AMERICAN PSYCHOLOGICAL ASSOCIATION

Psychological Assessment

© 2022 American Psychological Association
ISSN: 1040-3590

2022, Vol. 34, No. 12, 1138–1154
https://doi.org/10.1037/pas0001178

Assessing the Reliability of Single-Item Momentary Affective Measurements in Experience Sampling

Egon Dejonckheere[1, 2], Febe Demeyer[1], Birte Geusens[1], Maarten Piot[1], Francis Tuerlinckx[1], Stijn Verdonck[1], and Merijn Mestdagh[1]

[1] Faculty of Psychology and Educational Sciences, KU Leuven
[2] Department of Medical and Clinical Psychology, Tilburg University

Emotion researchers that use experience sampling methods (ESM) study how emotions fluctuate in everyday life. To reach valid conclusions, confirming the reliability of momentary emotion measurements is essential. However, to minimize participant burden, ESM researchers often use single-item measures, preventing a reliability assessment of people's emotion ratings. Furthermore, because emotions constantly change, checking reliability via conventional test–retest procedures is impractical, for it is impossible to separate measurement error from meaningful emotional variability. Here, drawing from classical test theory (CTT), we propose two time-varying test–retest adaptations to evaluate the reliability of single-item (emotion) measures in ESM. Following Method 1, we randomly repeat one emotion item within the same momentary survey and evaluate the discrepancy between test and retest ratings to determine reliability. Following Method 2, we introduce a subsequent, shortly delayed retest survey and extrapolate the size of test–retest discrepancies to the hypothetical instance where no time between assessments would exist. First, in an analytical study, we establish the mathematical relation between observed test–retest discrepancies and measurement error variance for both methods, based on common assumptions in the CTT literature. Second, in two empirical studies, we apply both methods to real-life emotion time series and find that the size of error in people's emotion ratings corresponds with almost a 10th of the scale, comprising around 27% of the total variability in participants' affective responses. Consequently, disregarding measurement error in ESM is problematic, and we encourage researchers to include a test–retest procedure in their future studies when relying on single-item measures.

Public Significance Statement
In research on emotions in daily life, measurement error is often ignored because emotions are assessed with a single item to reduce participant burden. We introduce two retests procedures to determine how reliable such emotion ratings are and show that measurement error variance is too substantial to simply disregard.

Keywords: ecological momentary assessment, measurement error, classical test theory, test–retest, daily life

Supplemental materials: https://doi.org/10.1037/pas0001178.supp

This article was published Online First September 8, 2022.

Egon Dejonckheere https://orcid.org/0000-0002-7219-7506
Febe Demeyer https://orcid.org/0000-0002-5736-0168
Birte Geusens https://orcid.org/0000-0001-5798-3245
Maarten Piot https://orcid.org/0000-0002-0118-5903
Francis Tuerlinckx https://orcid.org/0000-0002-1775-7654
Merijn Mestdagh https://orcid.org/0000-0001-5077-861X

The preparation of this article was supported by the Research Fund of KU Leuven (Grant C14/19/054). Merijn Mestdagh and Egon Dejonckheere are postdoctoral research fellows supported by the Fund for Scientific Research, Flanders (Grants 1256221N and 1210621N). Maarten Piot is a doctoral student supported by a KU Leuven technology transfer grant (C3/3H200318). The funding sources had no involvement in the data collection, analyses, or results reported in this article. The authors declare that they have no competing interests.

The authors made all data, materials, and code (R and MATLAB) publicly available on the Open Science Framework at https://osf.io/83cfk/. The preregistration of both protocols can be found on AsPredicted at https://aspredicted.org/blind.php?x=td2di8 (Protocol 1) and https://aspredicted.org/blind.php?x=6zy5y5 (Protocol 2).

Egon Dejonckheere, Febe Demeyer, and Birte Geusens contributed equally to this article and share first authorship.

Egon Dejonckheere played an equal role in conceptualization, funding acquisition, investigation, methodology, project administration, supervision, and writing of original draft. Febe Demeyer played an equal role in data curation, formal analysis, investigation, and writing of original draft. Birte Geusens played an equal role in data curation, formal analysis, investigation, visualization, and writing of original draft. Maarten Piot played an equal role in conceptualization, resources, software, and writing of review and editing. Francis Tuerlinckx played an equal role in conceptualization, funding acquisition, supervision, and writing of review and editing. Stijn Verdonck played an equal role in conceptualization, investigation, methodology, software, supervision, and writing of review and editing. Merijn Mestdagh played an equal role in conceptualization, formal analysis, funding acquisition, investigation, methodology, software, supervision, and writing of review and editing.

Correspondence concerning this article should be addressed to Egon Dejonckheere, Faculty of Psychology and Educational Sciences, KU Leuven, Tiensestraat 102, Leuven 3000, Belgium. Email: egon.dejonckheere@kuleuven.be

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

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

Mood and emotions are dynamic by definition (Frijda, 1988; Kuppens & Verduyn, 2017). When investigating affective experiences in the context of everyday life, researchers often study the dynamical qualities of people's emotions with the experience sampling method (ESM; Larson & Csikszentmihalyi, 1983). In an ESM study on emotions, participants are instructed to carry around a mobile device as they live their life that repeatedly invites them to rate their current emotional experience via momentary self-report surveys during their daily routine (Mestdagh & Dejonckheere, 2021). In order to reach robust conclusions about people's real-life emotional fluctuations and to truly advance our understanding of the temporal regularities that underlie our affective experiences, a first prerequisite is the reliable assessment of people's momentary emotions (Dejonckheere & Mestdagh, 2021).

Here, we rely on a time-varying variant of classical test theory (CTT; DeVellis, 2006; Lord et al., 1968) and propose two adapted test–retest methods to evaluate the reliability of single-item momentary affective assessments in ESM. First, we show analytically, using model assumptions that are common in the domain, how measurement error variance is related to observed test and retest scores when studying dynamical constructs. Next, in two different empirical studies, we apply each method to people's real-life emotion ESM data and determine how much measurement error is associated with people's repeated discrete emotion ratings. In a comprehensive analysis, we explore the role of affect type, investigate the moderating impact of different person-level covariates, and compare our approach with an established, but more model-based reliability estimation procedure in ESM.

Experience Sampling and Real-Life Emotion Dynamics

ESM is particularly suited to study how psychological phenomena such as emotions unfold over time (Mestdagh & Dejonckheere, 2021). Whether it is the question of how contextual antecedents elicit real-life emotional reactions (e.g., Heininga et al., 2019), how emotions predict each other over time (e.g., Pe & Kuppens, 2012), or how different emotion dynamics are indicative of mental well-being (e.g., Dejonckheere et al., 2018), a common denominator of many emotional ESM protocols is that they are interested in establishing the true (within-person) dynamical patterns of people's affective life.

Still, ESM also offers other advantages over more traditional data collection techniques. First, compared to lab research, the external or ecological validity of ESM data is unparalleled by other methods, as emotional assessments take place within personally relevant circumstances instead of artificial lab contexts (Verhagen et al., 2016). This results in the enhanced real-life generalizability of ESM findings (Santangelo et al., 2013). Second, compared to retrospective cross-sectional questionnaires, the repeated and momentary nature of ESM greatly reduces the confounding impact of recall biases (Scollon et al., 2009) and typically requires simpler cognitive operations (Trull & Ebner-Priemer, 2009). In this way, ESM assessments are thought to tap directly into participants' actual emotional experience, rather than being a remembered mental reflection.

Experience Sampling and Reliability

Despite these strengths, ESM also has disadvantages compared to more traditional ways of collecting data in affect research. In fact, we could consider these disadvantages a direct consequence of ESM's relative strengths, and they all potentially jeopardize the reliability of momentary emotion assessments in daily life studies (Dejonckheere et al., 2020). Following CTT (DeVellis, 2006; Lord et al., 1968), we understand the *reliability* of an emotion measure as the degree with which assessments under similar circumstances produce consistent scores when applied repeatedly to a person or an emotion that is assumed to remain invariant (Shrout & Lane, 2012). In CTT terms, the reliability of an emotion instrument indicates to what extent it can recover a participant's *true* or *latent emotion score* (denoted as θ), which is usually concealed by a certain amount of *measurement error* (denoted as ε; Dejonckheere & Mestdagh, 2021). These are random or systematic distortions that are inherent to the measurement process but that make people's observed or manifest emotion ratings (denoted as X) deviate from the true score (Bland & Altman, 1996; Schuurman & Hamaker, 2019). Thus, analytically, CTT decomposes a person's observed emotion rating in two terms (Guttman, 1945):

$$X = \theta + \varepsilon, \tag{1}$$

with $\varepsilon \sim N(0, \sigma_\varepsilon^2)$.

A first way in which the reliability of momentary emotion assessments in daily life studies could be compromised is that, because ESM researchers capitalize on the ecological validity of their data, they typically give in on controllability and standardization found in classic laboratory setups (Moskowitz & Young, 2006). In daily life research, environmental influences that randomly distort participants' emotional evaluations are entirely beyond the researcher's control. These random sources likely add more measurement error to people's responses, concealing changes in participants' true emotional experience. Second, because its repeated assessments make ESM quite an intensive data collection method, researchers aim to keep participant burden and careless responding to a minimum (Eisele et al., 2022; Jaso et al., 2021). Contrary to traditional cross-sectional questionnaires, this leads ESM researchers to opt for single-item momentary emotion constructs (Wilhelm & Schoebi, 2007). However, a critical downside of single-item constructs is that they are more prone to measurement error (Brown, 1910; Sarstedt & Wilczynski, 2009; Spearman, 1910).

Although these two disadvantages suggest that the reliability of momentary emotion assessments in ESM could be problematic, the confounding role of measurement error is often ignored in the affective time series literature (Schuurman & Hamaker, 2019). Nevertheless, it is of crucial importance to know to what extent measurement error distorts participants' observed emotion ratings. Specifically, at a measurement level, a reliability assessment is imperative to distinguish between meaningful intraindividual change in people's emotion time series and random fluctuations that are the result of measurement error (Schuurman et al., 2015). In turn, at a modeling level, ESM researchers need to know how these noisy emotion ratings produce biased model parameters or unreliable time series summaries. Here, it has been established that measurement error typically attenuates model coefficients and correlations toward zero, providing an underestimation of the true parameter value in the population (Spearman, 1904). Because measurement error reduces statistical power (i.e., it restricts the maximum magnitude of an effect that can be observed), ESM researchers require larger sample sizes to detect a target effect size (i.e., they need more data to distinguish noise from meaningful effects; Phillips & Jiang, 2016).

Finally, at the level of interpretation, these inaccurate model parameters and correlations may lead to unstable results and findings that are less likely to generalize (Cronbach et al., 1963; Parsons et al., 2019). Researchers' derived conclusions may be invalid, and in the context of hypothesis testing, lead to false-negative decisions (i.e., Type II errors; Kobayashi & Boase, 2012).

If this problematic cascade of measurement error remains undisclosed because ESM researchers do not evaluate the reliability of their momentary emotion measures, this inaction may jeopardize true scientific progress in the long run. Alternatively, establishing measurement error variance allows researchers to take action when the quality of their ESM data is below par (Dejonckheere & Mestdagh, 2021). For example, reliability estimates can statistically guide researchers in their decision to remove certain momentary assessments or participants from their analysis when they show an excessive amount of unreliable or careless responding (Jaso et al., 2021). As such, this procedure allows data cleaning not only on the basis of data quantity (e.g., poor participant compliance, which is common practice in the ESM literature; Trull & Ebner-Priemer, 2020) but also data quality (e.g., poor measurement reliability).

An Internal Consistency Multi–Item Reliability Approach

An evident reason why measurement error is often ignored in ESM lies in the method's strong reliance on single-item measures (Wilhelm & Schoebi, 2007). Not only are single-item constructs more measurement error-prone, with only a single assessment of a momentary emotional state, evaluating the reliability of participants' associated ratings using standard tools becomes impossible (Weidman et al., 2017). To resolve this issue, the traditional CTT framework would advise ESM researchers to design multi-item momentary emotion constructs in order to imitate the similar or parallel conditions needed to evaluate the consistency of an emotion measure (DeVellis, 2006). Here, a common practice in ESM is to average various discrete positive and negative emotion items into a positive affect (PA) and negative affect (NA) composite score, respectively. Although this approach is thought to create an affective construct that is less distorted by the measurement error of its individual emotion item assessments (Brown, 1910; Spearman, 1910), multi-item constructs in ESM are not always feasible or desired.

Regarding feasibility, we already explained that the burdensome nature of daily life research may put a limit on the number of items researchers can include in their momentary survey (Eisele et al., 2022). Regarding desirability, some emotion researchers may simply have interest in the temporal correlates of specific discrete emotions rather than those of affective aggregates (e.g., *feeling down*; Wichers et al., 2020). Relatedly, other researchers may want to investigate certain affect dynamic measures that exclusively rely on the ratings of individual emotions and not averages (e.g., *emotion differentiation*; Van der Gucht et al., 2019). In both cases, it is of crucial importance to have an indication about the reliability of these individual emotion ratings.

A Model-Based Single-Item Reliability Approach

It is clear that single-item emotion measures prevent ESM researchers to rely on CTT's internal consistency methods (e.g., Hu et al., 2016; Nezlek, 2017). However, another way in which the reliability of single-item emotion constructs may be established is through a modeling approach (e.g., Dejonckheere & Mestdagh, 2021; Schuurman et al., 2015). Following such a procedure, assumptions are made about the true emotion generating model to estimate the amount of measurement error in time series data.

For example, Schuurman et al. (2015) operationalized time series data as a latent autoregressive (AR) model (of order 1) with an innovation term, that is, an AR(1) model; see Equation (6) for exact model equation. While the innovation term in a latent AR(1) model roughly corresponds to the intensity of an emotional stimulus people react to at a given assessment, the AR coefficient represents how long an individual's emotion typically lingers after that stimulus was introduced. Because innovations resonate through subsequent assessments via this AR relation they are also called *dynamical noise*, and together with the AR parameter, this model aims to describe how people truly emotionally react and recover from the contextual stimuli they encounter (Dejonckheere & Mestdagh, 2021). Consequently, the remaining variance in people's emotion ratings that cannot be explained by this latent AR(1) process is considered measurement error variance, because it does not carry over from one assessment to the next, and hence remains specific to a particular measurement occasion (e.g., Bland & Altman, 1996).

Assuming that an AR(1) model indeed underlies people's true emotional responding, this model-based reliability approach estimates that approximately 30%–50% of the variability in people's observed emotion time series is the result of measurement error. Although this estimate was based on a small sample, it suggests that the reliability of single-item ESM measures at the within-person level is strikingly low, and that disregarding measurement error is far from inconsequential (Schuurman et al., 2015). Nevertheless, the question remains to what extent these latent AR(1) models correspond to our actual emotional life (e.g., Loossens et al., 2021). Despite their popularity in affective ESM research (Kuppens & Verduyn, 2017), AR models assume that the ebbs and flows of our emotions take on a very specific form, and in reality, some of the imposed model assumptions may not be readily met. Consequently, when the hypothesized underlying model does not adequately represent the actual regularities in people's emotional responding, the model will falsely categorize these changes as measurement error (Schuurman & Hamaker, 2019; see Supplemental Material 1, for a simulated graphical example) with a potential overestimation of this error variance as a result (Rhemtulla et al., 2020).

A Test–Retest Single-Item Reliability Approach

Alternatively, a time-varying version of CTT allows ESM researchers to estimate measurement error variance for single-item emotion measures without their commitment to make too many assumptions about the emotion-generating model. When adopting a test–retest procedure following traditional CTT, similar or parallel test conditions may be accomplished by evaluating the consistency of an emotion rating at different moments in time (Parsons et al., 2019; Shrout & Lane, 2012). Critically, however, a traditional test–retest approach is only feasible when people's true emotion score remains invariant across momentary assessments (Shrout & Lane, 2012), a premise that is easily violated in ESM, where researchers are primarily interested in real-life emotional change. In a typical ESM study, using a test–retest procedure to disentangle meaningful emotional change (e.g., due to exposure to

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

different contextual stimuli) from measurement error (e.g., due to momentary inattention, fatigue, or careless responding) is difficult because it is impossible to detect which of these sources is driving the variability in participants' observed scores (Steyer et al., 2015). Nevertheless, with a few adjustments to the standard study design, applying a test–retest procedure may become feasible, and here, we put forward two variations that should allow researchers to evaluate the amount of measurement error variance in single-item emotion time series. Each method has its strengths and weaknesses, based on the underlying assumptions they make.

In a first approach, ESM researchers may randomly repeat one emotion item within the same momentary survey (see Method 1 in Figure 1). Because the completion of a short questionnaire only takes a couple of minutes, the likelihood that something emotionally moving happens in between the test and retest item is almost nonexistent. Consequently, the assumption that participants' true emotion score remains invariant across assessments is quite plausible (although not definite), allowing researchers to conclude that differences in emotion ratings can almost certainly be attributed to measurement noise. A downside to this approach, however, is that participants may easily notice exact item repetitions within short questionnaires. This could introduce other biases in participants' responding (e.g., strengthen demand characteristics or social desirability tendencies; Nichols & Maner, 2008), where people become attentive of their answer pattern, striving for consistency instead of truthfully rating their current emotional state.

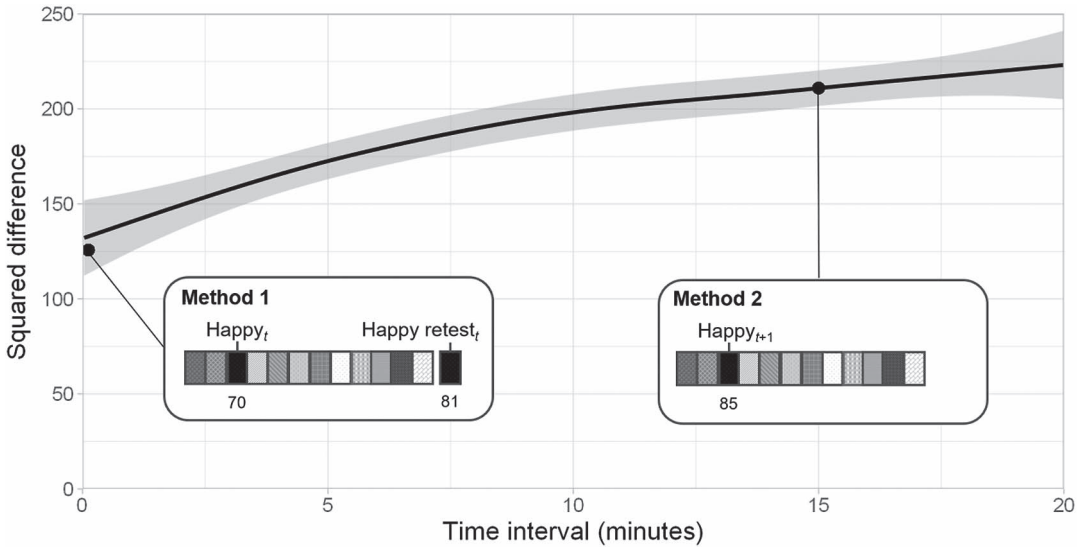
Therefore, in a second approach, ESM researchers could randomly vary the (short) interval widths between pairs of two repeated surveys (see Method 2 in Figure 1). Because all momentary emotion surveys are exactly identical, and no items are repeated within the same questionnaire, the risk of other biases distorting people's actual emotional responding may not be particularly elevated in this protocol. However, as more time between successive emotion ratings passes, the chance that emotions meaningfully fluctuate as a result of changing contextual input becomes substantial (Lowman et al., 2018; McArdle & Woodcock, 1997). Here, the assumption that people's true emotion scores did not change across assessments is less plausible, as a break between two momentary surveys might cause in-between real-life activities to change one's true latent emotional state. This assumption makes it more difficult to directly attribute differences in emotion ratings to measurement error. Nevertheless, because we can expect that momentary emotion assessments closer together in time are more similar than assessments further apart in time (i.e., we assume that people's proximate true emotion scores are more serially dependent; Santangelo et al., 2013), extrapolating the size of observed differences between two consecutive emotion ratings to the instance where there would be no time window between assessments may yield an (approximate) indication of the amount of measurement error associated with single-item momentary emotion ratings.

The Present Study

In this article, we will explore the potential of these two test–retest methods in emotional ESM. Our ultimate aim is for researchers to implement a test–retest procedure in their future ESM protocols, allowing them to evaluate the reliability of their momentary (emotion) measures. To this end, we will first, in an analytical study, demonstrate for each approach how measurement error variance can be determined on the basis of people's observed test and retest

Figure 1

Visual Representation of Our Two Test–Retest Adaptions in an ESM Context



Note. Each patterned rectangle refers to a unique emotion item within a momentary survey ($n = 12$), and we consider the (squared) difference between a pair of test and retest items to evaluate the measurement error variance in an individual's time series. With Method 1, we randomly repeat an emotion item within the same momentary survey t (e.g., *happy*). With Method 2, we randomly vary the length of the (short) time intervals between consecutive surveys t and $t + 1$. Here, we infer measurement error variance by considering the hypothetical instance where the time window between two consecutive assessments would be zero. ESM = experience sampling method.

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

ratings. Second, in two empirical studies, we will apply both methods to real-life emotional time series. Here, our goal is to evaluate the reliability of single-item momentary emotion measurements, with the hope that both approaches more or less reveal an equal amount of measurement error in participants' repeated emotion ratings. Specifically, we will conduct all analyses using a multilevel framework to take into account the hierarchical structure of the data (i.e., multiple momentary emotion ratings per person), estimating the measurement error variance in people's momentary emotion ratings (across emotion items and across individuals). Additionally, we will compare our test–retest adaptations with the estimation technique proposed by Schuurman et al. (2015) to evaluate the consistency with a modeling approach. We will also investigate the moderating impact of affect type and three relevant person-level covariates.

Analytical Study

Before applying our two test–retest methods to real-life emotion time series, we will draw from the CTT framework (Lord et al., 1968) to analytically show how researchers may establish measurement error variance from repeated test and retest scores in ESM. Specifically, we will demonstrate that the expected squared difference (i.e., Δ^2) between two identical emotion ratings provides an indication of the amount of measurement error variance that is present in people's emotional time series data. For each approach, we will make explicit the assumptions that underlie this mathematical relation.

Approach 1: Measurement Error From Test–Retest Ratings With an Immediate Retest

For the first test–retest method, we would analyze the squared discrepancy between a randomly selected test–retest emotion item pair within the same momentary emotion survey. Because both item evaluations are part of the exact same measurement occasion, their squared difference Δ^2 allows for a (virtually) direct estimation of participants' measurement error. If we assume no memory effects, the mathematical relation between measurement error variance and the expected Δ^2 between a test and an immediate retest rating is demonstrated in the following analytical equations. Applying CTT's general Equation (1) to a time-varying ESM context, we suppose that measurement X_t at time point t can be decomposed in a true latent emotion (θ_t) and a measurement error (ε_t):

$$X_t = \theta_t + \varepsilon_t. \tag{2}$$

The same can be said for the immediately repeated measurement X_t^{rep}, which is assumed to have the same underlying latent emotion θ_t because of the short time interval:

$$X_t^{\text{rep}} = \theta_t + \varepsilon_t^{\text{rep}}. \tag{3}$$

All measurement errors ε_t and $\varepsilon_t^{\text{rep}}$ are assumed to be independent and to come from the same underlying normal distribution with measurement error variance σ^2:

$$\varepsilon_t, \varepsilon_t^{\text{rep}} \sim N(0, \sigma_\epsilon^2). \tag{4}$$

The expected value of Δ^2, the squared difference between a test and an immediate retest rating, is given by:

$$\begin{aligned} E(\Delta^2) &= E[(X_t - X_t^{\text{rep}})^2]. \\ &= E[(\theta_t + \varepsilon_t - \theta_t - \varepsilon_t^{\text{rep}})^2] \\ &= E[(\varepsilon_t - \varepsilon_t^{\text{rep}})^2] \\ &= E[\varepsilon_t^2 + (\varepsilon_t^{\text{rep}})^2 - 2\varepsilon_t\varepsilon_t^{\text{rep}}] \\ &= E[\varepsilon_t^2] + E[(\varepsilon_t^{\text{rep}})^2] - 2E[\varepsilon_t\varepsilon_t^{\text{rep}}]. \end{aligned}$$

As both ε_t and $\varepsilon_t^{\text{rep}}$ are assumed to come from the same distribution with zero mean and variance σ_ϵ^2, it follows that $E[\varepsilon_t^2] = E[(\varepsilon_t^{\text{rep}})^2] = \sigma_\epsilon^2$. In addition, because of the assumed independence between ε_t and $\varepsilon_t^{\text{rep}}$, we may conclude that $E[\varepsilon_t\varepsilon_t^{\text{rep}}] = E[\varepsilon_t]E[\varepsilon_t^{\text{rep}}] = 0$. Therefore, we find that

$$E(\Delta^2) = \sigma_\epsilon^2 + \sigma_\epsilon^2 + 0 = 2\sigma_\epsilon^2. \tag{5}$$

As a result, to estimate measurement error variance σ_ϵ^2, we need to divide the expected value of the squared difference between a test rating X_t and an immediate retest rating X_t^{rep} by a Factor 2. However, this approach only provides an accurate reliability estimate if we assume that memory effects are absent, and participants do not deliberately pursue consistent ratings, remembering the score of their original test ratings. For Approach 2, chances are smaller that this assumption is violated.

Approach 2: Measurement Error From Test–Retest Ratings With a Delayed Retest

Here, compared to the first method, we investigate the expected value of the squared difference $\Delta^2(t)$ between a pair of test and retest ratings with a (short) time delay t. We start with a quite general assumption: The serial dependency between two true states at different moments in time decreases as more time elapses (Santangelo et al., 2013), and the measurement error is independent of time and true score (equivalent with the assumptions in the first approach). As a consequence, we can make two reasonable claims. First, the expected value of $\Delta^2(t)$ should increase with longer time intervals. Second, when we consider the hypothetical situation where the time interval between two consecutive measurement occasions is zero, then $E[\Delta^2(0)]$ should be equal to twice the measurement error variance, similar as in Approach 1. However, because these test and retest evaluations belong to two different momentary measurement occasions, the quantity $E[\Delta(0)^2]$ will be estimated based on extrapolation. Therefore, this approach provides a more indirect estimation of participants' measurement error.

In principle, the two claims made above hold for any reasonable time-varying extension of the traditional CTT. As such, these claims are independent of any specific underlying dynamic model, meaning that the test–retest method does not make any assumptions about the true dynamic emotion-generating model to determine the within-person reliability of people's affective responses. However, below, we will analytically proof for a specific model that our claims hold. The model we work with is the latent AR(1) model proposed by Schuurman et al. (2015), which we selected based on its popularity in affective ESM research (Kuppens & Verduyn, 2017). We could consider this model a time-varying version of the first approach (see Equation 2), appended with an additional AR(1) process at the latent level. Thus, similar to the first approach, we assume that measurement

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

X_t at time point t can be decomposed in a true latent emotion (θ_t) and a part measurement error (ε_t):

$$X_t = \theta_t + \varepsilon_t,$$

with $\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$. However, as an underlying change model for the true score θ, we assume an AR(1) process:

$$\theta_t = \alpha\theta_{t-1} + \eta_t, \tag{6}$$

with α reflecting a person's AR parameter and η_t the innovation noise, with $\eta_t \sim N(0, \sigma_\eta^2)$. At time point zero, it holds that $X_0 = \theta_0 + \varepsilon_0$ with $\theta_0 \sim N(0, \frac{\sigma_\eta^2}{1-\alpha^2})$, which is a stationary distribution.

The time unit in our model is arbitrary. Measurements can only be taken at regular time intervals (for $t = 0, 1, 2, 3, \ldots$) but the elapsed time is not specified. In order to bring elapsed time explicitly into the model, we may set $t = n \times \delta$, where δ is the minimal time difference between two consecutive measurements and $n = 0, 1, 2, 3, \ldots$. We will not do this explicitly to have a more direct link with the CTT from the first approach, but it should be kept in mind that the minimal time difference is chosen by the experimenter. Changing the minimal time difference will affect the value of the AR parameter α and the innovation variance σ_η^2, and hence also measurement error variance estimation (Schuurman & Hamaker, 2019).

Expressing θ_t as a function of θ_0 gives

$$\theta_t = \alpha^t\theta_0 + \sum_{k=0}^{t-1} \alpha^k \eta_{t-k}. \tag{7}$$

Using Equation (7), we can derive the following expression for $E[\Delta(0)^2]$, where the expectation is taken with respect to ϵ_t, η_t, and θ_0 (see Supplemental Material 2):

$$E[\Delta(t)^2] = E[(X_t - X_0)^2] = 2\sigma_\eta^2 \frac{1-\alpha^t}{1-\alpha^2} + 2\sigma_\varepsilon^2. \tag{8}$$

Therefore, by taking $t = 0$, we find that $E[\Delta(0)^2] = 2\sigma_\varepsilon^2$, which is the same result as in the first approach.

In a simulation study, we evaluated the performance of both methods under different assumption scenarios (e.g., Hu et al., 2016). The simulation results suggest that the performance of both methods improves when more test–retest pairs are included to the time series. Furthermore, an immediate test–retest method within the same momentary survey (Method 1) outperforms a test–retest method that relies on delayed assessment (Method 2). For a full overview of the simulation study, we refer to Supplemental Material 3.

Empirical Study

Now that we mathematically established that $\frac{E[\Delta(0)^2]}{2}$ should correspond to the variance in an ESM time series that is attributed to measurement error, we will implement each test–retest procedure in a real-life ESM study to establish the reliability of single-item momentary emotion ratings. We hypothesize that both methods should more or less converge on the amount of measurement error variance they determine, and we will compare our test–retest approaches with the model-based estimation procedure proposed by Schuurman et al. (2015). Additionally, we examine what factors impact the degree of measurement error in emotional time series data, investigating both the moderating role of (a) affect type and (b) person-level covariates.

Hypotheses

The Role of Affect Type

We will examine if the reliability for positive and negative emotion items is similar or whether measurement error distorts these emotion ratings differently. Here, an important factor is the restrictive mean–variance relation for bounded response scales (Mestdagh et al., 2018; Vanbelle & Lesaffre, 2018). This relation implies that depending on the mean of a number of ratings, the amount of associated variability one can observe in these scores is limited differently: As the mean approaches one of the scale end points (e.g., due to responses that are extremely high or low), the variance of the individual ratings becomes restricted (Eid & Diener, 1999). Because measurement error manifests as an important source of variance (Shrout & Lane, 2012), the degree with which it conceals participants' true emotional responding may depend on the mean levels of the associated scores (Mestdagh et al., 2018). In community samples, people usually experience little negative emotion, resulting in low negative emotion ratings and low average levels of NA (Kalokerinos et al., 2020). In contrast, positive emotions are more common, generating mean PA scores that are further away from the scale's boundaries. Consequently, we hypothesize that the amount of measurement error will be higher for positive emotion ratings than for negative ones, simply because a floor effect for NA scores limits the variability we can observe.

Person-Level Covariates

Similar to cross-sectional survey research (e.g., Ghiselli, 1963), it is plausible that the reliability of people's momentary emotion responses differs as a function of important person-level characteristics (Hu et al., 2016). Here, we explore the potential moderating role of three variables: social desirability, conscientiousness, and neuroticism. First, people who prefer to present themselves in a socially desirable way typically answer to a certain established favorable norm (Borkenau & Ostendorf, 1989), rather than to the varying actual events or situations they encounter. Consequently, although this tendency negatively impacts the validity of their repeated emotional evaluations (Nichols & Maner, 2008), their responses may be more consistent within and across assessments, resulting in less random measurement error. Second, individuals who are highly conscientious tend to comply with instructions more deliberately (John & Srivastava, 1999), completing questionnaires with greater care and precision. As such, we hypothesize that conscientious people will show less measurement error in their momentary emotion evaluations. Finally, neuroticism is associated with stronger instability and variability in (negative) affect (Houben et al., 2015). However, as Kalokerinos et al. (2020) point out, in accordance with Mestdagh et al. (2018), the link between neuroticism and emotional variability may again be partly explained by the structural relation between mean and variance in bounded affect ratings. Because neurotic people more frequently experience negative emotion, their elevated average levels of NA increase the amount of variability they can display in these negative emotional states (Mestdagh et al., 2018). A similar rationale can be applied to

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

measurement error: The reliability of momentary emotion ratings may be lower in highly neurotic people because their average levels of NA are more located toward the middle of the response scale. Another (more substantive) explanation for the positive link between neuroticism and measurement error is that neurotic people tend to have more difficulties accurately identifying and labeling their (negative) emotions (Thompson et al., 2015), producing more unreliable responses.

Materials and Method

Transparency and Openness

Both protocols were approved by the Social and Societal Ethics Committee of the KU Leuven (G-2019 12 1895). All reported hypotheses, data collection methods, and analysis plan were preregistered unless explicitly mentioned otherwise and can be retrieved online from AsPredicted (Protocol 1: No. 39405; Protocol 2: No. 39407; see Supplemental Material 4). To reproduce our results, the data and code (both in R and MATLAB) are available online from the Open Science Framework (https://osf.io/83cfk/). We report the determination of our sample size, all data exclusions (none), all manipulations, and all measures in the article.

Participants

To determine our sample size, we simulated a latent AR(1) model with measurement noise following the procedure outlined by Schuurman et al. (2015) based on the empirical PA and NA time series reported in the article of Dejonckheere et al. (2019; $n = 1{,}777$). This resulted in a model with a latent AR parameter of 0.45, an innovation error variance of 55.88, and a measurement error variance of 55.27. We imposed that each time series had 70 measurement occasions (which was the number of assessments we would aim for in both protocols), of which 35 occasions were original test ratings, 17 immediate retest ratings (after 1 s) and 18 delayed retest ratings (after 20 min). This power analysis indicated that a minimal sample size of 70 participants (per protocol) ensured a power of 80% to detect meaningful temporal differences between successive test and retest ratings (i.e., using a *t*-test with $\alpha = 0.05$, we were able to detect a significant difference in root-mean-square error between immediate and delayed test–retest discrepancies). To participate in a protocol, no specific inclusion criteria were imposed, except an age of 16 or older (to give autonomous informed consent) and speaking Dutch as a native language (since the study was entirely conducted in Dutch). As a stopping rule, our goal was to end data collection once we included 70 participants per protocol that had a compliance of at least 30% (Delespaul, 1995), but people that were already enrolled at that time could still complete the study. Because all participants met this requirement and no cases had to be excluded, we ended up with a final sample of 91 participants for Protocol 1 (14 males; $M_{age} = 22$, $SD_{age} = 6$) and 96 for Protocol 2 (15 males; $M_{age} = 21$, $SD_{age} = 7$). Both samples consisted of a balanced mix of students and other community members. All participants provided informed consent.

Procedure

Data collection took place between April and May 2020. Because of the COVID-19 pandemic, we contacted voluntary and eligible participants individually via phone call. In an introduction session, we explained the general aim and procedure of the study. To keep participants blind to our actual research questions, we framed the study goal as an attempt to determine different emotional trajectories in the context of daily life. After providing consent, we assisted participants in the installation of our in-house ESM app m-Path (https://www.m-path.io) on their personal smartphone. At the end of the introduction session, they completed the baseline questionnaires. Next, participants took part in one of two ESM protocols, answering short surveys about their current emotional state 10 times per day for 1 week. After 1 week of ESM, participants were contacted again for a short debriefing. They received information about their compliance, and hence, what their resulting compensation was (in research credits [$n = 58$] or a monetary reimbursement [$n = 129$]). Participants received €24 euro or 3 credits if they completed at least 75% of the momentary emotion assessments. A decrease of 5% or 15% was associated with a deduction of €2 or 1 credit, respectively. For Protocol 1, the average compliance was 92% ($SD = 8\%$), for Protocol 2 91% ($SD = 11\%$).

Protocol 1: Measurement Error From Test–Retest Ratings With an Immediate Retest. In the first protocol, we evaluated the test–retest reliability of single-item momentary emotion measurements by randomly repeating one emotion item within the same momentary survey (see Method 1 in Figure 1, for a visual representation). Adopting a time-contingent semirandom sampling approach, we divided participants' waking hours (9 a.m.–7 p.m.) into 10 equal blocks and randomly triggered a momentary emotion assessment within each block. In this way, we could representatively sample an entire day in participants' lives, without the assessments being too predictable in a way that they compromised natural behavior (Dejonckheere & Erbas, 2021). The average time interval between two consecutive assessments was 60 min ($SD = 21$ min). The average time to complete a momentary survey was 42 s ($SD = 58$ s).

Within each momentary emotion survey, both the order of the questions and the selection of the test–retest emotion item were determined completely at random. The presentation of the retest version always concluded the survey, but we guaranteed that participants had to rate at least six other items between the original and repeated emotion item in order to minimize (working) memory effects (Alvarez & Cavanagh, 2004).

Protocol 2: Measurement Error From Test–Retest Ratings With a Delayed Retest. In the second protocol, we evaluated the test–retest reliability of momentary emotion measurements by varying the size of the time interval between two identical emotion surveys (see Method 2 in Figure 1, for a visual representation). Here, we divided participants' waking hours (9 a.m.–7 p.m.) into five equal blocks and randomly triggered a first momentary emotion assessment within each block. After they completed this original survey, participants received a second contingent survey that was sent randomly in the following 20-min time interval (no lower limit was specified). In this way, participants could complete 10 assessments per day in total (five pairs of two linked emotion ratings with varying interval widths). The average time interval between two consecutive original assessments was 104 min ($SD = 44$ min). The average time interval between an original and contingent survey was 10 min ($SD = 6$ min, range = 1 s–20 min). The size of this 20-min time window was based on a trade-off between statistical and empirical concerns. On the one hand, we had to guarantee sufficient variation in observed retest intervals to allow adequate extrapolations. On the other hand, we aimed to limit the possibility

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

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

that people's true emotion score meaningfully changed, as previous research suggests that people usually quickly recover from emotional stimuli in daily life (Kalokerinos et al., 2017; Verduyn et al., 2009).

Measures

Within a momentary emotion survey, we randomly assessed the same 12 emotion items in both protocols. Their selection was based on the circumplex model of affect (Russell, 1980) to ensure a balanced assessment of different valence and arousal levels. Participants rated their momentary levels of six positive (*enthusiastic*, *cheerful*, *happy*, *pleased*, *relaxed*, *at ease*) and six negative (*scared*, *tense*, *frustrated*, *sad*, *bored*, *down*) emotions (*How ___ do you feel at this moment?*) on a continuous slider scale ranging from 0 (*not at all*) to 100 (*very much*). No other items were included in the survey. To minimize the possibility that differences in screen size would impact the amount of measurement error we established for each participant, we visualized the corresponding numerical value of their rating above the measurement scale (see Supplemental Material 5, for a screenshot of the rating interface). Next to these momentary ESM measures, we additionally collected baseline data about three important person-level characteristics: social desirability, conscientiousness, and neuroticism.

Social Desirability. To get an indication of participants' social desirability biases, we used the short version of the Marlowe–Crowne Social Desirability Scale (MC-SDS; Crowne & Marlowe, 1960). This 13-item questionnaire implicitly assesses both people's tendency to overreport desirable behavior (e.g., *I am always to admit it when I make a mistake.*), as well as to underreport undesirable behavior (e.g., *I sometimes feel resentful when I don't get my way.* [reversed]). Participants indicated to what extent they agreed with each statement on a 7-point Likert scale that ranged from 1 (*strongly disagree*) to 7 (*strongly agree*). Internal consistencies for this scale were .75 and .58 for Protocols 1 and 2, respectively. The average MC-SDS score was 4.25 ($SD = 0.77$) for Protocol 1 and 4.13 ($SD = 0.63$) for Protocol 2. Higher scores indicate higher social desirability tendencies.

Conscientiousness. To evaluate participants' levels of conscientiousness, we used the corresponding subscale of the Big Five Inventory (BFI-C; John & Srivastava, 1999). The Conscientiousness subscale consists of nine statements that assess (among other facets) how organized, persistent, diligent, and reliable a person is (e.g., *I see myself as someone who makes plans and follows through with them*; Costa et al., 1991). Participants rated the extent to which each item accurately described their personality on a 5-point Likert scale ranging from 1 (*strongly disagree*) to 5 (*strongly agree*). Internal consistencies for this subscale were .77 and .80 for Protocols 1 and 2, respectively. The mean BFI-C score in our sample was 3.50 ($SD = 0.62$) for Protocol 1 and 3.24 ($SD = 0.83$) for Protocol 2. Higher scores indicate higher conscientiousness levels.

Neuroticism. We also used the BFI (John & Srivastava, 1999) to determine participants' levels of neuroticism. The Neuroticism subscale (BFI-N) is made up of eight items that evaluate (among other aspects) people's tendency toward negative emotions, affective instability and reactivity, and rumination (e.g., *I see myself as someone who gets nervous easily.*). Again, each statement was answered on the same 5-point Likert scale. Internal consistencies for this subscale were .98 and .86 for Protocols 1 and 2, respectively. The mean BFI-N score was 3.04 ($SD = 1.11$) for Protocol 1 and 3.31 ($SD = 0.74$) for Protocol 2. Higher scores indicate higher neuroticism levels.

Statistical Analysis

All statistical analyses are conducted using a multilevel framework because it is a flexible and yet principled approach to deal with the nested and unbalanced data structure. Assumptions associated with this multilevel framework include the residuals of Δ^2 being normally distributed, these residuals being independent from each other, and homogeneity of variance. In the most simple form, to determine the average amount of measurement error in people's momentary mood measurements (across emotion items and individuals), we fitted the following random intercept model for Protocol 1:

$$\begin{aligned} \text{Level1:} \Delta_t^2 &= \beta_{0i} + \pi_{ti}, \\ \text{Level2:}\ \ \beta_{0i} &= \gamma_{00} + \kappa_{0i}. \end{aligned} \tag{9}$$

where Δ_{ti}^2 refers to the squared difference between the test and retest version of a randomly selected emotion item at assessment t (from 1 to 10) for individual i. At the first level, this outcome was modeled as a function of an intercept β_{0i} that was allowed to vary across participants, with γ_{00} at the second level referring to the fixed intercept and κ_0 to the random intercept with $\kappa_{0i} \sim N(0, \tau^2)$. Residual scores are represented by π_{ti} and are assumed to be normally distributed with $\pi_{ti} \sim N(0, \phi^2)$. Because γ_{00} represents the expected value of Δ^2, we derive measurement error variance by dividing this fixed intercept by a Factor 2 (and take the square root of this value to meaningfully interpret measurement error in relation to the actual range of the 0–100 measurement scale).

In a first extension of Equation (9) for Protocol 1, we additionally included Distance_{ti} as a predictor of interest to the model, which represents the number of other emotion items between the randomly selected test and retest item at measurement occasion t for individual i. In this way, we could check whether memory retrieval related to less measurement error variance. If memory effects play a role in Protocol 1, we expect to see that measurement error variance decreases as a function of the number of other interfering emotion ratings between test and retest. Its effect β_{1i} was also allowed to vary across participants, with γ_{00} at the second level referring to the fixed effect and κ_1 to the random effects with $\kappa_{1i} \sim N(0, \upsilon^2)$. Because the original version of this variable could only take values between 7 and 12 (due to the imposed intermission of at least six other items), we grand-mean centered this predictor for an intuitive interpretation of the model's intercept.

For Protocol 2, Δ_{ti}^2 in Equation (9) represents the squared difference in emotion ratings between a pair of two corresponding consecutive emotion measurements t (from 1 to 5) for individual i. Here, at the first level, we additionally included Time_{ti} as a predictor, which embodies the (absolute[1]) time interval in minutes between a pair of two consecutive surveys t for participant i. Its effect β_{1i} again varied across participants, with γ_{10} at the second level referring to

[1] In Supplemental Material 6, we additionally considered other time functions to model the squared differences between participants' test and delayed retest ratings (i.e., the natural logarithm of time, the square root of time, and a linear–quadratic version of time). Specifically, we performed a 10-fold blocked cross-validation procedure to evaluate which time function modeled these discrepancies best. The predictive accuracy of all temporal functions was more or less the same, leading us to report the results of the basic linear time function in the main article for the ease of interpretation. Using different time functions did not drastically impact our conclusions. Nevertheless, because all temporal functions showed comparable model fit, we also performed model averaging to obtain a measurement error variance estimate, and we report this average in the main article as well.

the fixed effect, indicating $E(\Delta^2)$, and κ_1 to the random effects with $\kappa_{1i} \sim N(0, \xi^2)$. To get an indication of the average measurement error variance, we divide the fixed intercept γ_{00} by a Factor 2 (and take the square root to come to an appropriate scale interpretation) because this represents the (hypothetical) situation where Time_{ti} would be zero (and no meaningful emotional change could have occurred).

The Role of Affect Type

Second, for each protocol, we investigated whether differences in measurement error could be observed between emotion items in function of their valence, positive or negative. We present the model structure for Protocol 1 below (for Protocol 2, a person-varying Time_{ti} variable is also included):

$$\text{Level 1}: \Delta^2_{ti} = \beta_{0i} + \beta_{1i}\text{Valence}_{ti} + \pi_{ti}, \quad (10)$$

$$\text{Level 2}: \beta_{0i} = \gamma_{00} + \kappa_{0i}$$
$$\beta_{1i} = \gamma_{10} + \kappa_{1i}.$$

At the first level, Valence_{ti} is a dummy variable that refers to differences between positive and negative emotion items, with the positive valence being the reference level. Its effect β_{1i} was allowed to vary across participants, with γ_{10} at the second level referring to the fixed effect, and κ_1 to the random effects with $\kappa_{1i} \sim N(0, \varsigma^2)$.[2]

Person-Level Covariates

Third, for both protocols, we examined whether three trait variables (separately) moderated the amount of measurement error variance in people's momentary emotion measurements (e.g., Hu et al., 2016). To this end, at the second level of the model in Equation (9), we introduced P_i, which is a grand-mean standardized person-level covariate score that either refers to participants' reported levels of social desirability, conscientiousness, or neuroticism. For each covariate, we investigated its effect γ_1 on the model's intercept. Below, the model structure for Protocol 1 is shown (for Protocol 2, an additional person-varying Time_{ti} variable is included at Level 1):

$$\text{Level 1}: \Delta^2_{ti} = \beta_{0i} + \pi_{ti}, \quad (11)$$

$$\text{Level 2}: \beta_{0i} = \gamma_{00} + \gamma_{01}P_i + \kappa_{0i}.$$

Exploratory Analyses

We ran a few additional analyses that were not preregistered but that were aimed at deepening our understanding of our confirmatory findings. First, in order to detect to what extent measurement error was a function of a rating's location on the bounded measurement scale (Mestdagh et al., 2018), we introduced a predictor Distance to the closest bound_{ti} to the model in Equation (9). This variable indicated the (absolute) numerical distance to the nearest scale end point (from 0 to 50) for the original test emotion item at measurement t for participant i. In both protocols, its effect β_{1i} was allowed to vary across participants, with γ_{10} at the second level referring to the fixed effect and κ_1 to the random effects with $\kappa_{1i} \sim N(0, \iota^2)$.

Second, we also included the $\text{Within-person variance}_i$ of each original test emotion item as a Level 2 predictor in the model of Equation (11) for both protocols. This allowed us to explore to what extent a more variable emotional life and stronger affective shifts relate to more measurement error. To this end, we investigated its effect γ_1 on the model's intercept, representing $E(\Delta^2)$. Additionally, we could calculate how much of the total variability in people's emotional responses could statistically be explained by measurement error variance.

Full Model

In a final model, to examine the robustness of our results, we introduced various focal predictors simultaneously. At the first level, in addition to the random intercept in the model of Equation (9), we either included Distance_{ti} for Protocol 1 or Time_{ti} for Protocol 2, together with Valence_{ti} and Distance to the closest bound_{ti}. At the second level, all predictors were allowed to vary randomly across participants, but only the intercept was a function of participants' grand-mean standardized levels of social desirability, conscientiousness, neuroticism, and within-person variability (simultaneously).

Comparison With a Model-Based Reliability Estimation

Finally, we aimed to anchor our approach in the current time series literature by comparing our obtained measurement error variance estimates for both protocols with the model-based reliability estimation procedure proposed by Schuurman et al. (2015). Following one of the approaches of Schuurman et al. (2015), we used a person-specific latent AR(1) model per emotion to estimate the amount of measurement error variance. For each person, we averaged the estimated measurement error variances across emotions, and correlated these person-level error aggregates with the average-emotion person-specific intercepts in Equation (9) for each protocol. We also compared the mean of their respective distributions.

Results

Table 1 presents a summarizing overview for the results of all the confirmatory multilevel models. First, the significant random intercept for Protocol 1 in Equation (9) indicates that the average squared difference between a random test and retest emotion item within the same momentary survey is 124.78, which corresponds with an estimated measurement error *SD* of 7.90 (i.e., $\sqrt{124.78/2}$; we take the square root for an intuitive interpretation). This means that, on average, on a continuous slider scale from 0 to 100, participants' observed and true emotion scores diverge around 8 measurement points.

Critically, for Protocol 2, the estimation of people's measurement error variance is of a similar magnitude, showing convergence across the two test–retest approaches. Here, we deduce measurement error variance by considering the hypothetical instance where the time interval between two successive assessments would be zero, which is represented by the significant random intercept of 143.00 in Model 1. This squared difference coincides with a measurement error *SD* of 8.46, meaning that, on a 0–100 visual slider, participants' observed

[2] We also used this model to examine whether differences existed between the 12 discrete emotion ratings separately. To this end, we ran a series of 66 different models per protocol, where the dummy variable represented a unique contrast between two emotion items (e.g., *happy* vs. *relaxed*). In this way, we could bypass the need to select an arbitrary reference category, and comprehensively test all 66 possible emotional contrasts. Due to the large number of models, we applied a false discovery rate procedure to control for false positive results (Benjamini & Hochberg, 1995; see Supplemental Material 7).

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

Table 1
Summary of All Confirmatory Multilevel Models

	Protocol 1					Protocol 2				
Predictor	Estimate	*SE*	*t*	95% CI	*p*	Estimate	*SE*	*t*	95% CI	*p*
1. Average measurement error										
Intercept	124.78	4.89	25.54	[115.20, 134.36]	<.001	143.00	15.59	9.17	[112.45, 173.55]	<.001
Time						5.55	0.84	6.62	[3.91, 7.19]	<.001
Intercept	126.43	9.41	13.44	[107.99, 144.88]	<.001					
Distance	6.68	2.85	2.35	[1.10, 12.26]	.019					
2. The role of affect type										
Intercept	124.51	10.02	12.43	[104.87, 144.16]	<.001	153.06	18.18	8.42	[117.42, 188.69]	<.001
Time						5.48	0.83	6.62	[3.86, 7.10]	<.001
Valence	4.69	12.06	0.39	[−18.96, 28.33]	.698	−17.80	8.99	−1.98	[−35.41, −0.18]	.048
3. Person-level covariates										
Intercept	124.88	4.88	25.59	[115.32, 134.63]	<.001	143.04	15.36	9.31	[112.93, 173.15]	<.001
Time						5.55	0.84	6.62	[3.91, 7.19]	<.001
Social desirability	−19.06	4.84	−3.94	[−28.55, −9.57]	<.001	−27.69	13.36	−2.07	[−53.88, −1.50]	.038
Intercept	125.05	4.88	25.60	[115.48, 134.63]	<.001	143.09	15.63	9.15	[112.45, 173.72]	<.001
Time						5.55	0.84	6.62	[3.90, 7.19]	<.001
Conscientiousness	−10.83	4.81	−2.25	[−20.26, −1.40]	.024	−5.16	13.71	−0.38	[−32.04, 21.71]	0.707
Intercept	125.07	4.88	25.65	[115.51, 134.63]	<.001	142.97	15.53	9.21	[112.54, 173.41]	<.001
Time						5.55	0.84	6.63	[3.91, 7.20]	<.001
Neuroticism	23.58	4.82	4.89	[14.13, 33.03]	<.001	−5.60	13.66	−0.41	[−32.38, 21.18]	.682

Note. CI = confidence interval; *SE* = standard error.

emotion ratings deviate, on average, around 8 points from their actual emotion score. When performing model averaging for the various time functions (see Supplemental Material 6), the mean extrapolated intercept is 105.86. This corresponds with a measurement error *SD* of 7.28 (95% CI [5.94, 8.54]).

In addition, the significant effect of *Time* for Model 1 in Protocol 2 indicates that the squared difference between two identical emotion ratings increases as a function of larger interval widths (see Figure 1, for a graphical representation of these results). This suggests that the likelihood that an emotional stimulus alters an individual's consecutive emotion rating increases as the window between the two assessments becomes larger. For Protocol 1, we establish a similar pattern for the significant predictor *Distance*. As the number of other emotion ratings between the test and retest item grows, measurement error increases (although the effect is rather small with only 2 points per item). This could indicate that an effect of memory decreases as participants have to rate more emotion items between the test and retest item.

The Role of Affect Type

With respect to differences between PA and NA items, both protocols provide mixed results (see Figure 2, for a graphical representation). For Protocol 1, Table 1 does not reveal significant differences in measurement error between these valences. For Protocol 2, however, momentary NA items (which are usually low in nonclinical samples) were associated with slightly less measurement error (an average difference of 3 points).

Person-Level Covariates

Third, we evaluated the moderating impact of participants' levels of social desirability, conscientiousness, and neuroticism, respectively. To this end, we examined the main effect of these person-level variables on the random intercept in Table 1. For Protocol 1, in line with our hypotheses, we observed significant moderations for all three characteristics. That is, individuals with stronger social desirability tendencies and higher levels of conscientiousness showed slightly less measurement error in their emotion ratings (a reduction of 3 and 2 points per *SD*, respectively). In contrast, more neurotic participants showed more measurement error in their responses (an increase of 3 points per *SD*). For Protocol 2, however, we only partly confirmed this pattern of results. Similar to Protocol 1, higher social desirability was associated with less measurement error (a decrease of 4 points per *SD*). Here, conscientiousness and neuroticism were unrelated to measurement error.

Exploratory Results

To get a deeper understanding of our confirmatory findings, we ran a couple of exploratory models of which the results are presented in Table 2. First, we explored whether the amount of measurement error associated with an emotion rating was a function of the specific location on the bounded measurement scale. In both protocols, we observed a significant and positive effect of *distance to the closest bound*, meaning that ratings closer to the scale end points were associated with less measurement error (i.e., floor or ceiling values; see Figure 3, for a real visual example). In both protocols, per 10 points closer to one of the bounds of the measurement scale, measurement error significantly decreased around 8 points.

Second, we explored whether stronger emotional variability was associated with more measurement error. In both protocols, a higher *within-person variance* of an emotion predicted more measurement error in that item.

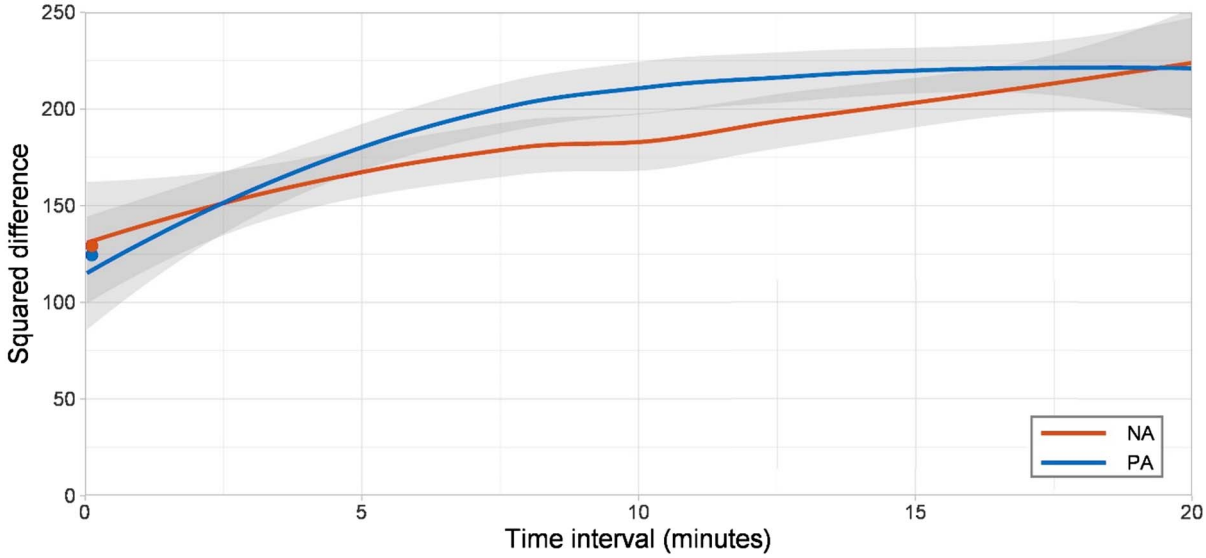
Full Model

In a final model, we included all relevant variables simultaneously to evaluate the robustness of our findings and to determine the most focal predictors (see Table 3). For Protocol 1, we only observed significant effects for *distance, distance to the closest bound*, and *within-person variance*. As the distance between the test and retest emotion items expanded, test ratings were placed more toward the middle of the scale, and people showed more variability in their

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

Figure 2

Nonsignificant (Protocol 1) and Significant (Protocol 2) Differences in Measurement Error Between Positive and Negative Emotion Items



Note. The gray band refers to the 95% confidence interval. PA = positive affect; NA = negative affect. See the online article for the color version of this figure.

emotional time series, measurement error variance increased. Similar to the basic model, we observed no significant differences between PA and NA. At the second level, the three person-level covariates did no longer moderate the amount of measurement error in participants' affective ratings when entered simultaneously.

For Protocol 2, we observed similar significant effects for *within-person variance* and *distance to the closest bound*. Stronger emotional variability and ratings closer to the scale's midpoint were related to more measurement error in participants' emotion ratings. In contrast to the basic model, an effect of affect type was no longer observed. At the second level, only people's neuroticism scores were related to measurement error.

Comparison With a Model-Based Reliability Estimation

Finally, the results of our comparison with the model-based measurement error estimates by Schuurman et al. (2015) are summarized in Table 4. As mentioned earlier, the test–retest measurement error *SD* determinations in both protocols converged around 8 points on a 0–100 measurement scale. Similarly, the model-based reliability estimations converged in both protocols but were both significantly higher than the test–retest ones (around 10 points on a 0–100 measurement scale; bootstrapped confidence intervals did not overlap). In terms of explained variance, measurement error made up around 27% of the total variability in people's affective responses following our test–retest approach, while this percentage increased to 42.5% following the model-based approach (which is of similar magnitude compared to the empirical application of the AR model reported in Schuurman et al., 2015). Nevertheless, both the test–retest and model-based measurement error estimations correlated considerably across participants in both protocols (r's $\geq .71$, p's $< .001$).

General Discussion

Based on an extension of the traditional CTT framework (DeVellis, 2006; Lord et al., 1968), the aim of this article was to introduce two

Table 2

Summary of All Exploratory Multilevel Models

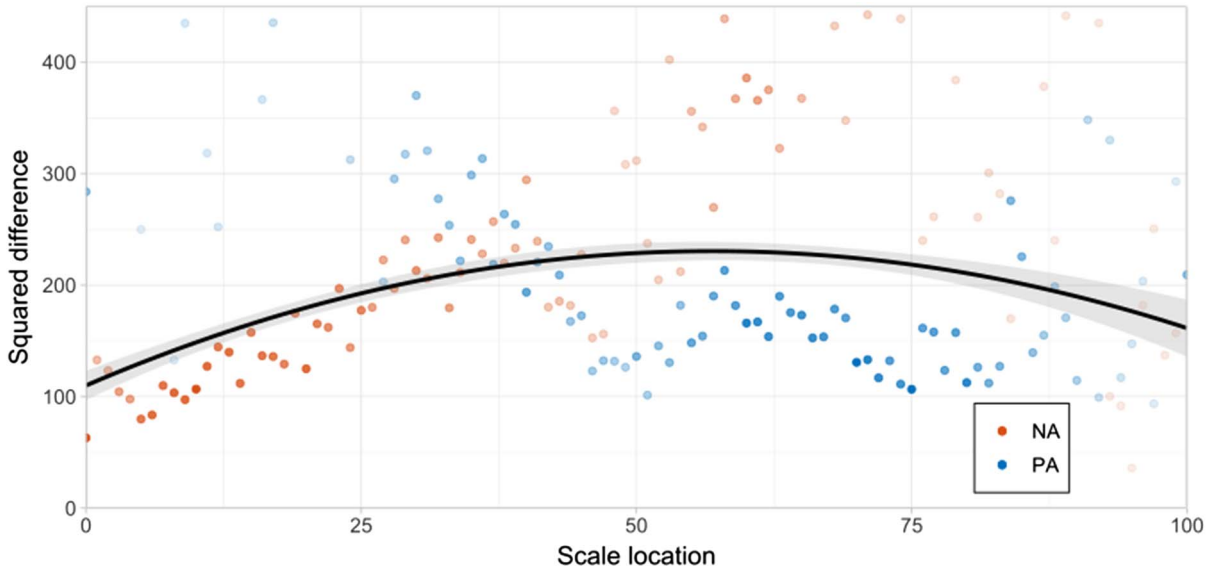
Predictor	Protocol 1					Protocol 2				
	Estimate	*SE*	*t*	95% CI	*p*	Estimate	*SE*	*t*	95% CI	*p*
1. Role of rating position										
Intercept	100.87	13.10	7.70	[75.19, 126.56]	<.001	120.81	19.69	6.13	[82.21, 159.41]	<.001
Time						5.47	0.84	6.54	[3.83, 7.11]	<.001
Distance to the closest bound	1.14	0.40	2.85	[0.36, 1.93]	.004	1.35	0.42	3.21	[0.52, 2.18]	.001
2. Role of emotional variability										
Intercept	125.87	4.81	26.17	[116.44, 135.29]	<.001	114.50	13.49	8.49	[88.06, 140.94]	<.001
Time						5.56	0.84	6.61	[3.91, 7.21]	<.001
Within-person variance	69.53	4.86	14.31	[60.01, 79.06]	<.001	28.68	2.25	12.77	[24.28, 33.08]	<.001

Note. CI = confidence interval; *SE* = standard error.

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

Figure 3

Squared Differences Between Test and Retest Items as a Function of Scale Location, for the PA (Blue) and NA Items (Red) in Protocol 2



Note. Emotion ratings closer to one of the scale's bounds are associated with less measurement error. The transparency of each ratings indicates the relative frequency of a selected scale position, with more opaque ratings representing higher frequencies. PA = positive affect; NA = negative affect. See the online article for the color version of this figure.

adapted test–retest methods that allow researchers to evaluate the reliability of time-varying single-item momentary affective measures in ESM (Shrout & Lane, 2012). In a first approach, we proposed to randomly repeat a single emotion item within the same momentary survey, which minimizes the possibility that the latent emotional intensity score meaningfully changes but runs the risk of introducing other systematic biases, such as memory effects or striving for consistency. In a second approach, we suggested to repeat the entire survey and to randomly vary the short time interval between two consecutive assessments. Here, as an approximation of measurement error variance, researchers should consider the hypothetical instance where a time window between two momentary surveys does not exist. Compared to the first approach, memory or consistency biases are less likely in this protocol, but the possibility that participants' retest score is altered by a meaningful event or situation cannot entirely be ruled out. In a comprehensive analysis, we (a) established on the basis of common assumptions in CTT that manifest test–retest discrepancies mathematically relate to measurement error variance and (b) implemented both approaches in two empirical ESM studies to estimate the actual measurement error variance that is associated with people's single-item momentary emotion ratings.

Crucially, in our empirical ESM studies, both approaches converged on the amount of measurement error they established (see intercepts in Figure 1). In absolute numbers, participants' observed ratings, on average, diverged around 8 measurement points from their true emotion scores in both protocols, which corresponds with almost a 10th of the 0–100 visual slider scale. To illustrate this with an (extreme) example, it is possible that even though an individual's true emotion score of 50 remains constant across two measurement occasions, we nevertheless observe a change from 34 to 66 in that

Table 3

Summary of the Full Multilevel Model

Predictor	Protocol 1					Protocol 2				
	Estimate	*SE*	*t*	95% CI	*p*	Estimate	*SE*	*t*	95% CI	*p*
Intercept	91.50	14.29	6.40	[63.48, 119.52]	.001	112.10	20.91	5.36	[71.10, 153.09]	<.001
Time						5.30	0.83	6.36	[3.67, 6.94]	<.001
Distance	6.59	2.82	2.34	[1.07, 12.11]	.019					
Valence	18.66	11.93	1.56	[−4.73, 42.06]	.118	−4.62	9.87	−0.47	[−23.96, 14.71]	.639
Distance to the closest bound	1.11	0.42	2.66	[0.29, 1.92]	.008	0.02	0.01	1.98	[0.00,0.04]	.048
Within-person variance	66.60	6.31	10.56	[54.23, 78.96]	.001	29.69	2.37	12.53	[25.05, 34.33]	<.001
Social desirability	−10.20	8.93	−1.14	[−27.70, 7.30]	.253	−12.15	10.44	−1.16	[−32.61, 8.31]	.244
Conscientiousness	−8.04	8.22	−0.98	[−24.14, 8.06]	.328	1.81	10.01	0.18	[−17.82, 21.43]	.857
Neuroticism	−6.82	8.23	−0.83	[−22.95, 9.32]	.407	−23.97	10.30	−2.33	[−44.15, −3.78]	.020

Note. CI = confidence interval; *SE* = standard error.

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

Table 4
Comparison Between Test–Retest and Model-Based Reliability Approaches

Approach	*M SD* [95% CI]	% total variance	*r* model approach
Test–retest approach			
Protocol 1—$\sqrt{(\Delta^2/2)}$ between test and retest item within the same survey	7.90 [7.29, 8.47]	26%	.72***
Protocol 2—$\sqrt{(\Delta^2/2)}$ between consecutive surveys with varying interval widths	8.46 [7.50, 9.38]	28%	.71***
Model-based approach			
Protocol 1—latent AR(1) model with white noise following Schuurman et al. (2015)	10.15 [9.66, 10.70]	40%	
Protocol 2—latent AR(1) model with white noise following Schuurman et al. (2015)	10.55 [10.05, 11.04]	45%	

Note. CI = confidence interval; these were retrieved via a bootstrap procedure (1,000 random resamples). AR = autoregressive.
*** $p < .001$.

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

person's manifest emotion intensity rating that is entirely the result of measurement error.[3] In relative numbers, we observed how measurement error made up around 27% of the total variance observed in people's emotion ratings. From a traditional CTT perspective, this percentage would correspond with a reliability of people's emotion ratings of about .73 at the within-person level (Shrout & Lane, 2012).

Although this percentage is smaller than the 30%–50% estimated by Schuurman et al. (2015), the size of this fraction illustrates that simply disregarding measurement error in ESM is far from inconsequential. At a measurement level, our findings underscore that accurately assessing affective states in the moment is challenging, and that overly relying on single-item measures paves the way for unreliable measurements (Sarstedt & Wilczynski, 2009). In turn, at a modeling level, these noisy emotion ratings may produce biased model parameters and unreliable time dynamic summaries. Specifically, measurement error typically weakens meaningful relations between variables, resulting in an underestimation of the true association in the population (Spearman, 1904). Consequently, at the level of interpretation, inaccurate model parameters and correlations may lead to ambiguous findings that are harder to replicate (Cronbach et al., 1963; Parsons et al., 2019). Theoretically, for example, this cascade has been put forward in emotion research to explain why many complex affect dynamic measures only show little association with psychological well-being compared to mean levels of PA and NA (and variability in these affective states; Dejonckheere et al., 2019). The estimation of these more complex dynamic summaries is much more prone to random distortions and measurement noise than more robust metrics such as the mean, which weakens their potentially meaningful relations with person-level well-being variables (Jaso et al., 2021; Mestdagh & Dejonckheere, 2021; Wendt et al., 2020). Practically, this chain of events constitutes an important barrier to effectively implement real-time therapeutic interventions in clinical care. Poor data quality may invalidate accurate model development, often undermining the adaptive and timely nature of real-life micro-interventions (e.g., Kim et al., 2019).

Within-Person and Between-Person Differences in Measurement Error

Although the amount of measurement error associated with a particular momentary assessment is expected to be completely random (Bland & Altman, 1996), we observed how the size of its distortion was both a function of item and person characteristics. First, regarding item characteristics, our results pointed toward a differentiating role of affect type in Protocol 2 (see Figures 2 and 3). Here, negative momentary emotion scores were associated with slightly less measurement error compared to positive ones. In search of an underlying rationale, exploratory analyses revealed that the typical (bounded) scale location of PA and NA item scores could be an explanation for this effect (Mestdagh et al., 2018; Vanbelle & Lesaffre, 2018). While ratings of positive feelings in healthy community samples are usually dispersed around the middle of the measurement scale (allowing considerable room for distortions), negative feelings are less common, which results in ratings closer to the lower scale end point (drastically restricting the size of possible distortions; Kalokerinos et al., 2020). Indeed, follow-up investigations in both protocols indicated that a rating's relative position to one of the bounds significantly impacted how much measurement error we observed. Momentary emotion ratings closer to one of the scale's end points (usually NA items) were found to be more reliable. This observation contradicts one of the common assumptions in CTT where researchers assume that random measurement errors are always normally distributed (e.g., Cappelleri et al., 2014). In the case of bounded measurements, our findings demonstrate that this assumption can never be met, as the degree with which measurement noise follows a normal distribution actually depends on the specific (bounded) scale location of the latent emotion score. That being said, compared to the absolute amount of measurement error variance, it should be emphasized that differences in measurement error between positive and negative emotion items and the effect of scale location were actually relatively small to be considered substantively meaningful (an average difference between PA and NA items of 3 points maximum on a 0–100 measurement scale and an average increase of 4 points in measurement error per 5 points closer to the scale's midpoint, respectively). Moreover, we did not establish a differential effect of affect type in Protocol 1.

More consistently, in both the single and full models of Protocols 1 and 2, we observed how the ratings of emotion items with a higher within-person variance suffered from more measurement error. Although this observation may seem trivial (because measurement error is an important source of within-person variance in psychological time series; Schuurman & Hamaker, 2019), exploratory follow-up analyses also revealed that measurement error positively correlated with the latent variability of an emotion item. Again, this association is typically not expected in standard reliability research, because CTT assumes that people's true scores and random errors are entirely independent (e.g., Cappelleri et al., 2014).

[3] If $\epsilon \sim N(0, \sigma_\epsilon^2)$, then the corresponding 95% confidence interval covers [0 – 1.96 × σ_ϵ; 0 + 1.96 × σ_ϵ], which ranges from 34 to 66.

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

Second, regarding person characteristics, results were less consistent, but we observed the most robust effects for social desirability and neuroticism. In Protocols 1 and 2, both the single person-level models demonstrated that participants who tended to present themselves in a more socially desirable way provided more reliable emotion responses. Still, although their responses were less contaminated by measurement error (e.g., because they consistently aimed to answer to a favorable norm; Borkenau & Ostendorf, 1989), we cannot rule out that this tendency not biased the truthfulness of their latent emotional responses (e.g., because they exaggerated their positive feelings; Dejonckheere & Bastian, 2021). As such, this pattern of results is a good illustration of the fact that adequate reliability is a necessary, but not sufficient condition for validity (Du, 2010). Nevertheless, in both protocols, social desirability no longer predicted individual differences in measurement error once its effect was evaluated in the full model, together with the other person-level characteristics.

For neuroticism, the single model in Protocol 1 showed how measurement error in people's emotion ratings was higher as they were more emotionally unstable (for the full model in Protocol 2, we observed a reverse association due to multicollinearity with people's within-person variance). In statistical terms, this observation nicely aligns with the previous finding that measurement error positively relates to an emotion's (latent) within-person variance, and that the size of distortion differs as a function of usual scale location (with neurotic people showing NA levels further away from the bound; Kalokerinos et al., 2020). In terms of a process explanation, highly neurotic people may have more difficulty to actually access their true latent emotion score due to reduced emotional insight (Thompson et al., 2015), which hinders the reliability of their corresponding emotion ratings. Having neurotic people take part in a similar ESM study where participants additionally rate other subjective experiences (e.g., job performance, motivation, or relational conflicts) may reveal whether reduced emotional clarity is indeed the underlying mechanism.

Finally, for conscientiousness, we only observed a significant predictive effect in the single model of Protocol 1. In line with our hypothesis, highly conscientious people provided emotion ratings that were more reliable. In Protocol 2, however, this effect was not replicated nor did it survive significance testing in both full models when the explanatory power of the other person-level covariates was taken into account.

Taken together, in terms of hypothesis testing, our research provides rather mixed results on the role of these person-level characteristics. In terms of effect sizes, however, we can again decisively conclude that their effects were rather small (compared to the absolute amount of measurement error variance), with an increase of maximum 4 points in measurement error on a 0–100 scale per added *SD* to the person-level covariate score. Again, this begs the question to what extent these person-level differences are truly worth considering when exploring the confounding role of measurement error in an ESM study on emotions.

Test–Retest Versus Model-Based Reliability Estimations

Finally, we observed how our test–retest approach yields explicit measurement error variance estimations that are considerably smaller than the more implicit ones we retrieved from the dynamic model-based approach proposed by Schuurman et al. (2015). Although these new, higher reliability estimates may seem promising, it is impossible to offer a conclusive verdict about which approach is the most accurate, as participants' true emotion scores will always remain unobserved. That being said, the reason why the model-based reliability approach likely yields measurement error estimates that are remarkably higher than the test–retest ones is because of the way its latent emotion-generating model is defined (Rhemtulla et al., 2020). Specifically, when the hypothesized underlying dynamic model cannot adequately capture the (complex) regularities in people's emotional responding, the model interprets these changes automatically as measurement error (Schuurman & Hamaker, 2019; see Supplemental Material 1, for a simulated graphical example). Taken together, different latent model operationalizations of measurement error produce different reliability estimates. An operationalization of measurement error variance that is not explicitly model-based does not make assumptions about the underlying dynamic process and will, therefore, always produce the same reliability estimates independent of the assumed latent process. Noteworthy, it is possible to extend an AR(1) model to handle retest data; however, assumptions regarding the underlying dynamic process will remain.

Recommendations for Future Research

When heavily relying on single-item constructs, we advise researchers to occasionally include a randomly selected retest item in the surveys of their future ESM protocol. It could be argued that answering an additional item further increases participant burden, but we believe that this small cost does not outweigh the possibility to check the consistency of people's momentary responses and to remove certain measurement occasions or participants if they present highly unreliable data (e.g., when the signal-to-noise ratio is smaller than 1; Dejonckheere & Mestdagh, 2021). To further fine-tune the trade-off between limiting participant burden and evaluating single-item reliability, we argue that it is not required to include a retest item every measurement occasion, but that it suffices to present them semirandomly throughout the protocol. On the one hand, this guarantees sampling representativeness and avoids temporal effects (Dejonckheere & Erbas, 2021). On the other hand, this softens certain participants' striving for consistency, as exact item repetitions are not always part of every measurement occasion. Furthermore, in the case of discrete emotions, given the absence of sizable differences in measurement error variance between items in our protocols, we believe that reliability evaluations for one emotion item may be appropriately generalized to another. Finally, in the case of multi-item (emotion) constructs, internal consistency procedures (e.g., Hu et al., 2016; Nezlek et al., 2017) may be preferred over the current test–retest approach, which makes the inclusion of retest items redundant.

Limitations

The current pattern of results should be evaluated in light of the following limitations that put a constraint on the generality of our findings. First, within affective ESM research, the present measurement error determinations cannot be easily applied to other single-item momentary affective constructs. Although the results in Supplemental Figure 7 suggest that there are little meaningful differences between various (narrow) discrete emotion items, our findings may not necessarily generalize to single-item affect measures

that intend to assess mood more broadly (e.g., *how positive/negative are you feeling right now?*; Sayed et al., 2021). Future research will need to investigate whether these ratings are associated with significantly more (e.g., because their diffuse nature sparks ambiguity) or less measurement error (e.g., because it is easier to remember the score of a single mood item). Relatedly, for other disciplines that rely on ESM, we cannot make claims about the amount of measurement error that is associated to the momentary ESM ratings of other psychological phenomena (e.g., psychopathological symptoms), assessed with other types of survey questions (e.g., 7-point Likert scales), in other participant groups (e.g., clinical samples with higher NA levels). We always advise researchers to establish the reliability of their single-item ratings for their own specific study, since this estimate is not inherently tied to the instrument under study but also depends on specific sample characteristics or test conditions (Hammersley, 1987; Thompson, 2003).

Second, it is worth reiterating that participants could see the corresponding numerical value of their rating when providing their answer on the slider scale. While this decision was intended to mitigate the potentially confounding impact of different mobile device screen sizes, this could have facilitated consistency in participants' ratings (especially in Protocol 1, where we observed how the amount of measurement error was a function of the number of other items between the test and retest emotion rating, but possibly also in Protocol 2). If so, these memory effects likely lead to a slight underestimation of the measurement error in people's emotion measurements. However, it is equally possible that exact item repetitions could have induced annoyance, reactivity, or confusion in some participants, introducing more measurement error in their emotional ratings. Future research will need to explore these possibilities.

Conclusion

Until now, due to a combination of different methodological challenges, measurement error of single-item measures has been largely ignored in the ESM literature. However, the present results demonstrate that the amount of measurement error variance in emotional time series is not negligible. Because reliable measures are the first important hurdle to take in order to make scientific progress in daily life research, we therefore encourage ESM researchers to introduce at least one randomly repeated retest item in the flow of their momentary surveys when they rely on single-item measures.

References

Alvarez, G. A., & Cavanagh, P. (2004). The capacity of visual short-term memory is set both by visual information load and by number of objects. *Psychological Science*, *15*(2), 106–111. https://doi.org/10.1111/j.0963-7214.2004.01502006.x

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B. Methodological*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bland, J. M., & Altman, D. G. (1996). Measurement error. *BMJ*, *312*(7047), Article 1654. https://doi.org/10.1136/bmj.312.7047.1654

Borkenau, P., & Ostendorf, F. (1989). Descriptive consistency and social desirability in self- and peer reports. *European Journal of Personality*, *3*(1), 31–45. https://doi.org/10.1002/per.2410030105

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, *3*(3), 296–322. https://doi.org/10.1111/j.2044-8295.1910.tb00207.x

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

Cappelleri, J. C., Jason Lundy, J., & Hays, R. D. (2014). Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures. *Clinical Therapeutics*, *36*(5), 648–662. https://doi.org/10.1016/j.clinthera.2014.04.006

Costa, P., Jr., McCrae, R., & Dye, D. (1991). Facet scales for agreeableness and conscientiousness: A revision of the NEO personality inventory. *Personality and Individual Differences*, *12*(9), 887–898. https://doi.org/10.1016/0191-8869(91)90177-D

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology*, *16*(2), 137–163. https://doi.org/10.1111/j.2044-8317.1963.tb00206.x

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, *24*(4), 349–354. https://doi.org/10.1037/h0047358

Dejonckheere, E., & Bastian, B. (2021). Perceiving social pressure not to feel negative is linked to a more negative self-concept. *Journal of Happiness Studies*, *22*(2), 667–679. https://doi.org/10.1007/s10902-020-00246-4

Dejonckheere, E., & Erbas, Y. (2021). How to design an ESM study? In I. Myin-Germeys & P. Kuppens (Eds.), *The open handbook of experience sampling methodology: A step-by-step to designing, conducting, and analyzing ESM studies* (pp. 33–69). Amazon Publishing.

Dejonckheere, E., & Mestdagh, M. (2021). On the signal-to-noise ratio in emotional time series. In C. Waugh & P. Kuppens (Eds.), *Affect dynamics* (pp. 131–152). Springer International Publishing. https://doi.org/10.1007/978-3-030-82965-0_7

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

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

Dejonckheere, E., Mestdagh, M., Kuppens, P., & Tuerlinckx, F. (2020). Reply to: Context matters for affective chronometry. *Nature Human Behaviour*, *4*(7), 690–693. https://doi.org/10.1038/s41562-020-0861-6

Delespaul, P. A. E. G. (1995). *Assessing schizophrenia in daily life*. Universitaire Pers Maastricht.

DeVellis, R. F. (2006). Classical test theory. *Medical Care*, *44*(Suppl. 11), S50–S59. https://doi.org/10.1097/01.mlr.0000245426.10853.30

Du, R. Y. (2010). *Research reliability and validity*. Wiley. https://doi.org/10.1002/9781444316568.wiem02008

Eid, M., & Diener, E. (1999). Intraindividual variability in affect: Reliability, validity, and personality correlates. *Journal of Personality and Social Psychology*, *76*(4), 662–676. https://doi.org/10.1037/0022-3514.76.4.662

Eisele, G., Vachon, H., Lafit, G., Kuppens, P., Houben, M., Myin-Germeys, I., & Viechtbauer, W. (2022). The effects of sampling frequency and questionnaire length on perceived burden, compliance, and careless responding in experience sampling data in a student population. *Assessment*, *29*(2), 136–151. https://doi.org/10.1177/1073191120957102

Epskamp, S., Cramer, A., Waldorp, L., Schmittmann, V., & Borsboom, D. (2012). Qgraph: Network visualizations of relationships in psychometric data. *Journal of Statistical Software*, *48*(4), 1–18. https://doi.org/10.18637/jss.v048.i04

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

Frijda, N. H. (1988). The laws of emotion. *American Psychologist*, *43*(5), 349–358. https://doi.org/10.1037/0003-066X.43.5.349

Ghiselli, E. E. (1963). Moderating effects and differential reliability and validity. *Journal of Applied Psychology*, *47*(2), 81–86. https://doi.org/10.1037/h0047177

Guttman, L. (1945). A basis for analyzing test–retest reliability. *Psychometrika*, *10*(4), 255–282. https://doi.org/10.1007/BF02288892

Hammersley, M. (1987). Some notes on the terms 'validity' and 'reliability.'. *British Educational Research Journal*, *13*(1), 73–81. https://doi.org/10.1080/0141192870130107

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

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

Hu, Y., Nesselroade, J. R., Erbacher, M. K., Boker, S. M., Burt, S. A., Keel, P. K., Neale, M. C., Sisk, C. L., & Klump, K. (2016). Test reliability at the individual level. *Structural Equation Modeling*, *23*(4), 532–543. https://doi.org/10.1080/10705511.2016.1148605

Jaso, B. A., Kraus, N. I., & Heller, A. S. (2021). Identification of careless responding in ecological momentary assessment research: From posthoc analyses to real-time data monitoring. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000312

John, O., & Srivastava, S. (1999). The big five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (pp. 102–138). Guilford Press.

Kalokerinos, E. K., Murphy, S. C., Koval, P., Bailen, N. H., Crombez, G., Hollenstein, T., Gleeson, J., Thompson, R. J., Van Ryckeghem, D. M. L., Kuppens, P., & Bastian, B. (2020). Neuroticism may not reflect emotional variability. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(17), 9270–9276. https://doi.org/10.1073/pnas.1919934117

Kalokerinos, E. K., Résibois, M., Verduyn, P., & Kuppens, P. (2017). The temporal deployment of emotion regulation strategies during negative emotional episodes. *Emotion*, *17*(3), 450–458. https://doi.org/10.1037/emo0000248

Kim, J., Marcusson-Clavertz, D., Yoshiuchi, K., & Smyth, J. M. (2019). Potential benefits of integrating ecological momentary assessment data into mHealth care systems. *BioPsychoSocial Medicine*, *13*(1), Article 19. https://doi.org/10.1186/s13030-019-0160-5

Kobayashi, T., & Boase, J. (2012). No such effect? The implications of measurement error in self-report measures of mobile communication use. *Communication Methods and Measures*, *6*(2), 126–143. https://doi.org/10.1080/19312458.2012.679243

Kuppens, P., & Verduyn, P. (2017). Emotion dynamics. *Current Opinion in Psychology*, *17*, 22–26. https://doi.org/10.1016/j.copsyc.2017.06.004

Larson, R., & Csikszentmihalyi, M. (1983). The experience sampling method. *New Directions for Methodology of Social & Behavioral Science*, *15*, 41–56. https://psycnet.apa.org/record/1984-00092-001

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), Article 6218. https://doi.org/10.1038/s41598-021-85320-4

Lord, F. M., Novick, M. R., & Birnbaum, A. (1968). *Statistical theories of mental test scores*. Addison-Wesley.

Lowman, G. H., Wood, D., Armstrong, B. F., Harms, P. D., & Watson, D. (2018). Estimating the reliability of emotion measures over very short intervals: The utility of within-session retest correlations. *Emotion*, *18*(6), 896–901. https://doi.org/10.1037/emo0000370

McArdle, J. J., & Woodcock, R. W. (1997). Expanding test–retest designs to include developmental time-lag components. *Psychological Methods*, *2*(4), 403–435. https://doi.org/10.1037/1082-989X.2.4.403

Mestdagh, M., & Dejonckheere, E. (2021). Ambulatory assessment in psychopathology research: Current achievements and future ambitions. *Current Opinion in Psychology*, *41*, 1–8. https://doi.org/10.1016/j.copsyc.2021.01.004

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

Moskowitz, D. S., & Young, S. N. (2006). Ecological momentary assessment: What it is and why it is a method of the future in clinical psychopharmacology. *Journal of Psychiatry & Neuroscience*, *31*(1), 13–20. https://psycnet.apa.org/record/2007-03393-002

Nezlek J. B. (2017). A practical guide to understanding reliability in studies of within-person variability. *Journal of Research in Personality*, *69*, 149–155. https://doi.org/10.1016/j.jrp.2016.06.020

Nezlek, J. B., Newman, D. B., & Thrash, T. M. (2017). A daily diary study of relationships between feelings of gratitude and well-being. *The Journal of Positive Psychology*, *12*(4), 323–332. https://doi.org/10.1080/17439760.2016.1198923

Nichols, A. L., & Maner, J. K. (2008). The good-subject effect: Investigating participant demand characteristics. *The Journal of General Psychology*, *135*(2), 151–165. https://doi.org/10.3200/GENP.135.2.151-166

Parsons, S., Kruijt, A. W., & Fox, E. (2019). Psychological science needs a standard practice of reporting the reliability of cognitive-behavioral measurements. *Advances in Methods and Practices in Psychological Science*, *2*(4), 378–395. https://doi.org/10.1177/2515245919879695

Pe, M. L., & Kuppens, P. (2012). The dynamic interplay between emotions in daily life: Augmentation, blunting, and the role of appraisal overlap. *Emotion*, *12*(6), 1320–1328. https://doi.org/10.1037/a0028262

Phillips, G. W., & Jiang, T. (2016). Measurement error and equating error in power analysis. *Practical Assessment, Research & Evaluation*, *21*, Article 9. https://doi.org/10.7275/0snn-zm67

Rhemtulla, M., van Bork, R., & Borsboom, D. (2020). Worse than measurement error: Consequences of inappropriate latent variable measurement models. *Psychological Methods*, *25*(1), 30–45. https://doi.org/10.1037/met0000220

Russell, J. (1980). A Circumplex Model of Affect. *Journal of Personality and Social Psychology*, *39*(6), 1161–1178. https://doi.org/10.1037/h0077714

Santangelo, P., Ebner-Priemer, U., & Trull, T. (2013). Experience sampling methods in clinical psychology. In J. S. Comer & P. C. Kendall (Eds.), *The Oxford handbook of research strategies for clinical psychology* (pp. 188–210). Oxford University Press.

Sarstedt, M., & Wilczynski, P. (2009). More for less? A comparison of single-item and multi–item measures. *Die Betriebswirtschaft*, *69*(2), 211–227. https://www.proquest.com/docview/750491216?pq-origsite=gscholar&fromopenview=true

Sayed, F., McGowan, A. L., Jovanova, M., Cosme, D., Kang, Y., Stanoi, O., Mucha, P., Ochsner, K. N., Bassett, D. S., Falk, E., & Lydon-Staley, D. M. (2021). *Momentary associations between affect and alcohol use in the daily lives of college students*. PsyArXiv. https://doi.org/10.31234/osf.io/a7cgu

Schuurman, N. K., & Hamaker, E. L. (2019). Measurement error and person-specific reliability in multilevel autoregressive modeling. *Psychological Methods*, *24*(1), 70–91. https://doi.org/10.1037/met0000188

Schuurman, N. K., Houtveen, J. H., & Hamaker, E. L. (2015). Incorporating measurement error in n = 1 psychological autoregressive modeling. *Frontiers in Psychology*, *6*, 1038. https://doi.org/10.3389/fpsyg.2015.01038

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.

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

Shrout, P. E., & Lane, S. P. (2012). Psychometrics. In M. R. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 302–320). The Guilford Press.

Spearman, C. (1904). 'General intelligence,' objectively determined and measured. *The American Journal of Psychology*, *15*(2), 201–293. https://doi.org/10.2307/1412107

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, *3*(3), 271–295. https://doi.org/10.1111/j.2044-8295.1910.tb00206.x

Steyer, R., Mayer, A., Geiser, C., & Cole, D. A. (2015). A theory of states and traits—Revised. *Annual Review of Clinical Psychology*, *11*(1), 71–98. https://doi.org/10.1146/annurev-clinpsy-032813-153719

Thompson, B. (2003). *Score reliability: Contemporary thinking on reliability issues*. Sage Publications. https://doi.org/10.4135/9781412985789

Thompson, R. J., Kuppens, P., Mata, J., Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Gotlib, I. H. (2015). Emotional clarity as a function of neuroticism and major depressive disorder. *Emotion*, *15*(5), 615–624. https://doi.org/10.1037/emo0000067

Trull, T. J., & Ebner-Priemer, U. W. (2009). Using experience sampling methods/ecological momentary assessment (ESM/EMA) in clinical assessment and clinical research: Introduction to the special section. *Psychological Assessment*, *21*(4), 457–462. https://doi.org/10.1037/a0017653

Trull, T. J., & Ebner-Priemer, U. W. (2020). Ambulatory assessment in psychopathology research: A review of recommended reporting guidelines and current practices. *Journal of Abnormal Psychology*, *29*(1), 56–63. https://doi.org/10.1037/abn0000473

Van der Gucht, K., Dejonckheere, E., Erbas, Y., Takano, K., Vandemoortele, M., Maex, E., Raes, F., & Kuppens, P. (2019). An experience sampling study examining the potential impact of a mindfulness-based intervention on emotion differentiation. *Emotion*, *19*(1), 123–131. https://doi.org/10.1037/emo0000406

Vanbelle, S., & Lesaffre, E. (2018). Modeling agreement on bounded scales. *Statistical Methods in Medical Research*, *27*(11), 3460–3477. https://doi.org/10.1177/0962280217705709

Verduyn, P., Delvaux, E., Van Coillie, H., Tuerlinckx, F., & Van Mechelen, I. (2009). Predicting the duration of emotional experience: Two experience sampling studies. *Emotion*, *9*(1), 83–91. https://doi.org/10.1037/a0014610

Verhagen, S. J., Hasmi, L., Drukker, M., van Os, J., & Delespaul, P. A. (2016). Use of the experience sampling method in the context of clinical trials. *Evidence-Based Mental Health*, *19*(3), 86–89. https://doi.org/10.1136/ebmental-2016-102418

Weidman, A. C., Steckler, C. M., & Tracy, J. L. (2017). The jingle and jangle of emotion assessment: Imprecise measurement, casual scale usage, and conceptual fuzziness in emotion research. *Emotion*, *17*(2), 267–295. https://doi.org/10.1037/emo0000226

Wendt, L. P., Wright, A. G. C., Pilkonis, P. A., Woods, W. C., Denissen, J. J. A., Kühnel, A., & Zimmermann, J. (2020). Indicators of affect dynamics: Structure, reliability, and personality correlates. *European Journal of Personality*, *34*(6), 1060–1072. https://doi.org/10.1002/per.2277

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

Wilhelm, P., & Schoebi, D. (2007). Assessing mood in daily life structural validity, sensitivity to change, and reliability of a short-scale to measure three basic dimensions of mood. *European Journal of Psychological Assessment*, *23*(4), 258–267. https://doi.org/10.1027/1015-5759.23.4.258

Received October 21, 2021
Revision received July 19, 2022
Accepted August 9, 2022 ■

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