To cite this article:

Revol, J.*, Ariens, S.*, Lafit, G., Adolf, J. K., & Ceulemans, E. (in press). Episode-contingent experience-sampling designs for accurate estimates of autoregressive dynamics. *Psychological Methods*. https://doi.org/10.1037/met0000758

Episode-contingent experience-sampling designs for accurate estimates of autoregressive dynamics

Jordan Revol* †[1], Sigert Ariens *[1], Ginette Lafit[2], Janne Adolf[1], and Eva Ceulemans[1]

[1]Research Group of Quantitative Psychology and Individual Differences, KU Leuven, Leuven, Belgium

[2]Methodology of Educational Sciences Research Group, KU Leuven, Leuven, Belgium

Note: this manuscript has been accepted for publication in *Psychological Methods*.

*These authors contributed equally to this work.

†Corresponding author: Jordan Revol, Faculty of Psychology and Educational Science, Research Group of Quantitative Psychology and Individual Differences, KU Leuven, Tiensestraat 102 - box 3713, Leuven 3000, Belgium. E-mail: jordan.revol@kuleuven.be

Abstract

Affect dynamics are often studied by means of first-order autoregressive (AR) modeling applied to intensive longitudinal data. A key target in these studies is the AR parameter, which is often tied conceptually to regulatory behavior in the affective process. The data is typically gathered using experience sampling methods, which are designed to pick up on fluctuations in affective variables as they evolve over time in naturalistic settings. In this manuscript, we compare classical time-contingent sampling designs to episode-contingent sampling designs, which initiate sampling when an emotional episode has been signalled. We define emotional episodes as periods where an affective process strays relatively far away from its mean. Compared to time-contingent designs, episode-contingent designs leverage on increased affective variability, which can have beneficial implications for the precision of the ordinary least squares AR effect estimator. Using an extensive simulation study, we attempt to delineate which characteristics of an episode-contingent design are important to consider, and how these characteristics are related to estimation benefits. We then turn to an empirical illustration, showing how an episode-contingent design can be implemented in practice. We also show that various patterns we expect based on the theoretical parts of the manuscript are recovered in the data. We conclude that episode-contingent designs can have marked benefits for the precision of the AR effect estimator, and discuss a number of challenges when it comes to implementing episode-contingent designs in practice.

Keywords: Episode-contingent design; experience sampling method; autoregressive modeling; affective processes.

1 Introduction

Affective science has seen a surge of research aiming to investigate the temporal properties of affective processes at the within-subjects level, by collecting intensive longitudinal data (ILD). This research tends to find that affective variables vary around a stable mean, or emotional home-base (Chow et al., 2005; Kuppens, Oravecz, and Tuerlinckx, 2010). This variability is usually assumed to arise due to both internal and external factors, which influence the affective system, and thus act as *perturbations*, pushing the affective system away from its mean value (Bylsma et al., 2008; Hamaker, 2012). *Emotion regulation* mechanisms manifest themselves by regulating such out-of-balance affective states back towards their mean value (Gross and Jazaieri, 2014; Kuppens and Verduyn, 2017; Kuppens and Verduyn, 2015; Suls et al., 1998).

To quantify such features, autoregressive (AR) models are commonly formulated and fit to the ILD (Vanhasbroeck et al., 2021). Such a model regresses a process measured at time t on its value at the preceding time point, $t-1$. The AR parameter is of particular interest in affect research (e.g. Bosley et al., 2019). It reflects how quickly the system regulates back to its mean. AR parameters close to 1 are assumed to be indicative of inefficient emotion regulation, since the time series can linger away from its mean value for long periods of time (Kuppens, Allen, and Sheeber, 2010, Kuppens and Verduyn, 2015, De Haan-Rietdijk et al., 2016). In contrast, if the AR parameter is close to 0, the effects of perturbations die out quickly (Hamilton, 1994), indicating that emotions are regulated more efficiently (Dejonckheere and Mestdagh, 2021). The AR parameter can be estimated via an ordinary least squares (OLS) regression (Hamilton, 1994; Lütkepohl, 2005). How well it can be estimated will depend on the ILD under study.

ILD are often gathered using the experience sampling method (ESM) or ecological momentary assessment (e.g., Ebner-Priemer and Trull, 2009). Here, researchers typically rely on smartphones to prompt participants to respond to questionnaires, often multiple times a day over the course of a number of weeks. A major advantage of such designs is that they allow researchers to gather data on the affective variables as they unfold in daily life (Scollon et al., 2009). When designing an ESM study, researchers are presented with a number of decisions, of which an important choice is *when to sample*. The method by which the measurements are taken is referred to as the *sampling design*

(Dejonckheere and Erbas, 2021; Van Berkel et al., 2017). Time-contingent (TC) designs appear to be the norm for researchers interested in affective ILD, including those primarily investigating regulatory ability (e.g. Pawluk et al., 2021; Bernstein et al., 2019; Wang et al., 2021; Funkhouser et al., 2021). TC designs sample based on the time elapsed since the previous measurement.

For researchers interested in obtaining accurate estimates of the AR parameter specifically, we argue that there may be ESM sampling designs that are more suitable for this purpose. It is known from the regression literature that the precision of the slope estimator depends on the variability of the predictor measurements, more variability is typically associated with more precise estimates (e.g., Greene, 2003, p. 46). For the AR models we consider, the predictor is the lagged criterion variable, which means that precise estimates of the AR parameter could be obtained in ESM designs that capture measurements during and following emotional episodes (defined as periods where affect strays relatively far away from its mean).

In general, we term any design which has a mechanism or *instrument* that signals the onset of an emotional episode and initiates sampling as an episode-contingent (EC) design. We already provide two examples of possible instruments which have received attention in the methodological literature. Firstly, *passive sensing* has become an active area of research (Niemeijer and Kuppens, 2024; Van Berkel et al., 2017; Pejovic et al., 2016). Here, researchers attempt to tie passively monitored smartphone data to the onset of emotional episodes (DaSilva et al., 2019; Spathis et al., 2019). Examples are the use of the GPS-location, measures of physical activity (e.g. gyroscope data), or smartphone usage. Such data has been tied to the onset of manic episodes in individuals with bipolar mood disorder, for instance (Beiwinkel et al., 2016; Palmius et al., 2016; Gruenerbl et al., 2014). Secondly, designs have been proposed where the participant is granted the freedom to initiate the sampling procedure, measurements are taken when the participant reports having experienced an emotionally relevant event (e.g. Myin-Germeys et al., 2018). In this case, the instrument signaling the onset of an emotional episode is thus participant initiation. Proponents of these types of designs argue that they (a) maximise the within-subjects variance in the outcome of interest, (b) protect the researcher from missing emotionally relevant episodes, and (c) increase the signal-to-noise ratio in the data (Reichert et al., 2020; Intille et al., 2007; Dejonckheere and Mestdagh, 2021).

Nevertheless, both the theoretical and statistical arguments in favor of using such EC sampling

designs have typically been presented heuristically, and little is known about which conditions should be met for such designs to indeed yield accurate estimates of AR effects. In this manuscript, we provide concrete evidence that capturing emotional episodes in the sampling design can have beneficial implications for the accuracy of the estimates of AR effects, and attempt to delineate which design characteristics are important to consider when designing an EC ESM study.

The manuscript is structured as follows. In section 2, we introduce the AR model and its estimation by means of OLS regression. We focus on how the *estimation precision* of the AR effect estimator depends on the variability of the predictor measurements. In section 3, we contrast TC and EC sampling designs and illustrate how EC sampling can increase the estimation precision of the AR effect estimator. In section 4, we report on an extensive simulation study where we vary a number of important design characteristics and assess their implications for estimation precision. We vary the precision of the instrument (the probability that the instrument correctly identifies the presence of an emotional episode), the number of follow-up measurements (measurements taken following the signalization of an emotional episode), the intensity of the episodes (how strongly the affective system should be perturbed for an instrument to identify a measurement as indicative of an episode) and the symmetry of sampling (whether the instruments identifies episode-measurements equally well on both sides of the mean). These design characteristics are studied in conjunction with data characteristics which are of relevance for applied research. We manipulate the number of available time points and the size of the AR effect generating the data. In section 5 we report on an empirical study where we implemented an EC design. We illustrate that various patterns we expect based on the theoretical parts of the manuscript are recovered in the data. Section 6 contains discussion where we summarise the simulation results and discuss some further implications for ESM design and modeling practices.

2 The AR model and parameter estimation

2.1 The AR model

The AR(1) model assumes that the values of the affective process at each time point, y_t, can be predicted based on the previous values of the process, y_{t-1}, with $t = 1, 2, \ldots, T$. T denotes the total

number of measurements of the affective process. In equations, the AR(1) model can be formulated as:

$$y_t = \beta_0 + \beta_1 y_{t-1} + v_t. \tag{1}$$

The slope β_1 is termed the AR effect, it quantifies the linear dependence between consecutive affective measurements. The measurements are assumed to be interleaved by equal time intervals[1]. Importantly, the AR effect is assumed to be smaller than 1 in absolute value for the affective process to display *stationarity* (see Hamilton, 1994, p. 45-47), implying that the statistical moments such as the mean and covariances do not change over time. The regression errors, also termed innovations in the AR modeling literature, are assumed to be independent and identically distributed (*i.i.d.*) according to a normal distribution with mean 0 and variance σ_v^2. A stationary AR(1) model implies that the affective measurements y_t display a normal distribution overall with mean $\mu_y = \frac{\beta_0}{1-\beta_1}$ and variance $\sigma_y^2 = \frac{\sigma_v^2}{1-\beta_1^2}$.

2.2 Estimation

Once the data have been collected, the AR(1) model can be fit and its parameters can be estimated by the OLS method. The choice for OLS is motivated by the fact that least squares estimation allows us to tie features of the data, such as predictor variances and time series lengths, to estimation precision via well-known and straightforward-to-understand equations (see Section 3). OLS is therefore ideal to explain how different patterns in the data tie to differences in estimation precision from a mechanistic point of view.

To estimate the parameters using OLS, one fits the standard regression model to the set of complete criterion and predictor pairs of measurements. Predictor measurements are obtained by appropriately lagging the criterion measurements. If the data are interleaved by equal time intervals, a set of T measurements, when appropriately lagged, will yield $T - 1$ pairs (termed *effective pairs* in the following) of predictor and criterion measurements, which are used to estimate the model parameters. The -1 stems from the fact that there is no lagged predictor measurement for the first

[1] More specifically, interleaved by time intervals of 1 (de Haan-Rietdijk et al., 2017)

time point. If the data are not interleaved by equal intervals, missing values are inserted to prevent regressing criterion on predictor measurements in the distant past[2] (see Table 1 for an illustration). This results in fewer effective pairs, and hence less available information for estimating β_1. To make the distinction between the total number of time points T and the total number of *effective pairs*, we use T^* to indicate the number of effective pairs.

Time point	y_t	y_{t-1}	y_t	y_{t-1}
1	2.1	NA	2.1	NA
2	3.4	2.1	3.4	2.1
3	1.8	3.4	NA	3.4
4	0.4	1.8	0.4	NA
5	1.1	0.4	1.1	0.4
	$T^* = 4$		$T^* = 2$	

Table 1: An illustration of the practice of lagging criterion variables appropriately. On the left, the data is gathered at $T = 5$ consecutive time points. Due to lagging, a single NA is inserted, yielding $T^* = 4$ effective pairs. The effective pairs are illustrated with yellow squares. On the right, no measurement is taken at time point 3, a missing value NA is inserted. Upon lagging, only 2 effective pairs remain.

The OLS estimator (see Wooldridge, 2012) of the AR effect is given by $\hat{\beta}_1 = r_{y_t y_{t-1}} \frac{s_{y_t}}{s_{y_{t-1}}}$, where s denotes the sample standard deviation (SD), and r the sample correlation. The OLS estimator of the intercept is given by $\hat{\beta}_0 = m_{y_t} - \hat{\beta}_1 m_{y_{t-1}}$, where m denotes the sample mean. In what follows, we focus on the AR effect estimator.

2.2.1 Estimation precision

An obtained point estimate of the AR effect will likely not equal the true underlying value, β_1, but will usually be somewhat larger or smaller due to *sampling variability*. This uncertainty is reflected in the sampling distribution, which any AR parameter point estimate is drawn from. This sampling distribution is asymptotically normal with mean β_1 and variance $\sigma^2_{\hat{\beta}_1}$ (Hamilton, 1994, p.215-216). The SD of this distribution, $\sqrt{\sigma^2_{\hat{\beta}_1}}$, is called the *standard error* (SE) of the AR estimator and we can take it as a measure of how precise our estimates are. Large SE's indicate that the point estimates would vary considerably from sample to sample.

In practice, the SE is also estimated from the data. The OLS estimator for the SE of $\hat{\beta}_1$ is given

[2]Failure to respect the requirement of equal intervals is known to lead to biased parameter estimates, especially for the AR effect (see de Haan-Rietdijk et al., 2017).

by:

$$\widehat{SE}(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}_v^2}{(T^* - 1)s_{y_{t-1}}^2}}. \tag{2}$$

The estimator of the SE is a function of the estimator of the innovation variance, $\hat{\sigma}_v^2$, the number of effective pairs, T^*, and the variance of the predictor measurements, $s_{y_{t-1}}^2$.

3 Episode-contingent sampling

Emotional experience varies in intensity and content over time. Both TC and EC designs attempt to pick up on these fluctuations, but they do so in different ways. In this section we present the classical TC design, and in what respects EC designs differ. First, we illustrate how TC and EC designs yield different structures of data. Then we illustrate why EC designs can yield relatively more precise estimates of the AR effect, through the implications of the data structure of EC designs for the variance of the predictor measurements. We close this section with a discussion of some important *design characteristics* which we will manipulate in the simulation study, and how we expect these manipulations to impact estimation precision.

3.1 Data structure

We use Figure 1 to illustrate TC and EC sampling of a hypothetical affective process. In panel A, the affective process is displayed in red as it evolves over time. The ESM researcher captures the process by collecting $T = 20$ measurements, illustrated by blue circles. For a TC design, these measurements are taken consecutively, each unit increment in time yields a measurement. In panel B the resulting $T^* = 19$ effective pairs are shown in the top.

The key idea behind an EC design is that we can capitalize on the fluctuations in the affective process by informing the sampling procedure to sample during and following *emotional episodes*. An *emotional episode* can be understood as a salient affective experience which is limited in its duration but pronounced in its intensity (see Urry, 2009). We define emotional episodes statistically as periods where the affective process strays far from its mean value (see also Ferrer et al., 2012). In

Figure 1, the threshold for an episode is illustrated by the dashed red lines, corresponding to certain quantiles of the overall distribution of affective states, which is displayed as a density curve in panel A. In practice, the onset of such episodes is to be picked up by an *instrument*. This instrument could monitor for episodes in both tails of the process (i.e. the sampling is symmetric) or only a single tail. In the example above, the sampling is symmetric.

With the onset of an episode signaled, the researcher starts to sample from the process. These measurements thus consist of an *initial episode measurement*, and a number of *follow-up* measurements. In the example in Figure 1, 4 follow-up measurements are taken. As episodes are per definition extreme and therefore rather rare, the measurements are no longer interleaved by equal time intervals. In this example, the EC design thus yields $T^* = 16$ effective pairs which can be used for estimation.

3.2 Variance of predictor measurements and estimation precision

The different data structures resulting from TC versus EC designs have some important implications for how precisely the AR effect can be estimated. As illustrated in Equation 2, the estimated SE depends on the variance of the predictor measurements, the estimated error variance, and the number of effective pairs. We discuss the implications of an EC design for each ingredient in turn (see equation 2.2.1).

Under a TC design the variance of the predictor measurements, $s^2_{y_{t-1}}$, will be close to the variance of the criterion, $s^2_{y_t}$, asymptotically they are equivalent. Under an EC design however, (initial) episode measurements are taken when the process is far away from its mean. Especially these initial episode measurements feature as predictor measurements, and as such are used to predict the later follow-up episode measurements. This implies that the variance of the predictor measurements under EC sampling will be larger than that obtained under TC sampling of the same AR process, at least if the sampling is symmetric. We illustrate the above in Figure 2. Here, an equal number of measurements are obtained by TC and EC sampling of an AR(1) process. The top parts of panels A and B display scatter plots of the effective pairs under TC and EC sampling respectively. The predictor y_{t-1} is displayed on the x-axis and the criterion y_t on the y-axis. The variance of the predictor measurements is illustrated indirectly in terms of deviations from the observed mean (red

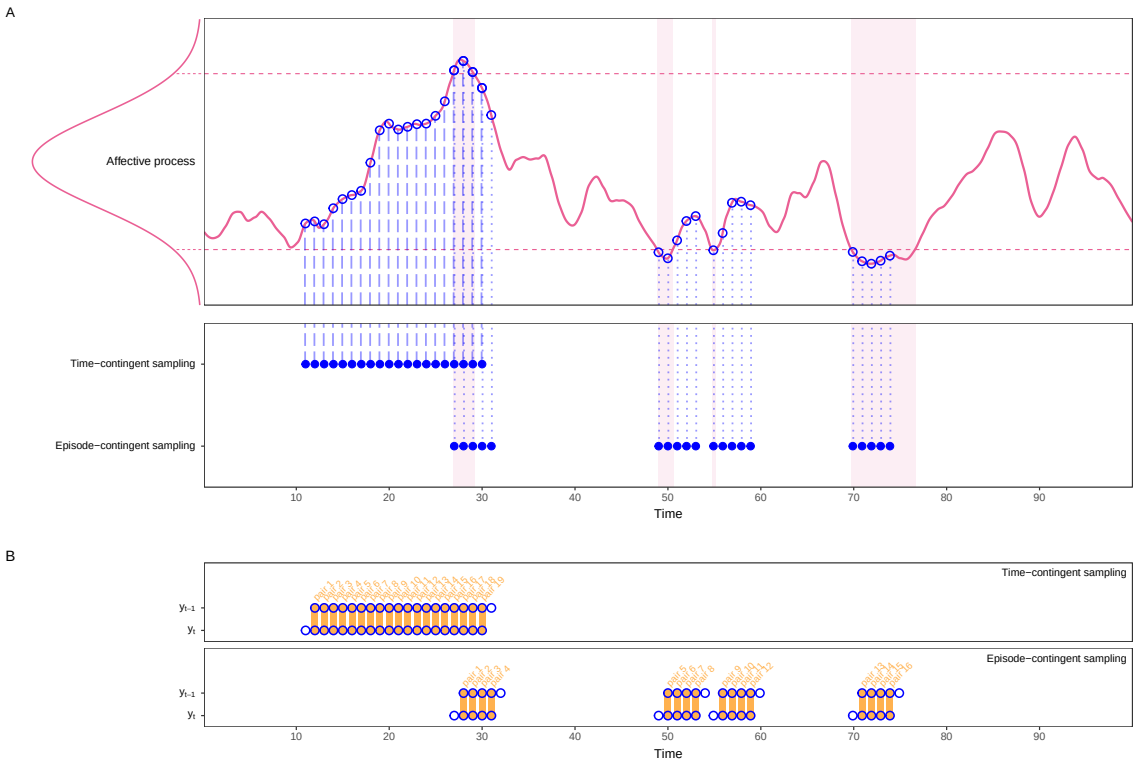

Figure 1: An illustration of TC and EC designs. In panel A, an example of an affective process is visualized, with its overall distribution on the left. Blue circles indicate when a sample is taken from the process. Under a TC design, the samples are taken consecutively. Under an EC design, samples are taken when an instrument identifies the onset of an emotional episode, the red dashed lines indicate this threshold. In panel B, the corresponding effective pairs under the TC and EC design are illustrated with yellow bars.

segments). The bottom parts of both panels show the variance directly. Clearly, EC sampling under these conditions increases the variance of the predictor measurements relative to TC sampling.

The second ingredient in the SE estimator, the estimated innovation variance, does not appear to change substantially between the two designs (yellow segments and bars in scatter and lollipop plots). This is important, because it means that the differences in SEs can be related to differences in the variance of the predictor measurements and numbers of effective pairs.

The final ingredient, the number of effective pairs, T^*, could be kept constant between TC and EC designs. If in addition both designs yield similar estimates of the innovation variance, differences in estimation precision would be driven primarily by differences in the variance of the predictor measurements. However, keeping the number of effective pairs constant would imply that

a researcher using EC sampling would have to collect more measurements T than a researcher using TC sampling[3], due to the implications of the data structure mentioned above. We argue that a more intuitive and practical comparison would be where researchers under TC and EC sampling have the same number of measurements T at their disposal. As a consequence an EC design yields fewer effective pairs than a TC design.

We focus on this comparison in the paper[4], in which case an EC design can be seen to trade effective pairs for variance of the predictor measurements, both T^* and $s^2_{y_{t-1}}$ are in the denominator of equation 2. An important question is under which conditions the benefits of capitalizing on the increased episode-related variability outweigh the costs of obtaining less effective pairs for estimation. Already from the example above, we appear to have a tentative indication that this can happen, the estimated SE under EC sampling appears smaller than that under TC sampling (see the annotations in Figure 2).

3.3 Sampling design characteristics

The above example suggests that EC designs may indeed have beneficial implications for the precision of the AR effect estimator. However, there are a number of *design characteristics* that can be varied within an EC design. To the extent these influence the variability of the predictor measurements and the number of effective pairs, they will likely impact how a potential benefit plays out.

As a first design characteristic, we can manipulate the *intensity* of the episodes. Specifically, we can vary the quantiles by which an episode is defined. The more extreme these quantiles, the further the obtained episode-samples lie away from the mean, increasing the variance of the predictor measurements and consequently AR estimator precision. In the context of the example in Figure 1, this would correspond to making the quantiles (dashed red lines) more extreme.

Secondly, we can manipulate the *number of follow-up measurements* taken after an initial episode measurement. On the one hand, including more follow-up measurements will increase the number of effective pairs implied under the design, which could benefit estimation precision. At the same time, these further follow-up measurements will likely not capitalize as strongly on the increased

[3]Using an EC design the researcher already has to sample for a longer period of time to have the same number of measurements, because episodes will only sometimes be present (see Figure 1).

[4]In Appendix 2, the simulation results are shown for the case where the number of effective pairs is kept constant between the sampling designs.

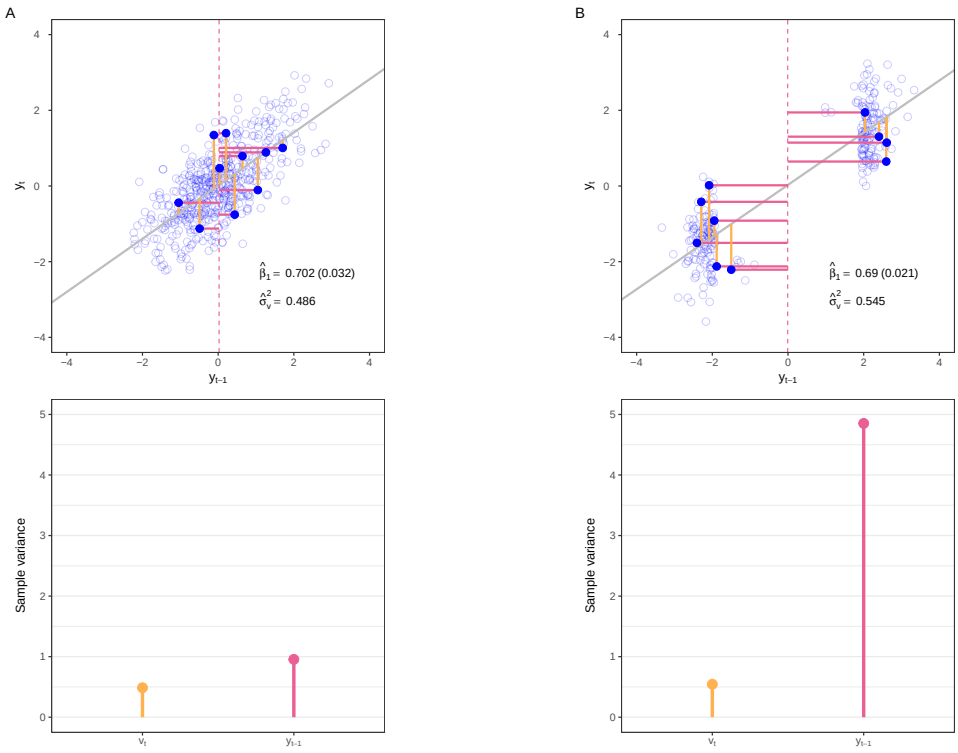

Figure 2: An illustration of the variance of the predictor measurements and estimated innovation variance under TC relative to EC designs is given. Panel A concerns TC sampling, whereas panel B concerns EC sampling. In both cases, a scatterplot is shown relating the predictor to the criterion measurements, together with the deviations of the predictor measurements from their mean (horizontal red line segments) and the regression residuals (vertical yellow line segments). Finally, the estimate of β_1 and its SE estimate are shown, together with the estimate of the residual variance. Below each scatterplot, a visualization of the residual variance and variance of the predictor measurements observed in each sampled dataset is given.

episode-related variability. How this trade-off will play out is to be investigated in the simulation study.

Thirdly, we can manipulate the *precision of the instrument*. We can vary the *precision* of the instrument by varying the ratio of correctly to correctly plus incorrectly identified initial episode measurements (see Fawcett, 2006). The more precise the instrument, the more the obtained measurements will consist of measurements tapping into episodes. We expect that the more precise the instrument, the more precise the estimates of the AR effect due to the relatively increased variance of the predictor measurements.

Finally, we can manipulate the *symmetry of sampling*. In the examples above, the instrument

identified episode-measurements lying in both tails of the process (two-sided symmetric sampling). However, it is conceivable that more episode measurements will be identified in one tail relative to the other (two-sided asymmetric sampling). In the extreme case where initial episode measurements lie only in one tail (one-sided sampling), this would likely *decrease* the variability of the predictor measurements to levels well below the variability expected under TC sampling.

4 Simulation study

We investigate the impact of the above design characteristics and two data characteristics (time series length and size of AR effect) by manipulating them in a simulation study. Specifically, we expect that characteristics which increase the variance of the predictor measurements without resulting in too few effective pairs will increase the precision of the AR effect estimator. We are primarily interested in how the design and data characteristics influence the precision of the AR effect estimator, but pay attention to biases throughout.

4.1 Simulation design and procedure

To emulate data drawn from different sampling designs, we start by simulating large T^+ data (we set $T^+ = 100000$) based on an AR(1) process. On these data we then apply the different sampling designs. Specifically, we select $T < T^+$ measurements from the simulated data using EC sampling and TC sampling[5]. The TC sampling design consists of simply using the first T measurements. For the *EC* sampling designs, we have a number of conditions (see below) which are used to identify an emotional episode. To implement them, we first classify the T^+ data into disjoint blocks of 5 consecutive measurements. This guarantees that we can allow at least one missing value in between the various blocks of EC measurements, since we allow for up to 3 follow-up episode measurements in the simulation design. For each block, we check whether the first (initial) measurement satisfies the conditions to be classified as indicative of an episode. If the conditions are met, this initial measurement and a number N_f of follow-up measurements within the block are selected, to be

[5]If the large T^+ data was not sufficiently long to obtain T measurements, as is possible for some more extreme design settings, a further set of large T^+ data was generated until T measurements were available for each sampling design.

included in the analysis. If the conditions are not met, the algorithm continues to the next block, and the conditions are again checked. With the data generated and sampled, appropriate missing values are inserted in between each set of EC measurements. All data were generated in R (R Core Team, 2023, ver 4.3.0) with the functionalities of dplyr (Wickham, François, et al., 2023, ver 1.1.2), data.table (Dowle and Srinivasan, 2023, ver 1.14.8), tidyr (Wickham, Vaughan, and Girlich, 2023, ver 1.3.0), and tibble (Müller and Wickham, 2023, ver 3.2.1).

4.2 Manipulations of sampling design characteristics

According to our previous elaborations, we manipulate the following sampling design characteristics:

(a) *Intensity of episode*: To manipulate the intensity, we vary which quantiles from the normal distribution implied under the AR(1) process are used to identify episodes. We consider two sets of quantile values, $q = \{.05, .95\}$ and $q = \{.1, .9\}$. For $q = \{.05, .95\}$, the instrument identifies relatively extreme measurements as indicative of episodes.

(b) *Number of follow ups*: We include $N_f = \{1, 2, 3\}$ follow-up episode measurements.

(c) *Precision of instrument*: We compare the case where the instrument has a perfect precision with two imperfect precision levels where $\frac{1}{3}$ or $\frac{2}{3}$ of the initial measurements are triggered by episodes, the rest being triggered by non-episodes. We also allow for a *random burst* condition where the instrument initiates sampling at random. The (average) precision under this sampling design condition naturally depends on the base rate of episodes and thus the epsiode-defining quantiles, resulting $\frac{1}{5}$ and $\frac{1}{10}$ of initial measurements being episode measurements on average. Comparing these conditions allows us to pinpoint gains in estimation precision as arising due to the increased episode-related variability.

(d) *Symmetry of sampling*: We manipulate the symmetry of sampling by including three conditions. For two-sided symmetric sampling, initial measurements come from both tails of the processes overall distribution with the same probability. For two-sided asymmetric sampling, $\frac{1}{3}$ of the initial episode measurements come from one and $\frac{2}{3}$ from the other tail (i.e., we discard a number of episode measurements from one tail), allowing us to assess the influence of some extent of asymmetry in the sampled episode-data. One-sided sampling corresponds to the case where all initial episode measurements come from the same tail.

4.3 Manipulations of data characteristics

The above manipulations of the sampling design are assessed in the context of two manipulations of the data characteristics:

(a) *Time series length*: We vary the number of time points T available to the researchers as $T = \{50, 100, 1000\}$. As mentioned above, we consider the case where researchers have an equal number of available measurements for each sampling design.

(b) *Effect sizes*: We vary the size of the AR parameter $\beta_1 = \{0, .2, .5, .8\}$, corresponding to no, low, moderate, and high AR effects for the generated time series. We do not manipulate other parameter values, fixing the intercept as $\beta_0 = 0$ and innovation variance as $\sigma_v^2 = 1$.

In total we have (4 x 3) manipulations of data characteristics interacting with (2 x 3 x 5 x 2) manipulations of design characteristics, yielding 1080 cells overall. Each cell consists of 1000 replicates, i.e. simulated datasets.

4.4 Analysis and outcome measures

For each simulated dataset, we first calculate a number of sample statistics. We calculate the sample variance of the predictor measurements and the sample autocorrelation of the process. Then we fit the AR(1) model to the simulated data relying on OLS regression as implemented in the lm() function in R. We obtain the parameter estimates and also calculate the number of effective pairs.

Across simulated data sets, we monitor bias and precision of the AR effect, intercept, and error variance estimators, and bias of the SE estimator of the AR effect. To evaluate estimator precision, we calculate the empirical variances of the estimates. To evaluate estimator bias, we rely on the notion of relative bias (see Hoogland and Boomsma, 1998). Relative bias expresses the (absolute) bias of an estimator in proportion to the true effect size. For an estimator $\hat{\beta}$, the observed relative bias $B(\hat{\beta})$ in percent can be defined as:

$$B(\hat{\beta}) = \frac{m_{\hat{\beta}} - \beta}{\beta} * 100, \tag{3}$$

where m denotes the mean over simulation replicates and β corresponds to the population level parameter. We can also calculate relative bias for the SE estimator in an analogous way, replacing

$m_{\hat{\beta}}$ with the average of the SE estimates over replications and β with the SD of the corresponding parameter point estimates over replications. The latter functions as an estimate of the population level SE (Hoogland and Boomsma, 1998).

Following the classical cut-offs for *acceptable* levels of bias in simulation studies, we consider a relative bias in the parameter estimates not exceeding 5% as acceptable, and for the SE estimates we take the conventional 10% threshold (see again Hoogland and Boomsma, 1998).

4.5 Results

For reasons of presentability, we illustrate the results in detail for the case where $T = 100$ and $\beta_1 = 0.5$, and focus on the precision of the AR estimator. We provide reference to the other results afterwards.

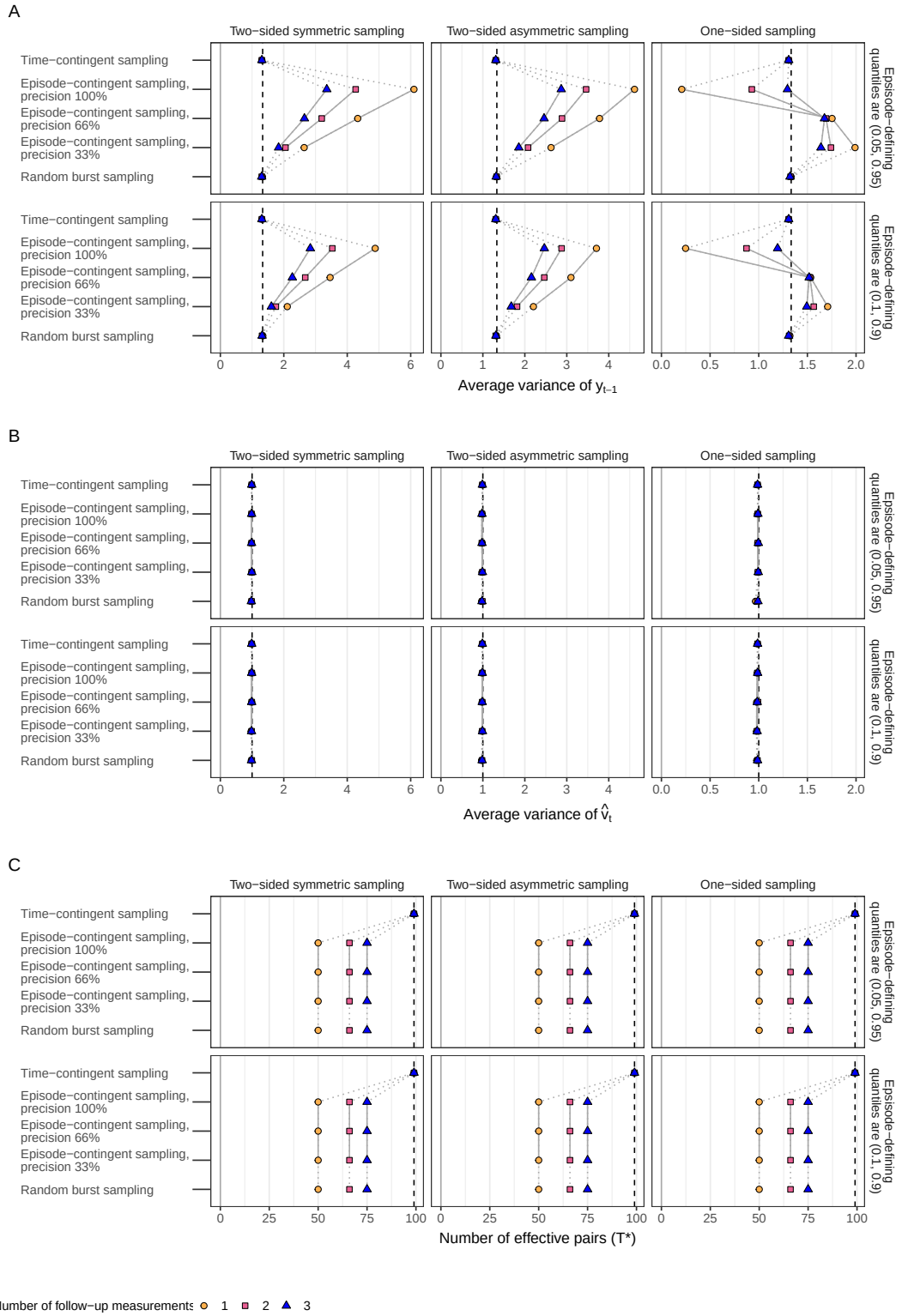

Figure 3: Average variances and numbers of effective pairs under the different sampling conditions. Panel A displays the average variance of the predictor measurements, panel B the average variance of the innovations and panel C the number of effective pairs. Dashed black lines mark the model-implied values under regular TC sampling as a reference. Per panel results for different symmetry-levels are presented across columns, and results for different episode-defining quantile values across rows. Different numbers of follow-up measurements are distinguished by point shape and color. The population-level AR effect size is 0.5. The number of time points is 100. The number of effective pairs of time points is 99 for the TC sampling condition. For the EC and random burst sampling conditions the number of effective pairs depends on the number of follow-up measurements: 50 for 1, 66 for 2, and 75 for 3 follow-up measurements.

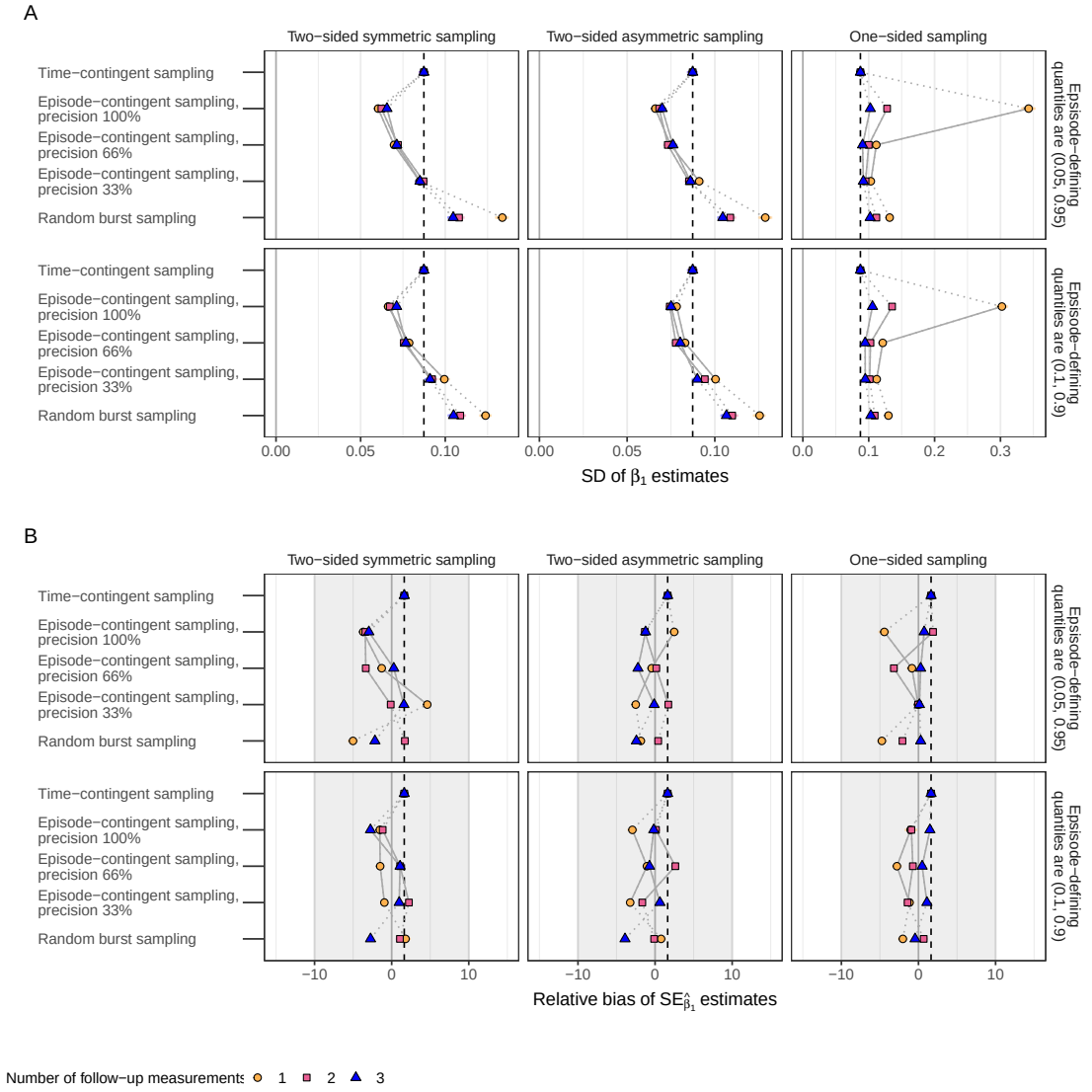

Figure 4: Accuracy of the AR point and SE estimates under the different sampling conditions. Panel A displays the SD of the AR point estimates and panel B the relative bias of the AR SE estimates. Dashed black lines mark the values under regular TC sampling as a reference. Per panel results for different symmetry-levels are presented across columns, results for different episode-defining quantile values across rows. Different numbers of follow-up measurements are distinguished by point shape and color. The population-level AR effect size is 0.5. The number of time points is 100. The number of effective pairs of time points is 99 for the TC sampling condition. For the EC and random burst sampling conditions the number of effective pairs depends on the number of follow-up measurements: 50 for 1, 66 for 2, and 75 for 3 follow-up measurements.

First, Figure 3 illustrates how the factors influencing the SE estimate depend on the design char-

acteristics. In each case, these factors are displayed over manipulations of the instrument precision including TC sampling as a reference (y-axis) and the number of follow-up episode measurements (colored shapes). The plots are further faceted by the symmetry of sampling (columns) and the intensity of the episodes (rows).

In panel A, the variance of the predictor measurements, averaged across replicates, is displayed. Compared to TC sampling, the variance of the predictor measurements can be seen to increase depending on the manipulated design characteristics in expected ways. This increase in variance is most pronounced when (1) a single follow-up is taken, (2) the instrument identifies more extreme episodes, and (3) the instrument is precise. Only for one-sided sampling condition do we see that the variance of the predictor measurements under EC sampling can be lower than that under TC sampling, particularly with a precise instrument and a single follow-up.

In panel B we see that the average estimated innovation variance appears to be the same across all sampling conditions. Panel C illustrates the differences in effective pairs. Naturally, EC sampling with a single follow-up yields the fewest effective pairs, and TC sampling the most.

Turning to the precision of $\hat{\beta}_1$, we see that the trade-off between effective pairs and variance of the predictor measurements seems to favor EC sampling in most conditions. In Figure 4 panel A, the SDs of AR parameter estimates across replicates are shown. For two-sided sampling we see that EC sampling yields in most cases more precise (less variable) estimates of the AR effect than TC sampling, but the gains depend strongly on the precision of the instrument. Particularly instruments with a precision of .33 sometimes yield negligible differences in precision compared to TC sampling, and can even yield estimates that are slightly less precise. The effect of the number of follow-up measurements seems to depend on the precision of the instrument, for precise instruments a single follow-up episode measurement yields the highest estimation precision, but differences are small. For imprecise instruments, including only a single follow-up measurement can detriment estimation precision, but again to small extents. With respect to one-sided sampling, we see that EC sampling yields less precise estimates of the AR effect compared to TC sampling, as expected from the decreased variance of the predictor measurements.

The random burst sampling condition has an equal number of effective pairs as the EC sampling conditions but does not capitalize as strongly on the increased variance of the predictor measurements

(see Figure 1). We indeed see that the random burst condition performs poorly relative to both TC and EC conditions.

With respect to the SE estimator, Figure 4 panel B shows that the SE estimator appears to display negligible levels of relative bias for all TC and EC sampling conditions. These relative biases can be considered acceptable, all well below the 10% threshold.

Appendices 1 and 2 contain additional figures. In Figure 9, we show that small sample biases (e.g. for $T = 50$) are apparent for the AR effect estimator under TC sampling, which is expected (see Maeshiro, 2000). Interestingly, we do not observe substantial small sample biases of the AR effect estimator under EC sampling. EC sampling actually seems to *mitigate* these small sample biases, especially with precise instruments. Next, Figures 10 and 11 show that the innovation variance and intercept estimator respectively do not show signs of substantial bias, regardless of our manipulations of design or data characteristics. However, both estimators do appear to show reduced precision under EC sampling relative to TC sampling (see Figure 12). Finally, Figure 13 provides the average sample statistics (variances and autocorrelations) resulting from the different sampling conditions, which we remark on in the discussion. Regarding the other settings of β_1, we report that higher AR effects appear to be estimated with higher precision, but did not find qualitative differences in the patterns of results compared to the $\beta_1 = .5$ case we present in the manuscript (see Figure 14).

In Appendix 2, we show that the results change rather gradually when the number of effective pairs T^* is kept constant between EC and TC designs. In this case, differences in the precision of the AR parameter estimator appear to be driven only by the differences in the variance of the predictor measurements, as expected, and are therefore more pronounced.

5 Empirical illustration

In this section we present an empirical example of an EC-ESM study. We have two goals in mind. Firstly we provide resources which can be used by applied researchers interested in designing an EC-ESM study. We use the m-Path software (Mestdagh et al., 2023) to implement the ESM study. We discuss the sampling design and the instrument employed, and provide the associated m-Path code. Furthermore, we provide a preprocessing report of the data (Revol et al., 2023), showing how

the data from an EC study can be processed using R to gauge data quality (e.g. to check that the sampling design was implemented as intended). These materials are included as supplementary materials.

Secondly, we aim to recover patterns related to the sample statistics and estimator precision which we expect from the prior sections of the manuscript. Our procedure is as follows. We first visualize the data to gauge the performance of our instrument. Due to the presence of non-stationarities in the data, the instrument did not perform perfectly for all participants and at all time points. To show that the results of our analyses depend on the performance of the instrument in expected ways, we perform a post-hoc categorization of episodes in terms of *instrument precision*. When estimating the AR effect, we pool information over all participants. We therefore detrend the data at the intraindividual level by subtracting time-varying means such that the measurements from all different individuals and episodes vary around 0.

In our three main analyses, we then show that: (A) We recover the theoretically expected decay within episodes in the empirical data. (B) On average, including the first follow-up measurement as a predictor leads to greater predictor variance compared to when one includes further follow-up measurements as predictors. (C) The average AR effect across individuals (which resembles the fixed AR effect in multilevel modeling) is estimated more precisely when one follow-up measurement is included relative to two or three follow-up measurements. For each analysis, we show that these results interact with instrument precision in expected ways.

We close by reflecting on our design choices and in particular on our operationalization of the instrument.

5.1 Design and participants

In this ESM study, we presented participants with self-report questionnaires about their affective state, context, experienced burden, and resilience. Here, we only report on and analyse the affect measurements. Affect was measured on a bipolar visual-analogue scale. The participants were asked "how do you feel right now" ("hoe voel je je op dit moment?") on a scale ranging from 0 to 100. The question had as anchors 0 "very negative" ("heel negatief), 50 "neutral" ("neutraal") and 100 "very positive" ("heel positief").

The ESM study is divided into 3 weeks. In the first week, participants are prompted 8 times per day (*regular beeps*); these beeps are 90 minutes apart. For all participants, the first regular beep was sent randomly in a 10-minute interval centered around 10:00. The data from this first week are used to calculate person-specific quantiles of the affect variable which are used as the *instrument* to identify episodes and send *follow-up beeps* in weeks 2 and 3. In these weeks, we also sent 8 regular beeps a day, and chose to identify episodes as regular beeps exceeding the 15th or 85th quantiles of the affect process in week 1. To limit burden, the last regular beep per day was not allowed to trigger follow-ups. We would thus expect that approximately 30% of the regular beeps (excluding the last regular beep each day) in weeks 2 and 3 were identified as initial episode measurements and lead to subsequent follow-up measurements. Whenever an episode was identified, we sent 4 follow-up beeps, separated by 10-minute intervals.

We targeted a student population of Dutch-speaking individuals, recruited through university channels and social media. There were no specific exclusion criteria, interested adults not connected with the university were also allowed to participate. In total, we collected data from 70 participants. Participants were awarded up to €60 for participating and were made aware that their compensation would depend on their compliance. Participants received the maximal compensation if they completed at least 75% of the regular beeps, with diminishing returns for lower compliance. The study was approved by the local ethics committee (SMEC, KU Leuven, file number: G-2021-2990-R2(AMD)).

5.2 Analyses

5.2.1 Data visualization and instrument performance

We visualized each participant's data by plotting the affect measurements as a function of time and calculated how often follow-up beeps were triggered, expecting that roughly 30% of the regular beeps are classified as initial episode measurements and lead to follow-up measurements. Based on these results (see appendix 3), we discovered that for many participants, the week-1 quantiles used as the instrument for signalling episodes did not appear to be representative of the quantiles of the process in weeks 2 and/or 3.

For many participants, non-stationarities appeared to be present[6]. Such non-stationarities have important implications for the performance of the instrument. For instance, if the process undergoes a shift from week 1 to weeks 2 and 3, the instrument can systematically misidentify regular beeps as initial episode measurements, leading to frequent false positives. Another possibility is that the interquantile range becomes more narrow in the second and third weeks compared to the first week, in which case episodes would rarely be triggered at all, leading to many false negatives. Such patterns can indeed be observed for many individuals in the time-series plots (Figure 18 to 23) in appendix 3.

Despite concerns about the performance of the instrument for some participants, we can leverage on the mis-identification of regular beeps as initial episode measurements to show that our results depend on the performance of the instrument in expected ways. As in our simulation design, we will use *instrument precision* as a measure of performance. Specifically, we perform a post-hoc classification of episodes as *precisely identified* versus *imprecisely identified* relative to time-varying quantiles we calculated to account for the non-stationarity.

To categorize episodes, we roll a moving window over the affect data (using only the regular beeps). The time-varying quantiles for the k'th regular beep are calculated using the kth regular beep and the 8 prior and 8 subsequent regular beeps[7]. If the instrument (the quantiles calculated in week 1) signalled that a regular beep was an initial episode measurement, we classified the episode as a *precisely* identified episode if the initial episode measurement also exceeded the time-varying quantiles at that time point, i.e. a true positive. An *imprecisely* identified episode is one where the initial episode measurement did not exceed the time-varying quantiles at that moment, i.e. a false positive. These time-varying quantiles and the associated categorization are shown in detail for a single participant in figure 5.

[6]These could arise due to changes in the process dynamics or measurement reactivity (e.g. Anvari et al., 2023)

[7]The window has a width of 17 regular beeps, and as such all measurements before the 9th regular beep and all measurements after the last $T - 9$th regular beep are not included in subsequent steps.

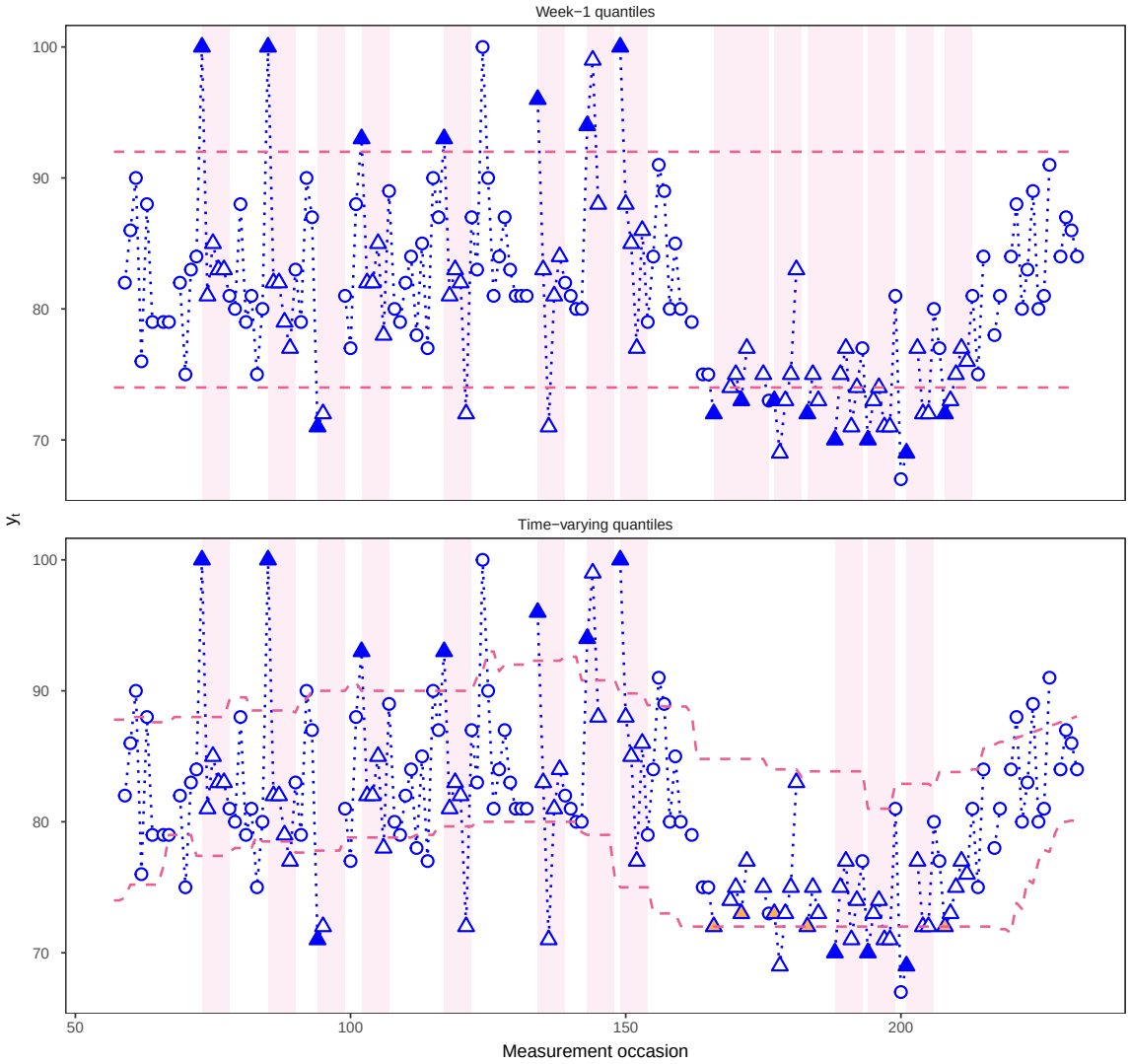

Figure 5: Visualization of the week 2 and 3 affect data for a single participant. The raw affect data is displayed as a function of the measurement occasion. Shapes distinguish regular beeps (circles) from episode measurements (triangles). We further distinguish initial episode measurements (filled triangles) from follow-up measurements (unfilled triangles). In the top panel, the week-1 quantiles used to identify episodes are shown with red horizontal dashed lines, and episodes are further illustrated using red shading. In the bottom panel, the time-varying quantiles used to categorize initial episode measurements as precisely or imprecisely identified are shown instead. Imprecisely identified episodes are shown as yellow rather than blue-filled triangles, and the shading is kept for only precisely identified episodes.

5.2.2 Detrending and episode selection

To render different episodes from different individuals comparable in mean level, we use the moving windows calculated above to detrend the data at the intraindividual level. For each participant, we subtract the means of the moving windows from the regular measurements. Specifically, we subtract the mean of the window in which the regular measurement is the middle-most (9th) measurement. If the regular beep was an initial episode measurement, the same time-varying mean was subtracted from the follow-up measurements of that episode.

We then merge the detrended data from all participants into a single, long time series and appropriately lag the data. For analysis (A), we simply use all available detrended data. For analyses (B) and (C), we retain all the episodes consisting of an initial episode measurement and 3 valid (i.e. consecutive non-missing) follow-up measurements. This selection ensures that there are an equal number of available episodes per N_f condition, which will be manipulated in these analyses. In total, this leads us to include 4364 episodes, of which 2708 were categorized as precisely identified and 1656 as imprecisely identified. This information is displayed per participant in Figure 17, appendix 3.

5.2.3 A: Episode-related decay

To graph the decay following episodes identified by the instrument, we condition on the episode being positive or negative[8] and plot the detrended affect data as a function of the episode measurement index (0 corresponding to an initial episode measurement, 4 to the final follow-up measurement in an episode). This is shown in Figure 6. The average conditional decays over participants are illustrated with thicker lines. Over individuals and episodes, this decay does indeed appear to be exponential as one would expect when specifying an AR(1) model. There is however inter-episode heterogeneity in this pattern. We also condition on the episodes being precisely or imprecisely identified. Comparing the panels in figure 6, we see that the intensity of the episodes and the subsequent decay appear to be less pronounced when only imprecisely identified episodes are included.

[8] i.e. exceeding the participants' upper or lower quantiles

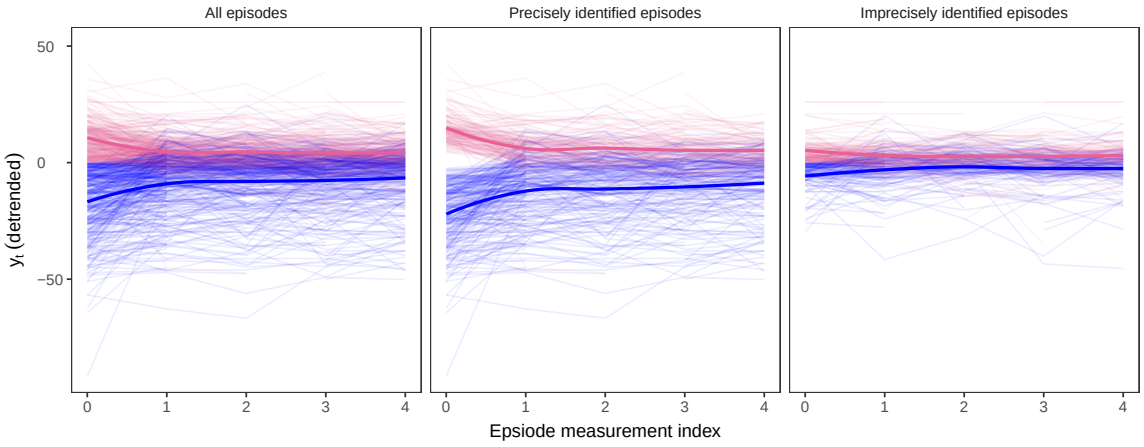

Figure 6: Demonstration of decay following episodes. The y-axis shows the detrended affect measurements. The x-axis shows the episode measurement index. Colors are used to distinguish positive and negative episodes. The average conditional decays over participants are illustrated with thicker lines. Panels are included to show how the patterns of decay depend on whether or not the episodes were precisely vs imprecisely identified, with the left-most plot showing the decay overall.

5.2.4 B: Episode-related predictor variance

As we argued in the manuscript, we expect the predictor variance to increase when using EC sampling. Moreover, the increase in variance should be most pronounced when a single follow-up is taken. To assess whether the predictor variances are on average greater in the case where only a single follow-up is included relative to two or more follow-ups, we calculate $s^2_{y_{t-1}}$ conditional on retaining $N_f = 1$, 2, or 3 follow-ups. We also compare whether the episodes were precisely or imprecisely identified, expecting to see that the differences in predictor variances are most pronounced when including precise compared to imprecisely identified episodes.

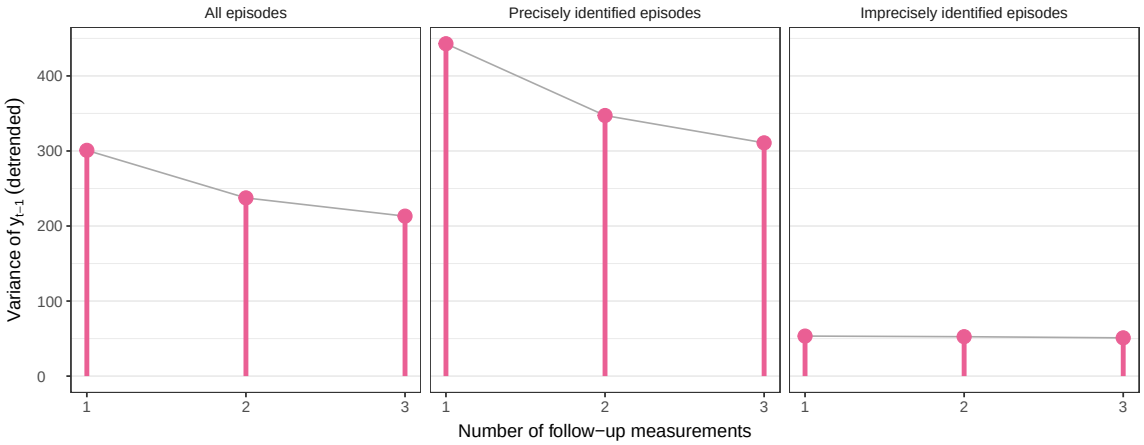

Figure 7: Predictor variances under EC sampling. The y-axis provides the variance, with the x-axis representing the N_f conditions. Panels are included to show how the predictor variances depend on whether or not the episodes were precise vs imprecisely identified, with the left-most plot showing the variances overall.

From Figure 7, it is evident that the predictor variance for an AR(1) model appears to be greatest on average for a single follow-up compared to the case where further follow-ups are included. Moreover, the predictor variance is much lower for imprecisely identified episodes compared to precisely identified episodes, and does not depend as strongly on N_f. This shows that differences in sample statistics can be tied to the performance of the instrument.

5.2.5 C: Episode-related estimator precision

To assess whether the average AR effect across individuals is estimated more precisely for the case where a single follow-up is included relative to more follow-ups (N_f manipulation), we require an estimate of sampling variability. To obtain this estimate from the empirical data while actively controlling the number of measurements T, we employ a bootstrapping procedure[9] where we draw 1000 bootstrapped data sets per N_f condition from the merged time series and estimate the average AR effect across individuals. The SD of these 1000 AR estimates yields the bootstrapped standard error (SE_{boot}).

To obtain a bootstrap data set, we sample 50 episodes with replacement from the set of all episodes, and then retain either 1, 2, or 3 follow-up measurements (N_f manipulation) per episode. We

[9]We used the boot R package (Canty and Ripley, 2022, ver 1.3-28.1) to perform the bootstrapping procedure.

then select the first 50 pairs (y_t,y_{t-1}) of these bootstrapped datasets. The first pair of each episode has NA as predictor due to the appropriate lagging. In this way, each bootstrapped dataset has $T = 50$, and conditions with fewer follow-ups have fewer effective pairs as in our simulation design. We also perform the bootstrapping procedure on the set of all precisely identified vs imprecisely identified episodes separately.

Overall, we expect that SE_{boot} will be smaller (i.e. the AR estimator will be more precise) when 1 follow-up is included relative to two or three follow-ups. Concerning the episode categorization, we expect that SE_{boot} will be smaller in the case where precisely identified episodes are analyzed relative to imprecisely identified episodes. This would allow us to tie differences in estimator precision to the performance of the instrument.

In Figure 8, we see that the average AR effect across the individuals appears to be estimated most precisely for the case where $N_f = 1$, as expected. Moreover, the SE's are smaller for all N_f conditions when precisely identified episodes are analyzed compared to imprecisely identified episodes, as expected.

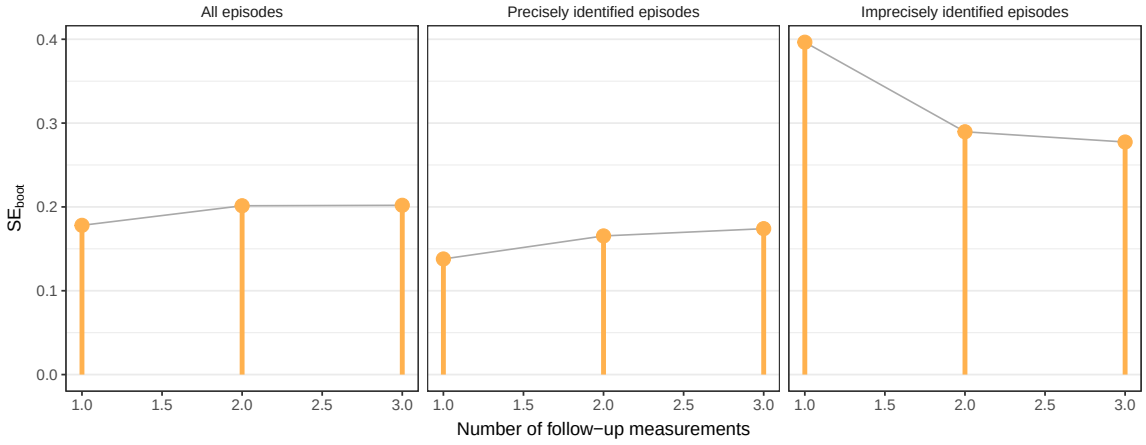

Figure 8: Bootstrapped standard errors (SE_{boot}) of the average AR effect across the individuals. The y-axis shows SE_{boot} over different N_f conditions (x-axis). Panels are included to show how SE_{boot} depends on whether or not the episodes were precisely or imprecisely identified, with the left-most plot showing SE_{boot} overall.

5.2.6 Discussion of empirical illustration

The three analyses above illustrate that patterns expected under EC sampling do appear to be recovered in the empirical data. In hindsight, however, our operationalization of the instrument might not have been the optimal choice for our data. Many participants displayed non-stationarity in their affect processes over time, rendering the week-1 quantiles relatively meaningless when it comes to identifying episodes at weeks 2 and 3. In our analysis, we controlled for this non-stationarity by allowing the quantiles to change over time, allowing us to categorize episodes as precisely vs imprecisely identified relative to the time-varying quantiles.

An important avenue for future research is that a similar technique could be applied to the sampling design itself. Rather than using data from a first subset of observations to calculate quantiles, one could use a moving window of previous observations to calculate the quantiles of the process. In this way, episodes could be identified relative to a more representative baseline state of the individual and retain their interpretations as salient affective experiences limited in duration but pronounced in intensity.

5.2.7 Limitations of empirical illustration

In this illustration, we performed an analysis on the group level, pooling information over different individuals. This approach is limited in the sense that the possibility of individual differences in the AR effects and innovation variances are ignored. Indeed, we pointed to the fact that we observed inter-episode heterogeneity in the episode-related decay, such heterogeneity might well be explained in part by interindividual differences. Secondly, the data from some participants have more weight in our analysis procedure, since the data from an individual with more detected episodes will be more likely to be drawn from during the bootstrapping procedure. Optimally, one would thus run an analysis on the intra-individual level, allowing us to control for individual differences in AR effects. However, we did not do this because the number of detected episodes, and therefore available data, was limited for multiple individuals (see Appendix 3). Fitting an AR(1) model to a time series of limited length can lead to inaccurate estimates (e.g., Marriott and Pope, 1954). Moreover, the main objective of the analysis was to illustrate how patterns expected from the theoretical sections are observed in a real data setting.

6 General discussion

The design of ESM studies of affect is a topic that has received attention in the recent methodological literature (e.g. Adolf et al., 2021; Eisele et al., 2022; Hasselhorn et al., 2023). In many cases, researchers employ AR models to quantify the regulatory dynamics in the time series of affect.

In this manuscript, we evaluated the merits of EC sampling designs when it comes to providing precise estimates of AR parameters. These designs use an instrument to inform the sampling procedure, measurements are taken during and following emotional episodes. In our simulation study, we generated data according to an AR(1) process, and proposed a sampling design which selects parts of these data where the dynamics will be most easily observed, i.e. during and following emotional episodes. Also from a more substantive perspective, researchers have emphasized that assessing affect during and following affective episodes provides relevant information for predicting outcomes of interest, such as divorce (Gottman and Levenson, 1999).

Our main result is that emotional episodes can indeed be capitalized on to yield more precise estimates of the AR effect. It is worth pointing out that this resonates with ideas reported in the theoretical literature: Shortly after the affective system has been strongly perturbed, the regulatory dynamics will be most apparent (Boker, 2002, p.415). In the manuscript we showed that EC sampling can yield more precise estimates through its implications for the variance in the obtained affective time series, specifically of the predictor measurements. Keeping the total number of measurements constant, EC designs where the increase in the variance of the predictor measurements outweighs the loss in effective pairs yield more precise estimates relative to TC designs.

This trade-off can be used to understand the results of our simulation study, where different design characteristics imply different variances of the predictor measurements and numbers of effective pairs. As a consequence, we observed the most pronounced benefits in terms of estimator precision for two-sided symmetric sampling with precise instruments, intense episodes, and the inclusion of a single follow-up episode measurement. Variations of these characteristics typically yielded fewer gains, but in many cases gains could still be observed. This paints an optimistic picture for the feasibility of EC designs, where researchers likely lack perfect instruments.

In what follows, we briefly elaborate on some implications and future directions. In particular,

we discuss the role of instruments, implications for participant burden, and implications for modeling practices. We close with some future directions which are relevant for the analysis of episode contingent data.

6.1 Instruments

As mentioned in the introduction, EC sampling designs can be implemented in different ways. An important choice in this respect is the choice of the instrument. In the manuscript, we imagined that we had such an instrument available, which allowed us to study how general properties of the instrument translate into estimation benefits. There are nevertheless a number of challenges when it comes to the choice of an instrument.

Integrating passive sensing data in the sampling design is one possible avenue to signal episodes. Passive sensing is often used to gain information on the participants' context, or sample based on the occurrence of events of interest (see for instance Van Berkel et al., 2017, Reichert et al., 2020). A challenge in this respect is how to select relevant contextual information which can be used to inform the sampling procedure (e.g. Pejovic et al., 2016). Some researchers suggest using predictable modalities, such as location, since these typically follow circadian rhythms (Mehrotra et al., 2015), deviations from which could in principle be used as markers for episodes. At the same time, an instrument based on deviations from very general patterns might not be a very precise instrument. Recently, researchers have reported promising results when it comes to identifying high-arousal (intense) psychological states using physiological data taken from wearables (Osotsi et al., 2020).

As mentioned in the introduction, another possible avenue to signal episodes is participant initiation. Here, researchers have expressed concerns that selection biases could occur (see Dejonckheere and Erbas, 2021, p.48). Participants could forget or otherwise fail to initiate sampling for a variety of reasons. Some clearly relevant episodes such as those associated with panic attacks or binge-eating can be missed simply because the participant has other things on their mind than initiating a questionnaire. The extent to which participants self-censor embarrassing events or otherwise display selective behavior in the cues they use to initiate sampling is unclear, but nevertheless has implications for the precision of the instrument and how representative the measurements are for the process of interest.

Finally, in our empirical example we used the quantiles of affect measurements during a first week of regular ESM sampling as an instrument to signal episodes in later weeks. Such an instrument uses the measurements of the affect process itself to identify relatively extreme process states. Despite the limitations of the instrument which we uncovered while analyzing this data, we could recover expected patterns related to the sample statistics and estimator precision under EC sampling. We also concluded that dealing with non-stationarities such as those induced by response shifts is an important challenge when considering this implementation of the instrument.

There is clearly important work ahead when it comes to using instruments to identify and sample emotional episodes. Nevertheless, our results show that the instrument need not be perfectly precise for gains in estimation precision to become apparent. Incorrectly sampling a portion of non-episodes is thus not of grave concern. Also with respect to asymmetry, although the largest gains are observed if the instrument identifies episodes in both tails of the process-implied overall distribution, an extent of asymmetry in sampling can still result in apparent benefits for estimation precision. It is also worth noting that creating adaptive ESM sampling designs is possible in accessible ESM software, such as the m-path application we used in our empirical illustration (Mestdagh et al., 2023). Such software offers a convenient route for implementing EC sampling designs in practice.

6.2 Implications for burden

A crucial challenge cited in many works on ESM design is how to mitigate participant burden while capturing as much relevant information for the research question (e.g. Eisele et al., 2022; Osotsi et al., 2020; Dejonckheere and Mestdagh, 2021). In one sense, EC sampling could accomplish precisely this. Under some conditions, fewer self-reports of affect are needed to yield the same estimation precision for the AR effect relative to TC designs. In another sense, EC sampling could increase participant burden. Firstly, sampling during episodes may be more burdensome than sampling in more typical states, as alluded to above. Secondly, an EC sampling design requires the researcher to monitor the process for a long enough time for a given number of episodes to manifest themselves, increasing the overall study duration. The implications of these trade-offs with respect to participant burden and data quality deserve further attention.

6.3 Modeling the process

In the manuscript, we conceived of the EC design as an interesting candidate for yielding more precise estimates of AR effects. Researchers may however be more interested in intercept or variance parameters. Although we did not find evidence that EC sampling yielded biased estimates of the intercept and residual variance parameter, they both appear to be estimated less precisely compared to TC sampling. EC sampling thus appears particularly well-tailored to research questions targeting AR parameters.

An important remark in this respect is that for TC sampling under an AR(1) model, the variances of predictor and criterion measurements are asymptotically equivalent. As such estimates of the AR effect and the lag-1 autocorrelation would coincide if a large number of measurements are available (see the equation of the AR parameter estimator in section 2.2.1). This no longer holds for the EC design, under which we demonstrated that the variance of the predictor measurements differs from the variance of the criterion measurements (see Figure 2). Crucially however, both TC and EC designs appear to yield accurate estimates of the AR effect despite the differences in the resulting ratio of the sample variances. This means that researchers using EC designs should - if they are interested in the AR effect - fit the AR(1) model to the data rather than calculate the autocorrelation descriptively (see also Figure 13). In this way, the researcher controls for the differences in sample statistics which emerge as a consequence of the EC data structure.

6.4 Future directions

The simulation studies in the manuscript were designed to manipulate features that are relevant to the OLS estimator of the AR effect. Nevertheless, other estimation related topics can be considered in future research. First, researchers may desire to fit continuous time models to episode contingent data. Next, researchers may want to explore more advanced methods of dealing with missing values than listwise deletion.

6.4.1 Continuous time methods

In this manuscript, we considered discrete-time AR models rather than their continuous-time counterparts (e.g. Driver and Voelkle, 2018). One may expect that fitting continuous time models will also benefit from EC data structures. We ran a limited version of the previous simulation study using the software package ctsem (Driver et al., 2017) to estimate the continuous time AR(1) model (see supplement C). In Figure 2, supplementary materials C, we show that with a high instrument precision, the estimates derived from fitting a continuous time model are biased upwards, a bias which becomes less severe as the episodes become less intense or more follow-ups are included. A possible reason is that continuous-time methods have a hard time dealing with data consisting of a lot of 'sudden jumps' (Loossens et al., 2021). Future research could use specifically designed simulation studies to investigate under which conditions continuous-time methods might benefit from EC sampling.

6.4.2 Missing data

A second area for future research is the use of methods which impute missing data, rather than relying on the listwise deletion of OLS. Popular software such as Mplus (Muthén and Muthén, 1998–2017) and ctsem (Driver et al., 2017) impute missing values by default through Kalman filtering (Kalman, 1960). There may be further estimation benefits associated with such procedures, though performance is likely to depend on the proportion of missing data (Asparouhov and Muthén, 2024) and the lengths of the blocks containing missing values.

7 Conclusion

We hope the results presented in this manuscript can be used by applied researchers seeking to design ESM studies aimed at quantifying regulatory dynamics in affective processes by means of AR(1) modeling. Informing the sampling procedure about the onset of emotional episodes is a challenge, but has the opportunity to yield marked benefits for estimating the AR effect.

Author Note

Intermediate results regarding this manuscript were presented at the International Meeting of the Psychometric Society (IMPS), held on 16/07/2024. Additionally, preliminary drafts of this manuscript were made accessible as preprints on PsyArXiv (https://osf.io/preprints/psyarxiv/wbu3x_v1). The current manuscript provides additional materials, analyses, and results beyond those presented in the intermediate steps.

The authors made the following contributions. Jordan Revol: Conceptualization, Methodology, Formal analysis, Software, Writing - Review & Editing. Sigert Ariens: Conceptualization, Methodology, Formal analysis, Writing - Original Draft Preparation, Writing - Review & Editing, Data Curation, Funding aquisition; Ginette Lafit: Conceptualization, Methodology, Formal analysis, Supervision, Writing - Review & Editing; Janne K. Adolf: Conceptualization, Methodology, Formal analysis, Visualization, Funding acquisition, Supervision, Writing – Review & Editing. Eva Ceulemans: Conceptualization, Methodology, Funding acquisition, Supervision, Writing – Review & Editing.

Conflicts of Interest

The authors declare that there were no conflicts of interest with respect to the authorship or the publication of this article.

Supplemental Material

The supplemental material for this paper is available at the GitLab repository https://gitlab.kuleuven.be/ppw-okpiv/researchers/u0148925/episode_contingent_study.

Funding

The research presented in this article was supported by research grants from the Fund for Scientific Research-Flanders (FWO; Project No. G0C9821N) awarded to E. Ceulemans and J. K. Adolf and

the Research Council of KU Leuven (C14/23/062; iBOF/21/090) awarded to E. Ceulemans, and (PDMT2/23/018) awarded to S. Ariens.

References

Adolf, J. K., Loossens, T., Tuerlinckx, F., & Ceulemans, E. (2021). Optimal sampling rates for reliable continuous-time first-order autoregressive and vector autoregressive modeling. *Psychological Methods*, *26*(6), 701.

Anvari, F., Efendić, E., Olsen, J., Arslan, R. C., Elson, M., & Schneider, I. K. (2023). Bias in self-reports: An initial elevation phenomenon. *Social Psychological and Personality Science*, *14*(6), 727–737.

Asparouhov, T., & Muthén, B. (2024). *Continuous time dynamic structural equation models* (Version 5, October 18). https://statmodel.com/download/CTRDSEM.pdf

Beiwinkel, T., Kindermann, S., Maier, A., Kerl, C., Moock, J., Barbian, G., Rössler, W., et al. (2016). Using smartphones to monitor bipolar disorder symptoms: A pilot study. *JMIR mental health*, *3*(1), e4560.

Bernstein, E. E., Curtiss, J. E., Wu, G. W., Barreira, P. J., & McNally, R. J. (2019). Exercise and emotion dynamics: An experience sampling study. *Emotion*, *19*(4), 637.

Boker, S. M. (2002). Consequences of continuity: The hunt for intrinsic properties within parameters of dynamics in psychological processes. *Multivariate Behavioral Research*, *37*(3), 405–422.

Bosley, H. G., Soyster, P. D., & Fisher, A. J. (2019). Affect dynamics as predictors of symptom severity and treatment response in mood and anxiety disorders: Evidence for specificity. *Journal for Person-Oriented Research*, *5*(2), 101.

Bylsma, L. M., Morris, B. H., & Rottenberg, J. (2008). A meta-analysis of emotional reactivity in major depressive disorder. *Clinical Psychology Review*, *28*(4), 676–691. https://doi.org/10.1016/j.cpr.2007.10.001

Canty, A., & Ripley, B. D. (2022). *Boot: Bootstrap r (s-plus) functions* [R package version 1.3-28.1].

Chow, S.-M., Ram, N., Boker, S. M., Fujita, F., & Clore, G. (2005). Emotion as a Thermostat: Representing Emotion Regulation Using a Damped Oscillator Model. *Emotion*, *5*(2), 208–225. https://doi.org/10.1037/1528-3542.5.2.208

DaSilva, A. W., Huckins, J. F., Wang, R., Wang, W., Wagner, D. D., & Campbell, A. T. (2019). Correlates of stress in the college environment uncovered by the application of penalized

generalized estimating equations to mobile sensing data. *JMIR mHealth and uHealth*, *7*(3), e12084.

de Haan-Rietdijk, S., Voelkle, M. C., Keijsers, L., & Hamaker, E. L. (2017). Discrete-vs. continuous-time modeling of unequally spaced experience sampling method data. *Frontiers in psychology*, *8*, 1849.

De Haan-Rietdijk, S., Gottman, J. M., Bergeman, C. S., & Hamaker, E. L. (2016). Get over it! a multilevel threshold autoregressive model for state-dependent affect regulation. *Psychometrika*, *81*, 217–241.

Dejonckheere, E., & Erbas, Y. (2021). Designing an experience sampling study. In I. Myin-Germeys & P. Kuppens (Eds.), *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; Ambulatory Methods Leuven; Leuven.

Dejonckheere, E., & Mestdagh, M. (2021). On the signal-to-noise ratio in real-life emotional time series. In *Affect dynamics* (pp. 131–152). Springer.

Dowle, M., & Srinivasan, A. (2023). *Data.table: Extension of 'data.frame'*. https://CRAN.R-project.org/package=data.table

Driver, C. C., Oud, J. H. L., & Voelkle, M. C. (2017). Continuous time structural equation modeling with R package ctsem. *Journal of Statistical Software*, *77*(5), 1–35. https://doi.org/10.18637/jss.v077.i05

Driver, C. C., & Voelkle, M. C. (2018). Hierarchical Bayesian continuous time dynamic modeling. *Psychological Methods*, *23*(4), 774–799. https://doi.org/10.1037/met0000168

Ebner-Priemer, U. W., & Trull, T. J. (2009). Ecological momentary assessment of mood disorders and mood dysregulation. *Psychological assessment*, *21*(4), 463.

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.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters*, *27*(8), 861–874.

Ferrer, E., Steele, J. S., & Hsieh, F. (2012). Analyzing the dynamics of affective dyadic interactions using patterns of intra-and interindividual variability. *Multivariate Behavioral Research*, *47*(1), 136–171.

Funkhouser, C. J., Kaiser, A. J., Alqueza, K. L., Carrillo, V. L., Hoffman, L. M., Nabb, C. B., Auerbach, R. P., & Shankman, S. A. (2021). Depression risk factors and affect dynamics: An experience sampling study. *Journal of psychiatric research*, *135*, 68–75.

Gottman, J. M., & Levenson, R. W. (1999). Rebound from Marital Conflict and Divorce Prediction. *Family Process*, *38*(3), 287–292. https://doi.org/10.1111/j.1545-5300.1999.00287.x

Greene, W. H. (2003). *Econometric analysis*. Pearson Education India.

Gross, J. J., & Jazaieri, H. (2014). Emotion, Emotion Regulation, and Psychopathology: An Affective Science Perspective. *Clinical Psychological Science*, *2*(4), 387–401. https://doi.org/10.1177/2167702614536164

Gruenerbl, A., Osmani, V., Bahle, G., Carrasco, J. C., Oehler, S., Mayora, O., Haring, C., & Lukowicz, P. (2014). Using smart phone mobility traces for the diagnosis of depressive and manic episodes in bipolar patients. *Proceedings of the 5th augmented human international conference*, 1–8.

Hamaker, E. L. (2012). Why researchers should think" within-person": A paradigmatic rationale. In M. Mehl & T. Connor (Eds.), *Handbook of research methods for studying daily life* (pp. 43–61). The Guilford Press.

Hamilton, J. D. (1994). *Time series analysis* (Vol. 2). Princeton New Jersey.

Hasselhorn, K., Ottenstein, C., & Lischetzke, T. (2023). Modeling careless responding in ambulatory assessment studies using multilevel latent class analysis: Factors influencing careless responding. *Psychological Methods*.

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods & Research*, *26*(3), 329–367.

Intille, S. S., Stone, A., Shiffman, S., Atienza, A., & Nebeling, L. (2007). Technological innovations enabling automatic, context-sensitive ecological momentary assessment. *The science of real-time data capture: Self-reports in health research*, 308–337.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, *82*(1), 35–45. https://doi.org/10.1115/1.3662552

Kuppens, P., Allen, N. B., & Sheeber, L. B. (2010). Emotional inertia and psychological maladjustment. *Psychological science*, *21*(7), 984–991.

Kuppens, P., Oravecz, Z., & Tuerlinckx, F. (2010). Feelings change: Accounting for individual differences in the temporal dynamics of affect. *Journal of personality and social psychology*, *99*(6), 1042.

Kuppens, P., & Verduyn, P. (2015). Looking at emotion regulation through the window of emotion dynamics. *Psychological Inquiry*, *26*(1), 72–79.

Kuppens, P., & Verduyn, P. (2017). Emotion dynamics. *Current Opinion in Psychology*, *17*, 22–26.

Loossens, T., Tuerlinckx, F., & Verdonck, S. (2021). A comparison of continuous and discrete time modeling of affective processes in terms of predictive accuracy. *Scientific reports*, *11*(1), 6218.

Lütkepohl, H. (2005). *New introduction to multiple time series analysis*. Springer, Berlin, Heidelberg.

Maeshiro, A. (2000). An illustration of the bias of ols for yt= λ y t-1+ u t. *The Journal of Economic Education*, *31*(1), 76–80.

Marriott, F., & Pope, J. (1954). Bias in the estimation of autocorrelations. *Biometrika*, *41*(3/4), 390–402.

Mehrotra, A., Vermeulen, J., Pejovic, V., & Musolesi, M. (2015). Ask, but don't interrupt: The case for interruptibility-aware mobile experience sampling. *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers*, 723–732.

Mestdagh, M., Verdonck, S., Piot, M., Niemeijer, K., Kilani, G., Tuerlinckx, F., Kuppens, P., & Dejonckheere, E. (2023). M-path: An easy-to-use and highly tailorable platform for ecological momentary assessment and intervention in behavioral research and clinical practice. *Frontiers in digital health*, *5*, 1182175.

Müller, K., & Wickham, H. (2023). *Tibble: Simple data frames*. https://CRAN.R-project.org/package=tibble

Muthén, L. K., & Muthén, B. O. (1998–2017). *Mplus user's guide* (Eighth). Los Angeles, CA, Muthén & Muthén.

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.

Niemeijer, K., & Kuppens, P. (2024). Emotion. In M. Mehl, M. Eid, C. Wrzus, G. Harari, & U. Ebner-Priemer (Eds.), *Mobile sensing in psychology: Methods and applications*. Guilford Press.

Osotsi, A., Oravecz, Z., Li, Q., Smyth, J., & Brick, T. R. (2020). Individualized modeling to distinguish between high and low arousal states using physiological data. *Journal of Healthcare Informatics Research*, *4*, 91–109.

Palmius, N., Tsanas, A., Saunders, K. E., Bilderbeck, A. C., Geddes, J. R., Goodwin, G. M., & De Vos, M. (2016). Detecting bipolar depression from geographic location data. *IEEE Transactions on Biomedical Engineering*, *64*(8), 1761–1771.

Pawluk, E. J., Koerner, N., Kuo, J. R., & Antony, M. M. (2021). An experience sampling investigation of emotion and worry in people with generalized anxiety disorder. *Journal of anxiety disorders*, *84*, 102478.

Pejovic, V., Lathia, N., Mascolo, C., & Musolesi, M. (2016). Mobile-based experience sampling for behaviour research. *Emotions and personality in personalized services: Models, evaluation and applications*, 141–161.

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Reichert, M., Giurgiu, M., Koch, E. D., Wieland, L. M., Lautenbach, S., Neubauer, A. B., von Haaren-Mack, B., Schilling, R., Timm, I., Notthoff, N., et al. (2020). Ambulatory assessment for physical activity research: State of the science, best practices and future directions. *Psychology of sport and exercise*, *50*, 101742.

Revol, J., Carlier, C., Lafit, G., Verhees, M., Sels, L., & Ceulemans, E. (2023). Preprocessing esm data: A step-by-step framework, tutorial website, r package, and reporting templates.

Scollon, N. C., Prieto, C.-K., & Diener, E. (2009). Experience sampling: Promises and pitfalls, strength and weaknesses. In E. Diener (Ed.), *Assessing well-being: The collected works of ed diener* (pp. 157–180). Springer Netherlands. https://doi.org/10.1007/978-90-481-2354-4_8

Spathis, D., Servia-Rodriguez, S., Farrahi, K., Mascolo, C., & Rentfrow, J. (2019). Passive mobile sensing and psychological traits for large scale mood prediction. *Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare*, 272–281.

Suls, J., Green, P., & Hillis, S. (1998). Emotional Reactivity to Everyday Problems, Affective Inertia, and Neuroticism. *Personality and Social Psychology Bulletin*, *24*(2), 127–136. https://doi.org/10.1177/0146167298242002

Urry, H. L. (2009). Using reappraisal to regulate unpleasant emotional episodes: Goals and timing matter. *Emotion*, *9*(6), 782.

Van Berkel, N., Ferreira, D., & Kostakos, V. (2017). The experience sampling method on mobile devices. *ACM Computing Surveys (CSUR)*, *50*(6), 1–40.

Vanhasbroeck, N., Ariens, S., Tuerlinckx, F., & Loossens, T. (2021). Computational models for affect dynamics. In C. E. Waugh & P. Kuppens (Eds.), *Affect dynamics* (pp. 213–260). Springer International Publishing. https://doi.org/10.1007/978-3-030-82965-0_10

Wang, X., Blain, S. D., Meng, J., Liu, Y., & Qiu, J. (2021). Variability in emotion regulation strategy use is negatively associated with depressive symptoms. *Cognition and emotion*, *35*(2), 324–340.

Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). *Dplyr: A grammar of data manipulation*. https://CRAN.R-project.org/package=dplyr

Wickham, H., Vaughan, D., & Girlich, M. (2023). *Tidyr: Tidy messy data*. https://CRAN.R-project.org/package=tidyr

Wooldridge, J. M. (2012). Basic regression analysis with time series data. In *Introductory econometrics: A modern approach* (5th ed., pp. 344–380). Cengage learning.

8 Appendices

8.1 Appendix 1: Bias and precision

Reference to Figures 9 to 14 can be found in the results section.

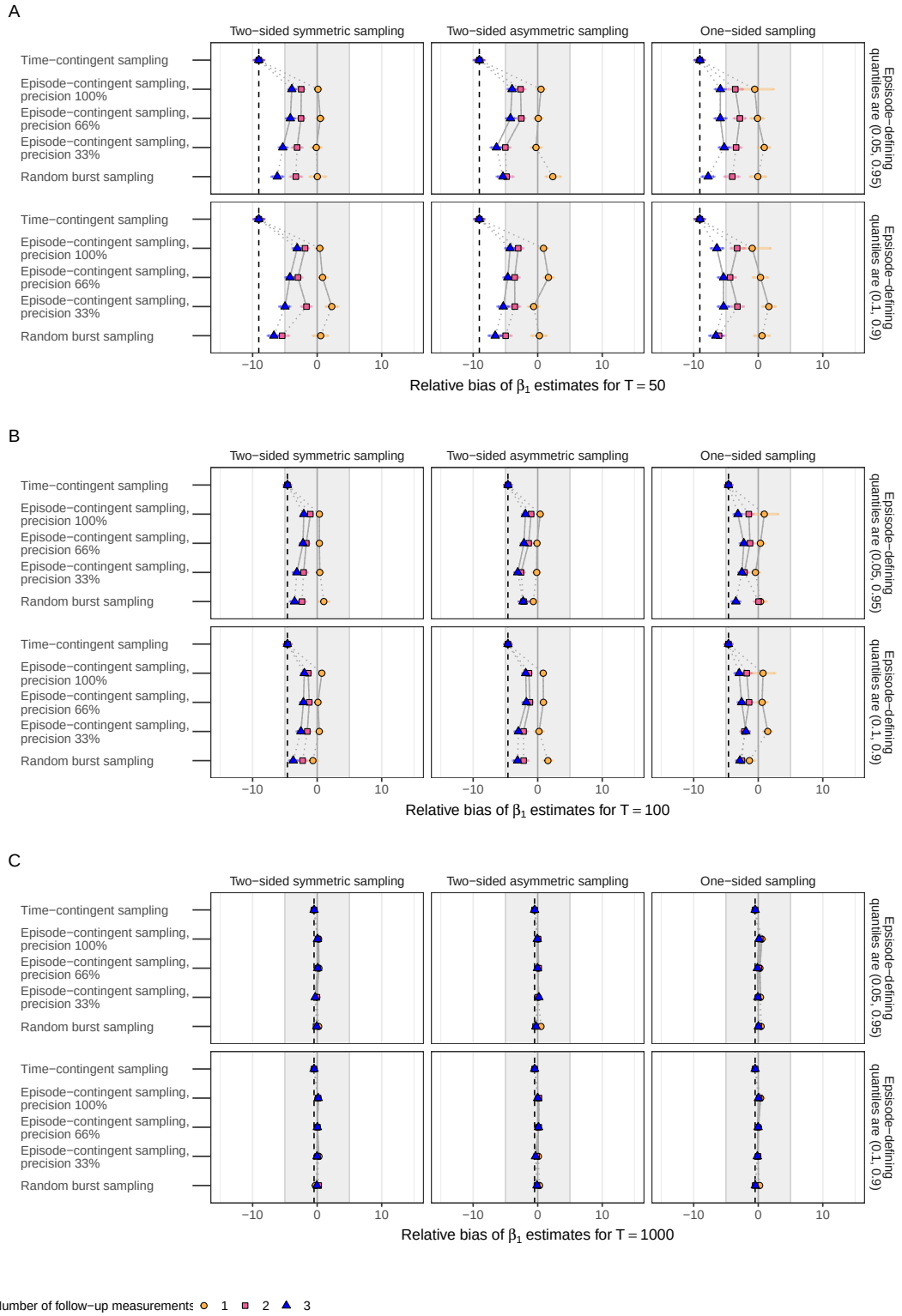

Figure 9: Relative biases of AR point estimates under the different sampling schemes for the different sample sizes. Panel A displays the relative bias for T = 50, panel B the relative bias for T = 100 and panel C the relative bias for T = 1000. Dashed black lines mark the values under regular TC sampling as a reference. Per panel results for different symmetry-levels are presented across columns, results for different episode-defining quantile values across rows. Different numbers of follow-up measurements are distinguished by point shape and color. The population-level AR effect size is 0.5.

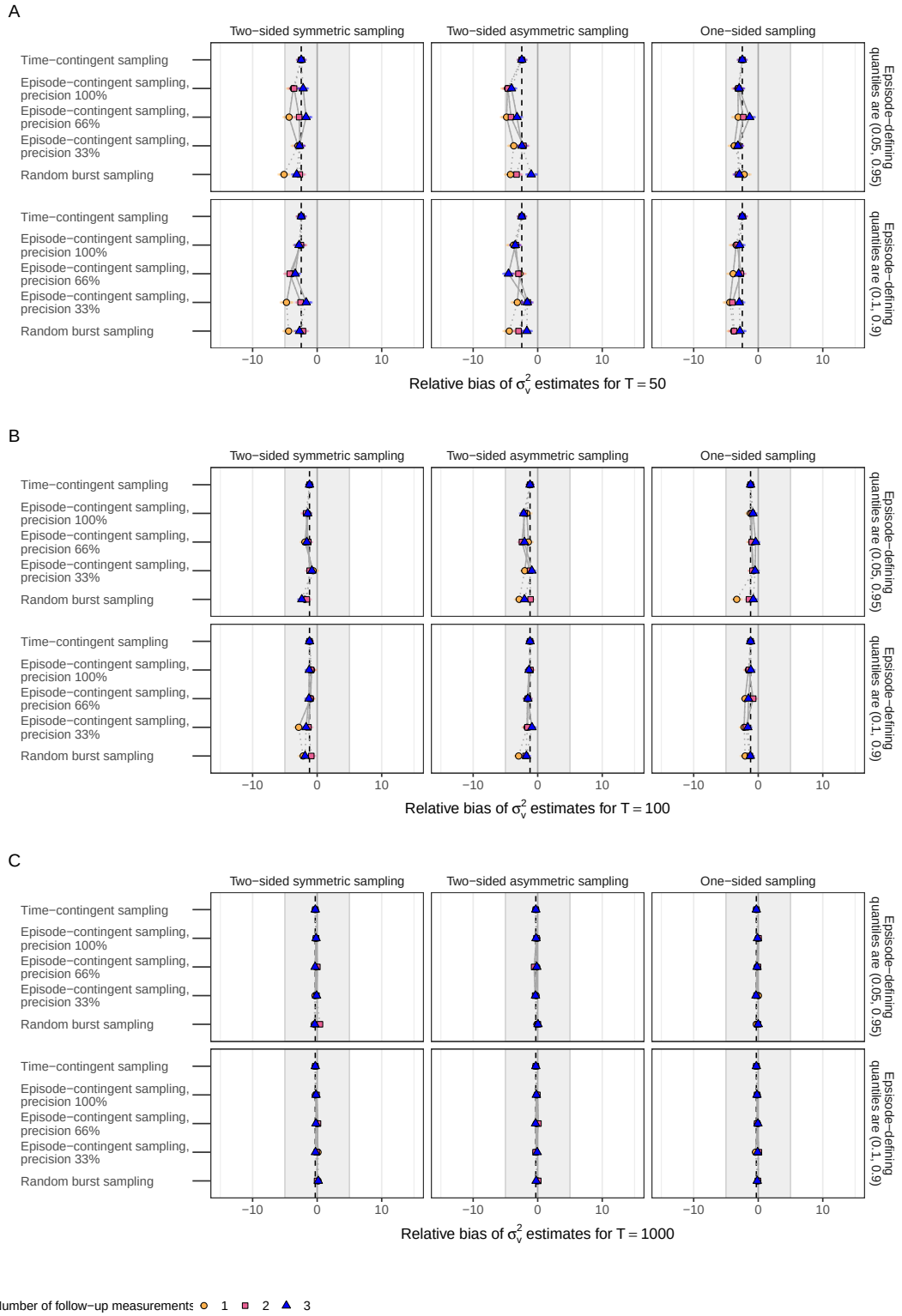

Figure 10: Relative biases of residual variance point estimates under the different sampling schemes for the different sample sizes. Panel A displays the relative bias for T = 50, panel B the relative bias for T = 100 and panel C the relative bias for T = 1000. Dashed black lines mark the values under regular TC sampling as a reference. Per panel results for different symmetry-levels are presented across columns, results for different episode-defining quantile values across rows. Different numbers of follow-up measurements are distinguished by point shape and color. The population-level AR effect size is 0.5.

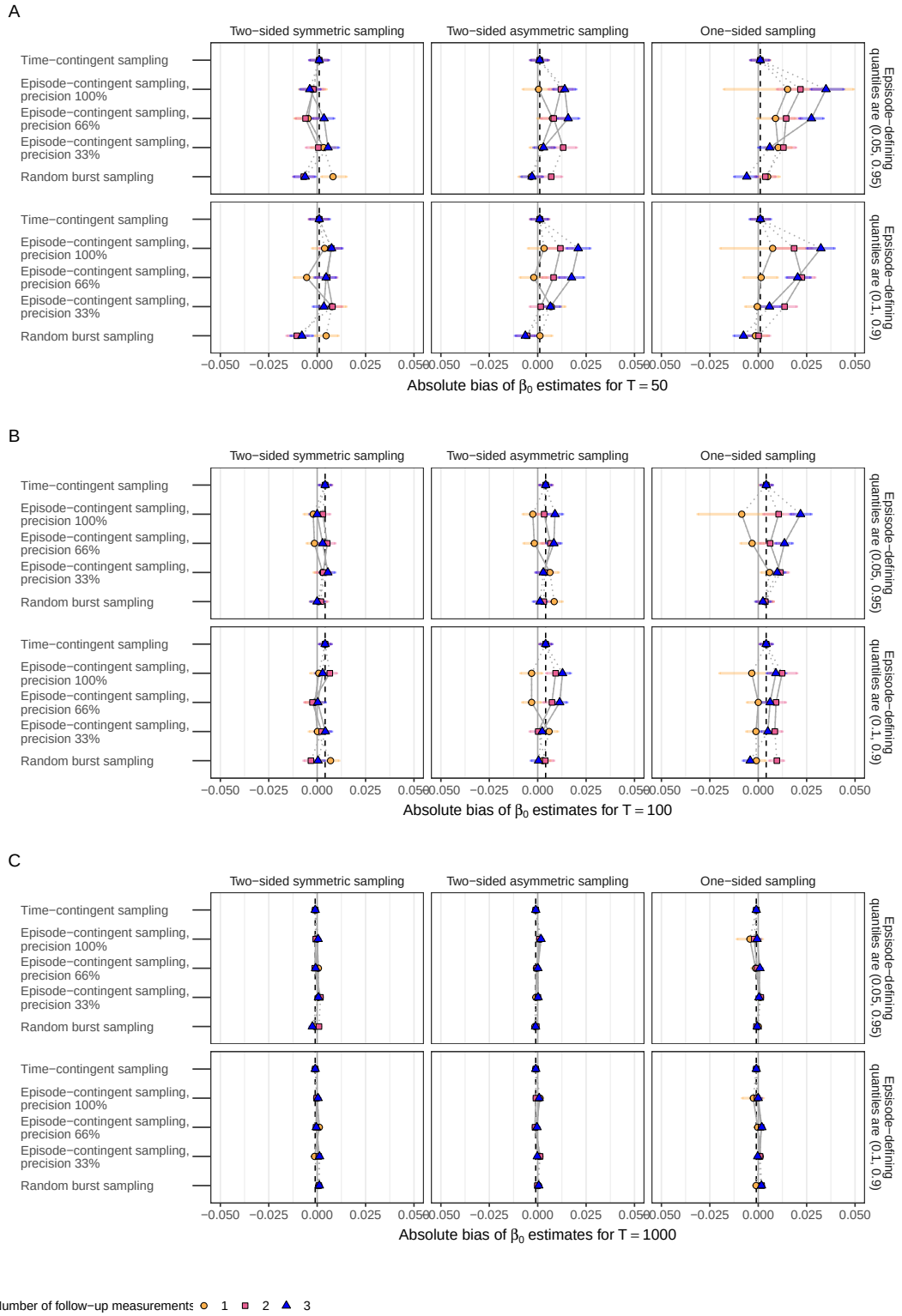

Figure 11: Absolute biases of intercept point estimates under the different sampling schemes and the different sample sizes. Panel A displays the bias for T = 50, panel B the bias for T = 100 and panel C the bias for T = 1000. Dashed black lines mark the values under regular TC sampling as a reference. Per panel results for different symmetry-levels are presented across columns, results for different episode-defining quantile values across rows. Different numbers of follow-up measurements are distinguished by point shape and color. The population-level AR effect size is 0.5.

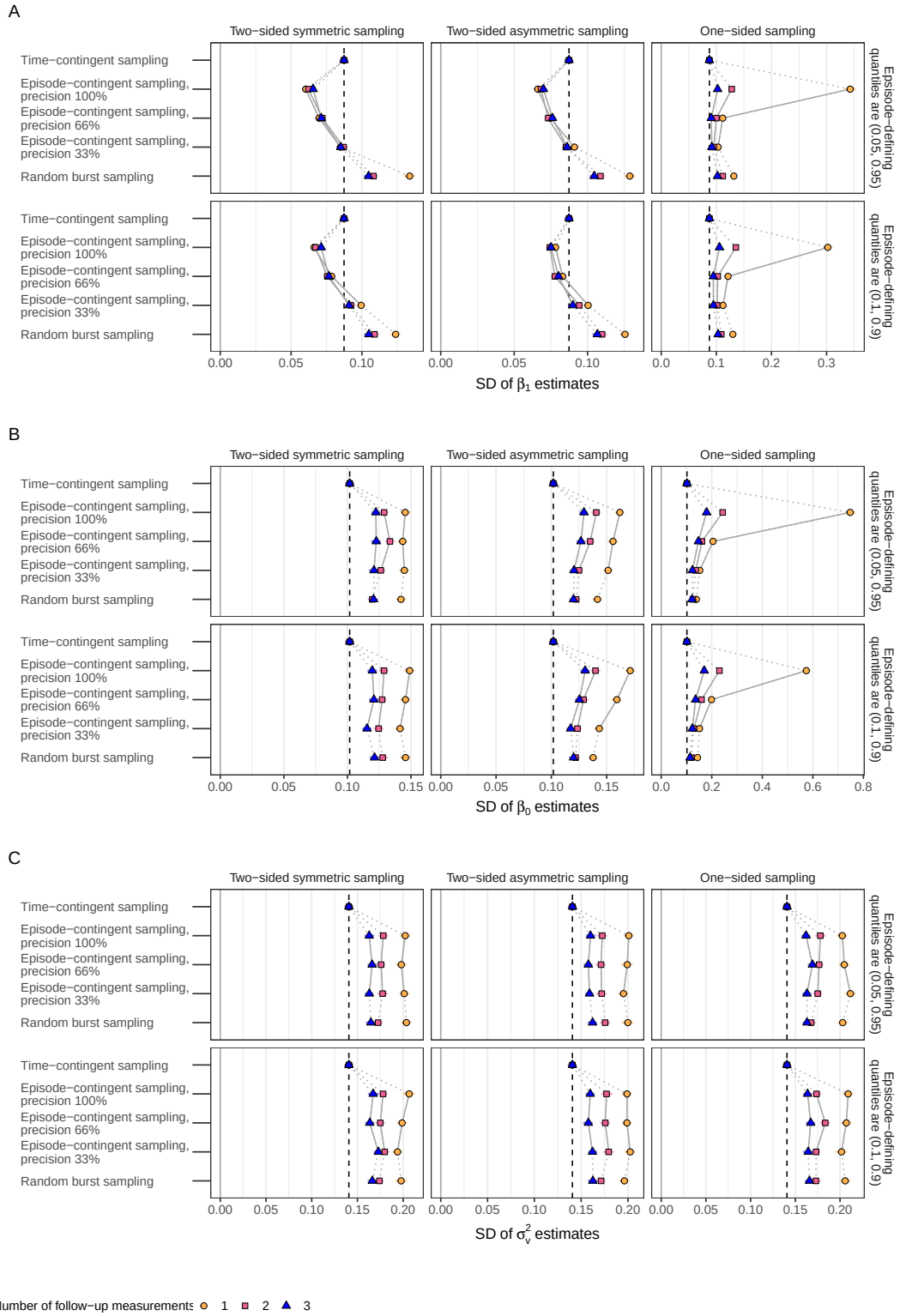

Figure 12: SDs of point estimates under the different sampling schemes for the different parameters. Panel A displays the SD of the AR estimates, panel B the SD of the intercept estimates and panel C the SD of the innovation variance estimates. Dashed black lines mark the values under regular TC sampling as a reference. Per panel results for different symmetry-levels are presented across columns, results for different episode-defining quantile values across rows. Different numbers of follow-up measurements are distinguished by point shape and color. The population-level AR effect size is 0.5. The number of time points is 100. The number of effective pairs of time points is 99 for the TC sampling scheme. For the EC and random burst sampling schemes the number of effective pairs depends on the number of follow-up measurements: It is 50 for 1, 66 for 2, and 75 for 3 follow-up measurements.

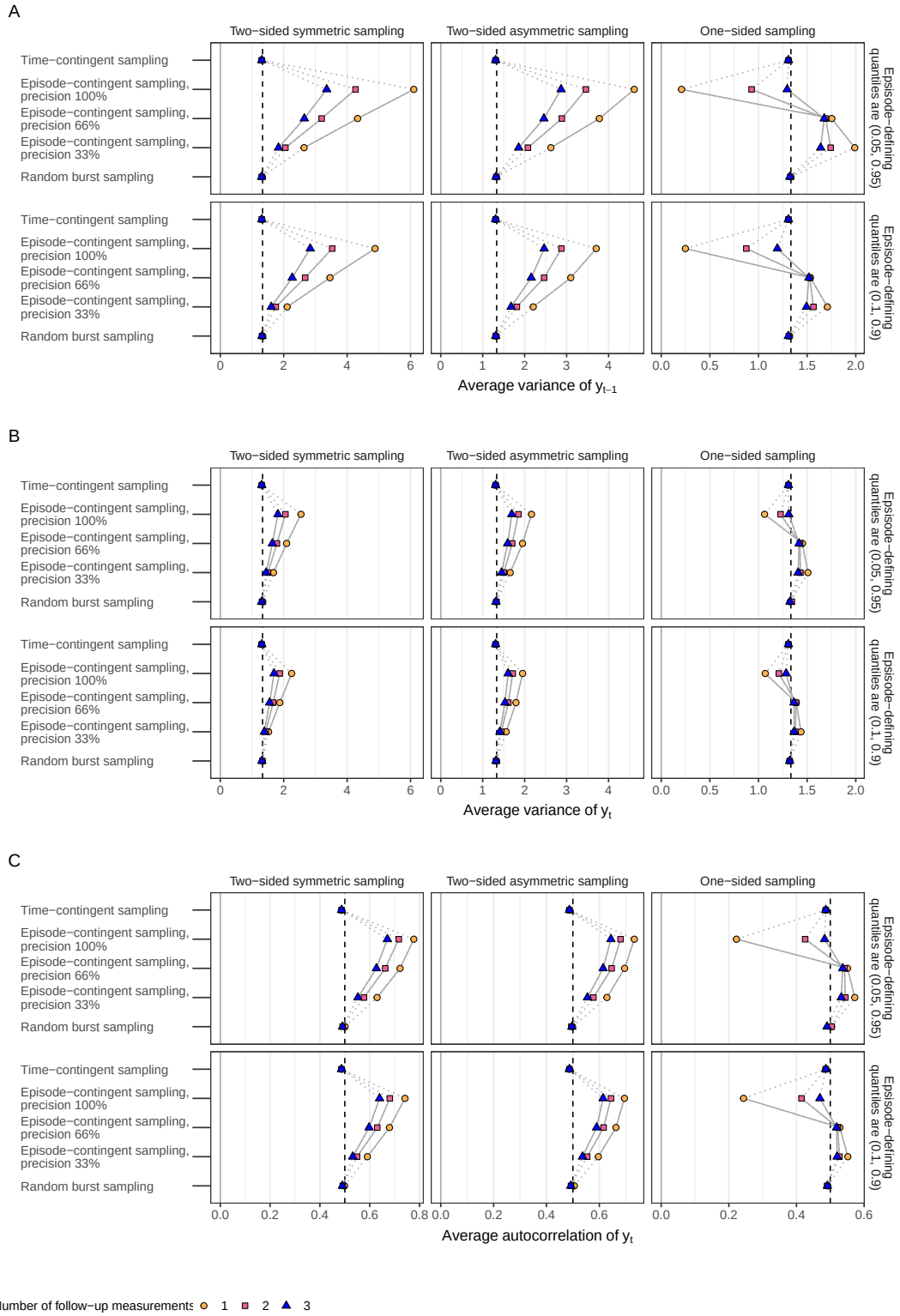

Figure 13: Average variances and autocorrelation under the different sampling schemes. Panel A displays the average variance of the predictor measurements, panel B the average variance of the criterion measurements and panel C the average autocorrelation of the criterion measurements. Dashed black lines mark the model-implied values under regular TC sampling as a reference. Per Panel results for different symmetry-levels are presented across columns, results for different episode-defining quantile values across rows. Different numbers of follow-up measurements are distinguished by point shape and color. The population-level AR effect size is 0.5. The number of time points is 100. The number of effective pairs of time points is 99 for the time-contingent sampling scheme. For the episode-contingent and burst sampling schemes the number of effective pairs depends on the number of follow-up measurements: It is 50 for 1, 66 for 2, and 75 for 3 follow-up measurements.

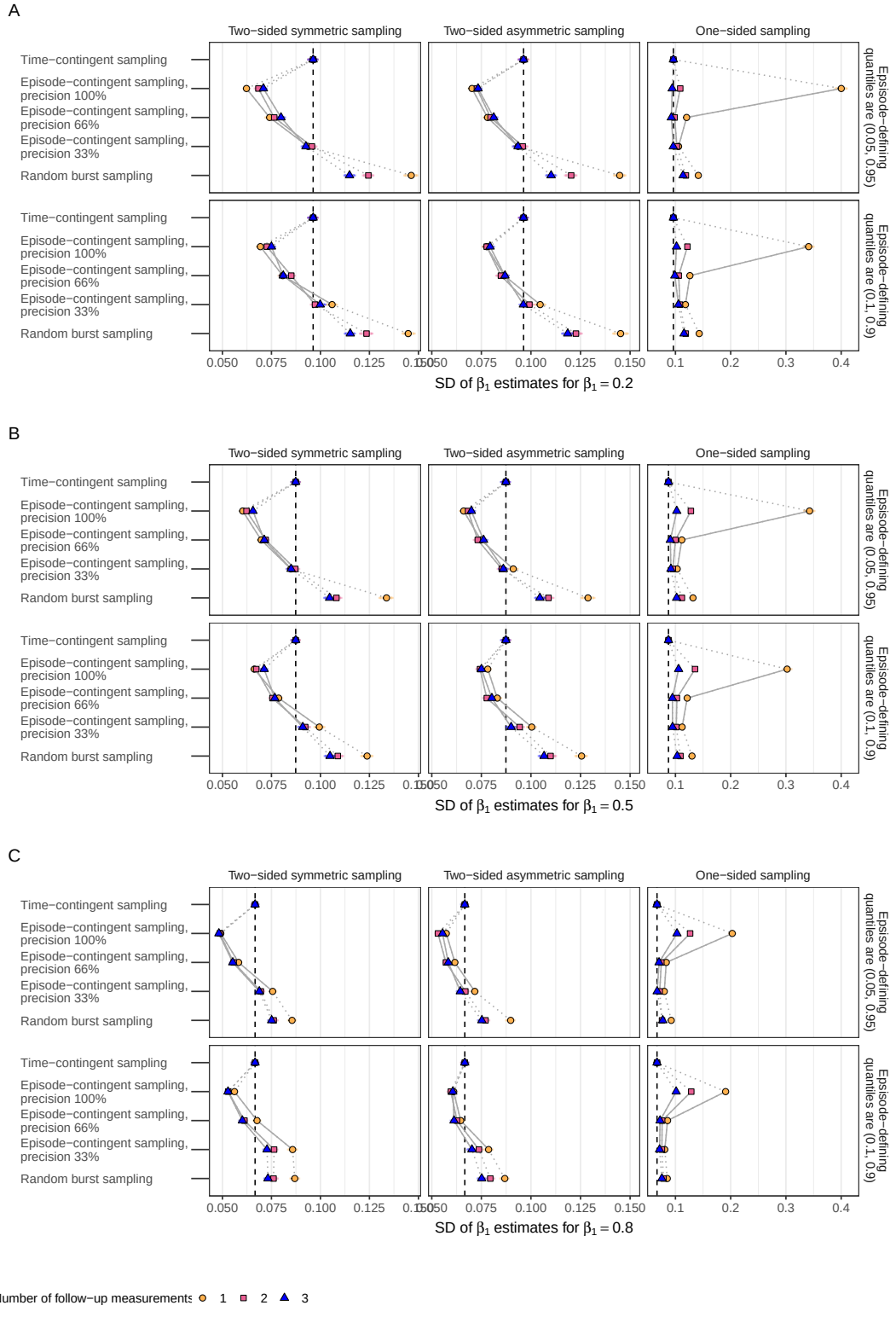

Figure 14: SDs of AR point estimates under the different sampling schemes for the different population level AR effect sizes. Panel A displays the SD for an AR effect of size 0.2, panel B for an AR effect of size 0.5 and panel C for an AR effect of size 0.8. Dashed black lines mark the values under regular TC sampling as a reference. Per panel results for different symmetry-levels are presented across columns, results for different episode-defining quantile values across rows. Different numbers of follow-up measurements are distinguished by point shape and color. The number of time points is 100. The number of effective pairs of time points is 99 for the TC sampling scheme. For the EC and random burst sampling schemes the number of effective pairs depends on the number of follow-up measurements: It is 50 for 1, 66 for 2, and 75 for 3 follow-up measurements.

8.2 Appendix 2: Fixing the number of effective pairs

Figures 15 and 16 provide simulation results analogous to those reported in the results section, but where the number of effective pairs T^* is kept constant between the various sampling designs.

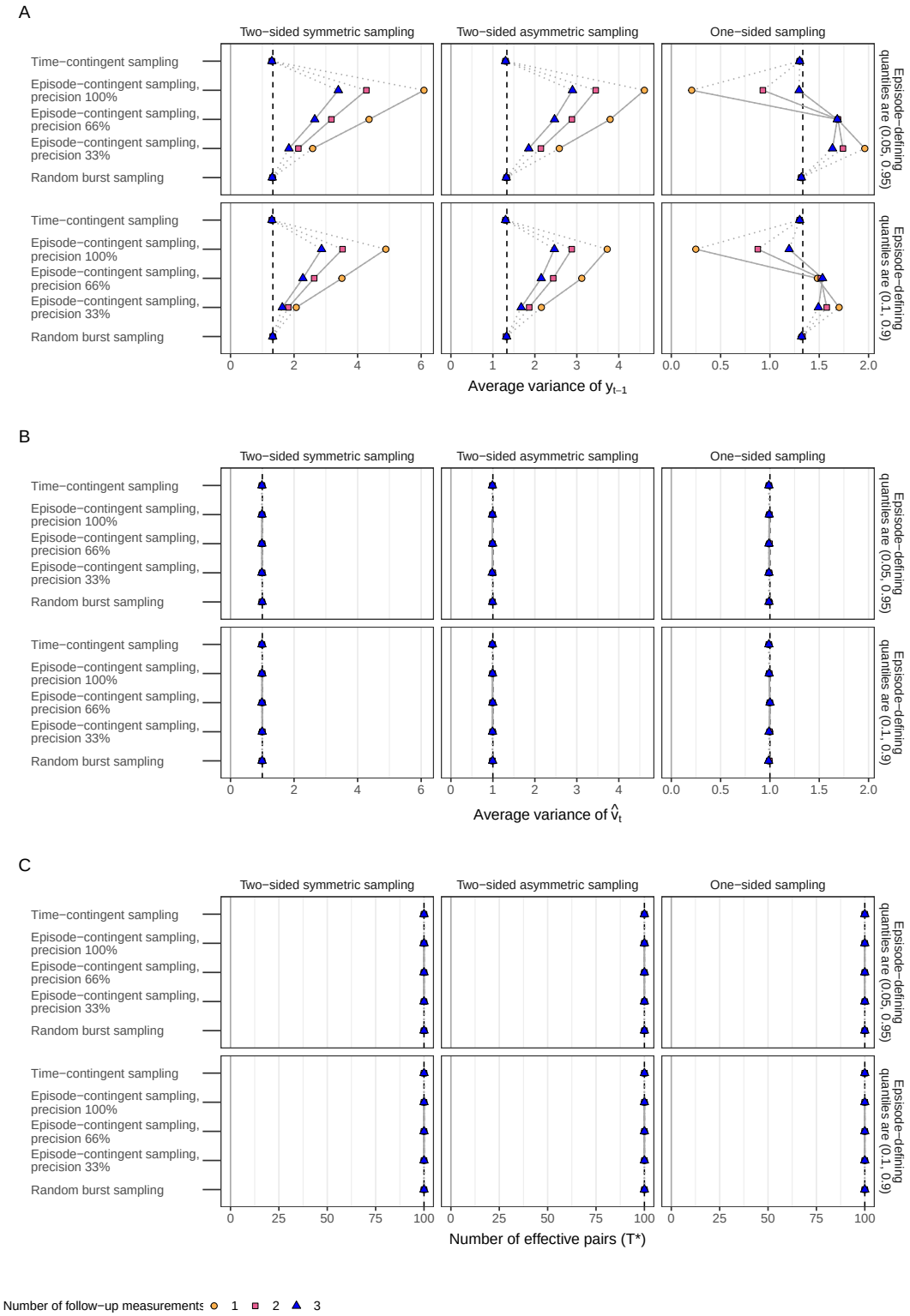

Figure 15: Figure 3 from manuscript with T = T*

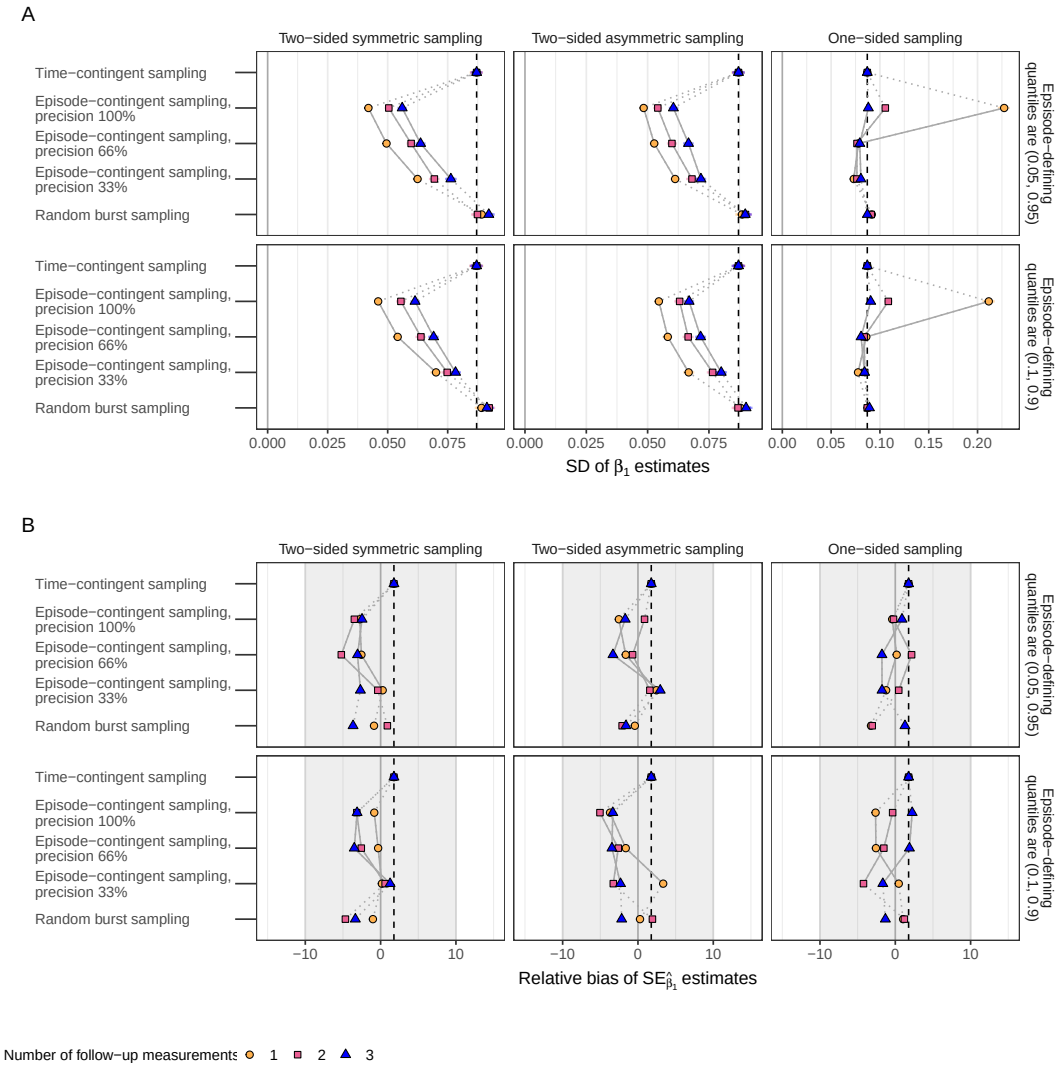

Figure 16: Figure 4 from manuscript with T = T*.

8.3 Appendix 3: Figures for empirical illustration

Figure 17 shows the number of episodes categorized as precisely and imprecisely identified per participant. Figures 18 to 23 show time-series plots for the week 2 and 3 data for all participants. The week-1 quantiles used as the instrument are shown in red lines.



Figure 17: Number of detected episodes in the data per participant. Colors distinguish precisely from imprecisely identified episodes. Per participant, a point is added to show the expected number of episodes given the week-1 quantiles (i.e. 30%) of the regular beeps in weeks 2 and 3 (excluding the last regular beep per day) should have been classified as initial episode measurements)

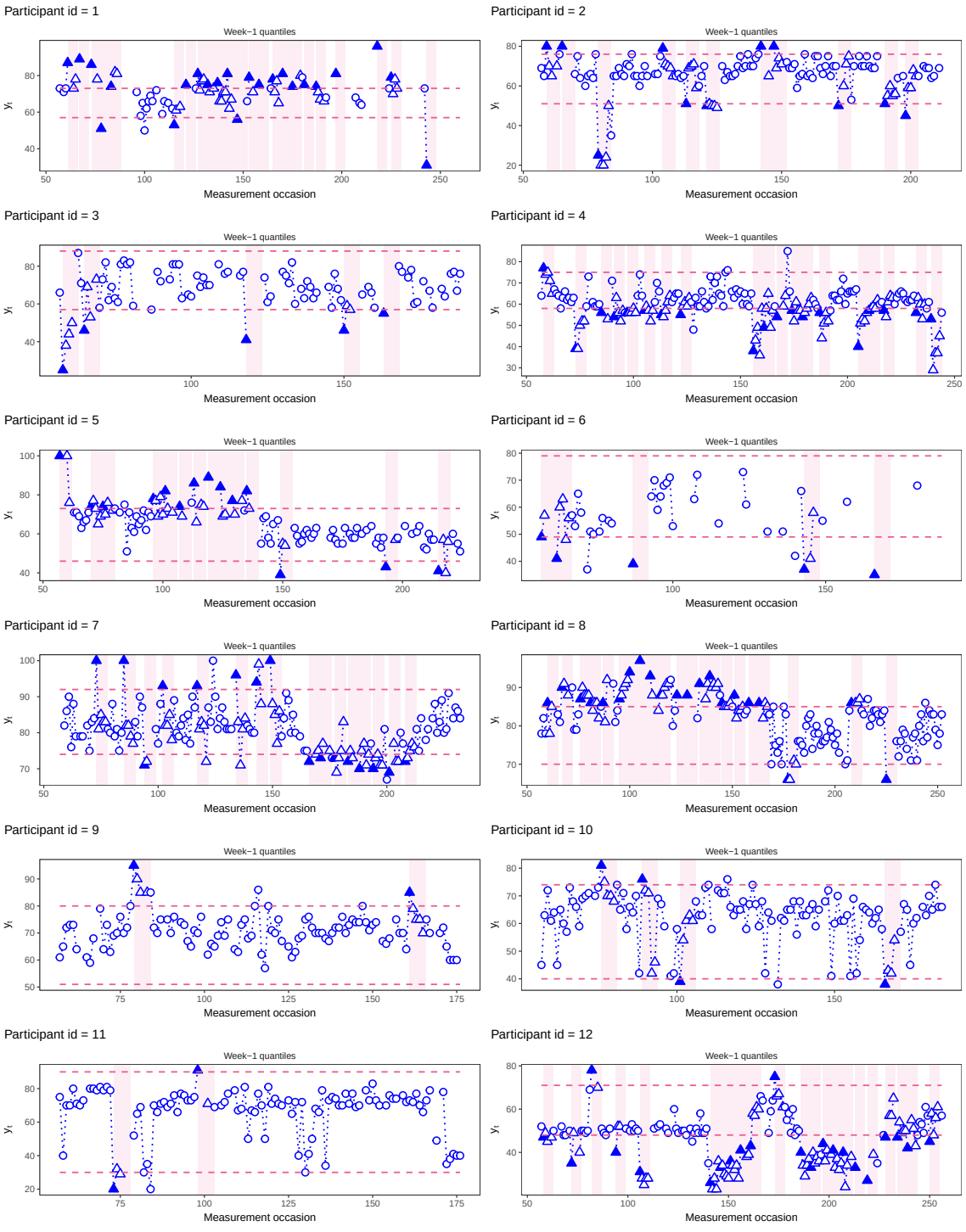

Figure 18: Top panel of figure 5 from the manuscript, but for participants 1 to 12.

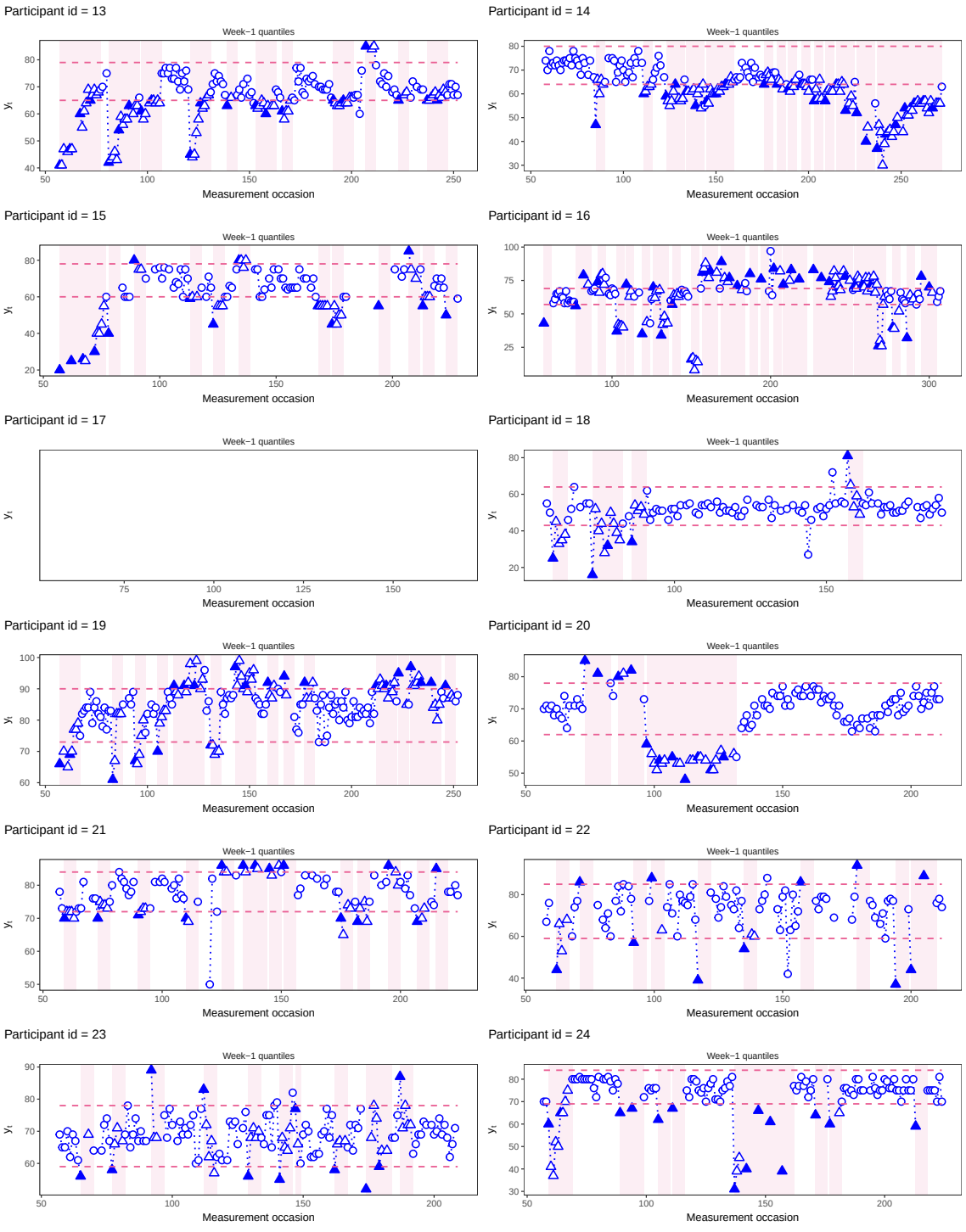

Figure 19: Top panel of figure 5 from the manuscript, but for participants 13 to 24.

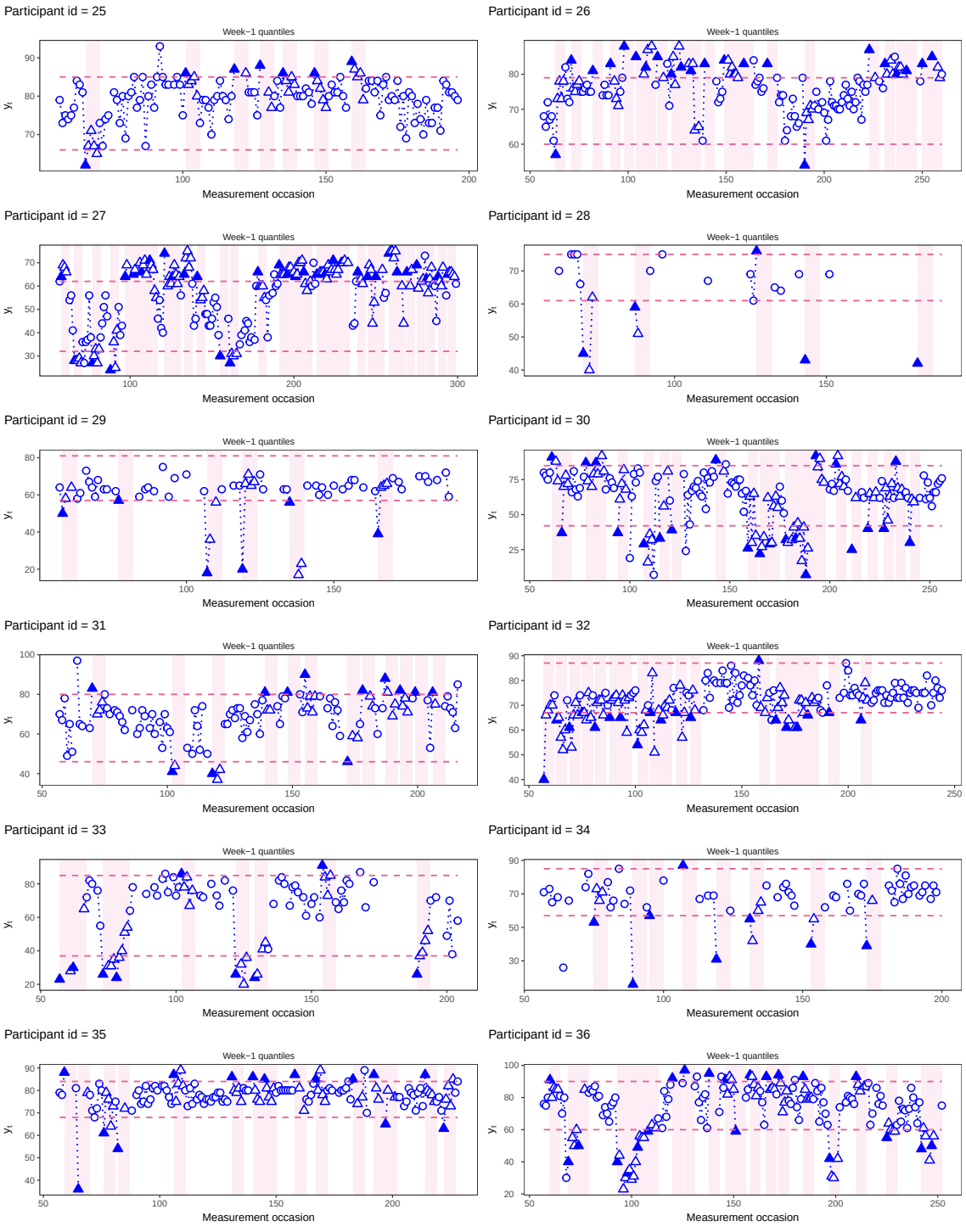

Figure 20: Top panel of figure 5 from the manuscript, but for participants 25 to 36.

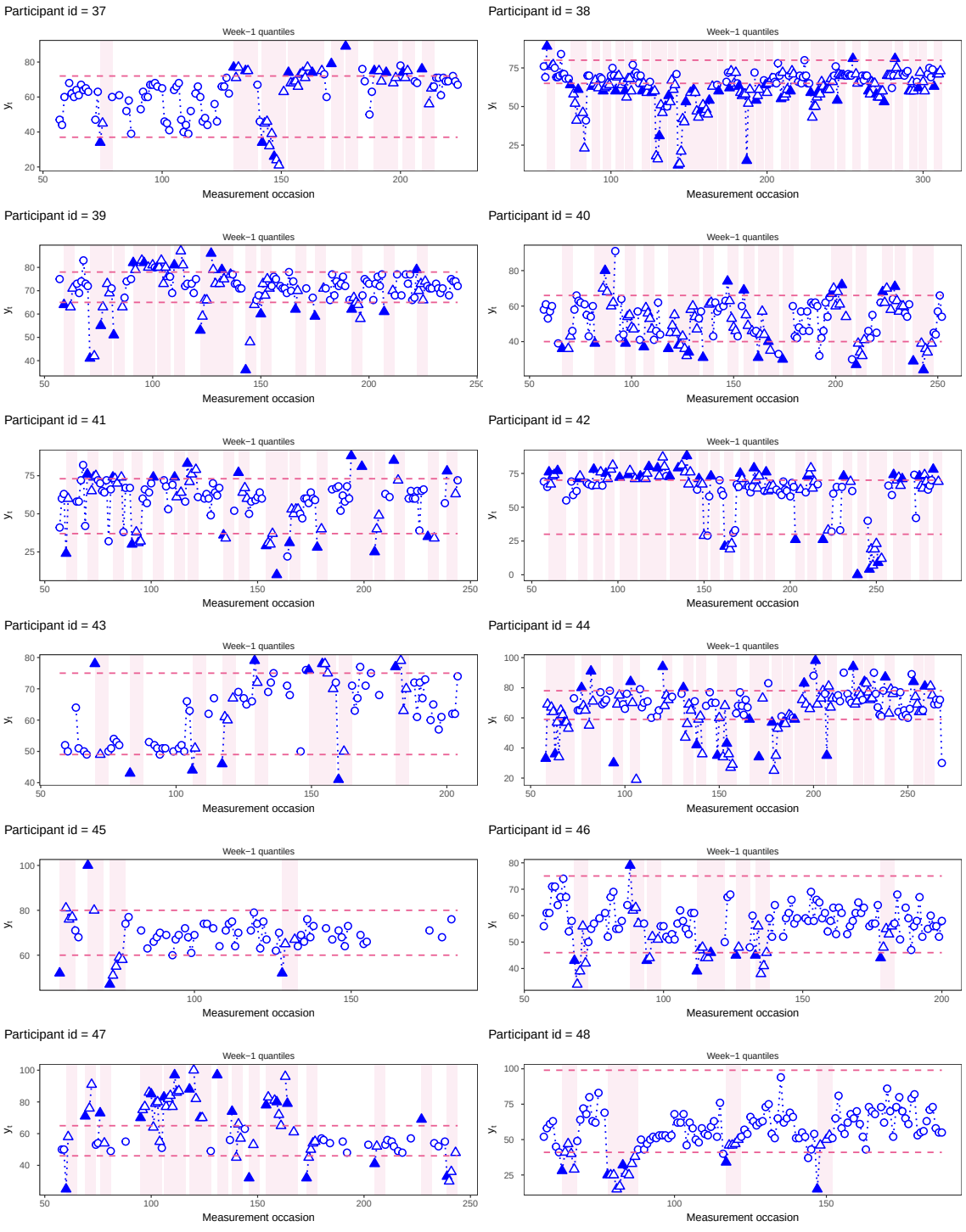

Figure 21: Top panel of figure 5 from the manuscript, but for participants 37 to 48.

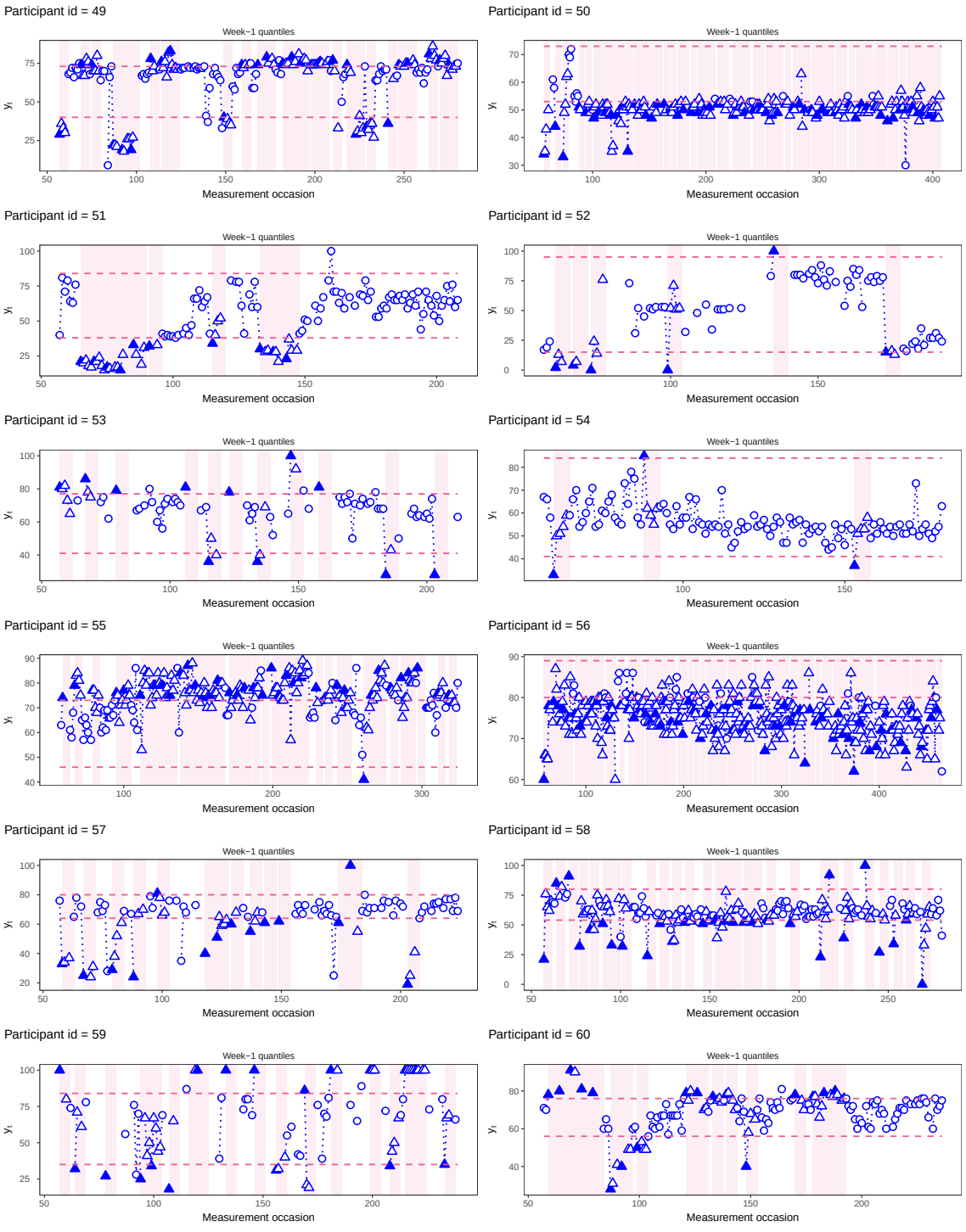

Figure 22: Top panel of figure 5 from the manuscript, but for participants 49 to 60.

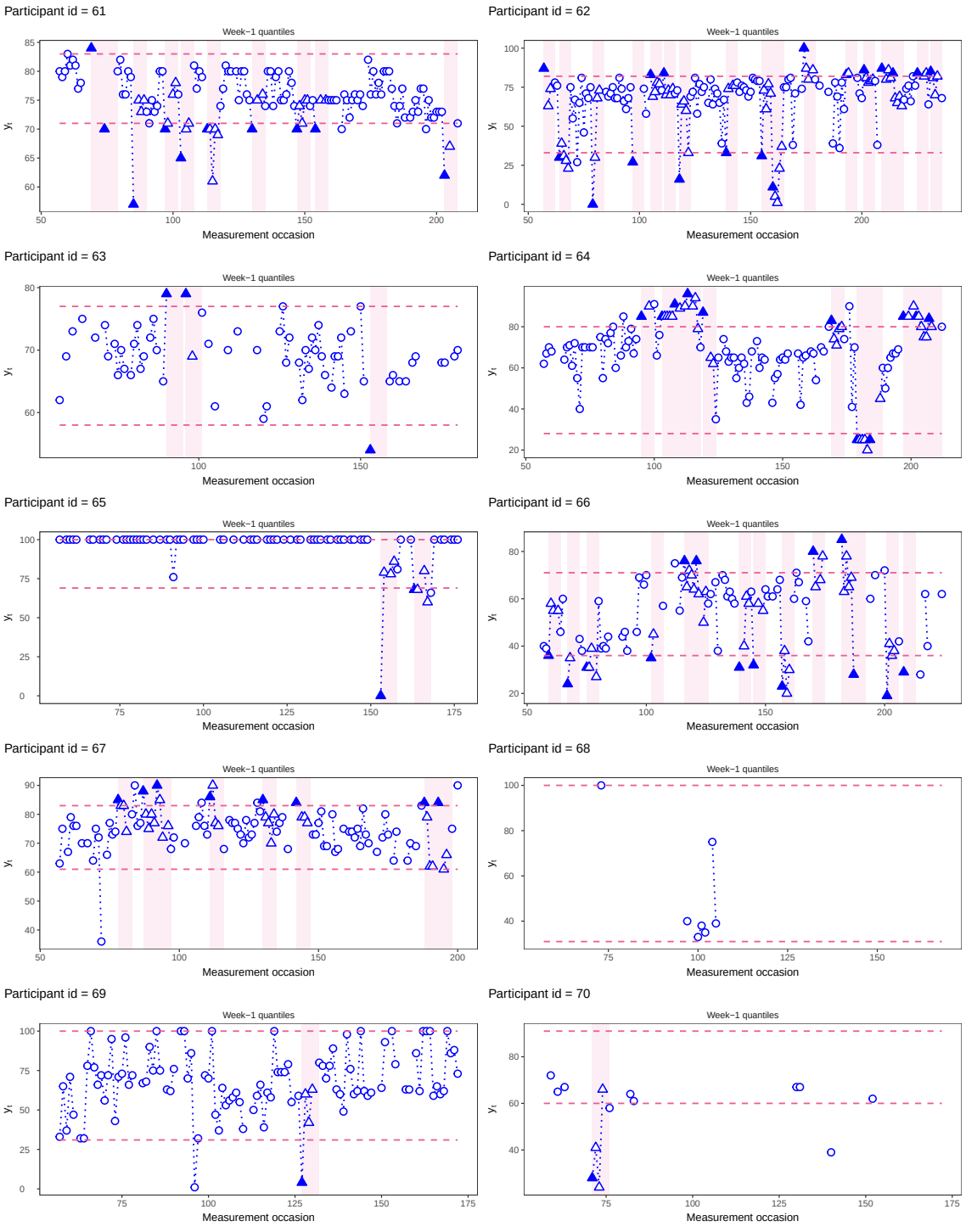

Figure 23: Top panel of figure 5 from the manuscript, but for participants 61 to 70.