clear effect of the size of the mean change: the larger the change, the lower the *ARL* values. Overall, the EWMA and CUSUM procedures consistently have a steeper *ARL* curve than the Shewhart procedure, with lower $ARL_1$ values. Furthermore, for negative affect, the $ARL_0$ values of the Shewhart procedure drop substantially, indicating that the procedure is affected by the distribution of the observed scores. While the Shewhart procedure performs almost uniformly worse than EWMA and CUSUM, the differences between the EWMA and CUSUM results are very small and their direction depends on the specific design cells. Because EWMA is on average slightly better than CUSUM and clearly outperforms Shewhart, we opted to focus on the EWMA procedure in the remainder of this section, for simplicity's sake.

From the EWMA results in Figure 7, we see that the $ARL_0$ values for PA are mostly close to 370, the nominal value indicated by the horizontal line. The $ARL_0$ values for NA are too conservative, however, with $ARL_0$ values above 370 for the smaller numbers of beeps per day (i.e., 1, 2, and cyclical) and when Phase I is rather short (i.e., 20 or 50 days).

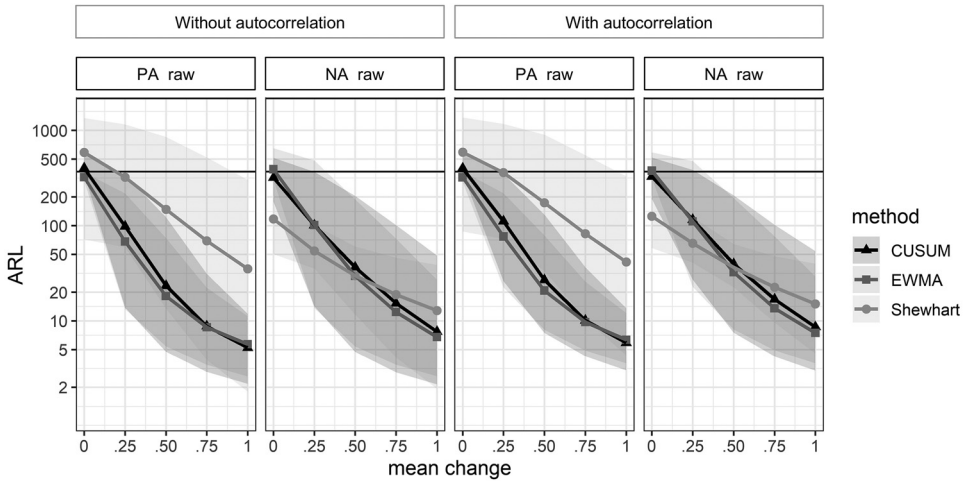
In turn, the $ARL_1$ values improve if the number of beeps per day increases, resulting in steeper *ARL* curves. The *ARL* curves for the cyclical pattern are most similar to the *ARL* curves of two beeps per day. This was expected given the impact of the averaging operation on the effect size of the mean change. The averaging operation also renders the very skewed distribution of the NA scores less skewed. This explains why for 10 beeps per day the difference between the *ARL* for PA and NA is substantially smaller than for one beep per day. The number of days in Phase I plays a further role, as the difference in the *ARL* for PA and NA becomes smaller as the number of Phase I days increase. The effect of autocorrelation is most notable for five and 10 beeps per day, in that the added benefit of monitoring more beeps per day decreases with the presence of autocorrelation. Finally, a longer Phase I leads to lower $ARL_1$ values. Including at least 50 days in Phase I is strongly recommended for five and 10 beeps per day, whereas at least 100 days is recommended for one and two beeps per day as well as for the cycle.

## Bivariate Results

Again, we focus on the results for the raw day averages, because analyzing AR(1)-residuals hardly improved the results, see https://osf.io/kv7hg/. Figure 8 shows the *ARL* curves of the Hotelling's $T^2$, MEWMA, and MCUSUM procedures, averaged over all "number of beeps per day" and "number of Phase I data" settings. The results and conclusions of the univariate study seem to largely generalize to the

**Figure 6**

*ARL of the Univariate SPC Procedures Applied to the Raw Day Averages, Averaged Over All "Number of Beeps per Day" and "Number of Phase I Data" Settings*

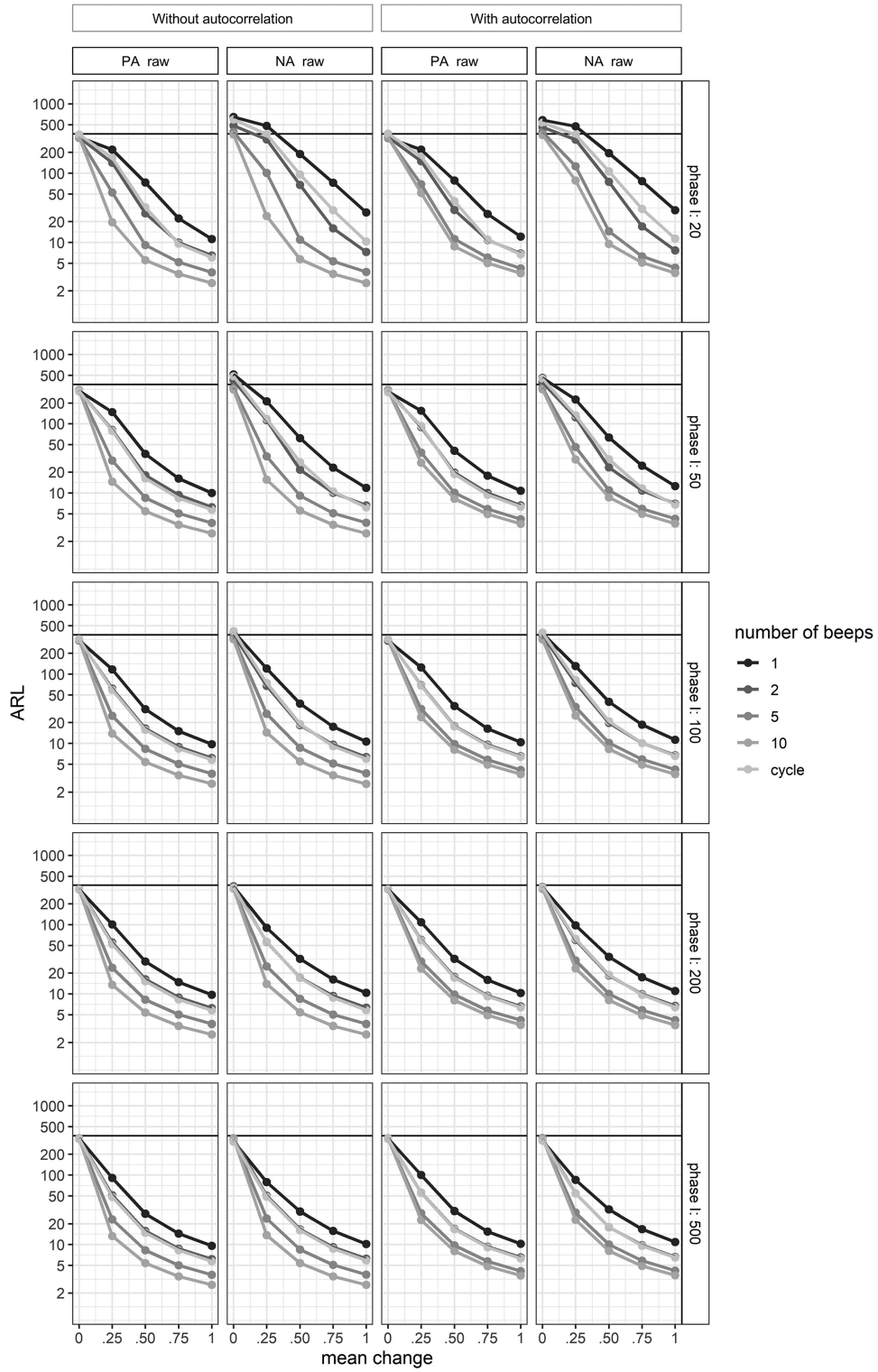



*Note*. The first two columns show the results without autocorrelation, the remaining columns show the results with autocorrelation. Within these settings, the first column shows the results for positive affect (PA), the next column for negative affect (NA). The *ARL* values are shown on a logarithmic scale and the horizontal black line shows the nominal $ARL_0$ value of 370. The shaded areas indicate the range of *ARL* values across all design cells. ARL = average run length; EWMA = exponentially weighted moving average; CUSUM = cumulative sum.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

**Figure 7**

*ARL Curves of the EWMA Procedure Applied to the Raw Day Averages, for Varying Number of Beeps per Day*

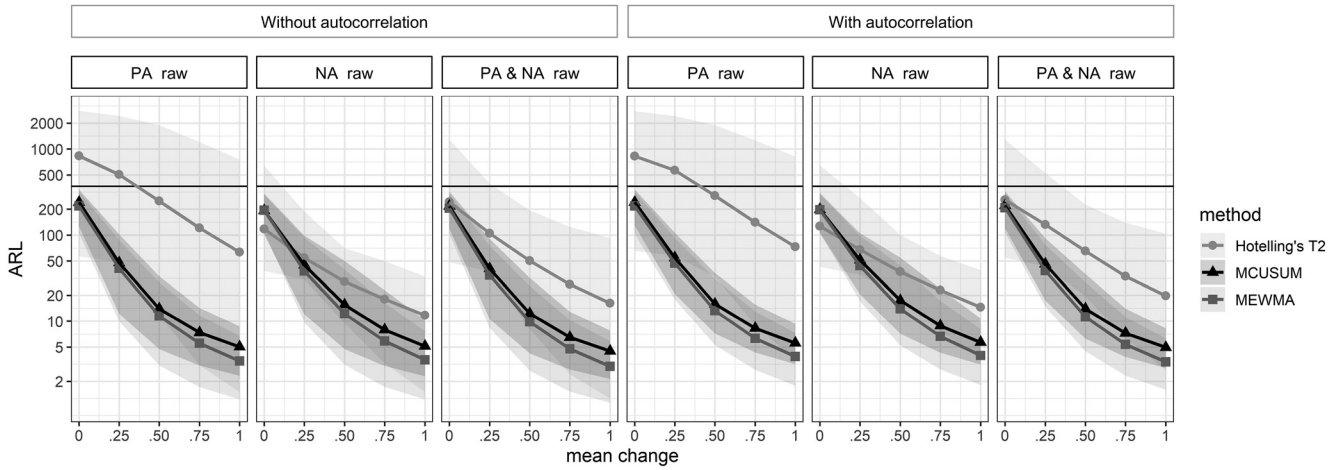



*Note*. The first two columns show the results without autocorrelation, the remaining columns show the results with autocorrelation. Within these settings, the first column shows the results for positive affect (PA), the next column for negative affect (NA). The rows indicate the number of days in Phase I. The *ARL* values are shown on a logarithmic scale and the horizontal black line shows the nominal $ARL_0$ value of 370. ARL = average run length; EWMA = exponentially weighted moving average.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

**Figure 8**

*ARL Curves of the Multivariate SPC Procedures Applied to the Raw Day Averages, Averaged Over All "Number of Beeps per Day" and "Number of Phase I Data" Settings*




*Note*. The first three columns show the results without autocorrelation, the remaining columns show the results with autocorrelation. Within these settings, the first column shows the results for positive affect (PA), the second column for negative affect (NA), and the third column for positive and negative affect (PA and NA). The *ARL* values are shown on a logarithmic scale and the horizontal black line shows the nominal $ARL_0$ value of 370. The shaded areas indicate the range of *ARL* values across all design cells. ARL = average run length; SPC = statistical process control; MCUSUM = multivariate cumulative sum; MEWMA = multivariate exponentially weighted moving average.

multivariate study. Like their univariate counterparts, the MEWMA and MCUSUM procedures consistently outperform the Hotelling's $T^2$ procedure. The MEWMA procedure seems to have more power at detecting changes than the MCUSUM procedure, with lower $ARL_1$ values. However, MEWMA also has more Type I errors with lower $ARL_0$ values. Because the $ARL_0$ differences are relatively small, we will focus on the MEWMA in the remainder of this section.

Figure 9 offers a more detailed overview of the MEWMA curves (the results for all design cells separately can be found at https://osf.io/kv7hg/). We see that the $ARL_0$ values are always too liberal, for both autocorrelated and not autocorrelated scores, although having a higher number of Phase I days clearly improves matters, probably due to better parameter estimation. In line with the univariate results, including more beeps per days also improves the *ARL* curves, with the *ARL* curves for the cyclical pattern again being most similar to the *ARL* curves of two beeps per day. Also, for the $ARL_1$, we observe that more Phase I days are needed than in the univariate case, as the results with 500 days are still visually better than those with 200 days.

## Discussion

Online methods that can accurately detect early warning signals of developing mood disorders in affective ESM data are much needed, as such methods would allow to intervene and to try to prevent an episode from occurring or to mitigate its severity. The family of SPC procedures that were initially developed for monitoring industrial production processes seem very promising tools, at least for detecting mean changes. We therefore recapitulated six well-known univariate and multivariate SPC procedures: Shewhart and Hotelling's $T^2$, EWMA and MEWMA, and CUSUM and MCUSUM, and illustrated their behavior on publicly available affective ESM data of a patient that relapsed into depression. We also investigated their performance on simulated data with typical affective ESM features. We first discuss the obtained results and use them to provide some recommendations. Next, we list a number of remaining challenges, that deserve attention in future research.

### Results and Recommendations

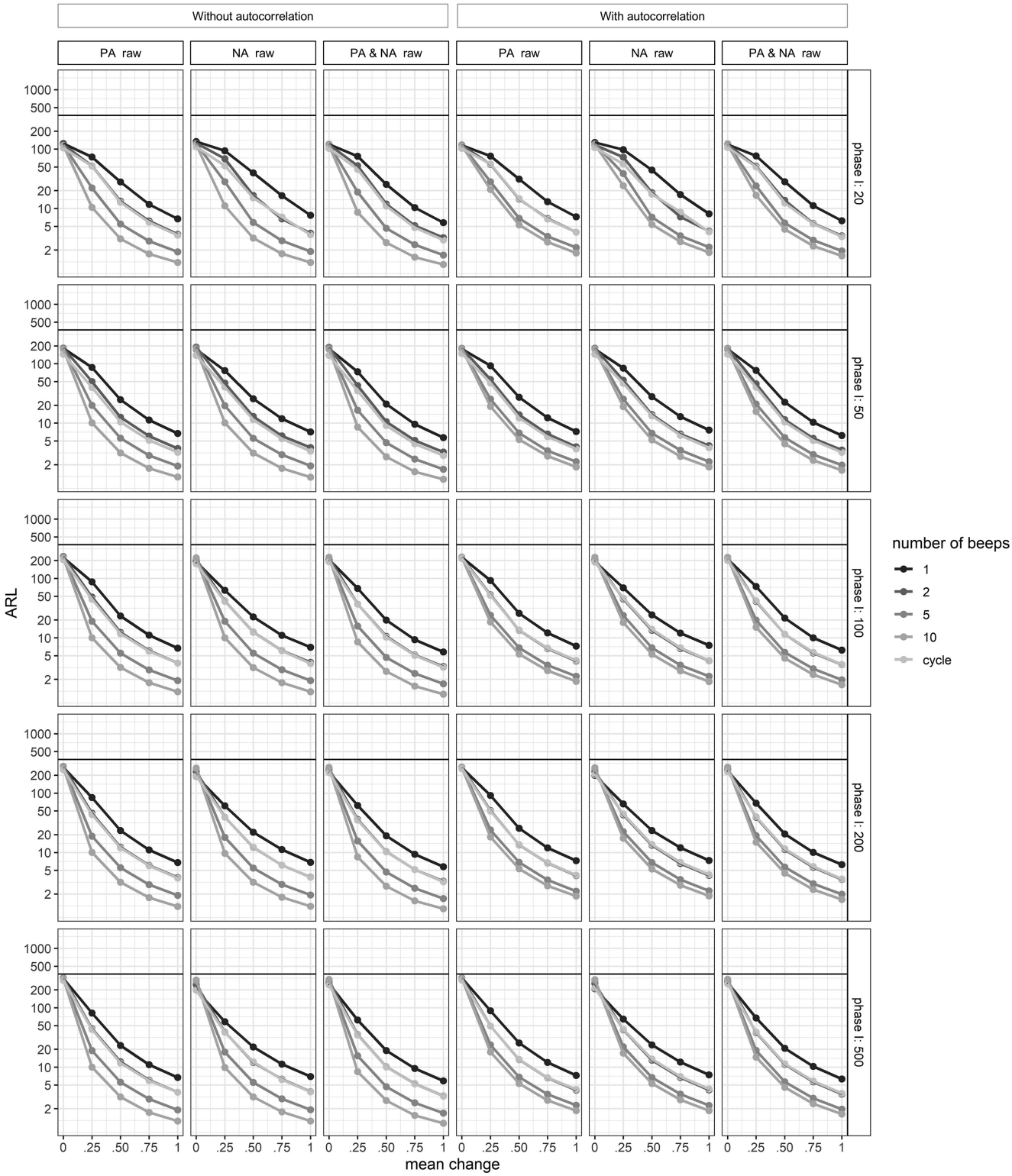
Analyzing the publicly available ESM data of a patient that relapsed into depression after antidepressant tapering (Groot, 2010; Wichers & Groot, 2016) showed that affective ESM data violate major assumptions of standard SPC procedures. Data distributions are skewed, scores are autocorrelated across time, and data were missing for a considerable number of measurement occasions. Importantly, these data features have been reported in great detail in affective ESM research (E. H. Bos et al., 2019; Eisele et al., 2020; Kuppens et al., 2012). As a solution to these violations, we proposed to compute and monitor the day averages rather than the scores at the individual measurement occasions. Next to rendering data distributions less skewed, decreasing autocorrelation in case of many beeps per day and mitigating missing data issues, this averaging operation comes with an attractive and important additional benefit: Effect size clearly increases because measurement error is averaged out, which boosts the performance of SPC procedures.

In the simulation study, we manipulated the size of the mean change, the distribution of the observed data, the presence of autocorrelation, the number of measurement occasions in Phase I, and the number of measurement occasions per day. The day averages of each generated data set were analyzed with all considered SPC procedures. We also investigated whether first fitting an AR(1) model to serially correlated day averages improves SPC performance. The simulation results indicate that the EWMA and CUSUM procedures, together with their multivariate counterparts, perform very similarly and clearly outperform the Shewhart and Hotelling's $T^2$

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

**Figure 9**

*ARL Curves of the MEWMA Procedure Applied to the Raw Day Averages, for Varying Number of Beeps per Day*




*Note*. The first three columns show the results without autocorrelation, the remaining columns show the results with autocorrelation. Within these settings, the first column shows the results for positive affect (PA), the second column for negative affect (NA), and the third column for positive and negative affect (PA and NA). The rows indicate the number of days in Phase I. The *ARL* values are shown on a logarithmic scale and the horizontal black line shows the nominal $ARL_0$ value of 370. ARL = average run length; MEWMA = multivariate exponentially weighted moving average.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

procedures, respectively, which is in line with previous research (Lowry et al., 1992; Montgomery & Mastrangelo, 1991; Roberts, 1959). Additionally, the (M)EWMA and (M)CUSUM procedures seem to be most robust to violations of the normality assumption, as they have shown to be in previous studies (Borror et al., 1999; Stoumbos & Reynolds, 2000; Stoumbos & Sullivan, 2002). Furthermore, results suggest that there is no benefit in monitoring the AR(1)-residuals as compared with the raw day averages. This makes sense as the autocorrelation was (partly) removed due to the beeping schedule and day averaging. Including more measurement occasions per day is advantageous because this increases the effect size of the mean change due to the day averaging operation, and thus lowers the $ARL_1$ values. However, the benefit of more measurement occasions only holds when all beeps are consistently responded to. If this is not the case, as with the cyclical pattern, the benefit may not be that large. Moreover, the day averaging operation renders skewed distributions less skewed. Finally, in the univariate case it is strongly recommended to include at least 50 days in Phase I for five and 10 beeps per day and at least 100 days for one and two beeps per day and the cyclical pattern. However, previous research has suggested an even larger number of Phase I data for the univariate procedures (Lu & Reynolds, 1999b, 2001). Even though it was indeed observed that adding more Phase I days slightly improved SPC performance, we argue that the improvement is too small when weighed against the additional burden on participants and researchers. For the multivariate case, much more Phase I data is needed. Even with 500 Phase I days, there was still room for improvement in the $ARL_0$ values.

## Future Directions

Although the results provided many valuable insights and recommendations, they also pinpointed some challenges and research directions which deserve further investigation. These challenges pertain to the interpretation of the *ARL* as performance measure, the type of change and to the typical characteristics of ESM data.

### *ARL*

**Interpretation of *ARL*.** In line with the SPC literature, we quantified the performance of the SPC procedures in terms of the *ARL*, the average run length across samples from the same Phase I and Phase II distributions. Although parsimonious and standard practice, only reporting the average disregards all other aspects of the run length distribution, which may however importantly qualify the performance interpretation. Indeed, it is known that the run length distribution is right skewed and has a large variance, when no change occurs (Figure 1b). This implies that false positives often occur very early in Phase II, although this is not apparent from the $ARL_0$ value. In case of a mean change, the run length distribution ideally not only has a low $ARL_1$, but also a small variance (Figure 1c), implying that this change will always be detected fast. The above observations imply that a therapist or researcher that is confronted with an out-of-control warning will still always be charged with the task of deciding whether this warning should be taken seriously or might be considered a false positive.

**Choice of $ARL_0$.** In the ESM application and simulation studies, the parameters of the SPC procedures were based on an $ARL_0$ of 370. However, if the cost of intervention is low and thus having more false positives is not problematic, the $ARL_0$ can be lowered making it easier for SPC procedures to detect changes (i.e., lower $ARL_1$ values). On the other hand, if the cost of intervention is very high, the $ARL_0$ can be increased. This decreases the probability of detecting false positives but also makes it more difficult to detect changes (i.e., higher $ARL_1$ values).

Future research is needed to cast the problem in a more general decision theoretical framework. In order to optimally set detection thresholds, we need to investigate what the costs and benefits of an early detection and a missed detection are. Most likely, such a decision theoretical analysis will lead to different recommendations depending on the disorder or problems under study.

### *Type of Change*

The simulation studies focused on detecting abrupt mean changes. Other choices could be considered however, with respect to the statistic as well as the speed of change (i.e., abrupt vs gradual).

**Focus on the Mean.** We opted for the mean for reasons of parsimony and because research has provided some indications that mean levels of affective states are often sufficient to predict which individuals are facing depressive symptoms (Dejonckheere, Mestdagh, et al., 2019), with other statistics yielding little additional information. Moreover, when generating the simulated data, we imposed that all other distributional characteristics remained unchanged across Phase I and II. However, empirical studies have shown that it may be possible to observe early warning signals in other statistics, such as the variance and (auto)correlation (Cabrieto et al., 2019; Wichers & Groot, 2016). Moreover, combinations of these changes can occur simultaneously or sequentially, which may both simplify or complicate detection. It thus remains to be established how SPC procedures perform in case a change occurs in a statistic other than the mean, or in settings with changes in multiple statistics (Crowder & Hamilton, 1992; Lu & Reynolds, 1999a; Reynolds & Stoumbos, 2001). Furthermore, the shape of the distribution may also change across phases, impacting SPC performance. This also remains to be investigated.

**Focus on Abrupt Changes.** In our studies, the mean changes were introduced abruptly at the start of Phase II. However, the development of early warning signs of a potential depressive episode may also be more gradual, slowly evolving from a smaller to a larger change over time. We do not expect the long-term detection of gradual changes to be a problem for SPC procedures (Chen & Nembhard, 2011; Sullivan & Woodall, 1996). However, it is reasonable to predict that gradual changes may not be detected immediately at their start. More research is thus needed to investigate such performance differences between abrupt and gradual changes.

### *Typical Characteristics of Affective ESM Data*

Although our results already shed some light on the effect of the number of variables, the number of Phase I days, the number of beeps per day, and missing data, important open questions and challenges remain.

**The Number of Variables.** Many ESM studies include a large number of monitored variables. In this article, we focused on univariate and bivariate data. Hence, how to optimally handle larger number of variables remains to be investigated. Two obvious options are to perform variable reduction or variable selection. Regarding variable reduction, one can work with average scores of all negative states and of all positive states as is often

done in affective ESM research (for example, Dejonckheere et al., 2018; Dejonckheere, Kalokerinos et al., 2019). This is likely to be a good option if the change signal is present in all variables to the same extent but perturbed by measurement error. However, another interesting option might be to apply principal component analysis on the variables, yielding a few orthogonal dimensions (Bulteel et al., 2014, 2018). Indeed, as demonstrated in Section *Hotelling's $T^2$ procedure*, SPC procedures may often be better at detecting changes when the variables are uncorrelated. Although extracting orthogonal principal components may reduce interpretability, we argue that this is not a major issue if the main goal is to detect changes as soon as possible rather than interpreting the monitored values. Variable selection is a valuable option in case researchers have a good hypothesis on which affective variables are of interest (see e.g., Smit et al., 2019). Furthermore, it remains to be established whether it is better to reduce variables to univariate data and apply univariate SPC methods, or to keep multiple variables and apply multivariate SPC procedures.

**The Number of Phase I Days.** Our results regarding the amount of Phase I data needed for optimal SPC performance are challenging for ESM research. Even for univariate applications, a large number of Phase I data is needed (i.e., 50 days), while for multivariate applications even more data is needed. Clearly, it is not trivial to obtain such large amounts of in-control Phase I data, though a number of recent studies have shown that collecting ESM data across many months is feasible (F. M. Bos et al., 2020; Dejonckheere et al., 2021; Helmich et al., 2020; Myin-Germeys et al., 2018; Olthof et al., 2020; Schreuder et al., 2020; Wichers et al., 2020; Wichers & Groot, 2016). A possible solution could be to use information from more standard 1- or 2-week ESM studies, that are regularly run to investigate between-person differences in affective dynamics and how they relate to other person-level characteristics (Dejonckheere et al., 2018; Dejonckheere, Kalokerinos, et al., 2019; Eisele et al., 2020; Houben et al., 2017). Specifically, we could pool the shorter ESM time series of healthy individuals with similar person-level characteristics as the to be monitored person, to obtain sufficient in-control data to compute control limits. A more sophisticated option would to be combine the pooled data from healthy individuals with a more limited amount of data from the individual under study, where the weight of the individual's data increases with the amount of data (see e.g., Maselyne et al., 2018; for a similar approach to monitoring pigs).

**The Number of Beeps Per Day.** Lastly, our results show that it is beneficial to include multiple measurement occasions per day and work with the day averages of the resulting data. However, considering the intrusiveness of responding to multiple beeps every day and the amount of Phase I data needed, is it reasonable to expect this from people for a longer period of time? Despite the benefits of having multiple beeps per day, it is also worth looking at alternative and especially less intrusive measurements. Other types of intensive longitudinal data are currently also being collected, such as sleep variables and passively obtained physiological data (see e.g., Hori et al., 2016; Minaeva et al., 2020). Future research can check whether monitoring these types of data, either separately or in combination with ESM data, also yields useful early warning signs.

**Missing Data.** Although our simulation results yield first indications that computing day averages may be a promising solution in case of missing data, further investigation is warranted. Indeed, we implemented a specific type of missingness in that the missingness of beeps was spread equally over the day. Further research could therefore look into different patterns of missingness. For example, when an individual is not doing so well, may this be for a day or for a longer period, he or she may be less inclined to respond to beeps. Answering patterns may also be time-dependent, in that morning beeps are more responded to than afternoon beeps or vice versa. Furthermore, the simulation results suggest that it is perhaps better to decrease the burden for participants and keep them motivated to respond to *all* given beeps throughout the day. This is in line with the effective sample size, in terms of the harmonic mean, decreasing due to unequal sample sizes. Responding to *all* five beeps per day yielded better SPC results than responding to only a subset of ten beeps per day (i.e., cycle).

### *Other Types of Data*

Although we focused on ESM data in this article, a wider range of applications and disciplines in the social and behavioral sciences can benefit from SPC procedures. For instance, many experimental studies include some physiological measures nowadays (e.g., Mauss et al., 2005; Meuret et al., 2008). Some studies have used change point detection methods or SPC like procedures to gain insight into the timing of participants' reactions to presented stimuli (Bulteel et al., 2014; Cabrieto, Tuerlinckx et al., 2018; Cabrieto et al., 2017; Hoover et al., 2012; Rosenfield et al., 2010). Within the clinically oriented field, technical advances (i.e., smartphones, wearable devices) are making it easier to collect intensive longitudinal data (i.e., active and passive) in individual's daily life and use those to implement interventions (Myin-Germeys et al., 2018; Torous et al., 2021). Smartphone sensing data, for example, can easily be harvested, and an increasing body of research is indeed suggesting that such passively collected data can aid in the understanding of behavioral patterns as well as contribute to interventions and treatments (Harari et al., 2016; Insel, 2018; Torous et al., 2021).

## Conclusion

SPC procedures are clearly promising for the detection of early warning signals of imminent mood disorders in affective ESM data. We provided some recommendations for optimizing SPC performance in this setting as well as a wide range of directions for future research.

## References

Adams, B. M., & Tseng, I. T. (1998). Robustness of forecast-based monitoring schemes. *Journal of Quality Technology*, *30*(4), 328–339. https://doi.org/10.1080/00224065.1998.11979869

Alwan, L. C. (1991). Autocorrelation: Fixed versus variable control limits. *Quality Engineering*, *4*(2), 167–188. https://doi.org/10.1080/08982119108918904

Alwan, L. C. (1992). Effects of autocorrelation on control chart performance. *Communications in Statistics. Theory and Methods*, *21*(4), 1025–1049. https://doi.org/10.1080/03610929208830829

Alwan, L. C., & Roberts, H. V. (1988). Time-series modeling for statistical process control. *Journal of Business & Economic Statistics*, *6*(1), 87–95.

Bagshaw, M., & Johnson, R. A. (1975). The effect of serial correlation on the performance of CUSUM Tests II. *Technometrics*, *17*(1), 73–80. https://doi.org/10.1080/00401706.1975.10489274

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Borror, C. M., Montgomery, D. C., & Runger, G. C. (1999). Robustness of the EWMA control chart to non-normality. *Journal of Quality Technology*, *31*(3), 309–316. https://doi.org/10.1080/00224065.1999.11979929

Bos, E. H., de Jonge, P., & Cox, R. F. A. (2019). Affective variability in depression: Revisiting the inertia-instability paradox. *British Journal of Psychology*, *110*(4), 814–827. https://doi.org/10.1111/bjop.12372

Bos, F. M., Snippe, E., Bruggeman, R., Doornbos, B., Wichers, M., & van der Krieke, L. (2020). Recommendations for the use of long-term experience sampling in bipolar disorder care: A qualitative study of patient and clinician experiences. *International Journal of Bipolar Disorders*, *8*(1), 38. https://doi.org/10.1186/s40345-020-00201-5

Bulteel, K., Ceulemans, E., Thompson, R. J., Waugh, C. E., Gotlib, I. H., Tuerlinckx, F., & Kuppens, P. (2014). DeCon: A tool to detect emotional concordance in multivariate time series data of emotional responding. *Biological Psychology*, *98*(1), 29–42. https://doi.org/10.1016/j.biopsycho.2013.10.011

Bulteel, K., Tuerlinckx, F., Brose, A., & Ceulemans, E. (2018). Improved insight into and prediction of network dynamics by combining VAR and dimension reduction. *Multivariate Behavioral Research*, *53*(6), 853–875. https://doi.org/10.1080/00273171.2018.1516540

Cabrieto, J., Adolf, J., Tuerlinckx, F., Kuppens, P., & Ceulemans, E. (2018). Detecting long-lived autodependency changes in a multivariate system via change point detection and regime switching models. *Scientific Reports*, *8*(1), 15637. https://doi.org/10.1038/s41598-018-33819-8

Cabrieto, J., Adolf, J., Tuerlinckx, F., Kuppens, P., & Ceulemans, E. (2019). An objective, comprehensive and flexible statistical framework for detecting early warning signs of mental health problems. *Psychotherapy and Psychosomatics*, *88*(3), 184–186. https://doi.org/10.1159/000494356

Cabrieto, J., Tuerlinckx, F., Kuppens, P., Grassmann, M., & Ceulemans, E. (2017). Detecting correlation changes in multivariate time series: A comparison of four non-parametric change point detection methods. *Behavior Research Methods*, *49*(3), 988–1005. https://doi.org/10.3758/s13428-016-0754-9

Cabrieto, J., Tuerlinckx, F., Kuppens, P., Hunyadi, B., & Ceulemans, E. (2018). Testing for the presence of correlation changes in a multivariate time series: A permutation based approach. *Scientific Reports*, *8*(1), 769. https://doi.org/10.1038/s41598-017-19067-2

Chang, Y. S. (2006). Effects of non-normality on the performance of univariate and multivariate CUSUM control charts. *Journal of the Korean Society for Quality Management*, *34*(4), 102–109.

Chen, S., & Nembhard, H. B. (2011). Multivariate Cuscore control charts for monitoring the mean vector in autocorrelated processes. *IIE Transactions*, *43*(4), 291–307. https://doi.org/10.1080/0740817X.2010.523767

Crosier, R. B. (1988). Multivariate generalizations of cumulative sum quality-control schemes. *Technometrics*, *30*(3), 291–303. https://doi.org/10.1080/00401706.1988.10488402

Crowder, S. V. (1987). A simple method for studying run-length distributions of exponentially weighted moving average charts. *Technometrics*, *29*(4), 401–407.

Crowder, S. V., & Hamilton, M. D. (1992). An EWMA for monitoring a process standard deviation. *Journal of Quality Technology*, *24*(1), 12–21. https://doi.org/10.1080/00224065.1992.11979369

Dejonckheere, E., Houben, M., Schat, E., Ceulemans, E., & Kuppens, P. (2021). The short-term psychological impact of the COVID-19 pandemic in psychiatric patients: Evidence for differential emotion and symptom trajectories in Belgium. *Psychologica Belgica*, *61*(1), 163–172. https://doi.org/10.5334/pb.1028

Dejonckheere, E., Kalokerinos, E. K., Bastian, B., & Kuppens, P. (2019). Poor emotion regulation ability mediates the link between depressive symptoms and affective bipolarity. *Cognition and Emotion*, *33*(5), 1076–1083. https://doi.org/10.1080/02699931.2018.1524747

Dejonckheere, E., Mestdagh, M., Houben, M., Erbas, Y., Pe, M., Koval, P., Brose, A., Bastian, B., & Kuppens, P. (2018). The bipolarity of affect and depressive symptoms. *Journal of Personality and Social Psychology*, *114*(2), 323–341. https://doi.org/10.1037/pspp0000186

Dejonckheere, E., Mestdagh, M., Houben, M., Rutten, I., Sels, L., Kuppens, P., & Tuerlinckx, F. (2019). Complex affect dynamics add limited information to the prediction of psychological well-being. *Nature Human Behaviour*, *3*(5), 478–491. https://doi.org/10.1038/s41562-019-0555-0

Delignette-Muller, M. L., & Dutang, C. (2015). fitdistrplus: An R package for fitting distributions. *Journal of Statistical Software*, *64*(4), 1–34. https://doi.org/10.18637/jss.v064.i04

Eisele, G., Vachon, H., Lafit, G., Kuppens, P., Houben, M., Myin-Germeys, I., & Viechtbauer, W. (2020). The effects of sampling frequency and questionnaire length on perceived burden, compliance, and careless responding in experience sampling data in a student population. *Assessment*. Advance online publication. https://doi.org/10.1177/1073191120957102

Groot, P. C. (2010). Patients can diagnose too: How continuous self-assessment aids diagnosis of, and recovery from, depression. *Journal of Mental Health (Abingdon, England)*, *19*(4), 352–362. https://doi.org/10.3109/09638237.2010.494188

Hackney, C., Darby, S. E., & Leyland, J. (2013). Modelling the response of soft cliffs to climate change: A statistical, process-response model using accumulated excess energy. *Geomorphology*, *187*, 108–121. https://doi.org/10.1016/j.geomorph.2013.01.005

Hamaker, E. L., & Wichers, M. (2017). No time like the present: Discovering the hidden dynamics in intensive longitudinal data. *Current Directions in Psychological Science*, *26*(1), 10–15. https://doi.org/10.1177/0963721416666518

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, *57*(2), 357–384. https://doi.org/10.2307/1912559

Harari, G. M., Lane, N. D., Wang, R., Crosier, B. S., Campbell, A. T., & Gosling, S. D. (2016). Using smartphones to collect behavioral data in psychological science: Opportunities, practical considerations, and challenges. *Perspectives on Psychological Science*, *11*(6), 838–854. https://doi.org/10.1177/1745691616650285

Harris, T. J., & Ross, W. H. (1991). Statistical process control for correlated observations. *Canadian Journal of Chemical Engineering*, *69*(1), 48–57. https://doi.org/10.1002/cjce.5450690106

Heininga, V. E., Dejonckheere, E., Houben, M., Obbels, J., Sienaert, P., Leroy, B., van Roy, J., & Kuppens, P. (2019). The dynamical signature of anhedonia in major depressive disorder: Positive emotion dynamics, reactivity, and recovery. *BMC Psychiatry*, *19*(1), 59. https://doi.org/10.1186/s12888-018-1983-5

Helmich, M. A., Wichers, M., Olthof, M., Strunk, G., Aas, B., Aichhorn, W., Schiepek, G., & Snippe, E. (2020). Sudden gains in day-to-day change: Revealing nonlinear patterns of individual improvement in depression. *Journal of Consulting and Clinical Psychology*, *88*(2), 119–127. https://doi.org/10.1037/ccp0000469

Hollenstein, T., Lichtwarck-Aschoff, A., & Potworowski, G. (2013). A model of socioemotional flexibility at three time scales. *Emotion Review*, *5*(4), 397–405. https://doi.org/10.1177/1754073913484181

Hoover, A., Singh, A., Fishel-Brown, S., & Muth, E. (2012). Real-time detection of workload changes using heart rate variability. *Biomedical Signal Processing and Control*, *7*(4), 333–341. https://doi.org/10.1016/j.bspc.2011.07.004

Hori, H., Koga, N., Hidese, S., Nagashima, A., Kim, Y., Higuchi, T., & Kunugi, H. (2016). 24-h activity rhythm and sleep in depressed outpatients. *Journal of Psychiatric Research*, *77*, 27–34. https://doi.org/10.1016/j.jpsychires.2016.02.022

Hotelling H. (1947). Multivariate quality control - Illustrated by the air testing of sample bombsights. In C. Eisenhart, M. W. Hastay, & W. A. Wallis (Eds.), *Techniques of statistical analysis* (pp. 111–184). MacGraw-Hill.

Houben, M., Claes, L., Vansteelandt, K., Berens, A., Sleuwaegen, E., & Kuppens, P. (2017). The emotion regulation function of nonsuicidal self-injury: A momentary assessment study in inpatients with borderline personality disorder features. *Journal of Abnormal Psychology*, *126*(1), 89–95. https://doi.org/10.1037/abn0000229

Houben, M., Van Den Noortgate, W., & Kuppens, P. (2015). The relation between short-term emotion dynamics and psychological well-being: A meta-analysis. *Psychological Bulletin*, *141*(4), 901–930. https://doi.org/10.1037/a0038822

Insel, T. R. (2018). Digital phenotyping: A global tool for psychiatry. *World Psychiatry*, *17*(3), 276–277. https://doi.org/10.1002/wps.20550

Jensen, W. A., Jones-Farmer, L. A., Champ, C. W., & Woodall, W. H. (2006). Effects of parameter estimation on control chart properties: A literature review. *Journal of Quality Technology*, *38*(4), 349–364. https://doi.org/10.1080/00224065.2006.11918623

Johnson, R. A., & Bagshaw, M. (1974). The effect of serial correlation on the performance of CUSUM Tests. *Technometrics*, *16*(1), 103–112. https://doi.org/10.1080/00401706.1974.10489155

Kalgonda, A. A., & Kulkarni, S. R. (2004). Multivariate quality control chart for autocorrelated processes. *Journal of Applied Statistics*, *31*(3), 317–327. https://doi.org/10.1080/0266476042000184000

Knoth, S. (2017). ARL numerics for MEWMA charts. *Journal of Quality Technology*, *49*(1), 78–89. https://doi.org/10.1080/00224065.2017.11918186

Knoth, S. (2020). spc: Statistical process control - Calculation of ARL and other control chart performance measures (0.6.4). https://cran.r-project.org/package=spc

Köksal, G., Kantar, B., Ula, T. A., & Testik, M. C. (2008). The effect of Phase I sample size on the run length performance of control charts for autocorrelated data. *Journal of Applied Statistics*, *35*(1), 67–87. https://doi.org/10.1080/02664760701683619

Kramer, H. G., & Schmid, L. V. (1997). Ewma charts for multivariate time series. *Sequential Analysis*, *16*(2), 131–154. https://doi.org/10.1080/07474949708836378

Kuppens, P., Allen, N. B., & Sheeber, L. B. (2010). Emotional inertia and psychological maladjustment. *Psychological Science*, *21*(7), 984–991. https://doi.org/10.1177/0956797610372634

Kuppens, P., Sheeber, L. B., Yap, M. B. H., Whittle, S., Simmons, J. G., & Allen, N. B. (2012). Emotional inertia prospectively predicts the onset of depressive disorder in adolescence. *Emotion*, *12*(2), 283–289. https://doi.org/10.1037/a0025046

Lafit, G., Adolf, J. K., Dejonckheere, E., Myin-Germeys, I., Viechtbauer, W., & Ceulemans, E. (2021). Selection of the number of participants in intensive longitudinal studies: A user-friendly Shiny app and tutorial to perform power analysis in multilevel regression models that account for temporal dependencies. *Advances in Methods and Practices in Psychological Science*, *4*(1), 1–24. https://doi.org/10.1177/2515245920978738

Lee, M. H., & Khoo, M. B. C. (2006). Optimal statistical designs of a multivariate CUSUM chart based on ARL and MRL. *International Journal of Reliability Quality and Safety Engineering*, *13*(05), 479–497. https://doi.org/10.1142/S0218539306002380

Longnecker, M. T., & Ryan, T. P. (1992). Charting correlated process data.

Lowry, C. A., Woodall, W. H., Champ, C. W., & Rigdon, S. E. (1992). A multivariate exponentially weighted moving average control chart. *Technometrics*, *34*(1), 46–53. https://doi.org/10.2307/1269551

Lu, C. W., & Reynolds, M. R., Jr. (1999a). Control charts for monitoring the mean and variance of autocorrelated processes. *Journal of Quality Technology*, *31*(3), 259–274. https://doi.org/10.1080/00224065.1999.11979925

Lu, C. W., & Reynolds, M. R., Jr. (1999b). EWMA control charts for monitoring the mean of autocorrelated processes. *Journal of Quality Technology*, *31*(2), 166–188. https://doi.org/10.1080/00224065.1999.11979913

Lu, C. W., & Reynolds, M. R., Jr. (2001). CUSUM charts for monitoring an autocorrelated process. *Journal of Quality Technology*, *33*(3), 316–334. https://doi.org/10.1080/00224065.2001.11980082

Lucas, J. M., & Saccucci, M. S. (1990). Exponentially weighted moving average control schemes: Properties and enhancements. *Technometrics*, *32*(1), 1–12. https://doi.org/10.1080/00401706.1990.10484583

Maselyne, J., Van Nuffel, A., Briene, P., Vangeyte, J., De Ketelaere, B., Millet, S., Van den Hof, J., Maes, D., & Saeys, W. (2018). Online warning systems for individual fattening pigs based on their feeding pattern. *Biosystems Engineering*, *173*, 143–156. https://doi.org/10.1016/j.biosystemseng.2017.08.006

Mastrangelo, C. M., & Montgomery, D. C. (1995). SPC with correlated observations for the chemical and process industries. *Journal of Quality Technology*, *11*(2), 79–89.

Mauss, I. B., Levenson, R. W., McCarter, L., Wilhelm, F. H., & Gross, J. J. (2005). The tie that binds? Coherence among emotion experience, behavior, and physiology. *Emotion*, *5*(2), 175–190. https://doi.org/10.1037/1528-3542.5.2.175

Mertens, K., Vaesen, I., Löffel, J., Ostyn, B., Kemps, B., Kamers, B., Bamelis, F., Zoons, J., Darius, P., Decuypere, E., De Baerdemaeker, J., & De Ketelaere, B. (2008). Data-based design of an intelligent control chart for the daily monitoring of the average egg weight. *Computers and Electronics in Agriculture*, *61*(2), 222–232. https://doi.org/10.1016/j.compag.2007.11.010

Mestdagh, M., Pe, M., Pestman, W., Verdonck, S., Kuppens, P., & Tuerlinckx, F. (2018). Sidelining the mean: The relative variability index as a generic mean-corrected variability measure for bounded variables. *Psychological Methods*, *23*(4), 690–707. https://doi.org/10.1037/met0000153

Meuret, A. E., Wilhelm, F. H., Ritz, T., & Roth, W. T. (2008). Feedback of end-tidal pCO2 as a therapeutic approach for panic disorder. *Journal of Psychiatric Research*, *42*(7), 560–568. https://doi.org/10.1016/j.jpsychires.2007.06.005

Minaeva, O., Booij, S. H., Lamers, F., Antypa, N., Schoevers, R. A., Wichers, M., & Riese, H. (2020). Level and timing of physical activity during normal daily life in depressed and non-depressed individuals. *Translational Psychiatry*, *10*(1), 259. https://doi.org/10.1038/s41398-020-00952-w

Montgomery, D. C. (2009). *Introduction to statistical quality control*. Wiley.

Montgomery, D. C., & Mastrangelo, C. M. (1991). Some statistical process control methods for autocorrelated data. *Journal of Quality Technology*, *23*(3), 179–193. https://doi.org/10.1080/00224065.1991.11979321

Myin-Germeys, I., Kasanova, Z., Vaessen, T., Vachon, H., Kirtley, O., Viechtbauer, W., & Reininghaus, U. (2018). Experience sampling methodology in mental health research: New insights and technical developments. *World Psychiatry*, *17*(2), 123–132. https://doi.org/10.1002/wps.20513

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

Nelson, B., McGorry, P. D., Wichers, M., Wigman, J. T. W., & Hartmann, J. A. (2017). Moving from static to dynamic models of the onset of mental disorder a review. *JAMA Psychiatry*, *74*(5), 528–534. https://doi.org/10.1001/jamapsychiatry.2017.0001

Noorossana, R., & Vaghefi, S. J. M. (2006). Effect of autocorrelation on performance of the MCUSUM control chart. *Quality and Reliability Engineering International*, *22*(2), 191–197. https://doi.org/10.1002/qre.695

Olthof, M., Hasselman, F., Strunk, G., van Rooij, M., Aas, B., Helmich, M. A., Schiepek, G., & Lichtwarck-Aschoff, A. (2020). Critical fluctuations as an early-warning signal for sudden gains and losses in patients receiving psychotherapy for mood disorders. *Clinical Psychological Science*, *8*(1), 25–35. https://doi.org/10.1177/2167702619865969

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Page, E. S. (1954). Continuous inspection schemes. *Biometrika*, *41*(1-2), 100–115. https://doi.org/10.2307/2333009

Pignatiello, J. J., Jr., & Runger, G. C. (1990). Comparisons of multivariate CUSUM charts. *Journal of Quality Technology*, *22*(3), 173–186. https://doi.org/10.1080/00224065.1990.11979237

Psarakis, S., & Papaleonida, G. E. A. (2007). SPC procedures for monitoring autocorrelated processes. *Quality Technology & Quantitative Management*, *4*(4), 501–540. https://doi.org/10.1080/16843703.2007.11673168

Qiu, P., & Li, Z. (2011). On nonparametric statistical process control of univariate processes. *Technometrics*, *53*(4), 390–405. https://doi.org/10.1198/TECH.2011.10005

Quesenberry, C. P. (1993). The effect of sample size on estimated limits for X and X control charts. *Journal of Quality Technology*, *25*(4), 237–247. https://doi.org/10.1080/00224065.1993.11979470

Reynolds, M. R., Jr., & Stoumbos, Z. G. (2001). Monitoring the process mean and variance using individual observations and variable sampling intervals. *Journal of Quality Technology*, *33*(2), 181–205. https://doi.org/10.1080/00224065.2001.11980066

Rigdon, S. E., Cruthis, E. N., & Champ, C. W. (1994). Design strategies for individuals and moving range control charts. *Journal of Quality Technology*, *26*(4), 274–287. https://doi.org/10.1080/00224065.1994.11979539

Roberts, S. W. (1959). Control chart tests based on geometric moving averages. *Technometrics*, *1*(3), 239–250. https://doi.org/10.1080/00401706.1959.10489860

Rosenfield, D., Zhou, E., Wilhelm, F. H., Conrad, A., Roth, W. T., & Meuret, A. E. (2010). Change point analysis for longitudinal physiological data: Detection of cardio-respiratory changes preceding panic attacks. *Biological Psychology*, *84*(1), 112–120. https://doi.org/10.1016/j.biopsycho.2010.01.020

Santos-Fernandez, E. (2016). Multivariate statistical quality control using R (Version 1.0.2). https://cran.r-project.org/web/packages/MSQC/MSQC.pdf

Schmid, W. (1995). On the run length of a Shewhart chart for correlated data. *Statistische Hefte*, *36*(1), 111–130.

Schreuder, M. J., Groen, R. N., Wigman, J. T. W., Hartman, C. A., & Wichers, M. (2020). Measuring psychopathology as it unfolds in daily life: Addressing key assumptions of intensive longitudinal methods in the TRAILS TRANS-ID study. *BMC Psychiatry*, *20*(1), 351. https://doi.org/10.1186/s12888-020-02674-1

Scrucca, L. (2004). qcc: An R package for quality control charting and statistical process control (Version 2.7). https://cran.r-project.org/package=qcc

Shewhart, W. A. (1931). *Economic control of quality of manufactured product*. Macmillan and Co., Ltd.

Silva, A. F., Sarraguça, M. C., Fonteyne, M., Vercruysse, J., De Leersnyder, F., Vanhoorne, V., Bostijn, N., Verstraeten, M., Vervaet, C., Remon, J. P., De Beer, T., & Lopes, J. A. (2017). Multivariate statistical process control of a continuous pharmaceutical twin-screw granulation and fluid bed drying process. *International Journal of Pharmaceutics*, *528*(1-2), 242–252. https://doi.org/10.1016/j.ijpharm.2017.05.075

Smit, A. C., Snippe, E., & Wichers, M. (2019). Increasing restlessness signals impending increase in depressive symptoms more than 2 months before it happens in individual patients. *Psychotherapy and Psychosomatics*, *88*(4), 249–251. https://doi.org/10.1159/000500594

Steel, Z., Marnane, C., Iranpour, C., Chey, T., Jackson, J. W., Patel, V., & Silove, D. (2014). The global prevalence of common mental disorders: A systematic review and meta-analysis 1980–2013. *International Journal of Epidemiology*, *43*(2), 476–493. https://doi.org/10.1093/ije/dyu038

Stoumbos, Z. G., & Reynolds, M. R., Jr. (2000). Robustness to non-normality and autocorrelation of individuals control charts. *Journal of Statistical Computation and Simulation*, *66*(2), 145–187. https://doi.org/10.1080/00949650008812019

Stoumbos, Z. G., & Reynolds, M. R., Jr. (2004). The robustness and performance of CUSUM control charts based on the double-exponential and normal distributions. In H.-J. Lenz & P.-T. Wilrich (Eds.), *Frontiers in statistical quality control* (pp. 79–100). Physica. https://doi.org/10.1007/978-3-7908-2674-6_6

Stoumbos, Z. G., & Sullivan, J. H. (2002). Robustness to non-normality of the multivariate EWMA control chart. *Journal of Quality Technology*, *34*(3), 260–276. https://doi.org/10.1080/00224065.2002.11980157

Sullivan, J. H., & Woodall, W. H. (1996). A comparison of multivariate control charts for individual observations. *Journal of Quality Technology*, *28*(4), 398–408. https://doi.org/10.1080/00224065.1996.11979698

Testik, M. C., Runger, G. C., & Borror, C. M. (2003). Robustness properties of multivariate EWMA control charts. *Quality and Reliability Engineering International*, *19*(1), 31–38. https://doi.org/10.1002/qre.498

Torous, J., Bucci, S., Bell, I. H., Kessing, L. V., Faurholt-Jepsen, M., Whelan, P., Carvalho, A. F., Keshavan, M., Linardon, J., & Firth, J. (2021). The growing field of digital psychiatry: Current evidence and the future of apps, social media, chatbots, and virtual reality. *World Psychiatry*, *20*(3), 318–335. https://doi.org/10.1002/wps.20883

Tracy, N. D., Young, J. C., & Mason, R. L. (1992). Multivariate control charts for individual observations. *Journal of Quality Technology*, *24*(2), 88–95. https://doi.org/10.1080/00224065.1992.12015232

Vanhatalo, E., & Kulahci, M. (2015). The effect of autocorrelation on the Hotelling T2 control chart. *Quality and Reliability Engineering International*, *31*(8), 1779–1796. https://doi.org/10.1002/qre.1717

Vasilopolous, A. V., & Stamboulis, A. P. (1978). Modification of control chart limits in the presence of data correlation. *Journal of Quality Technology*, *10*(1), 20–30. https://doi.org/10.1080/00224065.1978.11980809

Vigo, D., Thornicroft, G., & Atun, R. (2016). Estimating the true global burden of mental illness. *The Lancet. Psychiatry*, *3*(2), 171–178. https://doi.org/10.1016/S2215-0366(15)00505-2

Wardell, D. G., Moskowitz, H., & Plante, R. D. (1994). Run-length distributions of special-cause control charts for correlated processes. *Technometrics*, *36*(1), 3–17. https://doi.org/10.1080/00401706.1994.10485393

Western Electric. (1956). *Statistical quality control handbook*. Western Electric Co.

Wichers, M., & Groot, P. C. (2016). Critical slowing down as a personalized early warning signal for depression. *Psychotherapy and Psychosomatics*, *85*(2), 114–116. https://doi.org/10.1159/000441458

Wichers, M., Smit, A. C., & Snippe, E. (2020). Early warning signals based on momentary affect dynamics can expose nearby transitions in depression: A confirmatory single-subject time-series study. *Journal for Person-Oriented Research*, *6*(1), 1–15. https://doi.org/10.17505/jpor.2020.22042

Wittchen, H. U. (2012). The burden of mood disorders. *Science*, *338*(6103), 15. https://doi.org/10.1126/science.1230817

Woodall, W. H., & Adams, B. M. (1993). The statistical design of CUSUM charts. *Quality Engineering*, *5*(4), 559–570. https://doi.org/10.1080/08982119308918998

Woodall, W. H., & Montgomery, D. C. (2000). Using ranges to estimate variability. *Quality Engineering*, *13*(2), 211–217. https://doi.org/10.1080/08982110108918643

(*Appendices follow*)

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

# Appendix A

## MC1 Procedure

The MC1 procedure is a proposed multivariate extension of the CUSUM procedure by Pignatiello and Runger (1990). The procedure replaces the scales in the univariate CUSUM procedure by the corresponding vectors and accounts for the covariance of the different monitored variables. The MC1 vectors $\boldsymbol{C}_i^+$ are based on the norm of the cumulative sum and are thus defined as:

$$||\boldsymbol{C}_i^+|| = \left[\boldsymbol{C}_i^{+\prime}\ \hat{\Sigma}^{-1}\ \boldsymbol{C}_i^+\right]^{1/2}$$

where

$$\boldsymbol{C}_i^+ = \sum_{j=i-t_i^*+1}^{i} (\boldsymbol{x}_j - \hat{\boldsymbol{\mu}}_1).$$

Based on the MC1 vectors $\boldsymbol{C}_i^+$, we obtain the $MC1_i$ values that are monitored:

$$MC1_i = \max\left[0, ||\boldsymbol{C}_i^+|| - t_i^* K\right],$$

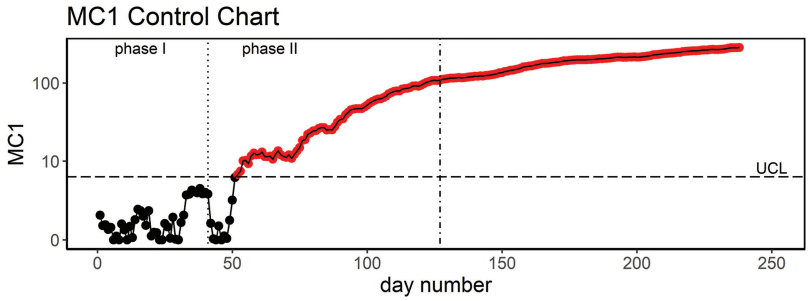
where

$$t_i^* = \begin{cases} t_{i-1}^* + 1 & \text{if } MC1_{i-1} > 0 \\ 1 & \text{if } MC1_{i-1} = 0 \end{cases}.$$

$t_i^* \geq 1$ denotes the number of measurement occasions since $MC1_{i-1}$ was equal to 0. When $MC1_{i-1}$ is equal to 0, $t_i^*$ is set to 1. The MC1 vectors $\boldsymbol{C}_i^+$ thus contains information from measurement occasion $(i - t_i^* + 1)$ up until measurement occasion $i$. The starting value $MC1_0$ is set to 0. The allowance parameter $K$ is set relative to the expected size of the mean change $\delta$. Figure A1 shows the MC1 control chart resulting from applying the MC1 procedure to the day averages of "restless" and "cheerful." The chart is very similar to the MCUSUM chart (see Figure 4).

**Figure A1**

*MC1 Chart of the Day Averages of "Restless" and "Cheerful"*




*Note.* The MC1 scores on the *y*-axis are shown on a logarithmic scale. Phase I consists of the first 41 days, the remaining days constitute Phase II, as indicated by the first dashed vertical line. The second dashed vertical line indicates the day of relapse (Day 127). The dashed horizontal line indicates the UCL. The red dots indicate the out-of-control measurement occasions that fall beyond the control limit.

(*Appendices continue*)

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

## Appendix B

## Relation CUSUM and MCUSUM Procedures

The MCUSUM procedure applied to univariate data produces the same results as the CUSUM procedure. The idea is the same, such that the standardized $x_i$ added to the previous upper CUSUM is compared to allowance parameter $K$. When the MCUSUM procedure is applied to univariate data, $Y_i^+$ becomes:

$$\begin{aligned} Y_i^+ &= \left[\left(C_{i-1}^+ + x_i - \hat{\mu}_1\right)'\left(\hat{\sigma}_1{}^2\right)^{-1}\left(C_{i-1}^+ + x_i - \hat{\mu}_1\right)\right]^{1/2} \\ &= \left[\left(C_{i-1}^+ + x_i - \hat{\mu}_1\right)^2\left(\hat{\sigma}_1{}^2\right)^{-1}\right]^{1/2} \\ &= \left(C_{i-1}^+ + x_i - \hat{\mu}_1\right)\hat{\sigma}_1{}^{-1} \\ &= C_{i-1}^+\hat{\sigma}_1{}^{-1} + \left(x_i - \hat{\mu}_1\right)\hat{\sigma}_1{}^{-1} \end{aligned}$$

If $Y_i^+ > K$, then

$$\begin{aligned} C_i^+ &= \left(C_{i-1}^+ + x_i - \hat{\mu}_1\right)\left(1 - \frac{K}{Y_i^+}\right) \\ &= \left(C_{i-1}^+ + x_i - \hat{\mu}_1\right)\left(1 - \frac{K}{\frac{\left(C_{i-1}^+ + x_i - \hat{\mu}_1\right)}{\hat{\sigma}_1}}\right) \\ &= \left(C_{i-1}^+ + x_i - \hat{\mu}_1\right) - \frac{K\left(C_{i-1}^+ + x_i - \hat{\mu}_1\right)}{\frac{\left(C_{i-1}^+ + x_i - \hat{\mu}_1\right)}{\hat{\sigma}_1}} \\ &= C_{i-1}^+ + x_i - \hat{\mu}_1 - K\hat{\sigma}_1 \\ &= C_{i-1}^+ + \hat{\sigma}_1\left(\left(x_i - \hat{\mu}_1\right)\hat{\sigma}_1{}^{-1} - K\right) \end{aligned}$$

The MCUSUM $C_i^+$ then becomes:

$$C_i^+ = \begin{cases} 0 & \text{if } Y_i^+ \leq K \\ C_{i-1}^+ + \hat{\sigma}_1\left(\left(x_i - \hat{\mu}_1\right)\hat{\sigma}_1{}^{-1} - K\right) & \text{if } Y_i^+ > K \end{cases}$$

As compared to the CUSUM procedure, in the MCUSUM procedure $C_{i-1}^+$ is divided by $\hat{\sigma}_1$ in $Y_i^+$. This is solved for in $C_i^+$ where $K$ is multiplied by $\hat{\sigma}_1$.

Received February 26, 2021
Revision received September 9, 2021
Accepted September 10, 2021 ■

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.