To cite this article:

De Pauw, F., Schat, E., Tuerlinckx, F., & Ceulemans, E. (in press). Reducing participant burden of ESM-based monitoring of mood disorder onset through variable sample size EWMA. *Behavior Research Methods*.

Reducing Participant Burden of ESM-Based Monitoring of Mood Disorder Onset through Variable Sample Size EWMA

Fien De Pauw, Evelien Schat, Francis Tuerlinckx, and Eva Ceulemans

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

Author Note

Fien De Pauw https://orcid.org/0009-0006-0748-891X

Evelien Schat https://orcid.org/0000-0003-1169-3984

Francis Tuerlinckx https://orcid.org/0000-0002-1775-7654

Eva Ceulemans https://orcid.org/0000-0002-7611-4683

Fien De Pauw, Eva Ceulemans, and Francis Tuerlinckx were supported by a research grant from the Research Council of KU Leuven (C14/23/062). Evelien Schat was supported by research grant 12AA826N from the Research Foundation Flanders (FWO). The R code of the simulation study is publicly available at https://osf.io/3rsw7. The dataset analyzed in this manuscript is publicly available at https://osf.io/j4fg8/. Correspondence concerning this article should be addressed to Fien De Pauw, Quantitative Psychology and Individual Differences, Department of Psychology and Education Sciences, KU Leuven, Tiensestraat 102 Box 3713, B-3000 Leuven, Belgium. Email: fien.depauw@kuleuven.be

Abstract

The exponentially weighted moving average (EWMA) procedure is a promising statistical process control technique for monitoring experience sampling method (ESM) data to detect imminent depressive episodes in real time, allowing for timely interventions. However, asking participants to complete multiple ESM questionnaires per day – relevant studies used a daily sample size of five or ten questionnaires – over extended periods is burdensome. We propose a variable sample size (VSS) EWMA approach in which the daily sample size depends on the extremity of participants' ESM scores on the previous days. VSS EWMA allows to keep the daily sample size and thus participant burden low when the individual is doing well. We evaluated the performance of VSS EWMA through a simulation study and an application to empirical ESM data. Our findings indicate that VSS EWMA improves the balance between detection speed and participant burden. Based on these results, we provide recommendations on how to tune VSS EWMA to different priorities and outline directions for future research.

Keywords: statistical process control, real-time monitoring, experience sampling method, adaptive sampling, participant burden

Reducing Participant Burden of ESM-Based Monitoring of Mood Disorder Onset through Variable Sample Size EWMA

Mood disorders such as major depression disorder are highly prevalent. They have a negative impact on the quality of life of the people suffering from it but also on their close environment and society at large (Vigo et al., 2016). Early detection of approaching mood disorder episodes is therefore important to be able to mitigate episode severity or prevent episodes from occurring (Moriarty et al., 2022; Smit et al., 2019; van Zoonen et al., 2014). Previous research has demonstrated that real-time monitoring of momentary affective experiences is promising, as meaningful changes can occur in these experiences more than a month before the onset of depression (Schreuder et al., 2024; Snippe et al., 2023). These results were obtained by assessing momentary affective experiences through the experience sampling method (ESM) and monitoring the resulting data using statistical process control (SPC).

In ESM studies, participants get prompted (i.e., receive 'beeps') several times per day, for multiple days, to report on their momentary experiences while going about their daily lives (Myin-Germeys & Kuppens, 2021). While most ESM studies are rather short (e.g., seven days; Wrzus & Neubauer, 2023), they can span from several weeks up to multiple months when monitoring possible onset of depression. Moreover, accurately detecting impending episodes typically requires sufficiently large daily sample sizes (i.e., number of beeps per day), with relevant long-term ESM studies using a sample size of 5 or 10 beeps per day (Groot, 2010; Snippe et al., 2023). This raises crucial questions regarding feasibility and participant burden (Helmich et al., 2021; Kunkels et al., 2023; van Genugten et al., 2020). The current study therefore aims to achieve an optimal balance between participant burden and onset detection performance by building on variable sample size SPC ideas to adapt the ESM sampling scheme across time.

Statistical Process Control

SPC (Montgomery, 2009) can be defined as the use of statistical procedures to monitor a process (e.g., an industrial production process, the spreading of a virus), to flag when it significantly changes (i.e., goes out of control) and intervention might be needed. This implies that one has to start by characterizing the data that one expects to observe when the process does not change and thus remains in control. This is usually done by specifying the mean and the standard deviation of the data that are generated by the in-control process. This can be done in two ways. A first option is to estimate the mean and the standard deviation using data collected during a baseline phase, also called Phase I, in which the process is assumed to be in control. A second option is to make theoretical assumptions about the process or to specify values based on historical data (i.e., data from earlier studies of the process of interest; Maselyne et al., 2018; Mertens et al., 2008). Based on the (estimated) mean and standard deviation of the in-control process, control limits are computed, which are respectively called the upper (UCL) and lower control limit (LCL).

Phase II then involves monitoring newly incoming data and comparing these against the control limits. As long as incoming data fall within the control limits, it is concluded that the data generating process remains in control and monitoring continues. Monitoring stops as soon as an incoming data point exceeds one of the control limits and an out-of-control warning is raised. Hence, if the process remains in control, the SPC procedure should ideally not report any out-of-control warnings since this would be a type I error (i.e., a false positive or false alarm). On the other hand, if the process does change, the SPC procedure ideally detects it as soon as possible. Failing to do so leads to a type II error (i.e., a false negative).

To study the performance of an SPC procedure, simulation studies are typically conducted. These simulations allow researchers to manipulate both the occurrence and the

size of the process change, which usually takes place at the start of Phase II. Performance is then quantified using the run length (RL), defined as the first Phase II measurement occasion on which an out-of-control warning is given. Run length should be high if no change occurs (equivalent to small type I error) and as low as possible once the process goes out of control (equivalent to a small type II error). In settings where a change occurs, the run length is also referred to as the detection speed (i.e., how fast a change is detected).

To illustrate how SPC can be applied for detecting the onset of depression, we use the fictitious person Ben, an individual at risk for depression, as an example. To characterize the in-control process and compute control limits, we can either collect baseline data when Ben is feeling well, or compute control limits based on the data of other healthy individuals (Schat et al., 2024; Schreuder et al., 2024). We take the former option and implement a Phase I period of 28 days, in which Ben is asked to complete 10 ESM questionnaires (also called beeps) per day about his momentary affective experiences. Based on the recommendations of Schat et al. (2023), we aggregate the 10 scores per day into day averages and monitor these instead of the individual scores. Schat et al. (2023) showed that monitoring day averages helps address violations of the assumptions underlying SPC (i.e., independence and normality), by reducing the skewness and autocorrelation inherent in ESM data, while also increasing the effect size of the change. The sample mean ($\hat{\mu}_{IC}$) and standard deviation ($\hat{\sigma}_{IC}$) of the 28 Phase I days are used to compute the control limits (i.e., UCL and LCL).

Phase II involves monitoring newly reported affective data – again 10 beeps per day aggregated into day averages – and comparing these against the control limits to determine whether and when Ben's affective experiences deviate from what is considered "normal" (i.e., go out-of-control), which may signal an impending depressive episode. Although many different SPC procedures exist, based on Schat et al. (2023), we use the exponentially weighted moving average (EWMA) procedure (Roberts, 2000). The EWMA procedure

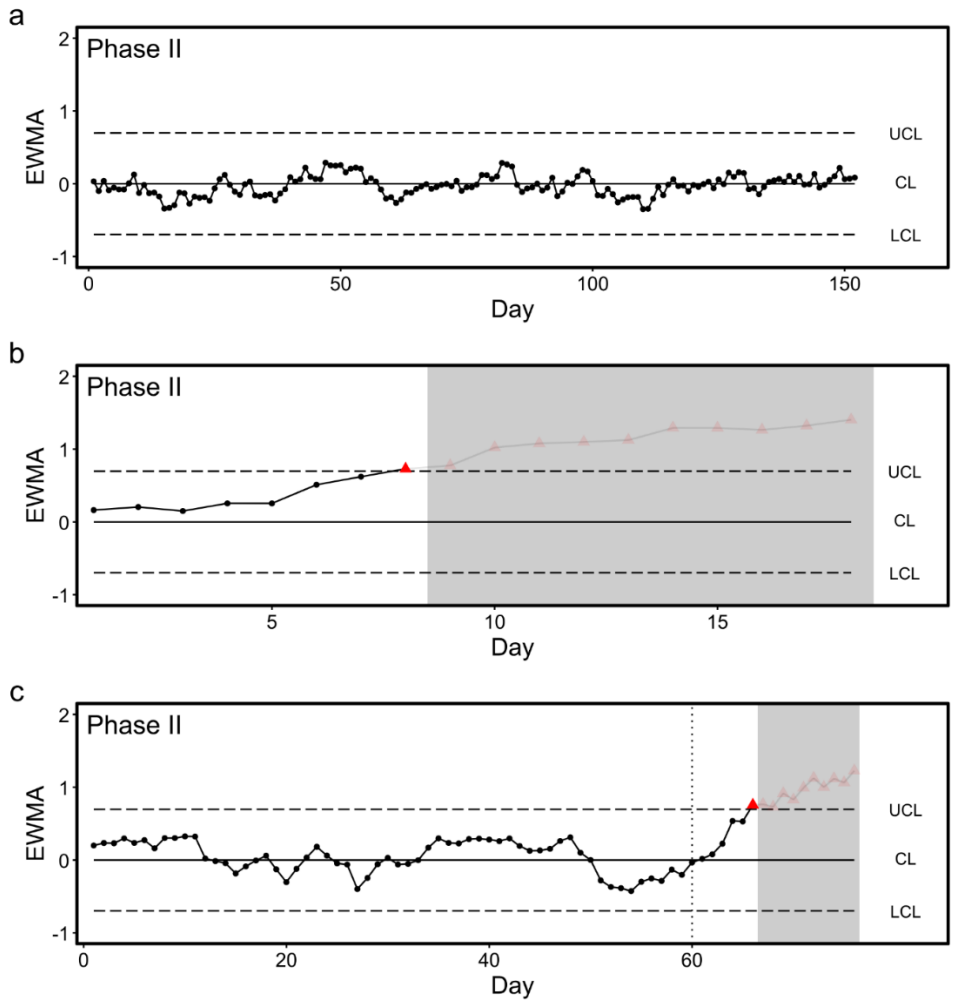
monitors the exponentially weighted moving averages of the day averages, obtained by combining the current day average and the EWMA score at the previous day (see *FSS EWMA Procedure)*. This procedure can be visualized in a control chart, in which Ben's EWMA scores are plotted against time. The control charts in Figure 1 showcase three possible scenarios: (1) no change occurs (i.e., Ben's affective experiences remain in control), (2) a change occurs immediately at the start of Phase II (i.e., Ben's affective experiences immediately shift due to the onset of a depressive state) and (3) Phase II starts with an in-control period, followed by the onset of depression after 60 Phase II days. The latter is the most realistic scenario in practical applications.

As explained before, the RL is usually described as the first measurement occasion in Phase II that is flagged as out of control, assuming that a potential change concurs with the start of Phase II. However, in this study we also introduce a scenario where the mean change is delayed (see Figure 1c). In this scenario, we will define the RL as the first out-of-control day after the change. This keeps the RL comparable across scenarios.

As can be seen in Figure 1, the RL strongly differs between the three scenarios. For the first scenario (no change; Figure 1a), there is no out-of-control warning during the 152 days of Phase II. This is a good result, since an out-of-control warning would be a false positive. In the second scenario (immediate change), the RL is 8. In the third scenario (delayed change), the RL after the change is 6, with no out-of-control warning occurring in the 60 days before the change.

Figure 1

EWMA Control Charts for Ben's Data for Three Scenarios.



Note. EWMA control chart for Ben (based on his affective experiences) if (a) there is no change, (b) there is a change immediately at the start of Phase II and (c) there is a change after 60 days in Phase II. The dashed horizontal lines depict the upper (UCL) and lower (LCL) control limits. The solid horizontal line is the estimated Phase I mean $\hat{\mu}_{IC}$, also referred to as the center line (CL). Every dot or triangle represents the EWMA score computed at a specific day. The black dots and red triangles depict the in-control and out-of-control scores respectively. The shaded background marks when monitoring would have stopped (after the first out-of-control warning). The dotted vertical line in chart (c) marks the introduction of the change. These charts are based on simulated data.

Participant Burden

In the example of Ben, monitoring is stopped when an out-of-control signal appears, to allow for intervention. Therefore, the number of beeps that Ben receives strongly depends on the scenario. When the change in the affective process occurs immediately at the start of Phase II (Figure 1b), Ben has to answer 80 questionnaires in total (i.e., from the beginning of Phase II) before receiving the first out-of-control warning. This is a considerable number, but comparable to what is asked for in typical ESM research (Wrzus & Neubauer, 2023). However, in the more realistic scenarios, where the change is delayed by 60 days (Figure 1c) or where no change occurs (Figure 1a), the number of beeps increases to 660 and 1520, respectively. This is a huge commitment for Ben, raising concerns regarding expected participant burden and compliance.

Previous research on participant burden and compliance in ESM studies indeed highlighted the impact of ESM study design choices, such as sampling frequency (i.e., the number of questionnaires per day) and study duration (Bos et al., 2020; Hasselhorn et al., 2022; Janssens et al., 2018; Moskowitz et al., 2009; Palmier-Claus et al., 2011; Santangelo et al., 2013; Stone et al., 2003; Vachon et al., 2019). However, the ESM protocols needed for monitoring the onset of depression exceed the typical durations of ESM studies. Longer-term ESM studies come with additional challenges and require thoughtful consideration of participant motivation and support strategies (Helmich et al., 2021; Palmier-Claus et al., 2011; Vachon et al., 2019). In such studies, active motivation by researchers can play a crucial role in maintaining a good compliance (Helmich et al., 2020) and even then, participating in these ESM studies remains a big commitment. This is further illustrated by Snippe et al. (2023), who reported that possible participants for their study often decided not to take part because of the expected burden, indicating that not all individuals are willing to

take on such a commitment. Participant burden is thus an important challenge for implementing real-time monitoring approaches in clinical practice.

The Current Study

Up until now, EWMA applications focusing on the detection of depression onset used a fixed sample size (e.g., 10 beeps per day), implying that the number of beeps per day remains constant and large, also on days during which not much is happening. This leaves one wondering whether the number of beeps per day cannot simply be reduced. Schat et al. (2023) indeed already demonstrated that this does not affect performance when the process remains in control. However, their results also show that when the process goes out of control, reducing the number of beeps per day significantly reduces detection speed. Interestingly however, these results point towards a potential solution, which would be to vary the number of beeps per day depending on how likely it is that the process is still in control that day.

A useful starting point to implement such a solution in the context of SPC is provided by Reynolds and Arnold (2001), who proposed an EWMA approach that varies the sample size based on how close incoming data are to the control limits. This method is called variable sample size (VSS) EWMA. The current study investigates how to use VSS EWMA with ESM data to monitor a single variable and examines its performance with regard to detection speed (i.e., RL) and participant burden (i.e., number of beeps sent per day). Note that the proposed approach is specifically intended for studies that are conducted with an early detection goal and may be less suitable for studies with different goals (e.g., understanding mood disorder dynamics).

In the remainder of this paper, we first outline the key principles of the traditional fixed sample size (FSS) EWMA procedure and then introduce the VSS EWMA approach.

Next, we evaluate FSS and VSS EWMA performance using a simulation study. Afterwards, we apply FSS and VSS EWMA to existing empirical data. Finally, we discuss the results and provide some directions for future research.

FSS and VSS EWMA procedures

FSS EWMA Procedure

The FSS EWMA procedure combines past and current information about the process into a single statistic, the exponentially weighted moving average, which is monitored over time (Lucas & Saccucci, 1990). The exponentially weighted moving average $EWMA_i$ at day i is defined as:

$$EWMA_i = \lambda x_i + (1 - \lambda) EWMA_{i-1}\,, \tag{1}$$

where x_i denotes the day average on day i, which is based on n beeps. Thus, the FSS EWMA procedure monitors a weighted sum of all day averages collected thus far, where the weight of a particular day average decreases over time (Roberts, 2000). The λ parameter ($0 < \lambda \leq 1$) represents the weight assigned to the current day average, thereby determining how fast the weight of past day averages decreases. In case $\lambda = 1$, the FSS EWMA method boils down to the Shewhart method (Shewhart, 1926), another SPC procedure that monitors only the current day average rather than a weighted sum of multiple day averages. As smaller values of λ are more suitable for detecting small mean changes, λ is usually taken from the interval $.05 < \lambda \leq .25$ (Montgomery, 2009). Previous studies on depression detection found that values between .05 and .15 work well (Schat et al., 2023; Smit et al., 2023; Smit & Snippe, 2023). In line with the study of Schat et al. (2023), we set λ to .1, the default of the qcc package in R.

When computing the EWMA score for the first day, $EWMA_{i-1}$ is set to the mean of the in-control distribution μ_{IC}. As indicated in the introduction, μ_{IC} and σ_{IC} (i.e., the standard

deviation of the in-control distribution) are either based on historical data, theoretical considerations or are estimated (i.e., $\hat{\mu}_{IC}$ and $\hat{\sigma}_{IC}$) through a baseline phase.

After computing the $EWMA_i$ value, it is compared to a set of control limits to determine whether or not the process goes out of control and to take appropriate action. Given λ, the upper and lower control limits are computed as a function of μ_{IC} and σ_{IC}:

$$UCL = \mu_{IC} + L_{EWMA}\,\sigma_{IC}\sqrt{\frac{\lambda}{2-\lambda}} \tag{2}$$

and

$$LCL = \mu_{IC} - L_{EWMA}\,\sigma_{IC}\sqrt{\frac{\lambda}{2-\lambda}}. \tag{3}$$

The parameter L_{EWMA} determines the width of the control limits. Its value depends on the chosen λ and ARL_0, which is the targeted average number of days before a false positive occurs (i.e., an out-of-control warning when the process remains in control in Phase II). The concept of ARL_0 will be discussed in more detail in the *Simulation Study* section. Following standard SPC practice, we use an ARL_0 of 370, meaning that a false positive should occur after 370 days on average (Montgomery, 2009). Given a λ of .1 and an ARL_0 of 370, the L_{EWMA} is set to 2.70, as determined using the R package 'spc' (Knoth, 2024).

In this paper, we opt to monitor standardized day averages:

$$z_i = \frac{x_i - \mu_{IC}}{\sigma_{IC}} \tag{4}$$

for reasons that will become clear in the next section (see *VSS EWMA Procedure*). Because of this standardization, the formulas for the control limits can be simplified to:

$$UCL = L_{EWMA}\sqrt{\frac{\lambda}{2-\lambda}} \tag{5}$$

and

$$LCL = -L_{EWMA}\sqrt{\frac{\lambda}{2-\lambda}}. \tag{6}$$

VSS EWMA Procedure

The VSS EWMA procedure differs from FSS EWMA in that the number of beeps differs across the days. In this paper, we will assume that there are two options: either a small number n_S or a large number n_L is sent, with $n_S = 5$ and $n_L = 10$. However, note that both n_S and n_L can be set to other values, as long as $n_S < n_L$. The values of n_S and n_L should best be selected based on the goals and characteristics of a specific study.

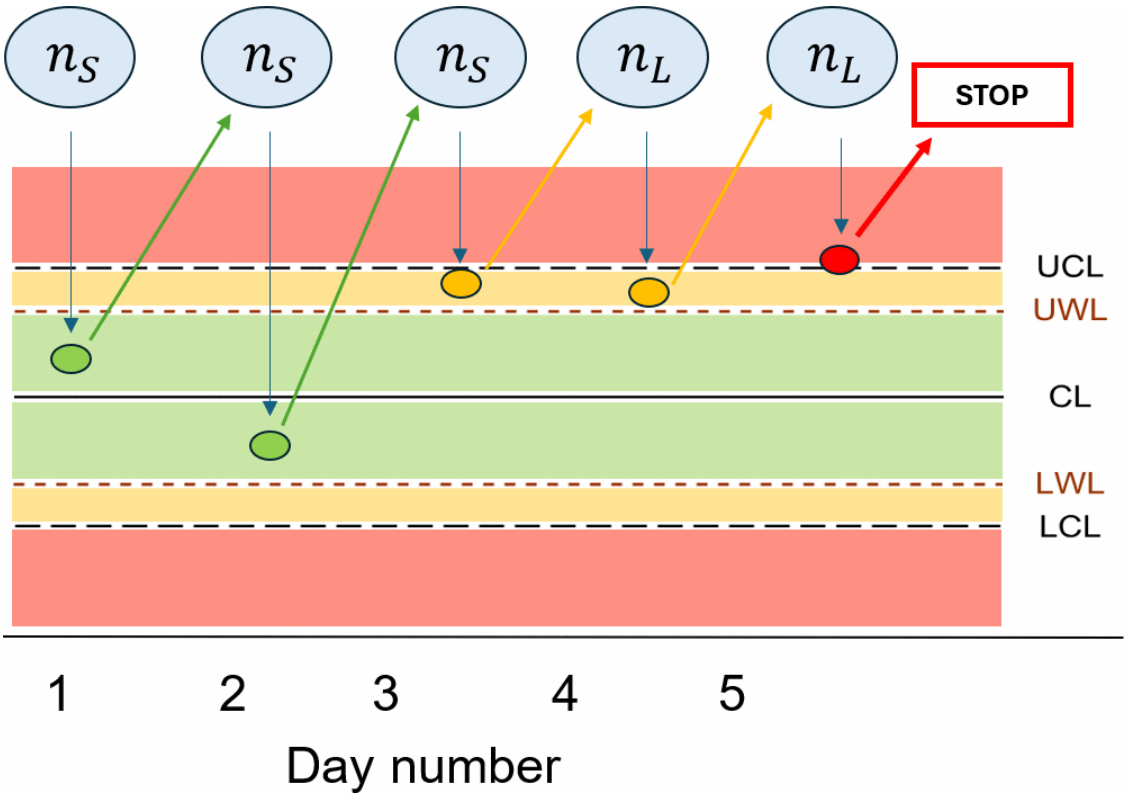
VSS EWMA sets the sample size for the next observation, that is, the number of beeps for the next day, based on the EWMA score of the current day (Reynolds & Arnold, 2001). Hence, when computing and interpreting a VSS EWMA score on a specific day, there are three possible decisions and related actions: (1) continue monitoring and use n_S the next day, (2) continue monitoring and use n_L the next day, or (3) stop monitoring and give an out-of-control signal. In line with FSS EWMA, the decision to continue (options 1 and 2) or stop monitoring (option 3) is still based on the control limits.

Deciding on which sample size to use (i.e., choosing between actions (1) or (2)) when monitoring continues requires introducing an additional set of limits. These limits are respectively called the upper (UWL) and lower warning limit (LWL). The warning limits divide the area between the control limits into two types of zones (see Figure 2), each associated with a different decision. Specifically, when $EWMA_i$ lies within the safety zone delineated by the UWL and LWL, it is quite unlikely that the process has gone out of control

and thus one can stick to n_S. An EWMA score that lies between the UWL and UCL or between the LWL and LCL (i.e., warning zone) might be indicative that the process has changed. If this is indeed the case, using a larger sample size the next day rather than the small sample size might speed up detection. This statement builds on Schat et al. (2023) who explained that the more beeps per day are used, the larger the effect size of the process change becomes and thus the faster it will be detected. This can be demonstrated with a simple example. Suppose the mean changes from 0 in Phase I to 0.2 in Phase II and the standard deviation remains 1 throughout both phases. This implies that the effect size, as measured through Cohen's d, amounts to .2 at the individual beep level, whereas it equals $0.2 * \sqrt{10}$ when using 10-beep day averages and $0.2 * \sqrt{5}$ when using 5-beep day averages. When varying both sample sizes over the course of the monitoring period, the effect size will thus be somewhere in between the effect size of always using 5 beeps and always using 10 beeps, implying that changes should be detected faster than with FSS EWMA with $n = 5$.

Figure 2

Visualization of the Three VSS EWMA Decisions



Note. The green area depicts the safety zone. The yellow area depicts the warning zone. The red area depicts the out-of-control zone.

Importantly, when using different sample sizes across the days, we have to rethink the standardization procedure that led to the control limits specified in Equations 5 and 6. For FSS EWMA, using this standardization is optional and mainly serves ease of interpretation. Indeed, with FSS EWMA every incoming day average is computed from the same number of individual beeps and thus can be considered a draw from the same sampling distribution, as long as the process remains in control. In contrast, when using VSS EWMA, standardization becomes essential. Even when the process remains in control, the number of beeps which are used to compute the day averages can vary across the days, implying that the day averages

are now drawn from different sampling distributions. As illustrated in Figure 3, the two in-control sampling distributions have the same mean but different standard deviations, which can be respectively computed as $\frac{\sigma_{IB}}{\sqrt{n_S}}$ and $\frac{\sigma_{IB}}{\sqrt{n_L}}$, where σ_{IB} represents the standard deviation of the measurements at individual beeps (i.e., before computing day averages).

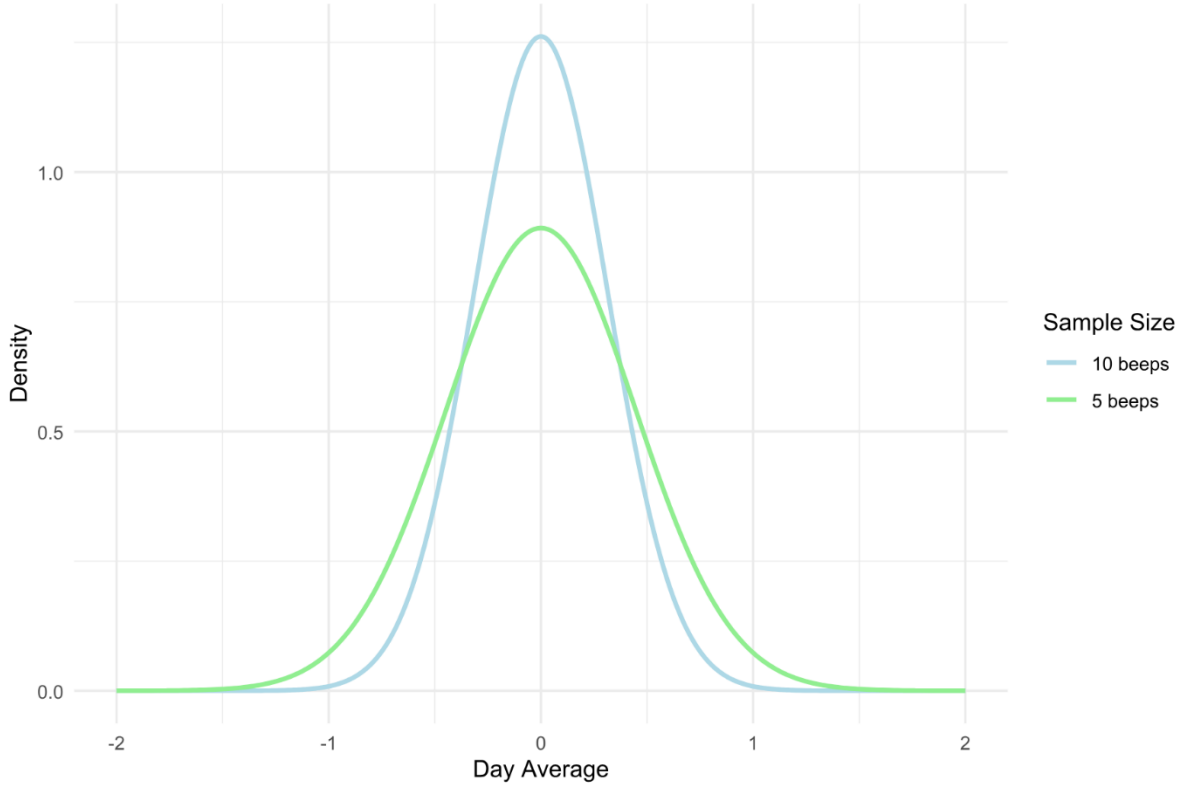
Ignoring the difference in variability when computing the EWMA scores and using limits that are either based on always using 5 beeps per day or on always using 10, would compromise VSS EWMA performance. In case the control limits are based on 10-beep day averages, the control limits will be quite narrow, and day averages from a 5-beep day will too often lead to an EWMA score that exceeds the limits, increasing the type I error rate. On the other hand, the control limits based on 5-beep day averages will be rather wide, and day averages from a 10-beep day will almost never lead to an EWMA score that exceeds the limits, implying a too low type I error rate and too low power to detect actual process changes. The differences in sampling distribution can however be accounted for when standardizing the day averages:

$$z_i = \frac{x_i - \mu_{IC}}{\sigma_{IB}/\sqrt{n_i}}, \tag{7}$$

where n_i is the sample size used on day i. This formula is thus equivalent to Equation 4, with the σ_{IC} replaced by either $\frac{\sigma_{IB}}{\sqrt{n_S}}$ or $\frac{\sigma_{IB}}{\sqrt{n_L}}$. After standardizing the day averages, the control limits in Equations 5 and 6 can again be adopted. When taking into account the variability differences as in Equation 7, VSS EWMA can be expected to have an appropriate type I error rate, while yielding greater sensitivity to changes and thus enhancing detection speed, compared to FSS EWMA with $n = 5$ and sending less beeps in total than FSS EWMA with $n = 10$.

Figure 3

Theoretical Normal Density Curves of Day Averages Computed from Sample Sizes 5 and 10.



Note. This density plot shows the distributions of day averages computed from 5 (green) and 10 (blue) beeps drawn from a standard normal distribution.

Illustration of VSS EWMA for the case of Ben

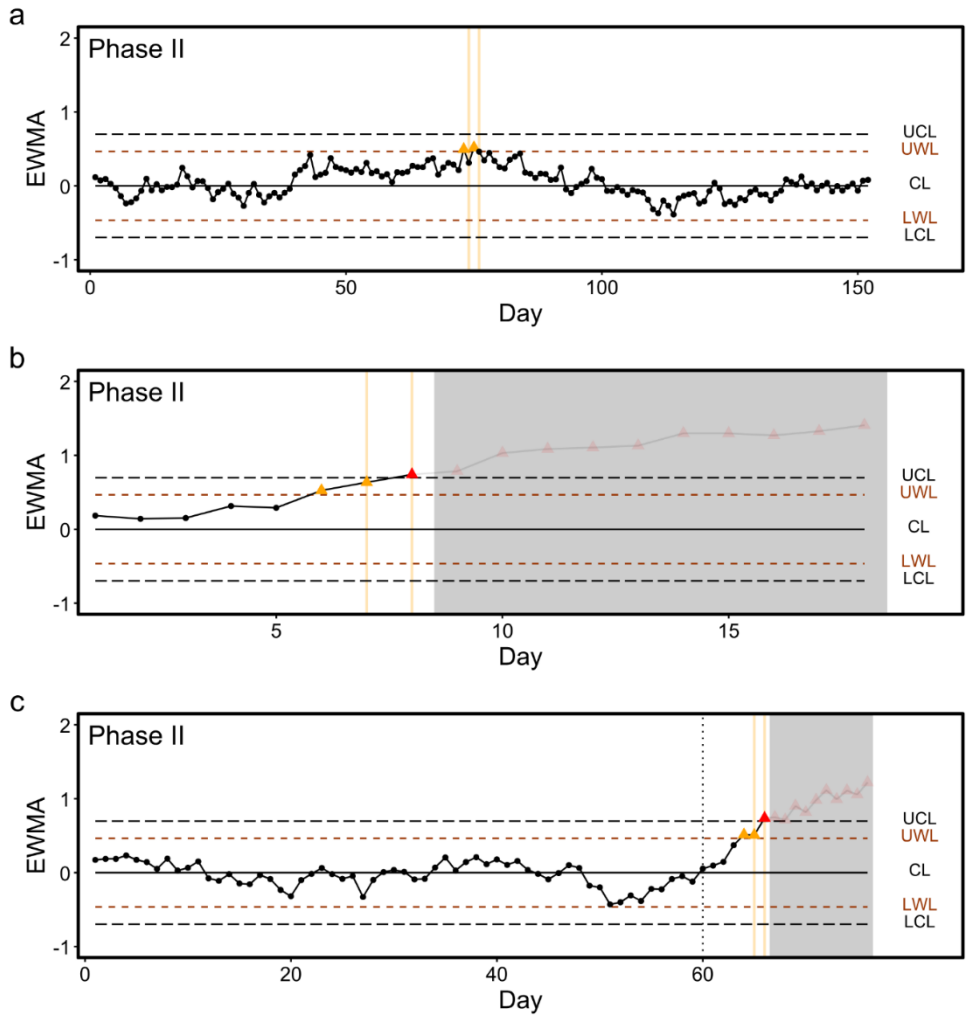
To apply VSS EWMA to the case of Ben, we have to decide where to position the warning limits. To the best of our knowledge, there are no specific recommendations for VSS EWMA approaches. In the case of the Shewhart procedure (an alternative to the EWMA procedure which essentially boils down to setting lambda equal to 1), control limits are set at three sigma (i.e., $3\sigma_{IC}$) from the in-control mean and warning limits are usually set at two sigma (i.e., $2\sigma_{IC}$; Montgomery, 2009). Building on this idea, we set the warning limit for our VSS EWMA approach to 2/3rd of the distance from the center line to the control limits.

Figure 4 depicts Ben's data when monitored using the VSS EWMA procedure (see Figure 1 for the FSS EWMA procedure). The n_S is set to five beeps per day and the n_L is set to 10. In the first scenario (no change; Figure 4a), there is no out-of-control warning, as was also the case in Figure 1a. However, to get these results, only 5 beeps were needed per day, with the exception of two days on which the sample size was increased to 10. This results in a total number of beeps of 770, compared to 1520 with the FSS EWMA procedure in Figure 1a, or approximately a 50% reduction. As can be seen in Figure 4b, the first out-of-control warning in the second scenario (immediate change) would be found on day 8 of Phase II. Ben's affect remains between the warning limits for the first 5 days of Phase II – each time eliciting a small sample size the next day – and exceeds the upper warning limit on day 6 and 7 – thus eliciting an increase in sample size on the next day – before going out of control on day 8. In Phase II Ben has thus received 5 beeps for 6 days and 10 beeps for 2 days, resulting in a total number of beeps of 50, which is considerably less than the 80 beeps needed by FSS EWMA with $n = 10$, or approximately a 40% reduction (Figure 1b).

For the third scenario (delayed change), a RL of 6 can be observed in Figure 4c. Ben's affect remains between the warning limits for the first 63 days of Phase II – each time eliciting a small sample size the next day – and exceeds the upper warning limit on day 64 and 65 – thus eliciting an increase in sample size on the next day – before going out of control on day 66. Throughout Phase II Ben has thus received 5 beeps for 64 days and 10 beeps for 2 days, resulting in a total number of beeps of 340. So again, VSS EWMA yields the same RL as FSS EWMA with $n = 10$, but the number of beeps dropped from 660 to 340, again approximately a 50% reduction.

Figure 4

VSS EWMA Control Chart for Ben's Data for Three Scenarios.



Note. VSS EWMA control chart for Ben if (a) there is no change, (b) there is a change immediately at the start of Phase II and (c) there is a delayed change after 60 days in Phase II. The black dashed horizontal lines depict the upper (UCL) and lower (LCL) control limits, the dark orange dashed horizontal lines depict the upper (UWL) and lower (LWL) warning limits. The solid horizontal line is the estimated Phase I mean $\hat{\mu}_{IC}$, also referred to as the center line (CL). Phase II consisted of 152 days. Every dot or triangle represents the EWMA score computed at a specific day. The black dots denote observations where $EWMA_i$ falls between the warning limits, prompting a small sample size the following day. The orange triangles indicate observations where $EWMA_i$ exceeds a warning limit but remains within the

control limits, prompting a larger sample size the next day. The red triangles mark the out-of-control warnings, where monitoring would stop. The orange vertical lines depict the days on which 10 beeps were sent. The shaded background marks the range of scores where normally monitoring would have stopped (after the first out-of-control warning). The dotted vertical line in chart (c) marks the introduction of the change. These charts are made with simulated data.

Simulation Study

Goal

In the example of Ben, both EWMA approaches had equal RL performance and thus equal detection speed. However, this comparison was based on one time series and thus one RL per scenario only. To draw more valid conclusions, it is standard practice in SPC research to repeatedly (e.g., 1000 times) simulate time series for each of the three scenarios, obtain RL values for each simulated time series and compute the average run length or ARL per scenario. As explained in section *FSS EWMA Procedure*, the observed ARL should be close to the targeted one (which often equals 370) when simulating scenario's where the underlying process remains in-control in Phase II (Montgomery, 2009); the ARL value in this scenario is typically labeled ARL_0. In scenarios where a change occurs in the underlying process (i.e., the process goes out-of-control), the ARL is denoted as ARL_1. The ARL_1 value represents the average number of days from the onset of a change to its detection and is ideally as low as possible.

In this section, we thus investigate the ARL_0 and ARL_1 performance of FSS EWMA with 10 beeps per day and VSS EWMA with $n_S = 5$ and $n_L = 10$ across all three change scenarios, by running an extensive simulation study. Next to ARL (i.e., detection speed), we

examine participant burden, operationalized as the average sample size (i.e., average number of beeps per day) until the first out-of-control signal. To give a comprehensive picture of the tradeoff between detection speed and participant burden, we include FSS EWMA with 5 beeps per day as well.

Moreover, for VSS EWMA we include different positions for the warning limits, as their position will also impact performance. Specifically, we expect that the closer the warning limits are to the control limits (resp. in-control mean or centerline), the more the VSS EWMA performance will resemble that of FSS EWMA with 5 beeps (resp. 10 beeps). For scenario A (no change), the ARL_0 of all procedures is expected to be close to 370, while the average sample size will strongly vary. Scenario B (immediate change) should yield a low ARL_1 for all procedures, but we expect procedures that send more beeps to do a bit better. Scenario C (delayed change) is particularly informative for evaluating the efficacy of the variable sample size approach: the average sample size should remain low while the process is still in control and increase only once a change occurs, ideally leading to a low ARL_1 and an average sample size close to 5. This study was not preregistered. The R code for the simulation study can be found at https://osf.io/3rsw7.

Design

Data Generation

All simulated datasets contain 10,000 Phase II days with 10 beeps per day. The scores at these beeps are drawn randomly from a normal distribution with a standard deviation of 1, while systematically manipulating two main data characteristics[1]:

[1] Additionally, we varied the sample size. We simulated FSS EWMA with 2, 3, 4, 5 and 10 beeps, as well as VSS EWMA with n_S (i.e., small sample size) of 2, 3, 4 or 5 beeps per day and n_L (i.e., large sample size) constant at 10 beeps per day. To keep the Results section digestible, we focus solely on the n_S= 5 results in the main text. The results for the other sample sizes can be found in the supplemental material.

(1) The size of the mean change, from 0 to 1 in steps of 0.1.

(2) The timing of the introduction of the mean change: immediately (scenario B) or after 60 days (scenario C).[2]

We simulate 10,000 replicates for each combination of size of the mean change and timing of the mean change. Hence, a total of 11*2*10,000 = 220,000 datasets are generated.

Note that we chose not to simulate a Phase I period, implying that we assume the ideal scenario in which the mean and standard deviation of the in-control distribution are already known and can be used to compute the control limits for use in Phase II. This ideal scenario allows us to isolate the effect of the different FSS and VSS procedures on detection speed and average sample size, without having to consider the potential effects of the different sample sizes used on the Phase I estimation of μ_{IC} and σ_{IC}.

Data Analysis

Each simulated dataset is analyzed five times, applying:

(1) FSS EWMA with a sample size of 5 (n_S)

(2) FSS EWMA with a sample size of 10 (n_L)

(3) VSS EWMA which alternates between 5 (n_S) and 10 (n_L), with warning limits at 2/3rd of the distance from the center line to the control limits

(4) Same as (3), but with warning limits at 1/2nd of the distance from the center line to the control limits

(5) Same as (3), but with warning limits at 1/3rd of the distance from the center line to the control limits

[2] We also inspected introducing the change after 30 days in our simulations. The results are reported in the supplemental material.

On days where only five beeps are needed (i.e. for the FSS EWMA method with 5 beeps and on 5-beep days of the VSS EWMA method), we select the first five simulated beeps of that day.

Results

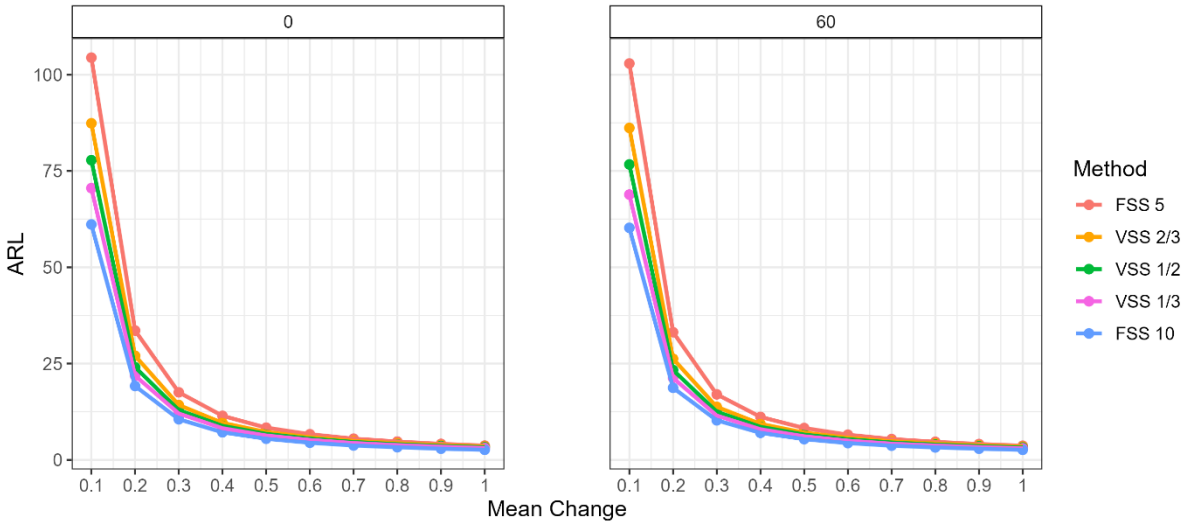
Detection Speed

We start by assessing detection speed performance through inspecting the ARL values for the different scenarios. Specifically, we investigate whether the ARL_0 values for scenario A are close to the envisioned ARL_0 of 370 and the ARL_1 values for scenario B and C decrease steeply as the size of the change increases. However, by including scenario C (delayed mean change), an additional complication shows up: premature out-of-control scores, which are a special case of false positives. These premature out-of-control scores are all the out-of-control signals that occur before the change has actually been introduced. To make sure that our comparisons are fair, we first check whether the chance of observing a false positive result is equal for the five compared methods. This is indeed the case, as the proportion of false positives was around 12.2% across all methods - for a delay of 60 days, an n_S of 5 and an n_L of 10. All datasets for which at least one of the five methods had a false positive result were deleted before analyzing the results.

All ARL_0 values were very close to the targeted value of 370 as they ranged from 361.1 to 375.6. The ARL_1 results for scenario B and C in Figure 5 also show the expected pattern: there is a clear impact of the size of the mean change, as ARL_1 values decrease with increasing mean changes. The best and worst ARL_1 performances are observed for the FSS EWMA method with 10 and 5 beeps, respectively. Consequently, the VSS EWMA curves are situated between the FSS EWMA curves, where warning limit positions closer to the control limits yield higher ARL_1 values due to less frequent sample size increases.

Figure 5

ARL_1 curves of the FSS and VSS EWMA Procedure for n_S of 5, n_L of 10 for No Delay and a Delay of 60 Days



Note. When the mean change is delayed with 60 days, the RL is determined as the first out-of-control signal after the introduction of the change. The ARL_1 values thus depict how many days it took on average to signal the change once it had occurred.

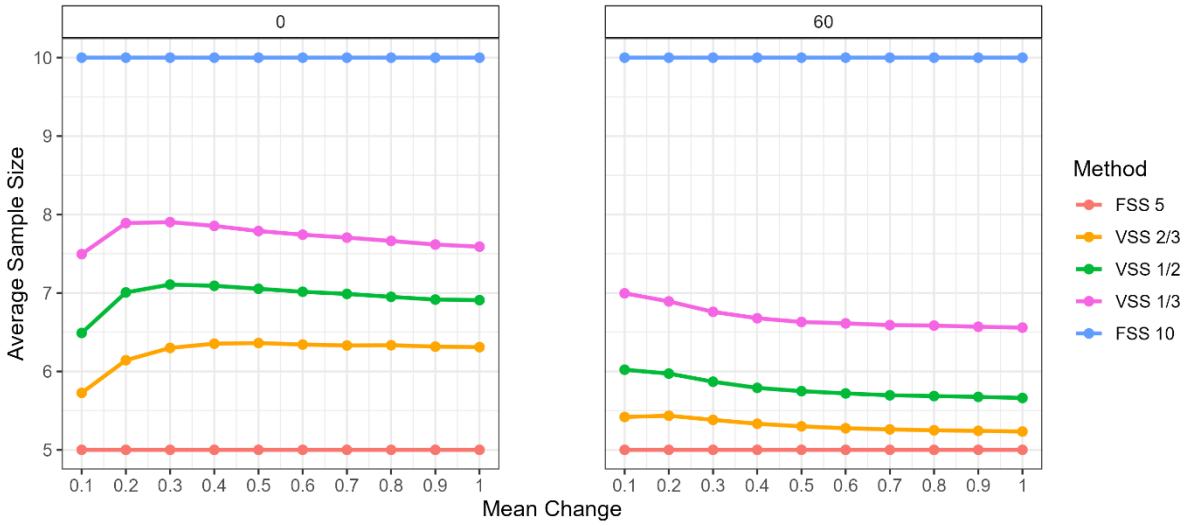
Participant Burden

To asses daily participant burden, we inspect the average sample size results in Figure 6. The FSS EWMA methods with 5 and 10 beeps per day obviously have constant average sample size values equal to their respective sample sizes, which are also the minimum and maximum average sample size values observed. Again, the VSS EWMA procedure performs between these two extremes, with the relative position depending on the position of the warning limits (i.e., warning limits closer to the control limit yield lower average sample

sizes), the timing of the mean change (i.e., delay yields lower average sample sizes), and the size of the mean change (i.e., smaller mean changes result in higher average sample sizes because they are harder to detect, thus extending the monitoring period and increasing the number of days the large sample size n_L is used). While small interaction effects can also be observed (e.g., larger impact of mean change in scenario B (immediate change) than in scenario C (delayed change)), they explain only a small portion of the variance.

Figure 6

Average Sample Size Results for the FSS and VSS EWMA Procedure for n_S of 5, n_L of 10 for No Delay and a Delay of 60 Days.



Combined performance

After separately inspecting the detection speed and participant burden of the FSS and VSS EWMA procedures, it makes sense to inspect how favorable the VSS EWMA tradeoff between participant burden and detection performance actually is. Therefore, we examine the normalized ARL_1 and average sample size (in the formula displayed as ASS) results, which were computed as follows:

$$Normalized\ ARL_1 = \frac{ARL_1(X) - ARL_1(FSS\ 5)}{ARL_1(FSS\ 10) - ARL_1(FSS\ 5)} \tag{8}$$

and

$$Normalized\ Average\ Sample\ Size = \frac{ASS(X) - ASS(FSS\ 5)}{ASS(FSS\ 10) - ASS(FSS\ 5)}\ , \tag{9}$$

where X is the specific method we want to evaluate. The normalized values show us how much of the difference between the performance of FSS EWMA with 5 and 10 beeps can be overcome by using VSS EWMA. The results are shown in Figure 7.

Figure 7

Normalized ARL_1 and Average Sample Size Results for the FSS and VSS EWMA Procedure for n_S of 5, n_L of 10 for No Delay and a Delay of 60 Days.

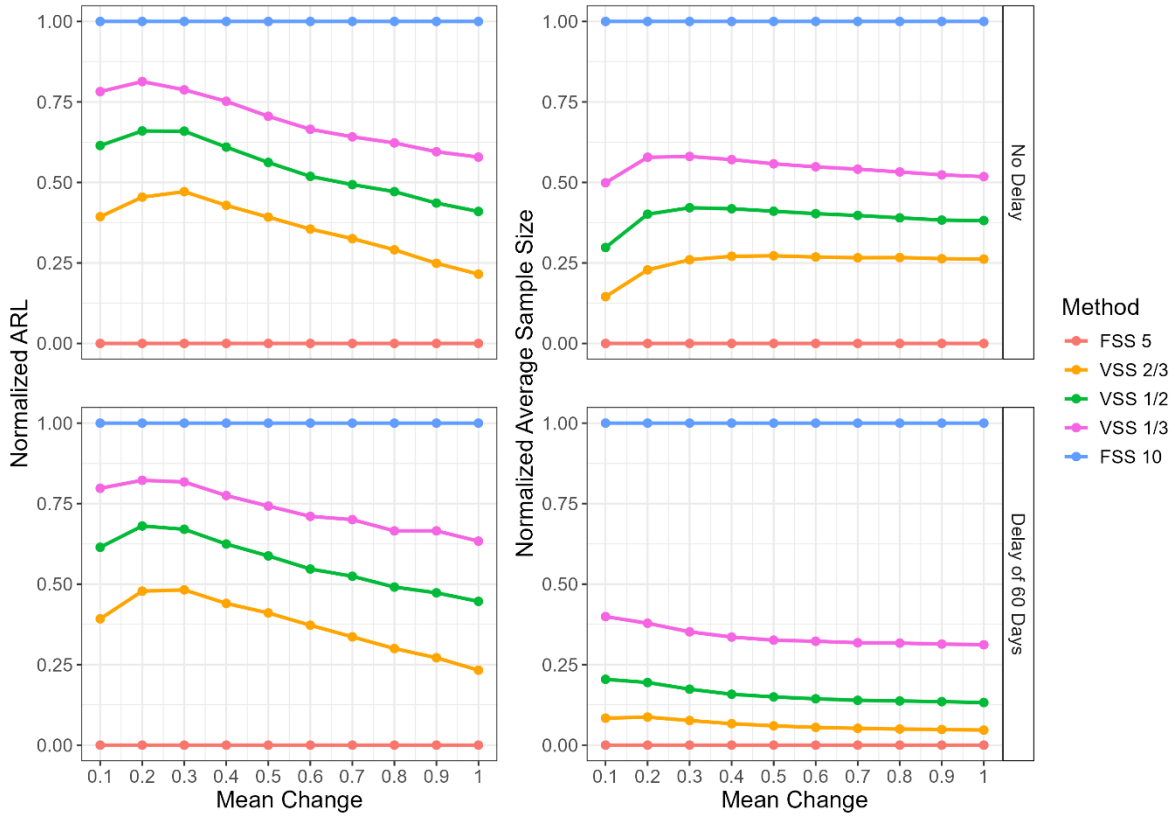


Figure 7 shows that the proportional gain in participant burden (i.e., average sample size) from using VSS EWMA instead of FSS EWMA with 10 beeps is bigger than the proportional loss in detection speed (i.e., ARL_1). For instance, for a mean change of 0.5, moving from FSS EWMA with 10 beeps to VSS EWMA with warning limits at 1/2nd of the distance between the control limits and the center line, proportionally reduces the participant burden by 85%, while the detection speed is proportionally slowed down by only 41% in case of a 60-day delay. Even without a delay, this effect persists: participant burden drops proportionally by 59% and detection speed slows proportionally by 44%.

Empirical Application

To illustrate the potential benefits of VSS EWMA, we retrospectively analyze single-subject data of Groot (2010). These data are publicly available at https://osf.io/j4fg8/. The participant in this study was a 57-year-old male with a history of major depressive disorder, who underwent a dose reduction of antidepressants after using them for 8.5 years. The study was divided into three periods: (1) a baseline period of 28 days, (2) a double-blind period of 98 days in which dose reduction took place, and (3) a follow-up period of 113 days. First indications of a relapse into depression were observed on day 127 of the study. Since dose reduction started on day 42, we defined Phase I as the first 41 days and Phase II as the remaining 198 days.

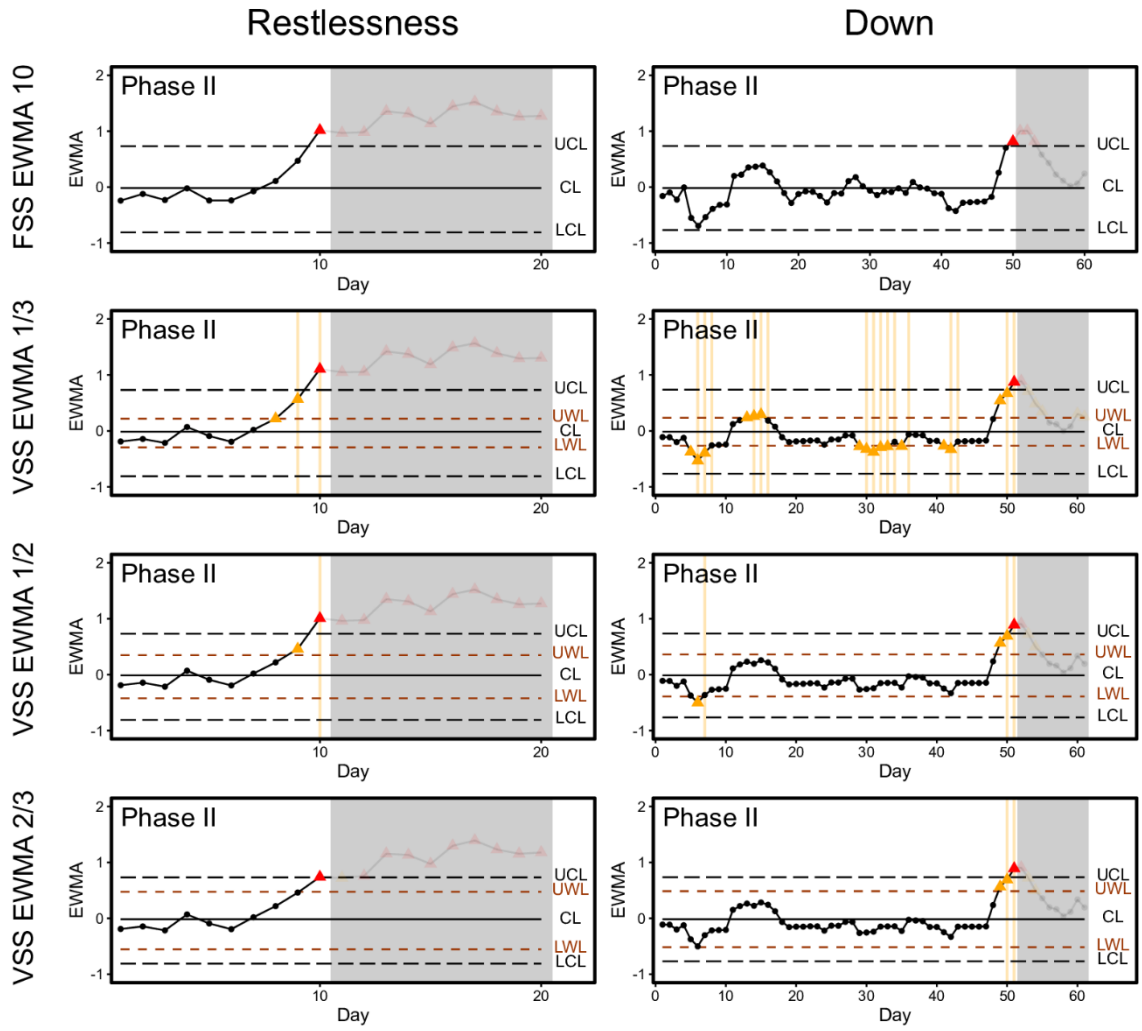
Our analysis focuses on momentary affect, which was assessed ten times a day. Specifically, based on Smit et al. (2023), we selected two items that showed distinct responses to the antidepressant dose reduction, corresponding to our scenario B (immediate change) – item 'restlessness', a prodrome – and scenario C (delayed change) – item 'down', a close proxy for core symptoms of depression. We analyzed these two items four times, to compare the performance of VSS EWMA (alternating between 5 and 10 beeps) using all three warning limits with FSS EWMA (using all 10 beeps). The warning limit positions and the values of n_S and n_L thus remain identical to the simulation study. However, the application differs from the simulation study in that the size and timing of the mean change is unknown, and in that some of the beeps were missed. Specifically, on average only 62% of the beeps were responded to in Phase II, reducing the average sample size of FSS EWMA. Similarly, on days where the VSS EWMA would send only five beeps, beeps 2, 4, 6, 8 and 10 of that day were selected, regardless of whether they were answered or not, implying that the average sample size of VSS EWMA can be lower than 5. We selected those specific beep numbers to mimic a realistic scenario, in which the beeps would be spread evenly over the

course of the day.

The analysis results shown in Figure 8 provide further evidence of VSS EWMA's efficacy in reducing participant burden, without impacting detection performance. Specifically, the 'restlessness' scores very quickly go out of control once tapering has started (i.e., an immediate change), resulting in a FSS EWMA RL of 10 and an average sample size of 8.2. VSS EWMA has the same RL, but an average sample size ranging from 4.4 to 5.3, depending on the warning limit position. Regarding the delayed change in the variable 'down' , FSS EWMA signals a first out-of-control score on day 50 of Phase II, with an average sample size of 7. VSS EWMA has a RL of 51 and an average sample size ranging from 3.7 to 4.6.

Figure 8

Application of FSS EWMA with 10 Beeps and VSS EWMA with Three Different Warning Limit Positions to 'Restlessness' and 'Down' Data from Groot (2010).



Note. The black dashed horizontal lines depict the upper (UCL) and lower (LCL) control limits, the dark orange dashed horizontal lines depict the upper (UWL) and lower (LWL) warning limits. The solid horizontal line is the estimated Phase I mean $\hat{\mu}_{IC}$, also referred to as the center line (CL). Every dot or triangle represents the EWMA score computed at a specific day. For FSS EWMA, the black dots denote observations that fall within the control limits. For VSS EWMA, the black dots denote observations where $EWMA_i$ falls between the

warning limits, prompting a small sample size the following day. The orange triangles indicate observations where $EWMA_i$ exceeds a warning limit but remains within the control limits, prompting a larger sample size the next day. The red triangles mark the out-of-control warnings, where monitoring would stop. The orange vertical lines depict the days on which 10 beeps would have been sent in the VSS EWMA conditions. For FSS EWMA, 10 beeps would have been sent every day. The shaded background marks the range of scores where normally monitoring would have stopped (after the first out-of-control warning).

Discussion

Previous research indicated that SPC procedures show promise in the early detection of mood disorder episodes, but require frequent sampling and extended ESM monitoring periods to yield accurate results (Schat et al., 2023; Schreuder et al., 2024; Snippe et al., 2023). These requirements place a substantial burden on the monitored individuals, which limits the practical use of SPC procedures in real-world settings. To address this issue, we focused on daily sample size (i.e., number of beeps per day), a factor repeatedly linked to burden (Bos et al., 2019; Hasselhorn et al., 2022; Janssens et al., 2018; Santangelo et al., 2013; Stone et al., 2003). Specifically, we investigated if participant burden can be reduced by using VSS EWMA (Reynolds & Arnold, 2001). With VSS EWMA, a smaller number of beeps per day is sent and this is only increased at the most crucial moments, without losing (too much) detection performance. We tested this approach through a simulation study and an application to empirical ESM data. We first discuss our results and the recommendations that follow. Then, we elaborate on directions for future research.

Results and Recommendations

Our findings suggest that using VSS EWMA indeed makes participating in long-term ESM studies more feasible, as it improves the balance between participant burden and detection speed. Specifically, we found that the gain in participant burden (conceptualized as average sample size) is proportionally stronger than the loss in detection speed (in terms of ARL) when moving from FSS EWMA with 10 beeps to a VSS EWMA approach that selects each day whether five or ten beeps will be sent the next day. This finding was supported by the empirical application, where equal detection speed was achieved with a lower average sample size when using VSS EWMA instead of FSS EWMA with 10 beeps.

The VSS EWMA has a number of tuning constants and the simulation results showed that these tuning constants matter, as it leads to different average sample size and ARL tradeoffs. The crucial tuning constants are the positioning of the warning limits and the specific sample sizes for n_S and n_L. Regarding the positioning of the warning limits, we found that warning limits closer to the control limits consistently yield lower average sample sizes and higher ARL than warning limits closer to the center line. The ideal warning limit positioning thus depends on the priorities of the researcher or the monitored individual. When the emphasis is on reducing participant burden, a warning limit closer to the control limits is to be preferred, whereas maintaining fast detection is best served by setting the warning limits closer to the center line. Regarding the choice of n_S and n_L, the results in the main text as well as those in the supplemental material (see https://osf.io/3rsw7) show that the value chosen for n_S directly affects the average sample size and ARL. When one prioritizes the reduction of participant burden over detection speed, using smaller n_S values is to be preferred. To facilitate optimal tailoring of VSS EWMA, we provide an R script (available at https://osf.io/3rsw7) that allows researchers to explore different settings of the tuning constants and evaluate the resulting tradeoffs between participant burden and detection speed.

We end this recommendation section by emphasizing that VSS EWMA is designed to trade-off detection speed and participant burden in real-time monitoring contexts. Our findings do not generalize to studies with a different goal. For example, when studying mood disorder dynamics, reducing the sample size might cause important nuances to be lost.

Directions for Future Research

While our findings are encouraging, they also reveal some important remaining challenges and directions for future research. First, our simulation study did not include a Phase I period, as doing so would raise questions regarding which sampling scheme should be used in Phase I for the different methods. For example, we could always use n_S in Phase I for both FSS and VSS EWMA, regardless of the sampling scheme in Phase II or use the respective n_L values, to obtain the most accurate estimates for the mean and standard deviation. In addition, differences in the number of beeps sent per day in Phase I would impact the RL and average sample size as well, complicating comparisons. Because this was the first study of VSS EWMA in ESM research, we have decided to evaluate the method under ideal conditions that may not fully reflect the reality of ESM studies. This should be taken into account when further developing this approach for practical applications.

Second, in our study we kept the large sample size n_L constant at 10 beeps, for simplicity's sake. However, there is no reason to do so and both the n_S and n_L can be varied to find the ideal balance between participant burden and detection speed. In theory, these sample sizes can be as low or high as one wishes – n_L can even exceed 10 to further speed up detection. The meta-analysis of Wrzus and Neubauer (2023) reports daily sample sizes of up to 81 scheduled beeps per day. However, meta-analyses also show that such high daily sample sizes are rarely paired with long monitoring periods (Kaurin et al., 2023; Wrzus & Neubauer, 2023). More specifically, none of the 477 studies included in the meta-analysis of

Wrzus and Neubauer (2023) combined a daily sample size of 10 beeps or more with an ESM study duration longer than 14 days. This suggests that using lower sample sizes might be recommended for the ESM-based monitoring of mood disorder onset. Supporting evidence comes from Groot (2010) and Snippe et al. (2023), who respectively reported compliance rates of 62% with 10 beeps per day over 238 days and 88% with five beeps per day over 4 months of monitoring. Based on these findings, we could conclude that a sample size of five beeps per day is feasible, although in the context of VSS EWMA the question remains if this implies that n_L should be set to five – meaning the sample size should generally be lower than five and never exceed it – or whether occasional days with more than five beeps might also be feasible, thus using five beeps as the n_S value.

Third, in line with Reynolds and Arnold (2001), we varied the daily sample size dichotomously, in that either n_S or n_L beeps were sent. However, in practice it could be useful to implement a more gradual and flexible strategy, using multiple thresholds rather than a single pair of warning limits and thus also more than two different sample sizes. Such a strategy could reduce the potential disruption caused by sudden (large) shifts in sample size and potentially decrease the average sample size further, with minimal impact on the sensitivity to changes in the monitored process.

Fourth, we have restricted ourselves to monitoring one variable with VSS EWMA. For some applications, in which there is only one variable of interest, this is sufficient. For example, restlessness is often monitored as a prodromal symptom for depression (Smit et al., 2019; Smit & Snippe, 2023). However, when multiple variables should be monitored simultaneously, this raises questions regarding how the sample size should be selected. Possible solutions are using a multivariate VSS EWMA method or monitoring a weighted sum of the variables.

Fifth, the VSS EWMA method entails that sample size – and thus burden – is increased when the ESM data indicate that the individual is feeling worse. However, burdening someone more when they are feeling worse may increase distress (Bos et al., 2019, 2020; Coenen et al., 2025; van Genugten et al., 2020). For example, Bos et al. (2020) reported that frequent confrontation with their mood caused some participants to feel more negative, worried and restless, especially when they were already feeling more depressed or (hypo)manic. It is therefore important to investigate how participants experience such changes in sample size and whether they should be informed about the adaptive sampling strategy. Moreover, future research on participant experience could examine the impact of this adaptive sampling strategy on compliance. One could argue that by reducing participant burden, VSS EWMA would also improve compliance. This should be further investigated.

Sixth, we assumed that participant burden increases linearly, with each additional beep contributing equally to the perceived burden. However, this may not reflect participants' subjective experiences. For example, an increase from 3 to 4 beeps per day may be perceived as less burdensome than an increase from 5 to 6 beeps per day. Moreover, participant burden in ESM studies might be further reduced by manipulating other factors than the daily sample size. For instance, reducing the number of items that participants have to rate per beep might also impact perceived burden (Eisele et al., 2022), as well as planned missingness designs (Rein et al., 2023) and adaptive or personalized item selection (Coenen et al., 2025; Olthof et al., 2023; Schneider et al., 2024).

Lastly, in the simulation study, we have assumed perfect compliance, meaning that all beeps were answered. However, in practice, ESM data often contain missingness (McLean et al., 2017; Roth, 1994; Vachon et al., 2019; van Genugten et al., 2020). In the empirical application these missing data were ignored under the assumption that they were missing completely at random (MCAR) – missingness is unrelated to the process under study. This

assumption may be unrealistic, as missingness could also depend on the affective state of the participant (e.g., answering less beeps when feeling down), which is referred to as missing not at random (MNAR). Such mechanisms could have a substantial impact on SPC performance. Assessing the impact of different missingness mechanisms on SPC performance is thus an important direction for future research, both for SPC procedures in general and for the proposed VSS EWMA method.

Conclusion

VSS EWMA is a promising approach to reduce participant burden in the context of ESM-based monitoring for mood disorder onset, as it decreases the average number of beeps participants receive per day with a proportionally smaller impact on detection speed.

Declarations

Funding

Fien De Pauw, Eva Ceulemans, and Francis Tuerlinckx were supported by a research grant from the Research Council of KU Leuven (C14/23/062). Evelien Schat was supported by research grant 12AA826N from the Research Foundation Flanders (FWO).

Conflicts of interest

All authors declare that they have no conflict of interest.

Ethics approval

Not applicable.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The dataset analyzed in this manuscript is publicly available at https://osf.io/j4fg8/.

Code availability

The R code of the simulation study is publicly available at https://osf.io/3rsw7.

Open Practices Statement

Materials and analysis code are available at https://osf.io/3rsw7. The data used in the empirical application are available at https://osf.io/j4fg8/. This study was not preregistered.

References

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

Bos, F. M., Snippe, E., Bruggeman, R., Wichers, M., & Van Der Krieke, L. (2019). Insights of patients and clinicians on the promise of the experience sampling method for psychiatric care. *Psychiatric Services*, *70*(11), 983–991. https://doi.org/10.1176/appi.ps.201900050

Coenen, T., Maerevoet, M., Chen, S., De Brouwer, M., Van Hoecke, S., Koster, E. H., Vanden Abeele, M. M., & Bombeke, K. (2025). Requirements and concerns of individuals remitted from depression for an early relapse detection mHealth app: Focus group study. *JMIR mHealth and uHealth*, *13*, e67141. https://doi.org/10.2196/67141

Eisele, G., Vachon, H., Lafit, G., Kuppens, P., Houben, M., Myin-Germeys, I., & Viechtbauer, W. (2022). The effects of sampling frequency and questionnaire length on perceived burden, compliance, and careless responding in experience sampling data in a student population. *Assessment*, *29*(2), 136–151. 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*, *19*(4), 352–362. https://doi.org/10.3109/09638237.2010.494188

Hasselhorn, K., Ottenstein, C., & Lischetzke, T. (2022). The effects of assessment intensity on participant burden, compliance, within-person variance, and within-person relationships in ambulatory assessment. *Behavior Research Methods*, *54*(4), 1541–1558. https://doi.org/10.3758/s13428-021-01683-6

Helmich, M. A., Olthof, M., Oldehinkel, A. J., Wichers, M., Bringmann, L. F., & Smit, A. C. (2021). Early warning signals and critical transitions in psychopathology: Challenges and recommendations. *Current Opinion in Psychology, Psychopathology*, *41*, 51–58. https://doi.org/10.1016/j.copsyc.2021.02.008

Helmich, M. A., Snippe, E., Kunkels, Y. K., Riese, H., Smit, A. C., & Wichers, M. (2020). *Transitions in Depression (TRANS-ID) Recovery: Study protocol for a repeated intensive longitudinal n = 1 study design to search for personalized early warning signals of critical transitions towards improvement in depression*. PsyArXiv. https://doi.org/10.31234/osf.io/fertq

Janssens, K. A. M., Bos, E. H., Rosmalen, J. G. M., Wichers, M. C., & Riese, H. (2018). A qualitative approach to guide choices for designing a diary study. *BMC Medical Research Methodology*, *18*(1), 140. https://doi.org/10.1186/s12874-018-0579-6

Kaurin, A., King, K. M., & Wright, A. G. C. (2023). Studying personality pathology with ecological momentary assessment: Harmonizing theory and method. *Personality Disorders: Theory, Research, and Treatment*, *14*(1), 62–72. https://doi.org/10.1037/per0000596

Knoth, S. (2024). *spc: Statistical process control – calculation of ARL and other control chart performance measures* (Version 0.7.1) [Computer software]. https://cran.r-project.org/web/packages/spc/index.html

Kunkels, Y. K., Smit, A. C., Minaeva, O., Snippe, E., George, S. V., van Roon, A. M., Wichers, M., & Riese, H. (2023). Risk ahead: Actigraphy-based early-warning signals of increases in depressive symptoms during antidepressant discontinuation. *Clinical Psychological Science*, *11*(5), 942–953. https://doi.org/10.1177/21677026221148101

Lucas, J. M., & Saccucci, M. S. (1990). Exponentially weighted moving average control schemes: Properties and enhancements. *Technometrics*, *32*(1), 1–12. https://doi.org/10.2307/1269835

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, Advances in the Engineering of Sensor-Based Monitoring and Management Systems for Precision Livestock Farming*, *173*, 143–156. https://doi.org/10.1016/j.biosystemseng.2017.08.006

McLean, D. C., Nakamura, J., & Csikszentmihalyi, M. (2017). Explaining system missing: Missing data and experience sampling method. *Social Psychological and Personality Science*, *8*(4), 434–441. https://doi.org/10.1177/1948550617708015

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

Montgomery, D. C. (2009). *Introduction to statistical quality control* (6. ed). Wiley.

Moriarty, A. S., Meader, N., Snell, K. I. E., Riley, R. D., Paton, L. W., Dawson, S., Hendon, J., Chew-Graham, C. A., Gilbody, S., Churchill, R., Phillips, R. S., Ali, S., & McMillan, D. (2022). Predicting relapse or recurrence of depression: Systematic review of prognostic models. *The British Journal of Psychiatry: The Journal of Mental Science*, *221*(2), 448–458. https://doi.org/10.1192/bjp.2021.218

Moskowitz, D. S., Russell, J. J., Sadikaj, G., & Sutton, R. (2009). Measuring people intensively. *Canadian Psychology*, *50*(3), 131–140.

Myin-Germeys, I., & Kuppens, P. (2021). *The open handbook of experience sampling methodology: A step-by-step guide to designing, conducting, and analyzing ESM studies*. Center for Research on Experience Sampling and Ambulatory Methods Leuven (REAL).

Olthof, M., Hasselman, F., Aas, B., Lamoth, D., Scholz, S., Daniels-Wredenhagen, N., Goldbeck, F., Weinans, E., Strunk, G., Schiepek, G., Bosman, A. M. T., & Lichtwarck-Aschoff, A. (2023). The best of both worlds? General principles of psychopathology in personalized assessment. *Journal of Psychopathology and Clinical Science*, *132*(7), 808–819. https://doi.org/10.1037/abn0000858

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

Rein, M. T., Vogelsmeier, L. V. D. E., & Bolsinova, M. (2023). *Assessing life satisfaction in everyday life: An example of scale development for ecological momentary assessment*. PsyArXiv. https://doi.org/10.31234/osf.io/47bry

Reynolds, M. R., & Arnold, J. C. (2001). EWMA control charts with variable sample sizes and variable sampling intervals. *IIE Transactions*, *33*(6), 511–530. https://doi.org/10.1080/07408170108936850

Roberts, S. W. (2000). Control chart tests based on geometric moving averages. *Technometrics*, *42*(1), 97–101. https://doi.org/10.1080/00401706.2000.10485986

Roth, P. L. (1994). Missing data: A conceptual review for applied psychologists. *Personnel Psychology*, *47*(3), 537–560. https://doi.org/10.1111/j.1744-6570.1994.tb01736.x

Santangelo, P. S., Ebner-Priemer, U. W., & Trull, T. J. (2013). Experience sampling methods in clinical psychology. In J. S. Comer & P. C. Kendall (Eds.), *The Oxford Handbook*

of Research Strategies for Clinical Psychology (p. 0). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199793549.013.0011

Schat, E., Tuerlinckx, F., Schreuder, M. J., De Ketelaere, B., & Ceulemans, E. (2024). Forecasting the onset of depression with limited baseline data only: A comparison of a person-specific and a multilevel modeling based exponentially weighted moving average approach. *Psychological Assessment*, *36*(6–7), 379–394. https://doi.org/10.1037/pas0001314

Schat, E., Tuerlinckx, F., Smit, A. C., De Ketelaere, B., & Ceulemans, E. (2023). Detecting mean changes in experience sampling data in real time: A comparison of univariate and multivariate statistical process control methods. *Psychological Methods*, *28*(6), 1335–1357. https://doi.org/10.1037/met0000447

Schneider, S., Junghaenel, D. U., Smyth, J. M., Fred Wen, C. K., & Stone, A. A. (2024). Just-in-time adaptive ecological momentary assessment (JITA-EMA). *Behavior Research Methods*, *56*(2), 765–783. https://doi.org/10.3758/s13428-023-02083-8

Schreuder, M. J., Kuppens, P., Schat, E., de Jonge, P., Hartman, C. A., & Ceulemans, E. (2024). Warning signals for mental health problems in at-risk young adults may be informed by momentary emotions reported by the general population: A novel application of the principles of statistical process control. *Mental Health Science*, *2*(3), e72. https://doi.org/10.1002/mhs2.72

Schreuder, M. J., Schat, E., Smit, A. C., Snippe, E., & Ceulemans, E. (2024). Monitoring emotional intensity and variability to forecast depression recurrence in real time in remitted adults. *Journal of Consulting and Clinical Psychology*, *92*(8), 505–516. https://doi.org/10.1037/ccp0000871

Shewhart, W. A. (1926). Quality control charts. *The Bell System Technical Journal*, *5*(4), 593–603. https://doi.org/10.1002/j.1538-7305.1926.tb00125.x

Smit, A. C., Schat, E., & Ceulemans, E. (2023). The exponentially weighted moving average procedure for detecting changes in intensive longitudinal data in psychological research in real-time: A tutorial showcasing potential applications. *Assessment*, *30*(5), 1354–1368. https://doi.org/10.1177/10731911221086985

Smit, A. C., & Snippe, E. (2023). Real-time monitoring of increases in restlessness to assess idiographic risk of recurrence of depressive symptoms. *Psychological Medicine*, *53*(11), 5060–5069. https://doi.org/10.1017/S0033291722002069

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

Snippe, E., Smit, A. C., Kuppens, P., Burger, H., & Ceulemans, E. (2023). Recurrence of depression can be foreseen by monitoring mental states with statistical process control. *Journal of Psychopathology and Clinical Science*, *132*(2), 145–155. https://doi.org/10.1037/abn0000812

Stone, A., Broderick, J., Schwartz, J., Shiffman, S., Litcher-Kelly, L., & Calvanese, P. (2003). Intensive momentary reporting of pain with an electronic diary: Reactivity, compliance, and patient satisfaction. *Pain*, *104*(1), 343–351. https://doi.org/10.1016/S0304-3959(03)00040-X

Vachon, H., Viechtbauer, W., Rintala, A., & Myin-Germeys, I. (2019). Compliance and retention with the experience sampling method over the continuum of severe mental disorders: Meta-analysis and recommendations. *Journal of Medical Internet Research*, *21*(12), e14475. https://doi.org/10.2196/14475

van Genugten, C. R., Schuurmans, J., Lamers, F., Riese, H., Penninx, B. W., Schoevers, R. A., Riper, H. M., & Smit, J. H. (2020). Experienced burden of and adherence to

smartphone-based ecological momentary assessment in persons with affective disorders. *Journal of Clinical Medicine*, *9*(2), 322. https://doi.org/10.3390/jcm9020322

van Zoonen, K., Buntrock, C., Ebert, D. D., Smit, F., Reynolds, C. F., Beekman, A. T., & Cuijpers, P. (2014). Preventing the onset of major depressive disorder: A meta-analytic review of psychological interventions. *International Journal of Epidemiology*, *43*(2), 318–329. https://doi.org/10.1093/ije/dyt175

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

Wrzus, C., & Neubauer, A. B. (2023). Ecological momentary assessment: A meta-analysis on designs, samples, and compliance across research fields. *Assessment*, *30*(3), 825–846. https://doi.org/10.1177/10731911211067538