Psychometrika (2026), **0**, 1–32
doi:10.1017/psy.2026.10133

PSYCHOMETRIC SOCIETY

THEORY AND METHODS

Single-Item Reliability for Intensive Longitudinal Data: A Three-Level Linear Mixed-Effect Autoregressive Modeling Approach

Tzu-Yao Lin[1,2], Francis Tuerlinckx[1] and Sophie Vanbelle[3]

[1]Research Group of Quantitative Psychology and Individual Differences, Faculty of Psychology and Educational Sciences, KU Leuven, Leuven, Belgium; [2]Department of Methodology and Statistics, Faculty of Health, Medicine and Life Sciences, Maastricht University, Maastricht, Netherlands; [3]Department of Methodology and Statistics, Faculty of Psychology and Neuroscience, Maastricht University, Maastricht, Netherlands

Corresponding author: Tzu-Yao Lin; Email: tzu-yao.lin@kuleuven.be

(Received 2 December 2025 UTC; revised 18 June 2026 UTC; accepted 24 June 2026 UTC)

Abstract

With the advancement of technology, intensive longitudinal data (ILD) are increasingly collected to study dynamic, moment-to-moment psychological and physiological processes. Despite their widespread use, the issues of measurement error and reliability are often inadequately addressed, particularly in the presence of serial correlation and complex multilevel structures. This article introduces a three-level linear mixed-effects model that explicitly accounts for nested structures (moments within days within persons) and incorporates two distinct autoregressive processes for moment- and day-level serial correlations, partitioning measurement error from within-person variability. Building on this model, we delineate two approaches to defining reliability: a variance partitioning approach and a correlation approach. This framework yields a multifaceted set of reliability coefficients, enabling researchers to distinguish between-person from within-person reliability, and to assess measurement agreement across different time scales (e.g., moment-to-moment and day-to-day). We further discuss the relationship between the proposed reliability measures and those derived from simpler models. We apply our model to study the reliability of a single negative affect item in an ILD study in which students were measured multiple times a day for several days before and after receiving their exam results.

Keywords: autoregressive process; longitudinal data; measurement error; multilevel model

1. Introduction

The analysis of intensive longitudinal data (ILD) is a rapidly growing area in behavioral and social sciences (Ariens et al., 2020; Bolger & Laurenceau, 2013; Hamaker, 2025; Hamaker et al., 2015; Walls & Schafer, 2006). Its primary aim is to understand processes and behaviors as they unfold over time. Unlike traditional longitudinal or panel data, which typically involve a few evenly spaced observations (e.g., annually over a couple of years), or time series studies that focus on a single unit, ILD are collected from multiple units at many time points (e.g., tens, hundreds, or sometimes even thousands). ILD studies are characterized by a relatively high temporal resolution compared to the time frame of the study (e.g., hourly assessments during daytime over a week), allowing researchers to closely track changes

F.T. and S.V. shared last authorship.

© The Author(s), 2026. Published by Cambridge University Press on behalf of Psychometric Society.
This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

Downloaded from https://www.cambridge.org/core. 13 Nov 2026 at 03:02:29, subject to the Cambridge Core terms of use.

over time and analyze patterns and dynamics that might be missed in studies with less frequent measurements (Walls & Schafer, 2006).

In psychology, ILD are often collected using methods, such as the experience sampling method (ESM; Tuerlinckx et al., 2025), ecological momentary assessment (EMA; Stone & Shiffman, 1994), or ambulatory assessment (AA; Trull & Ebner-Priemer, 2020), which capture real-time data on individuals' experiences, behaviors, and physiological states in naturalistic settings. For example, researchers in the life sciences and psychology can nowadays easily collect massive amounts of data in real-time using devices used to monitor health behaviors (e.g., medication adherence and physical activity), psychological states (e.g., mood and stress), and physiological responses (e.g., heart rate and sleep patterns).

An example of an ILD study that we will further analyze in detail in this article is the Exam Results Study by Kalokerinos et al. (2019; see also the next section). In this study, a sample of first-year university students self-reported their negative affect on a visual analog scale multiple times per day for nine consecutive days, some days before and after receiving their exam results (which is, for most students, a stressful and important event).

Based on the example, we can describe some key features of ILD that have implications for the way the data are being analyzed. First, it inherently possesses a multilevel structure. For instance, measurements (Level 1) might be repeatedly collected within the same day (Level 2), with multiple such assessment days for each individual (Level 3). Consequently, modeling ILD necessitates the distinction and estimation of variance components associated with these different hierarchical levels, such as between-person, between-day (or within-person), and within-day variances. Second, due to the high temporal resolution typical of ILD, successive observations are often correlated. This characteristic contrasts with traditional modeling approaches that often assume independence among observations.

Despite the rising popularity of studies with ILD, evaluating the reliability of measurements in ILD is often neglected or carried out inappropriately. This holds both for the reliability of measurements with single items (Dejonckheere et al., 2022; Schuurman & Hamaker, 2019; Schuurman et al., 2015) and for measurements with multiple indicators (Brose et al., 2020; Calamia, 2019; Castro-Alvarez et al., 2025; Marcoulides, 2019). The use of a single item to measure constructs in ILD is driven by the need to minimize participant burden from frequent assessments (Dejonckheere et al., 2022; Wilhelm & Schoebi, 2007). Obviously, a critical liability of single-item measures is their heightened susceptibility to measurement error (Sarstedt & Wilczynski, 2009). Although any observed score is unavoidably tainted by measurement error, averaging over several observed scores tends to improve the signal-to-noise ratio and thus reliability. However, this is not an option for single-item measures, and therefore, the evaluation of reliability is all the more important.

In addition, the two aforementioned key characteristics of ILD studies (i.e., the multilevel nature and time dependency) provide a further challenge to the accurate evaluation of reliability. First, addressing the multilevel structure is essential for accurate reliability estimation in ILD. There are approaches to estimate reliability at each level of the multilevel structure, such as generalizability theory (GT; Cranford et al., 2006; Lin et al., 2025; Nezlek, 2017; Shrout & Lane, 2012b; ten Hove et al., 2022) and multilevel SEM (Castro-Alvarez, Tendeiro, de Jonge, et al., 2022; Geldhof et al., 2014; Koch et al., 2025; Lai, 2021; Nacke & Mayer, 2025; Xiao et al., 2023). However, in practice, many researchers still resort to grabbing a single global reliability metric (e.g., Cronbach's alpha) or only consider between-person indices, improperly ignoring the within-person variance. This oversight can lead to inaccurate and potentially misleading estimates of reliability (Calamia, 2019; Castro-Alvarez et al., 2025; Schuurman & Hamaker, 2019; Vogelsmeier et al., 2024).

Second, estimating reliability in (intensive) longitudinal studies is particularly challenging because conventional methods (e.g., test–retest) presume that reliability remains static over time. However, in ILD, this assumption is often violated due to the dynamic nature of the constructs being measured and the presence of serial correlation among repeated measurements. Consequently, the serial correlation among repeated measurements is often neglected when estimating reliability (Marcoulides, 2019; Molenberghs et al., 2007; Revelle & Condon, 2019; Vangeneugden et al., 2004, 2005). While this

Downloaded from https://www.cambridge.org/core. 13 Nov 2026 at 03:02:29, subject to the Cambridge Core terms of use.

oversight might have a negligible impact on estimating the internal consistency of multi-item measures (e.g., Castro-Alvarez et al., 2025), it can lead to biased reliability estimates for the test–retest reliability of single-item measures, because the systematic variance due to temporal dependency is incorrectly attributed to either true score variance or measurement error. Therefore, establishing and reporting appropriate reliability metrics is an imperative step before utilizing ILD measures in practice. It is essential to develop and refine methods for estimating reliability that are tailored to the complex, dynamic, multilevel data structures inherent in intensive longitudinal studies.

To address these issues, reliability can no longer be a single coefficient (Castro-Alvarez et al., 2025; Tisak & Tisak, 1996; Vangeneugden et al., 2004); instead, it should be assessed at multiple levels and across different time scales. In this article, two distinct yet related approaches will be used to define reliability coefficients in ILD: the correlation approach (e.g., Bartko, 1966; Lin et al., 2025; Lord et al., 1968; Shrout & Fleiss, 1979; Vangeneugden et al., 2004, 2005) and the variance partitioning approach (e.g., Castro-Alvarez, Tendeiro, de Jonge, et al., 2022; Hunter, 2026; Laenen et al., 2007; Laenen et al., 2009b; Schuurman & Hamaker, 2019; Xiao et al., 2023). The first approach views reliability as the consistency or agreement of measurements across different occasions (e.g., measurements of different persons and measurements of the same person at different time points). The variance partitioning approach conceptualizes reliability as a variance ratio, addressing the extent to which an instrument can distinguish systematic or target variance from measurement error. We will formally introduce these two approaches in the later sections.

Our multifaceted approach results in various reliability measures that capture the properties of measurements from two distinct perspectives: (1) between-person versus within-person level and (2) stability across different time spans. For the first perspective, at the between-person level, reliability quantifies how well the measurement instrument distinguishes among individuals, separating true trait differences from measurement error. At the within-person level, it assesses how well the instrument tracks an individual's true state fluctuations over time. The second perspective concerns temporal similarity (i.e., test–retest reliability) across different time lags, such as moment-to-moment within the same day, day-to-day, or even longer intervals. While the relation between the correlation and variance partitioning approaches is not clear at first sight because of differences in starting points and equations, we will analyze their interrelatedness and provide an integrated framework.

To sum up, the goal of this article is to define and estimate various reliability measures for the single-item measurements in ILD, taking into account its multilevel structure and serial correlation among repeated measurements. To achieve this, we propose to use a general three-level linear mixed-effect model (LMM) with the autoregressive process to appropriately distinguish different sources of variance. Based on our illustration example (the Exam Results Study), specifically, these variance components may come from between-person differences, within-person differences across days, serial correlation at the day level and the moment level, and measurement error. Building upon this model, we then define and interpret reliability at different hierarchical levels and time spans by the correlation and variance partitioning approaches. Furthermore, we will demonstrate what the reliability coefficients will be when our approach is applied to a number of commonly used models in (intensive) longitudinal research that are special cases of our general three-level model, including the two-level LMM with/without autoregressive processes and the three-level LMM without autoregressive processes. This systematic evaluation will reveal how different model specifications determine the resulting single-item reliability coefficients and will provide researchers with a comprehensive understanding of how to assess reliability in various ILD contexts. Finally, these approaches will be illustrated on the Exam Results Study using a Bayesian framework.

2. Motivating example: The Exam Results Study

In the Exam Results Study, Kalokerinos et al. (2019) investigated the dynamics of affect of 101 Belgian first-year psychology students at KU Leuven (87 women; age: M = 18.6; SD = 1.5) before and after they received their first-semester exam results (covering five courses). The primary focus of the original study

Downloaded from https://www.cambridge.org/core. 13 Nov 2026 at 03:02:29, subject to the Cambridge Core terms of use.

was on how emotion regulation differs across individuals. Such emotion regulation strategies need to be employed when receiving these exam results. These results are particularly significant for freshman students, because the first year is considered to be a selection year in the Belgian university system. Thus, receiving the exam results is a highly personally relevant and real-life event compared to experimental studies of emotional regulation.

In this study, each student (level 3) was tracked for nine consecutive days (level 2) using experience sampling. Within each day, a momentary survey was administered using a stratified random interval scheme from 10 a.m. to 10 p.m. (level 1). The total number of measurement times is 90. The exam results were released on the third day between the 21st and 28th measurement times. Students self-reported their affect on a visual analog scale ranging from 0 (= not at all) to 100 (= very much) for various items related to specific emotions and emotion-regulation strategies. In this article, we focus on the single item measuring negative affect: "Please indicate how negative you are feeling right now." The average compliance rate, defined as the percentage of completed momentary assessments for each student, was generally high, with a mean of 90.8% (SD = 7.4%, min = 64.4%, max = 100%).

A researcher collecting these data or a critical consumer of this research may wonder: (1) To what extent can persons be distinguished from one another over an extended period of time? (2) To what extent would repeating the measurement at exactly the same given moment yield a similar result? and (3) How does the agreement between measurements change across varying time intervals? These questions illustrate different reliability assessments that can be computed for this data set.

3. Three-level linear mixed-effect model with autoregressive processes

Univariate LMMs have been widely used in ILD studies to model single-item measurements while taking into account the multilevel structure and the serial correlation (Lodewyckx et al., 2011; Ma et al., 2020; Nezlek, 2007; Vansteelandt & Verbeke, 2016; Verbeke & Molenberghs, 2000; Walls & Schafer, 2006). As an analysis model for the Exam Results Study, we propose to use the model proposed by Ma et al. (2020). This is a three-level LMM along with autocorrelated innovations at both the moment and day level.[1]

Let y_{ijk} be the observed measurement (e.g., a negative affect rating) for person $i = 1, \ldots, N$ on day $j = 1, \ldots, D$ at moment $k = 1, \ldots, M$, assuming a balanced design.[2] The model is defined by the following three levels:

$$\text{level 1 (moment): } y_{ijk} = \mu_{ij} + \omega_{ijk} + \varepsilon_{ikj} \tag{3.1}$$

$$\text{level 2 (day): } \mu_{ij} = \mu_i + v_{ij} + d_{ij} \tag{3.2}$$

$$\text{level 3 (person): } \mu_i = \beta_0 + s_i. \tag{3.3}$$

At level 1, the observation y_{ijk} is modeled as three components. First, μ_{ij} is a mean state of person i on day j. Second, $\omega_{ijk} \sim \mathcal{N}(0, \tau_m^2)$ is a moment-to-moment serial process effect. It captures within-day serial dependence and is modeled as an AR(1) process, where τ_m^2 is the stationary variance of the process. The covariance structure of ω's on the same day is $\mathrm{Cov}(\omega_{ijk}, \omega_{ijk'}) = \tau_m^2 \phi_m^{|k-k'|}$,[3] where $|\phi_m| < 1$ is the

[1] We generalize the model of Ma et al. (2020) by separating the moment-level autocorrelated innovations from measurement error. Ma et al. (2020) keep both together because they consider both to be errors.

[2] For simplicity, we assume a balanced design (i.e., everyone has the same number of measurements), but the model specification can be extended to unbalanced designs. In our practice, we apply the Bayesian framework and handle unbalanced or missing data by treating the missing observations as parameters to be estimated.

[3] There are two extensions regarding the serial process possible: (a) Unequal time spans between moments are allowed: $\mathrm{Cov}(\omega_{ijk}, \omega_{ijk'}) = \tau_m^2 \phi_m^{|t_k - t_{k'}|}$, where t_k and $t_{k'}$ represent the exact time of each moment k and k', respectively. (b) The serial correlation function can be assumed more generally as $\rho_i(|k-k'|)$, which is a decreasing function of time lag, satisfying $\rho_i(0) = 1$ and $\rho_i(\infty) = 0$. The AR(1) process is a special case of this general form, $\rho_i(|k-k'|) = \phi_m^{|k-k'|}$ with $|\phi_m| < 1$. For a more general serial correlation structure, we refer to Verbeke and Molenberghs (2000).

Downloaded from https://www.cambridge.org/core. 13 Nov 2026 at 03:02:29, subject to the Cambridge Core terms of use.

autoregressive parameter, indicating how much carry-over there is from the previous moment to the current one. Third, $\varepsilon_{ijk} \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ is the momentary measurement error. The errors are assumed to be mutually independent, and the error variance σ_ε^2 (similarly for τ_m^2) is assumed to be homogeneous across all time points (moments and days) and persons.

At level 2, the daily mean state μ_{ij} is also decomposed into three components. First, μ_i is the mean state of person i. Second, $v_{ij} \sim \mathcal{N}(0, \tau_d^2)$ represents a day-level AR(1) serial process across consecutive days, where τ_d^2 is the stationary variance of the serial process. The covariance structure of v's is $\mathrm{Cov}(v_{ij}, v_{ij'}) = \tau_d^2 \phi_d^{|j-j'|}$, where $|\phi_d| < 1$ is the autoregressive parameter for the day level. By specifying the autoregressive process, similar to level 1, we account for the systematic carry-over effect from one day to the next. Third, $d_{ij} \sim \mathcal{N}(0, \sigma_d^2)$ is a random effect for day (nested within person i). For simplicity, the variance of the day-specific deviations, σ_d^2, is assumed to be homogeneous across days and persons.[4]

Finally, at level 3, the person-mean μ_i is modeled as the sum of β_0, an overall grand mean (a fixed effect) for the population, and $s_i \sim \mathcal{N}(0, \sigma_s^2)$, a random effect for person i. Note that all three levels are assumed to be mutually independent.

Combining Equations (3.1)–(3.3) yields the single-equation form of the model:

$$y_{ijk} = \beta_0 + s_i + d_{ij} + v_{ij} + \omega_{ijk} + \varepsilon_{ijk}. \tag{3.4}$$

3.1. Marginal model and covariance structure

In general, Equation (3.4) can be expressed as a vectorized expression for each person i as follows:

$$\boldsymbol{y}_i = \boldsymbol{X}_i \boldsymbol{\beta}_i + \boldsymbol{Z}_i^{(s)} \boldsymbol{s}_i + \boldsymbol{Z}_i^{(d)} \boldsymbol{d}_i + \boldsymbol{Z}_i^{(v)} \boldsymbol{v}_i + \boldsymbol{\omega}_i + \boldsymbol{\varepsilon}_i, \quad i = 1, \ldots, N, \tag{3.5}$$

where $\boldsymbol{y}_i = (y_{i11}, \ldots, y_{i1M}, y_{i21}, \ldots, y_{iDM})^\top$ is a vector with all observed measurements for person i. $\boldsymbol{X}_i$, $\boldsymbol{Z}_i^{(s)}$, $\boldsymbol{Z}_i^{(d)}$, and $\boldsymbol{Z}_i^{(v)}$ are the design matrices[5] for the fixed effect (vector) $\boldsymbol{\beta}_i = \beta_0$ and the person random effect (vector) $\boldsymbol{s}_i = s_i$, person random effect (vector) $\boldsymbol{d}_i = (d_i, \ldots, d_D)^\top$, and the day-level serial effect (vector) $\boldsymbol{v}_i = (v_{i1}, \ldots, v_{iD})^\top$, respectively. $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i^{(s)}$ are a column vector of ones with length DM, and $\boldsymbol{Z}_i^{(d)}$ and $\boldsymbol{Z}_i^{(v)}$ are $\underset{D\times D}{\boldsymbol{I}} \otimes \underset{M\times 1}{1}$, where $\boldsymbol{I}$ is an identity matrix with diagonal elements equal to one, and zero elsewhere, and 1 is a column vector of ones. The operator $\otimes$ denotes the Kronecker product. Under the balanced design assumption, everyone has the same number of measurements, so the design matrices $\boldsymbol{X}_i$, $\boldsymbol{Z}_i^{(s)}$, $\boldsymbol{Z}_i^{(d)}$, and $\boldsymbol{Z}_i^{(v)}$ are the same for all persons i. The remaining vectors $\boldsymbol{\omega}_i = (\omega_{i11}, \ldots, \omega_{i1M}, \omega_{i21}, \ldots, \omega_{iDM})^\top$ and $\boldsymbol{\varepsilon}_i = (\varepsilon_{i11}, \ldots, \varepsilon_{i1M}, \varepsilon_{i21}, \ldots, \varepsilon_{iDM})^\top$ collect moment-level serial effects and measurement errors, respectively. Compared to a standard LMM formulation, Equation (3.5) explicitly distinguishes the serial processes $\boldsymbol{v}_i$ and $\boldsymbol{\omega}_i$ from the measurement error $\boldsymbol{\varepsilon}_i$, to stress that the fluctuation of within-person observations could come from two personal dynamical processes instead of coming from the measurement random errors (Diggle, 1988; Ma et al., 2020; Verbeke & Molenberghs, 2000).

[4] This assumption can be relaxed to allow heterogeneous day variance (i.e., σ_{dj}^2) to account for potential day-specific variability, such as modeling different variances for weekdays versus weekends. In practice, however, we may impose some constraints on σ_{dj}^2's to avoid potential identification issues. For instance, we can assume two different variances for weekdays and weekends or assume a linear or quadratic trend across days.

[5] Unlike the conventional notation for LMMs (Diggle, 1988; Verbeke & Molenberghs, 2000), we explicitly split the design matrix $\boldsymbol{Z}$ into three parts, $\boldsymbol{Z}^{(s)}$, $\boldsymbol{Z}^{(d)}$, and $\boldsymbol{Z}^{(v)}$, to distinguish the design matrix for the person-level random effect s_i and the day-related effects d_{ij}'s and v_{ij}'s. This distinction is helpful when discussing the variance components and reliability definitions later.

Downloaded from https://www.cambridge.org/core. 13 Nov 2026 at 03:02:29, subject to the Cambridge Core terms of use.

The total (co)variance of the marginal model (Equation (3.5)), denoted $\boldsymbol{V} = \text{Var}(\boldsymbol{y}_i)$, is the sum of the covariance matrices of the random components:

$$\begin{aligned}\boldsymbol{V} &= \text{Var}(\boldsymbol{y}_i) \\ &= \boldsymbol{Z}_i^{(s)} \text{Var}(\boldsymbol{s}_i) \boldsymbol{Z}_i^{(s)\top} + \boldsymbol{Z}_i^{(d)} \text{Var}(\boldsymbol{d}_i) \boldsymbol{Z}_i^{(d)\top} + \boldsymbol{Z}_i^{(v)} \text{Var}(\boldsymbol{v}_i) \boldsymbol{Z}_i^{(v)\top} + \text{Var}(\boldsymbol{\omega}_i) + \text{Var}(\boldsymbol{\varepsilon}_i) \\ &= \boldsymbol{\Sigma}_S + \boldsymbol{\Sigma}_D + \boldsymbol{\Sigma}_v + \boldsymbol{\Sigma}_\omega + \boldsymbol{\Sigma}_E.\end{aligned} \tag{3.6}$$

Each component in this decomposition represents a distinct source of variation: $\boldsymbol{\Sigma}_S = \sigma_s^2 \underset{DM\times DM}{\boldsymbol{J}}$, where $\boldsymbol{J}$ is a matrix with all elements equal to 1, represents the person-specific covariance matrix, accounting for the between-person variation between their effects; $\boldsymbol{\Sigma}_D = \sigma_d^2 \underset{D\times D}{\boldsymbol{I}} \otimes \underset{M\times M}{\boldsymbol{J}}$ represents the day random effect variance; $\boldsymbol{\Sigma}_v = \tau_d^2 \underset{D\times D}{\boldsymbol{H}_d} \otimes \underset{M\times M}{\boldsymbol{J}}$ represents the covariance between day-level serial process, following the AR(1) structure; $\boldsymbol{\Sigma}_\omega = \tau_m^2 \underset{D\times D}{\boldsymbol{I}} \otimes \underset{M\times M}{\boldsymbol{H}_m}$ represents the covariance between moment-level serial process; and, lastly, $\boldsymbol{\Sigma}_E = \underset{D\times D}{\boldsymbol{I}} \otimes \sigma_\varepsilon^2 \underset{M\times M}{\boldsymbol{I}} = \sigma_\varepsilon^2 \underset{DM\times DM}{\boldsymbol{I}}$.

The distinction between the day-related components, $\boldsymbol{\Sigma}_D$ and $\boldsymbol{\Sigma}_v$, becomes clearer when examining their underlying structures at level 2 (see Equation (3.2)) and examining their respective covariance matrix structures:

$$\text{Var}(\boldsymbol{d_i}) = \sigma_d^2 \underset{D\times D}{\boldsymbol{I}} = \begin{pmatrix} \sigma_d^2 & 0 & \cdots & 0 \\ 0 & \sigma_d^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_d^2 \end{pmatrix} \quad \text{and} \quad \boldsymbol{H}_d = \begin{pmatrix} 1 & \phi_d & \cdots & \phi_d^{D-1} \\ \phi_d & 1 & \cdots & \phi_d^{D-2} \\ \vdots & \vdots & \ddots & \vdots \\ \phi_d^{D-1} & \phi_d^{D-2} & \cdots & 1 \end{pmatrix}.$$

$\boldsymbol{H}_d$ depicts the autocorrelation of daily fluctuation based on the AR(1) process. In contrast, $\text{Var}(\boldsymbol{d_i})$ captures the residual variance of these daily fluctuations that is not captured by the serial correlation.

Similarly, the difference between $\boldsymbol{\Sigma}_\omega$ and $\boldsymbol{\Sigma}_E$, both at level 1 (see Equation (3.1)), can be understood by examining their covariance matrix structures:

$$\boldsymbol{H}_m = \begin{pmatrix} 1 & \phi_m & \cdots & \phi_m^{M-1} \\ \phi_m & 1 & \cdots & \phi_m^{M-2} \\ \vdots & \vdots & \ddots & \vdots \\ \phi_m^{M-1} & \phi_m^{M-2} & \cdots & 1 \end{pmatrix} \quad \text{and} \quad \text{Var}(\boldsymbol{\varepsilon}_{ij}) = \sigma_\varepsilon^2 \underset{M\times M}{\boldsymbol{I}} = \begin{pmatrix} \sigma_\varepsilon^2 & 0 & \cdots & 0 \\ 0 & \sigma_\varepsilon^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_\varepsilon^2 \end{pmatrix}.$$

In this context, $\boldsymbol{H}_m$ depicts the autocorrelation of momentary fluctuation based on the AR(1) process, while $\text{Var}(\boldsymbol{\varepsilon}_{ij})$ captures the residual variance at each moment that is not explained by the serial correlation or other variation accounted for in the model.

Finally, the total variance of observations for a specific person i, conditional on that person, can be expressed as

$$\boldsymbol{V}_i = \text{Var}(\boldsymbol{y}_i \mid i) = \boldsymbol{\Sigma}_D + \boldsymbol{\Sigma}_v + \boldsymbol{\Sigma}_\omega + \boldsymbol{\Sigma}_E, \tag{3.7}$$

for $i = 1, \ldots, N$. Conditioning on i emphasizes that the focus is solely on person i, thereby removing the person variance component in Equation (3.7). This distinction is fundamental for defining reliability later: the total variance $\boldsymbol{V}$ will be used when discussing the between-person reliability, while the person-specific variance $\boldsymbol{V}_i$ will be used when discussing the within-person reliability.

4. Reliability for intensive longitudinal studies

Before discussing how to define reliability for ILD, we briefly revisit the concept of reliability in the context of classical test theory (CTT). We will introduce two complementary approaches to define and interpret reliability: the variance partitioning and correlation approaches. We then introduce the

Downloaded from https://www.cambridge.org/core. 13 Nov 2026 at 03:02:29, subject to the Cambridge Core terms of use.

reliability coefficients under the LMM framework defined in Section 3 based on these two approaches, extending from the existing literature.

4.1. Reliability in CTT

The concept of reliability in psychometrics originates from CTT (Lord et al., 1968). In CTT, it is assumed that the observed score y_{it} for individual i at time[6] t can be decomposed into an individual true score $T_i = \mathrm{E}[y_{it}]$ and a measurement error ε_{it}, that is, $y_{it} = T_i + \varepsilon_{it}$. Assuming that the true score and error are uncorrelated and that errors are mutually independent, the variance of the observed score can be partitioned as $\sigma_y^2 = \sigma_T^2 + \sigma_\varepsilon^2$, where σ_T^2 is the true score variance, and σ_ε^2 is the measurement error variance. Consequently, reliability is defined as the squared correlation (or coefficient of determination) between the observed score and the true score across the population:

$$\rho_{yT}^2 = \mathrm{Cor}(y_{it}, T_i)^2 = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_\varepsilon^2}, \tag{4.1}$$

which simplifies to the ratio of the true score variance to the total observed score variance.

This definition can be seen as an instance of a more general principle: the *variance partitioning approach*. Under this approach, the total variance of the observations is separated into a systematic and an unsystematic part, and then reliability is defined as a ratio of the systematic variance over the total variance (Cranford et al., 2006; Cronbach, 1972; Ghiselli et al., 1981; Guttman, 1945; Laenen et al., 2007; Laenen et al., 2009a, 2009b; Lord et al., 1968; McDonald, 1999; McGraw & Wong, 1996; Schuurman & Hamaker, 2019; Shrout & Lane, 2012a; ten Hove et al., 2022). Systematic variance refers to score changes that are attributed to identifiable and often predictable influences—such as stable individual differences or an orderly pattern change in some trend. In contrast, unsystematic variance stems from random, unpredictable fluctuations, including noise characterized by a complete absence of order. This unsystematic variance constitutes the primary source of measurement error. Thus, the general formula for reliability under this approach is

$$R = \frac{\sigma_{\text{sys}}^2}{\sigma_{\text{sys}}^2 + \sigma_{\text{unsys}}^2} = 1 - \frac{\sigma_{\text{unsys}}^2}{\sigma_{\text{total}}^2}, \tag{4.2}$$

where the subscripts sys and unsys indicate the systematic and unsystematic variance components, respectively. Within the CTT framework, the systematic variance is the true score variance, and the unsystematic variance is the error variance. Consequently, $R = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_\varepsilon^2} = \rho_{yT}^2$. Conceptually, this coefficient quantifies the ability of a measurement instrument to differentiate between individuals despite the presence of measurement errors across the population. Assume two persons, A and B, and the observed score of A is higher than that of B. The higher the reliability, the more confidence we have in inferring that person A's true score is also larger than person B's true score.

Alternatively, there is a second approach to define reliability, denoted as the *correlation approach*. Under this approach, one defines reliability as a correlation between two parallel measurements of the same individuals (e.g., repeated administrations of the same test or administrations of strictly parallel forms) (Bartko, 1966; Lin et al., 2025; Lord et al., 1968; Molenberghs et al., 2007; Shrout & Fleiss, 1979; Vangeneugden et al., 2004, 2005). Because the true score and its variance are usually unavailable, researchers often define reliability based on parallel measurements (real or hypothetical), say y_{it} and $y_{it'}$, $t \neq t'$. Under CTT, parallel measurements of the same individual mean having equal true score ($T_{it} = T_{it'}$, i.e., tau-equivalent) and error variance ($\mathrm{Var}(\varepsilon_{it}) = \mathrm{Var}(\varepsilon_{it'})$). The reliability coefficient is

[6] Because the focus of this article is longitudinal data, we will introduce the CTT with an index for time.

Downloaded from https://www.cambridge.org/core. 13 Nov 2026 at 03:02:29, subject to the Cambridge Core terms of use.

defined as the correlation between two parallel measurements across the population, that is,

$$\rho_{t,t'} = \mathrm{Cor}(y_{it}, y_{it'}) = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_\varepsilon^2} = \rho_{yT}^2. \tag{4.3}$$

This correlation is also called intraclass correlation (ICC; Bartko, 1966; McGraw & Wong, 1996; Shrout & Fleiss, 1979).[7] The viewpoint of the correlation approach emphasizes the "agreement"[8] between any pairs of replicates from the same individual across the population. Reliability is called test–retest reliability if the parallel measurements y_{it} and $y_{it'}$ are obtained by replications over time.[9] The correlation approach has a clear and direct interpretation because it tells us how closely two measurements agree.

Comparing Equations (4.2) and (4.3), it is clear that the two approaches lead to the same reliability coefficient under CTT, that is, $\rho_{yT}^2 = \rho_{t,t'} = R$. The variance partitioning approach is further generalized to study designs beyond CTT through the GT (Brennan & Kane, 1977a, 1977b; Cronbach, 1972),[10] while the correlation approach keeps the same name. To underscore that the correlation and variance partitioning approach are two sides of the same coin, Bartko (1966) stated that a reliability coefficient, when defined as a ratio of variances estimated through a linear model, is only valid when it can be interpreted as a form of correlation coefficient. This interpretation ensures that the reliability reflects the degree of similarity between repeated measurements rather than being a purely statistical artifact. However, we have shown that both approaches do not always lead to the same reliability coefficients in the context of multilevel data structures (Lin et al., 2025).

4.2. Reliability in ILD: Challenges

Extending the concept of single-item reliability to ILD is not straightforward for several reasons. First, in CTT or GT as in multilevel reliability studies, the true score is assumed to be a stable attribute of an individual, constant over time or across different measurement occasions (Cranford et al., 2006; Nezlek, 2017; Shrout & Lane, 2012a). In contrast, in ILD, the true score of individuals may vary over time due to situational factors, temporal dynamics, or other influences. This within-person fluctuation in true scores is a key feature of ILD, and it often reflects meaningful patterns of change that researchers are interested in. Second, this within-person fluctuation also generates new opportunities, because it allows us to distinguish the within-person variance from the between-person variance. Third, intensive longitudinal studies often involve additional sources of variability (e.g., day effects) that cannot be accounted for in CTT. Therefore, the traditional framework or methods for multilevel reliability analysis may not

[7] Note that the reliability coefficients proposed in this study conceptually could belong to the broad family of ICCs. Specifically, they represent the expected correlation between two (or more) measurements of the same subject (or same class) under specific conditions. However, the term "ICC" has been defined in the literature through various lenses—ranging from the strict correlation approach (e.g., Bartko, 1966; Lin et al., 2025; Shrout & Fleiss, 1979; Vangeneugden et al., 2004, 2005) to variance partitioning approaches (e.g., McGraw & Wong, 1996; ten Hove et al., 2022) in classical ANOVA models or more complex designs. Given this definitional heterogeneity, and to avoid confusion with traditional ICCs often associated with inter-rater reliability or simple clustering, we have primarily referred to our indices as reliability coefficients. Nevertheless, they share the same fundamental mathematical interpretation as ICCs derived within the correlation-based framework of linear mixed models.

[8] Following the framework of ICCs proposed by McGraw and Wong (1996) or the correlation approach proposed by Lin et al. (2025) and Vangeneugden et al. (2005), reliability coefficients can be categorized as (relative) consistency or (absolute) agreement. Consistency concerns preserving the relative ranking of targets, thus ignoring any systematic nuisance variances. In contrast, agreement focuses on the absolute values of the measurements, thus taking into account all sources of variance. In this article, we focus on the agreement type of reliability coefficients, because we want to know the stability or similarity of the measurements across time, which requires taking into account all sources of variance.

[9] If the replicates t and t' indicate two different items from the same scale, then the correlation is called internal consistency. If the replicates are taken by two different raters, then it is called interrater reliability (Revelle & Condon, 2018; Vangeneugden et al., 2004).

[10] In GT, the reliability coefficients are called the generalizability coefficient (Cronbach, 1972) and the index of dependability (Brennan & Kane, 1977a, 1977b), depending on whether reliability is meant to reflect consistency or agreement.

Downloaded from https://www.cambridge.org/core. 13 Nov 2026 at 03:02:29, subject to the Cambridge Core terms of use.

adequately capture the complexity of ILD data, and new approaches are needed to define and estimate reliability in that context.

Reliability for (intensive) longitudinal data can be defined either using the correlation approach or the variance partitioning approach. For example, in the context of the single-item measures, Vangeneugden et al. (2004, 2005) formulated reliability via the correlation approach, while Laenen et al. (2007, 2009a, 2009b) utilized the variance partitioning approach under the LMM framework. However, these earlier works did not explicitly distinguish reliability across different hierarchical levels, which is crucial for modern ILD studies. Notably, other studies have successfully employed the variance partitioning approach to define reliability in intensive longitudinal designs (see Leertouwer et al., 2025, for a comprehensive guide).[11] For instance, Schuurman and Hamaker (2019) and Schuurman et al. (2015) defined single-item reliability at both the between- and within-person levels within the multilevel measurement error autoregressive (ML-MEAR) model. However, their framework primarily considered a single AR(1) process across time within a two-level structure (e.g., repeated measurements nested within persons). While their model could theoretically be extended to accommodate a three-level structure (i.e., moments nested within days nested within persons), their reliability definitions yield static point estimates that do not explicitly account for how reliability fluctuates or decays over varying temporal lags. Furthermore, Dejonckheere et al. (2022) proposed a model-free approach utilizing specific immediate- and delayed-retest designs to empirically isolate measurement error variance, which is then plugged back into the classical reliability formula. While highly innovative, Dejonckheere et al.'s (2022) approach requires a specialized data collection design that is absent in standard, pre-existing ILD datasets like ours.

To address these challenges and bridge the methodological gap, the present study focuses specifically on formalizing single-item test–retest reliability within a dynamic LMM framework (recall Equation (3.5)). Although the two approaches (correlation vs. variance partitioning) lead to the same reliability coefficients within CTT (and other simple situations), this is not necessarily the case for more complex data structures (Lin et al., 2025). Therefore, safely interpreting reliability in ILD applications requires understanding how these two approaches might diverge mathematically under complex true-score structures. In the following, we will formally derive reliability coefficients based on the correlation approach first and on the variance partitioning approach. For ease of distinction, the notation ρ will denote reliability defined through the correlation approach, while the notation R will correspond to the one constructed by the variance partitioning approach.

4.3. Correlation approach to reliability for LMM

The correlation approach has been used before to define reliability for longitudinal data and for multilevel data. Vangeneugden et al. (2004, 2005) proposed a general formula for reliability in longitudinal data using the correlation approach in the two-level LMM with one serial process, but they did not discuss how to distinguish the reliability at different hierarchical levels. On the other hand, Lin et al. (2025) further defined reliability coefficients for multilevel data at the different hierarchical levels, that is, between-person reliability and within-person reliability, but they did not consider any serial process. Thus, in this section, we will define reliability coefficients for a three-level model with serial correlation at two different time scales (day and moment).

[11] Note that McDonald's omega-like coefficients (Castro-Alvarez, Tendeiro, de Jonge, et al., 2022; Fuller-Tyszkiewicz et al., 2017; Geldhof et al., 2014; Hu et al., 2016; Lai, 2021; McDonald, 1999; Xiao et al., 2023) defining internal consistency for multiple-item measures also belong to the variance partitioning approach. However, these are typically discussed within the structural equation modeling (SEM) framework. Since our focus is on single-item measures, we do not cover SEM-based approaches here. For a detailed review of different reliability definitions in ILD, see Castro-Alvarez et al. (2025).

Downloaded from https://www.cambridge.org/core. 13 Nov 2026 at 03:02:29, subject to the Cambridge Core terms of use.

4.3.1. Between-person reliability

Following the correlation approach (Lin et al., 2025; Vangeneugden et al., 2004, 2005), the between-person test–retest reliability ρ^{B} is defined as the correlation between two measurements y_{ijk} and $y_{ij'k'}$ of the same person i at two moments k and k' on days j and j' across the population. For conciseness, we define the time differences $\Delta_{jj'} = |j - j'|$ and $\Delta_{kk'} = |k - k'|$. Based on our three-level LMM model with two AR(1) processes (Equation (3.4)), the general formula for this correlation is (the full derivation is provided in Appendix A)

$$\rho^{\mathrm{B}}(\Delta_{jj'}, \Delta_{kk'}) = \mathrm{Cor}(y_{ijk}, y_{ij'k'}) = \frac{\sigma_s^2 + \delta_{jj'}\sigma_d^2 + \tau_d^2\phi_d^{\Delta_{jj'}} + \delta_{jj'}\tau_m^2\phi_m^{\Delta_{kk'}}}{\sigma_s^2 + \sigma_d^2 + \tau_d^2 + \tau_m^2 + \sigma_\varepsilon^2}, \tag{4.4}$$

where the superscript B indicates the between-person level and $\delta_{jj'}$ is the Kronecker delta equal to 1 only if two moment measures are taken on the same day (i.e., $j = j'$) and equal to 0 otherwise. The within-day serial correlation process does not play a role if the two moments considered are from different days $j \neq j'$. Thus, $\delta_{jj'}$ determines whether σ_d^2 and $\tau_m^2\phi_m^{\Delta_{kk'}}$ appear in the numerator. Furthermore, $\Delta_{jj'}$ and $\Delta_{kk'}$ play vital roles in regulating the influence of the autoregressive effects ϕ_d and ϕ_m. Note that the two measurements y_{ijk} and $y_{ij'k'}$ are not necessarily parallel (or tau-equivalent) in the CTT sense, because they may have different true scores due to the presence of within-person dynamic patterns. However, $\rho^{\mathrm{B}}(\Delta_{jj'}, \Delta_{kk'})$ is still reflecting a crucial aspect of reliability: the agreement between two measurements across the population. Next, we will discuss several specific cases of the general formula.

Between-person moment-to-moment reliability: Suppose that researchers are interested in the between-person reliability across different moments (e.g., morning vs. afternoon) within a day, that is, the agreement of the measurements at two different moments $k \neq k'$ within the same day j across the population. Because $\delta_{jj} = 1$ and $\Delta_{jj'} = 0$, the between-person moment-to-moment reliability coefficient of Equation (4.4) simplifies to

$$\rho^{\mathrm{B}}_{\mathrm{M}}(\Delta_{kk'}) = \mathrm{Cor}(y_{ijk}, y_{ijk'}) = \frac{\sigma_s^2 + \sigma_d^2 + \tau_d^2 + \tau_m^2\phi_m^{\Delta_{kk'}}}{\sigma_s^2 + \sigma_d^2 + \tau_d^2 + \tau_m^2 + \sigma_\varepsilon^2}, \tag{4.5}$$

where the subscript M indicates moment-to-moment. This reliability is a function of the time lag between the two moments $\Delta_{kk'}$: The reliability decreases as $\Delta_{kk'}$ increases, because the two moments are more separated in time and thus the correlation between them is lower.

It might initially appear counterintuitive that day-related variances (σ_d^2 and τ_d^2) persist in the numerator when computing moment-to-moment reliability. This occurs because assessing reliability between two moments strictly on the same day ($j = j'$) means these two measurements inherently share the exact same daily environmental characteristics and structural trajectory. Therefore, without conditioning on a specific fixed day, this shared daily variability fundamentally contributes as systematic variance, enhancing their overall similarity.

Between-person day-to-day reliability: Similarly, if researchers are interested in the reliability across different days (such that $\Delta_{jj'} > 0$ and $\delta_{jj'} = 0$) across the population, the between-person day-to-day reliability coefficient is

$$\rho^{\mathrm{B}}_{\mathrm{D}}(\Delta_{jj'}) = \mathrm{Cor}(y_{ijk}, y_{ij'k'}) = \frac{\sigma_s^2 + \tau_d^2\phi_d^{\Delta_{jj'}}}{\sigma_s^2 + \sigma_d^2 + \tau_d^2 + \tau_m^2 + \sigma_\varepsilon^2}, \tag{4.6}$$

where the subscript D indicates day-to-day. In this case, k and k' can be either the same or different, but because the measurements are located on different days ($j \neq j'$), they are disconnected by the moment-level AR trajectory. As a result, the decayed moment-level covariance $\tau_m^2\phi_m^{\Delta_{kk'}}$ drops out of the numerator. Similarly, because they are sampled from different days, they no longer share the same day effect d_{ij}; hence, the day variance σ_d^2 also vanishes from the numerator. Nevertheless, the day-level autoregressive component $\tau_d^2\phi_d^{\Delta_{jj'}}$ remains, reflecting that the two measurements are still systematically bound by the

Downloaded from https://www.cambridge.org/core. 13 Nov 2026 at 03:02:29, subject to the Cambridge Core terms of use.

overarching day-level AR trajectory. The day-to-day reliability thus functions exclusively via the lag between days $\Delta_{jj'}$, continuously decreasing as $\Delta_{jj'}$ increases.

Both Equations (4.5) and (4.6) reflect the reliability based on a single observation at different time points. However, in practice, researchers often use the average of multiple observations to represent a more stable score for a specific period. For example, researchers may use the average of moments within a day to represent the daily score for further analysis. When this is the case, they are wondering about the reliability of the composite score $\bar{y}_{ij\cdot} = \frac{1}{M}\sum_{k=1}^{M} y_{ijk}$. Thus, there is another form of day-to-day reliability based on the average of moments within a day. In this case, the day-to-day reliability coefficient becomes

$$\rho^{\mathrm{B}}_{\overline{\mathrm{D}}}(\Delta_{jj'}) = \mathrm{Cor}(\bar{y}_{ij\cdot}, \bar{y}_{ij'\cdot}) = \frac{\sigma_s^2 + \tau_d^2 \phi_d^{\Delta_{jj'}}}{\sigma_s^2 + \sigma_d^2 + \tau_d^2 + \frac{\tau_m^2}{M} + \frac{\sigma_\varepsilon^2}{M}}, \tag{4.7}$$

where the subscript $\overline{\mathrm{D}}$ indicates the reliability for day-to-day based on the average of moments. Compared to Equation (4.6), the moment-level variance components τ_m^2 and σ_ε^2 are now divided by M in the denominator because of the averaging operation, which leads to a smaller denominator and thus a larger result. This makes sense because the average of several measurements is at least as reliable as a single measurement.

Between-person stability and instantaneity: As can be seen in Equations (4.5) and (4.6), the temporal autocorrelation (decaying across time lags) between two measurements is an important component of the aforementioned reliability coefficients. Because of the presence of serial correlation and time lag in the reliability coefficients, it fits well in the distinction between long-term stability and short-term instantaneity, depending on the time lag, as defined by Boyle et al. (1995), Cattell (1964), and Revelle and Condon (2019). *Stability* considers distant time points to evaluate the reliability over the whole measurement period or a further longer period, while *instantaneity*[12] considers two close time points to evaluate reliability immediately or in a short period.[13]

Conceptually, we defined stability and instantaneity as the two extremes of a time lag spectrum in intensive longitudinal studies. Stability represents a test–retest correlation (Equation (4.4) or (4.6)) when the time lag is very large such that the influence of the autoregressive effect ϕ_d (and also ϕ_m) completely disappears, or when, mathematically speaking, the time lag goes to infinity, that is, $\phi_d^{\Delta_{jj'}} \to 0$ as $\Delta_{jj'} \to \infty$. As a result, the between-person stability coefficient is defined as

$$\rho^{\mathrm{B}}_{\infty} = \lim_{\Delta_{jj'}\to\infty} \rho^{\mathrm{B}}_{D}(\Delta_{jj'}) = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_d^2 + \tau_d^2 + \tau_m^2 + \sigma_\varepsilon^2}, \tag{4.8}$$

where the subscript ∞ indicates stability. This coefficient can be alternatively defined as the correlation between the current measurement y_{ijk} and a hypothetical very distant future measurement $y_{i\infty\infty}$ from the same person i and then across the whole population, that is, $\rho^{\mathrm{B}}_{\infty} = \mathrm{Cor}(y_{ijk}, y_{i\infty\infty})$. The numerator only consists of the between-person variance component σ_s^2. All other sources of variance only appear in the denominator. This coefficient measures the degree to which individuals maintain their relative standing on the measured construct over an extended period, regardless of the specific time interval between assessments. Although the time lag to define stability is, in principle, infinitely long, for a substantive interpretation, it makes sense to think about it as "a very long time period in the context of the study." Or stated differently, the time lag at which the autocorrelation has effectively vanished. To assess the stability of personality items, we may think of a period of several years. In contrast, if we are measuring mood, a very long time period can be as short as weeks.

[12] Cattell (1964) and Revelle and Condon (2019) originally use the term *dependability*, but this article adopts *instantaneity* to avoid confusion with the index of dependability in GT (Brennan & Kane, 1977a, 1977b).

[13] A universal definition of the duration of a long or short time period cannot be given in general, but depends on the context (Revelle & Condon, 2019).

Downloaded from https://www.cambridge.org/core. 13 Nov 2026 at 03:02:29, subject to the Cambridge Core terms of use.

On the contrary, instantaneity is defined as the test–retest reliability (Equation (4.4) or (4.5)) when the time lag between two measurements is very short such that the influence of the autoregressive effect is maximal, or mathematically speaking, the time lag goes to zero, that is, $\phi_m^{\Delta_{kk'}} \to 1$ as $\Delta_{kk'} \to 0$. Thus, the between-person instantaneity coefficient is defined as

$$\rho_0^{\mathrm{B}} = \lim_{\Delta_{kk'} \to 0} \rho_M^{\mathrm{B}}(\Delta_{kk'}) = \frac{\sigma_s^2 + \sigma_d^2 + \tau_d^2 + \tau_m^2}{\sigma_s^2 + \sigma_d^2 + \tau_d^2 + \tau_m^2 + \sigma_\varepsilon^2}, \tag{4.9}$$

where the subscript 0 indicates instantaneity. This coefficient can be alternatively defined as the correlation between the current measurement y_{ijk} and a hypothetical parallel measurement y'_{ijk} from the same person i on the same day j at the same moment k, the only difference between the two measurements being the realized measurement error. Thus, $\rho_0^{\mathrm{B}} = \mathrm{Cor}(y_{ijk}, y'_{ijk})$. It shows that all variance components except the measurement error variance σ_ε^2 appear in the numerator. This coefficient quantifies the extent to which individuals maintain their relative standing on the measured construct over a very short period (as defined within the context of the study).

4.3.2. Within-person reliability

Within-person reliability focuses on the structural similarity of measurements within the same person instead of across the population of persons. We start again with a general formula for the within-person reliability coefficient ρ^{W}, defined as the correlation between two measurements y_{ijk} and $y_{ij'k'}$ for a specific arbitrary person i (the full derivation is provided in Appendix A):

$$\rho^{\mathrm{W}}(\Delta_{jj'}, \Delta_{kk'}) = \mathrm{Cor}(y_{ijk}, y_{ij'k'} \mid i) = \frac{\delta_{jj'}\sigma_d^2 + \tau_d^2\phi_d^{\Delta_{jj'}} + \delta_{jj'}\tau_m^2\phi_m^{\Delta_{kk'}}}{\sigma_d^2 + \tau_d^2 + \tau_m^2 + \sigma_\varepsilon^2}, \tag{4.10}$$

where the superscript W indicates the within-person level. As before in Equation (4.4), the general formula is a function of the time lags $\Delta_{jj'}$ and $\Delta_{kk'}$. However, unlike the between-person reliability, the person variance σ_s^2 disappears from both numerator and denominator because we consider a specific person i.

Again, moving across several time lags, we can define the within-person instantaneity, moment-to-moment reliability, and day-to-day reliability for single and average scores as follows:

$$\rho_0^{\mathrm{W}} = \lim_{\Delta_{jj'} \to 0,\, \Delta_{kk'} \to 0} \mathrm{Cor}(y_{ijk}, y_{ij'k'} \mid i) = \frac{\sigma_d^2 + \tau_d^2 + \tau_m^2}{\sigma_d^2 + \tau_d^2 + \tau_m^2 + \sigma_\varepsilon^2}, \tag{4.11}$$

$$\rho_{\mathrm{M}}^{\mathrm{W}}(\Delta_{kk'}) = \mathrm{Cor}(y_{ijk}, y_{ijk'} \mid i) = \frac{\sigma_d^2 + \tau_d^2 + \tau_m^2\phi_m^{\Delta_{kk'}}}{\sigma_d^2 + \tau_d^2 + \tau_m^2 + \sigma_\varepsilon^2}, \tag{4.12}$$

$$\rho_{\mathrm{D}}^{\mathrm{W}}(\Delta_{jj'}) = \mathrm{Cor}(y_{ijk}, y_{ij'k'} \mid i) = \frac{\tau_d^2\phi_d^{\Delta_{jj'}}}{\sigma_d^2 + \tau_d^2 + \tau_m^2 + \sigma_\varepsilon^2}, \tag{4.13}$$

$$\rho_{\mathrm{D}}^{\mathrm{W}}(\Delta_{jj'}) = \mathrm{Cor}(\bar{y}_{ij\cdot}, \bar{y}_{ij'\cdot} \mid i) = \frac{\tau_d^2\phi_d^{\Delta_{jj'}}}{\sigma_d^2 + \tau_d^2 + \frac{\tau_m^2}{M} + \frac{\sigma_\varepsilon^2}{M}}. \tag{4.14}$$

All within-person reliability coefficients are derived from the conditional correlations, which result in formulas from which the person variance σ_s^2 is removed. Note that the within-person stability is undefined because it only makes sense for individual differences.

4.4. Variance partitioning approach to reliability for LMM

Reliability coefficients in the context of longitudinal data can also be defined based on the variance partitioning approach (Laenen et al., 2007; Laenen et al., 2009a, 2009b; Schuurman & Hamaker, 2019; Schuurman et al., 2015). There are two distinct (between-person) reliability coefficients worth noticing,

Downloaded from https://www.cambridge.org/core. 13 Nov 2026 at 03:02:29, subject to the Cambridge Core terms of use.

R_{T} (Laenen et al., 2007) and R_{Λ} (Laenen et al., 2009b). Both reliability coefficients are defined in a similar form as Equation (4.2), but in a matrix version:

$$R = 1 - \frac{f(\boldsymbol{\Sigma}_{\text{unsys}})}{f(\boldsymbol{\Sigma}_{\text{total}})},$$

where $\boldsymbol{\Sigma}_{\text{unsys}}$ is the unsystematic part of the total covariance matrix $\boldsymbol{\Sigma}_{\text{total}}$ and $f(\cdot)$ is a function, which can be either the trace for $R_T = 1 - \frac{\text{tr}(\boldsymbol{\Sigma}_{\text{unsys}})}{\text{tr}(\boldsymbol{\Sigma}_{\text{total}})}$ (Laenen et al., 2007) or the determinant for $R_{\Lambda} = 1 - \frac{|\boldsymbol{\Sigma}_{\text{unsys}}|}{|\boldsymbol{\Sigma}_{\text{total}}|}$ (Laenen et al., 2009b), to summarize the covariance matrix into a scalar.

For both cases (i.e., R_T and R_{Λ}), an important consideration is how to partition the variance components into a systematic and an unsystematic part. This classification depends on the research question. For example, Laenen et al. (2007, 2009a, 2009b) argue that, based on their model of interest,[14] the systematic part should only include the stable individual differences, that is, the person-related covariance matrix, while the unsystematic part should include all other variance components, including the serial covariance matrix and measurement error covariance matrix. However, this partition is not absolute. In some ILD studies, temporal dynamics, such as daily or momentary fluctuations captured by AR(1) processes, represent meaningful, structured patterns and could therefore be considered as systematic variance. For example, if the research scope is restricted to daily trends (e.g., daily blood pressure level), then the day effect as well as the day-level serial effect should be considered as part of the systematic variance, while the moment-level serial effect should be treated as unsystematic variance. Conversely, if the research focus is on the dynamic process itself (e.g., momentary mood regulation or emotional inertia), then the moment serial effect should be considered as part of the systematic variance.

Depending on three different partitions of the variance components, we can construct three subtypes of R_{T}[15] based on our three-level LMM with two AR(1) processes (Equation (3.5)). As before, we distinguish between-person reliability and within-person reliability for each subtype. This leads to five reliability coefficients based on variance partitioning. In the next section, we will explain how these coefficients are related.

4.4.1. Between-person reliability

As shown in Equation (3.6), the total variance matrix of the observation $\boldsymbol{V}$ is decomposed into five variance components: the person-related covariance matrix $\boldsymbol{\Sigma}_S$, the day-related covariance matrices ($\boldsymbol{\Sigma}_D$ and $\boldsymbol{\Sigma}_v$), moment-related covariance matrix $\boldsymbol{\Sigma}_{\omega}$, and measurement error covariance matrix $\boldsymbol{\Sigma}_E$.

Reliability for momentary scores: If the momentary score is the measurement of interest, then the question is to which extent this score can be differentiated from other momentary scores (from the same person or from other persons). In this case, all variance components related to person, day, and moment are considered systematic. Only random measurement error ($\boldsymbol{\Sigma}_E$) is treated as unsystematic. The reliability is defined as

$$R^{\mathrm{B}}_{\mathrm{TM}} = 1 - \frac{\text{tr}(\boldsymbol{\Sigma}_E)}{\text{tr}(\boldsymbol{V})} = \frac{\sigma_s^2 + \sigma_d^2 + \tau_d^2 + \tau_m^2}{\sigma_s^2 + \sigma_d^2 + \tau_d^2 + \tau_m^2 + \sigma_\varepsilon^2}, \tag{4.15}$$

where the subscript TM indicates reliability based on the trace for the momentary score, and the superscript B indicates the between-person level. $R^{\mathrm{B}}_{\mathrm{TM}}$ reflects the belief that even though momentary states fluctuate, these fluctuations are part of a meaningful dynamic process. Especially in objective

[14] The model of Laenen et al. (2007, 2009a, 2009b) is a two-level LMM with covariates and a single AR(1) process. As a result, the total variance structure has only three distinct parts: the person-related covariance matrix, the serial covariance matrix, and the measurement error covariance matrix. This is slightly different from our case, as can be seen in Equation (3.6), where we have five components.

[15] Three subtypes of R_{Λ} can also be constructed in a similar manner to what we propose next. However, we focus our discussion on R_{T} because the resulting coefficients naturally compare to the reliability coefficients defined by the correlation approach and possess an easier interpretation.

Downloaded from https://www.cambridge.org/core. 13 Nov 2026 at 03:02:29, subject to the Cambridge Core terms of use.

physiological measures, such as heart rate or cortisol levels, the moment-to-moment fluctuations can be systematic and informative about an individual's current state.

The coefficient $R^{\mathrm{B}}_{\mathrm{TM}}$ is equivalent to the between-person instantaneity coefficient defined by the correlation approach: $R^{\mathrm{B}}_{\mathrm{TM}} = \rho^{\mathrm{B}}_{0}$. This equivalence tells that both reliability coefficients can be interpreted in two different ways: (1) to which extent can the momentary score be used to differentiate between individuals across the population (based on the variance partitioning approach) and (2) the level of agreement between individual momentary scores over an extremely short period of time (based on the correlation approach). This dual interpretation highlights the multifaceted nature of reliability in ILD.

Reliability of daily scores: When the target of measurement is the score at the day level, the moment-to-moment fluctuations within a day are considered unsystematic noise. In this case, the systematic variance includes only the person- and day-level components ($\boldsymbol{\Sigma}_S, \boldsymbol{\Sigma}_D, \boldsymbol{\Sigma}_v$), while the unsystematic variance includes the moment-level serial process ($\boldsymbol{\Sigma}_\omega$) and measurement error ($\boldsymbol{\Sigma}_E$). The reliability is defined as

$$R^{\mathrm{B}}_{\mathrm{TD}} = 1 - \frac{\mathrm{tr}(\boldsymbol{\Sigma}_\omega + \boldsymbol{\Sigma}_E)}{\mathrm{tr}(\boldsymbol{V})} = \frac{\sigma_s^2 + \sigma_d^2 + \tau_d^2}{\sigma_s^2 + \sigma_d^2 + \tau_d^2 + \tau_m^2 + \sigma_\varepsilon^2}, \tag{4.16}$$

where the subscript TD indicates reliability for the daily score.

The coefficient $R^{\mathrm{B}}_{\mathrm{TD}}$ is equivalent to the between-person moment-to-moment reliability $\rho^{\mathrm{B}}_{\mathrm{M}}(\Delta_{kk'})$ for a large within-day time lag $\Delta_{kk'}$ (such that the moment-level autoregressive effect ϕ_m vanishes). Consequently, this reliability coefficient quantifies: (1) to which extent the daily score can be used to differentiate between individuals across the population (based on the variance partitioning approach) and (2) the level of agreement between the observed scores over a long interval within the same day (based on the correlation approach). In other words, $R^{\mathrm{B}}_{\mathrm{TD}}$ reflects within-day stability.

Reliability of person (trait) scores: Lastly, if the goal is to measure a stable, person-level trait, all sources of within-person variance (daily fluctuations, momentary dynamics, and measurement error) can be considered unsystematic. The systematic variance is confined to the stable between-person component ($\boldsymbol{\Sigma}_S$). This perspective is common when researchers believe temporal dynamics are nuisance factors rather than meaningful signals. The reliability is defined as

$$R^{\mathrm{B}}_{\mathrm{TP}} = 1 - \frac{\mathrm{tr}(\boldsymbol{\Sigma}_D + \boldsymbol{\Sigma}_\omega + \boldsymbol{\Sigma}_v + \boldsymbol{\Sigma}_E)}{\mathrm{tr}(\boldsymbol{V})} = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_d^2 + \tau_d^2 + \tau_m^2 + \sigma_\varepsilon^2}, \tag{4.17}$$

where the subscript TP indicates reliability of the personal trait-like score. This reliability coefficient should be considered when any day-to-day and moment-to-moment fluctuations, together with their serial correlation, are considered as error and do not truly explain the difference in individual traits. This is reasonable, for example, when considering that serial patterns arise from carry-over effects, such as memory effects or practice effects, which can occur in studies involving self-report measurements.

This coefficient is equivalent to the between-person stability coefficient $R^{\mathrm{B}}_{\mathrm{TP}} = \rho^{\mathrm{B}}_{\infty}$ and indicates that this reliability coefficient quantifies: (1) to which extent the trait-like score can differentiate between individuals across the population (based on the variance partitioning approach) and (2) the level of agreement between trait-like scores within individuals over an extended period (based on the correlation approach).

4.4.2. Within-person reliability

The same variance partitioning logic can be applied at the within-person level. In this case, however, not the total variance $\boldsymbol{V}$ will be partitioned, but the individual-level variance $\boldsymbol{V}_i$ (see Equation (3.7)). $\boldsymbol{V}_i$ is composed of four parts: the day-related covariance matrices ($\boldsymbol{\Sigma}_D$ and $\boldsymbol{\Sigma}_v$), the moment-related covariance matrix $\boldsymbol{\Sigma}_\omega$, and the measurement error covariance matrix $\boldsymbol{\Sigma}_E$. Depending on the measurement of interest (e.g., a momentary score or a daily score), there are two subtypes of R_{T} at the

Downloaded from https://www.cambridge.org/core. 13 Nov 2026 at 03:02:29, subject to the Cambridge Core terms of use.

within-person level:

$$R^{\mathrm{W}}_{\mathrm{TM}} = 1 - \frac{\mathrm{tr}(\boldsymbol{\Sigma}_E)}{\mathrm{tr}(\boldsymbol{V}_i)} = \frac{\sigma_d^2 + \tau_d^2 + \tau_m^2}{\sigma_d^2 + \tau_d^2 + \tau_m^2 + \sigma_\varepsilon^2}, \tag{4.18}$$

$$R^{\mathrm{W}}_{\mathrm{TD}} = 1 - \frac{\mathrm{tr}(\boldsymbol{\Sigma}_\omega + \boldsymbol{\Sigma}_E)}{\mathrm{tr}(\boldsymbol{V}_i)} = \frac{\sigma_d^2 + \tau_d^2}{\sigma_d^2 + \tau_d^2 + \tau_m^2 + \sigma_\varepsilon^2}, \tag{4.19}$$

where the superscript W indicates the within-person level. $R^{\mathrm{W}}_{\mathrm{TM}}$ focuses on the momentary score within a person, so the unsystematic part only includes the measurement error covariance matrix $\boldsymbol{\Sigma}_E$. This coefficient is equivalent to the within-person instantaneity coefficient $R^{\mathrm{W}}_{\mathrm{TM}} = \rho^{\mathrm{W}}_0$. $R^{\mathrm{W}}_{\mathrm{TD}}$, focusing on the daily score within a person, considers both the moment-related covariance matrix $\boldsymbol{\Sigma}_\omega$ and measurement error covariance matrix $\boldsymbol{\Sigma}_E$ as the unsystematic part. This coefficient is equivalent to the within-person moment-to-moment reliability $\rho^{\mathrm{W}}_{\mathrm{M}}(\Delta_{kk'})$ when the time lag $\Delta_{kk'}$ within a day is large enough such that the moment-level autoregressive effect ϕ_m diminishes.

Both coefficients $R^{\mathrm{W}}_{\mathrm{TM}}$ and $R^{\mathrm{W}}_{\mathrm{TD}}$ quantify: (1) to which extent the measurement of interest (at the moment or day level) can differentiate between different time points within the same person from the variance partitioning viewpoint and (2) the level of agreement between target scores (momentary or daily score) within a person over a short period from the correlation viewpoint.

A further step could be to define $R^{\mathrm{W}}_{\mathrm{TP}}$ as the reliability coefficient for the within-person trait-like score. However, as the explanation in words already suggests, this coefficient cannot be determined. The reason is obvious: the person-level variance σ_s^2 is removed in both numerator and denominator when we condition on a specific person i. Consequently, there is no systematic variance component left in the numerator if we consider all other variance components as the unsystematic part. Similarly, we could not define the within-person stability coefficient using the correlation approach.

To summarize, we collect the aforementioned reliability coefficients for the three-level LMM with two AR(1) processes in a single row ("M1") in Table 1 for the between-person level and in Table 2 for the within-person level, respectively. From left to right, the time lag of interest widens, ranging from instantaneity to moment-to-moment fluctuations to day-to-day fluctuations to stability. It can be seen that the reliability coefficients defined by the variance partitioning approach are equivalent to those defined by the correlation approach under some conditions. This equivalence highlights that the correlation and variance partitioning approaches represent two distinct perspectives on the same underlying concept of reliability, which can, in some cases, yield identical results.

4.5. Comparing reliability coefficients for different models

To better understand the relationships among various reliability coefficients, we consider four related models with varying degrees of complexity, where simpler models are nested within the more complex ones. The four models for comparison are:

M1: a three-level LMM with two AR(1) processes as defined before in Equation (3.4);
M2: a three-level LMM but without any serially correlated processes: $y_{ijk} = \beta + s_i + d_{ij} + \varepsilon_{ijk}$;
M3: a two-level LMM with one AR(1) process at the moment level: $y_{it} = \beta + s_i + \omega_{it} + \varepsilon_{it}$;
M4: a two-level LMM without any serial process: $y_{it} = \beta + s_i + \varepsilon_{it}$.

Models M1 and M2 assume a three-level data structure (moments nested within days, and days nested within persons), while models M3 and M4 assume a two-level structure by ignoring one level (moments nested within persons). For models M3 and M4, we have combined the two time indices j and k into a single time index t, where $t = 1, \ldots, D \times M$. Thus, the time lag between any pair of measurements is denoted as $\Delta_t = |t - t'|$. In addition, model M3 is also the main model (but without person covariates)

Downloaded from https://www.cambridge.org/core. 13 Nov 2026 at 03:02:29, subject to the Cambridge Core terms of use.

Table 1. Comparison of between-person reliability coefficients across four models

Between- person	Time span			
	Instantaneity	Moment-to-moment	Day-to-day	Stability
M1: Three-level LMM with two AR(1)	$\rho_0^B = R_{TM}^B = \frac{\sigma_s^2+\sigma_d^2+\tau_d^2+\tau_m^2}{\sigma_s^2+\sigma_d^2+\tau_d^2+\tau_m^2+\sigma_\varepsilon^2}$	$\rho_M^B(\Delta_{kk'}) = \frac{\sigma_s^2+\sigma_d^2+\tau_d^2+\tau_m^2\phi_m^{\Delta_{kk'}}}{\sigma_s^2+\sigma_d^2+\tau_d^2+\tau_m^2+\sigma_\varepsilon^2}$ $\rho_M^B(\Delta_{kk'}^+) = R_{TD}^B = \frac{\sigma_s^2+\sigma_d^2+\tau_d^2}{\sigma_s^2+\sigma_d^2+\tau_d^2+\tau_m^2+\sigma_\varepsilon^2}$	$\rho_D^B(\Delta_{jj'}) = \frac{\sigma_s^2+\tau_d^2\phi_d^{\Delta_{jj'}}}{\sigma_s^2+\sigma_d^2+\tau_d^2+\tau_m^2+\sigma_\varepsilon^2}$	$\rho_\infty^B = R_{TP}^B = \frac{\sigma_s^2}{\sigma_s^2+\sigma_d^2+\tau_d^2+\tau_m^2+\sigma_\varepsilon^2}$
M2: Three-level LMM	$\rho_0^B = \rho_M^B = R_{TM}^B = R_{TD}^B = \frac{\sigma_s^2+\sigma_d^2}{\sigma_s^2+\sigma_d^2+\sigma_\varepsilon^2}$		$\rho_D^B = \rho_\infty^B = R_{TP}^B = \frac{\sigma_s^2}{\sigma_s^2+\sigma_d^2+\sigma_\varepsilon^2}$	
M3: Two-level LMM with one AR(1)	$\rho_0^B = R_{TM}^B = \frac{\sigma_s^2+\tau_t^2}{\sigma_s^2+\tau_t^2+\sigma_\varepsilon^2}$	$\rho_T^B(\Delta_{tt'}) = \frac{\sigma_s^2+\tau_t^2\phi^{\Delta_{tt'}}}{\sigma_s^2+\tau_t^2+\sigma_\varepsilon^2}$		$\rho_\infty^B = R_{TP}^B = \frac{\sigma_s^2}{\sigma_s^2+\tau_t^2+\sigma_\varepsilon^2}$
M4: Two-level LMM (= CTT)	$\rho = R = \frac{\sigma_s^2}{\sigma_s^2+\sigma_\varepsilon^2}$			

Note: ρ = reliability based on the correlation approach; R = reliability based on the variance partitioning approach; superscripts B = between-person level and W = within-person level; subscripts 0 = instantaneity, M = moment-to-moment; D = day-to-day, T = test–retest, ∞ = stability, TM = momentary scores, TD = daily score, TP = personal trait scores; $\Delta_{kk'}$ = time lag between moments, $\Delta_{kk'}^+$ = a large time lag between moments such that the moment-level autoregressive effect ϕ_m diminishes, $\Delta_{jj'}$ = time lag between days, and Δ_t = time lag between moments/measurements.

Downloaded from https://www.cambridge.org/core. 13 Nov 2026 at 03:02:29, subject to the Cambridge Core terms of use.

discussed in Laenen et al. (2007, 2009b) and Vangeneugden et al. (2004, 2005). Model M4 is equivalent to the traditional CTT model where the true score T_i is defined as $T_i = \beta + s_i$.

Tables 1 and 2 summarize the reliability coefficients for the between-person and within-person levels, respectively. There are some key findings that can be derived from these two tables. First, comparing models M1 and M3 to M2 and M4, respectively, the difference is the inclusion of serial correlation. The models with serial correlation have more comprehensive reliability coefficients compared to those which do not account for any serial correlation. These permit the distinction between instantaneity, moment-to-moment, day-to-day, and stability. Second, because of the three-level structure in model M2, we can still distinguish between the moment-to-moment reliability ρ_M and day-to-day reliability ρ_D. Nevertheless, both reliabilities are constant because there is no serial correlation in the model. Therefore, regardless of which pair of measurements is selected and used in the correlation, the reliability remains the same. For example, let us focus on the correlation between the measurement of person i on day j and moment k, y_{ijk}, and correlate this with the measurement one moment later (i.e., $y_{ij,\,k+1}$) or two moments later (i.e., $y_{ij,\,k+2}$). It turns out that in model M2, $\rho_M(\Delta_{kk'} = 1) = \text{Cor}(y_{ijk}, y_{ij,\,k+1}) = \frac{\sigma_s^2+\sigma_d^2}{\sigma_s^2+\sigma_d^2+\sigma_\varepsilon^2} = \text{Cor}(y_{ijk}, y_{ij,\,k+2}) = \rho_M(\Delta_{kk'} = 2)$. Third (and as said above), some reliabilities related to stable traits cannot be defined (or always be 0) for the within-person level (Table 2). This is a necessary consequence of conditioning on the person, which removes the between-person variance that defines these coefficients.

In addition, it is theoretically illuminating to demonstrate how the reliability coefficients derived from our simplified models seamlessly connect to existing reliability formulations in the literature. For the three-level model without autocorrelation (M2), the data structure conceptually mirrors the three-level GT framework often used for multiple-item ESM data when items are modeled as nested within time points[16] (in contrast to our moments nested within days at the lower levels), which are in turn nested within persons (e.g., Nezlek, 2017; Nezlek & Gable, 2001; Shrout & Lane, 2012a). If we constrain their GT framework to evaluate the reliability of a single item measured at a single time point, the between-person reliability formula (e.g., Equation 10 in Shrout & Lane, 2012a) is mathematically identical to our between-person reliability without autocorrelation (ρ^{B}_{∞} from Table 1). Similarly, their within-person reliability formula (e.g., Equation 11 in Shrout & Lane, 2012a) shares the exact same form as our within-person reliability (ρ^{W}_{0} from Table 2).

Likewise, the two-level model with a single AR(1) process (M3) corresponds closely to the ML-MEAR model proposed by Schuurman and Hamaker (2019), albeit they formulate it using a state-space representation.[17] Under the assumption of constant model parameters, the between-person reliability in the ML-MEAR framework (Equation 7 in Schuurman & Hamaker, 2019) equals our ρ^{B}_{∞} derived under M3. Analogously, their within-person reliability (Equation 9 in Schuurman & Hamaker, 2019) is equivalent to our ρ^{W}_{0}.

Finally, the traditional reliability in CTT (model M4) can be seen as the most basic special case for the more complex models (M1, M2, and M3) when we ignore the serial correlation and multilevel data

[16] It is crucial to distinguish between nested and crossed items. In typical multiple-item ILD studies, the exact same items are administered at every measurement occasion, meaning items are fully crossed with time points (e.g., as correctly modeled in Cranford et al., 2006). Treating such inherently crossed items as nested within time points—whether due to hierarchical modeling approaches (e.g., Nezlek & Gable, 2001) or planned missing data (e.g., Nezlek, 2017)—inappropriately confounds item-specific variance with residual error. This inevitably biases the estimated variance components and the resulting reliability coefficients.

[17] The original ML-MEAR framework by Schuurman and Hamaker (2019) was designed for multivariate time series (allowing for cross-lagged effects between multiple items) and permitted person-specific AR(1) parameters and measurement error variances as random effects. However, if their model is restricted to a univariate (single-item) setting and the AR(1) parameter and error variance are homogeneous across individuals, their state-space model is mathematically equivalent to the linear mixed-effects formulation of M3 (see Ho et al., 2006; Piepho & Ogutu, 2007; Tsimikas & Ledolter, 1997).

Downloaded from https://www.cambridge.org/core. 13 Nov 2026 at 03:02:29, subject to the Cambridge Core terms of use.

Table 2. Comparison of within-person reliability coefficients across four models

Within-person	Time span			
	Instantaneity	Moment-to-moment	Day-to-day	Stability
M1: Three-level LMM with two AR(1)	$\rho_0^W = R_{TM}^W = \frac{\sigma_d^2+\tau_d^2+\tau_m^2}{\sigma_d^2+\tau_d^2+\tau_m^2+\sigma_\varepsilon^2}$	$\rho_M^W(\Delta_{kk'}) = \frac{\sigma_d^2+\tau_d^2+\tau_m^2\phi_m^{\Delta_{kk'}}}{\sigma_d^2+\tau_d^2+\tau_m^2+\sigma_\varepsilon^2}$ $\rho_M^W(\Delta_{kk'}^+) = R_{TD}^W = \frac{\sigma_d^2+\tau_d^2}{\sigma_d^2+\tau_d^2+\tau_m^2+\sigma_\varepsilon^2}$	$\rho_D^W(\Delta_{jj'}) = \frac{\tau_d^2\phi_d^{\Delta_{jj'}}}{\sigma_d^2+\tau_d^2+\tau_m^2+\sigma_\varepsilon^2}$	–
M2: Three-level LMM	$\rho_0^W = \rho_M^W = R_{TM}^W = R_{TD}^W = \frac{\sigma_d^2}{\sigma_d^2+\sigma_\varepsilon^2}$		–	
M3: Two-level LMM with one AR(1)	$\rho_0^W = R_{TM}^W = \frac{\tau_t^2}{\tau_t^2+\sigma_\varepsilon^2}$	$\rho_T^W(\Delta_{tt'}) = \frac{\tau_t^2\phi^{\Delta_{tt'}}}{\tau_t^2+\sigma_\varepsilon^2}$		–
M4: Two-level LMM (= CTT)	–			

Note: Same notations in Table 1.

Downloaded from https://www.cambridge.org/core. 13 Nov 2026 at 03:02:29, subject to the Cambridge Core terms of use.

structure. Thus, the traditional reliability (Equation (4.1)) is always a constant and cannot capture the rich and dynamic nature of ILD.

It is worth noting that while it is theoretically possible to compute reliability coefficients from simplified models (e.g., ignoring a level of nesting or assuming zero autocorrelation), we strongly advise against choosing a model solely for the sake of obtaining a "better" or specific reliability value. Reliability estimates are only meaningful when derived from a model that adequately represents the data bonding structure. For instance, applying a two-level model to three-level data (ignoring day-level clustering) or neglecting substantial serial correlation may result in biased variance component estimates, which in turn propagate into biased reliability coefficients. Therefore, the selection of the reliability formula (M1–M4) should follow the design of the study and the underlying data structure.

5. Illustration using the Exam Results Study

In this section, we apply our developed reliability coefficients to the Exam Results Study (Kalokerinos et al., 2019). As indicated in the introductory section on the Exam Results Study, there are three distinct questions of interest regarding reliability, each answered by specific coefficients derived from our models: (1) To what extent does the single-item measure reflect a stable, trait-like tendency that distinguishes persons from one another? This is addressed by the stability coefficient (e.g., $\rho^{\mathrm{B}}_{\infty} = R^{\mathrm{B}}_{\mathrm{TP}}$). (2) To what extent would repeating the measurement at the exact same given moment yield a similar result? This is addressed by the instantaneity coefficients at both the between- and within-person levels (e.g., $\rho^{\mathrm{B}}_{0} = R^{\mathrm{B}}_{\mathrm{TM}}$ and $\rho^{\mathrm{W}}_{0} = R^{\mathrm{W}}_{\mathrm{TM}}$). Finally, (3) how does the agreement between measurements change across varying time intervals? This is addressed by evaluating the reliability functions over varying time lags (e.g., $\rho^{\mathrm{B}}_{\mathrm{M}}$, $\rho^{\mathrm{B}}_{\mathrm{D}}$, $\rho^{\mathrm{W}}_{\mathrm{M}}$, and $\rho^{\mathrm{W}}_{\mathrm{D}}$).

To investigate these questions, we apply the three-level model with autoregressive processes at two levels (Equation (3.4)) to analyze the single-item negative affect measurement. More information about the sample and the measurement procedure can be found in Section 2 and the original paper (Kalokerinos et al., 2019). In the current analysis, all students have been included.

5.1. Methods

Model parameters and reliability coefficients were estimated using Bayesian inference for two reasons. First, the three-level model with two serially correlated processes is not a standard model that can be easily fit using non-Bayesian software. Second, we want to quantify the uncertainty of the reliability coefficients. Both objectives are easier to reach in a Bayesian inferential framework.

The fitting procedure is as follows: we started from a three-level LMM model without any AR(1) structure. Then, we added AR(1) structures to the moment level and day level sequentially if the MCMC sampler converged well and the parameters for the AR(1) process were significantly different from zero. We used R (version 4.5.0; R Core Team, 2025) with the cmdstanr package (version 0.9.0; Gabry et al., 2025) as an interface of Stan (version 2.36.0; Stan Development Team, 2025a, 2025b) to fit the model. Missing data were handled by setting them as unknown parameters to be estimated. We ran four MCMC chains with 64,000 iterations each, a thinning factor of 8, and considering the first 32,000 iterations as warm-up. The total number of post-warm-up draws across four chains was therefore 16,000.[18] The various reliability coefficients were derived afterward based on the posterior samples of the

[18] While Stan's No-U-Turn-Sampler (NUTS) is typically highly efficient, requiring few iterations for standard models, our three-level model with dual autoregressive processes yields a complex posterior geometry. This can result in strong autocorrelations among parameter samples (e.g., between τ_m and σ_ε). Therefore, the large number of iterations was empirically necessary to ensure a sufficient effective sample size (ESS) (see Table 3 and the convergence diagnostics in the OSF repository). Although discarding information, a thinning factor of 8 was applied solely to reduce data storage requirements and post-processing time. Furthermore, applied researchers utilizing this approach should carefully inspect convergence diagnostics (e.g., $\hat{R}$ and ESS) and adjust their warm-up and total iteration counts accordingly based strictly on their specific dataset. Notably,

Downloaded from https://www.cambridge.org/core. 13 Nov 2026 at 03:02:29, subject to the Cambridge Core terms of use.

Table 3. Posterior summary statistics for the parameters of the three-level model applied to the Exam Results Study

Parameter	Posterior mean (posterior SD)	Posterior 95% HDI	$\hat{R}$	ESS
Fixed effect				
β	19.48 (1.26)	[17.04, 21.54]	1.00	14,677.32
Autoregressive effect				
ϕ_d	0.57 (0.08)	[0.46, 0.71]	1.00	3,771.06
ϕ_m	0.45 (0.08)	[0.32, 0.58]	1.01	297.65
Standard deviation				
σ_s	10.39 (1.60)	[7.95, 12.56]	1.00	6,030.31
σ_d	1.49 (1.08)	[0.13, 3.55]	1.00	3,104.66
τ_d	11.17 (0.88)	[9.98, 12.69]	1.00	6,313.74
τ_m	11.16 (0.84)	[9.96, 12.71]	1.01	283.68
σ_ε	10.80 (1.04)	[8.82, 12.10]	1.01	269.47

Note: $\hat{R}$: convergence diagnostic; ESS: effective sample size; β: grand mean; ϕ_d and ϕ_m: autoregression effect of AR(1) process for day and moment levels; σ_s, σ_d, σ_ε: SDs of person random effect, day random effect, and measurement error; and τ_d and τ_m: stationary SD of AR(1) process for day and moment levels.

model parameters. The Stan code and R code for data processing, model fitting, and result visualization are available on the Open Science Framework (OSF) at https://osf.io/hnkuc. The original data for the Exam Results Study can be accessed from the EMOTE database or from the authors' OSF repository.[19]

5.2. Results

The posterior means, posterior standard deviations, and 95% highest density intervals (HDIs) of the posterior distribution of the key model parameters are summarized in Table 3. Convergence was achieved for all parameters for our proposed model M1 (Equation (3.4)), with R-hat ($\hat{R}$) values below 1.015 (Vehtari et al., 2021) and a minimal effective sample size of 269.5 for the measurement error SD, σ_ε. Trace plots and autocorrelation plots of the key parameters confirmed good mixing and convergence (see the results on OSF).

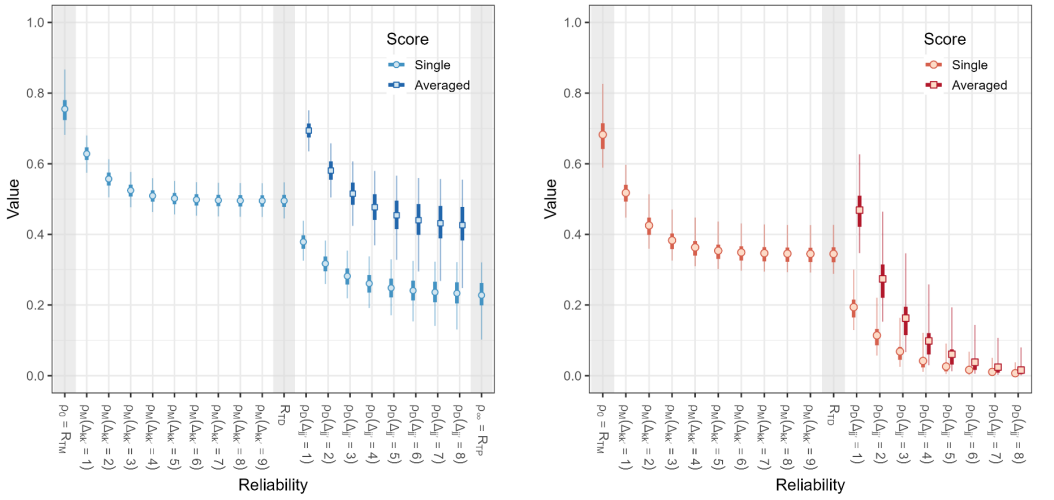
The estimated between-person standard deviation $\hat{\sigma}_s$ equals 10.39, suggesting substantial variability in negative affect levels among individuals. The day-level standard deviation ($\hat{\sigma}_d = 1.49$) is estimated to be relatively small, indicating that day-to-day fluctuations in negative affect are less pronounced as compared to between-person differences $\hat{\sigma}_s$ and the estimated measurement error $\hat{\sigma}_\varepsilon = 10.80$. This also means that there was only little variation between days, after accounting for carry-over effects. The measurement error quantifies the size of the average uncorrelated momentary deviations. Because there are no serial dependencies, it is a pure noise process. However, the day-to-day variation is also influenced by the serial correlation process, which has a stationary standard deviation of $\hat{\tau}_d = 11.17$. The stationary standard deviation of the moment-to-moment serial correlation process is of the same order of magnitude (i.e., $\hat{\tau}_m = 11.16$).

The relatively large magnitudes of the stationary standard deviations are partly due to the moderately sized autoregressive parameters, both for day-level ($\hat{\phi}_d = 0.57$) and moment-level ($\hat{\phi}_m = 0.45$). For

the number of warm-up iterations does not need to match the number of sampling iterations, and optimizing this ratio can save considerable computation time.

[19]The EMOTE database is available at https://emotedatabase.com/datasets/2, and the original OSF repository can be found at https://osf.io/bmaf2.

Downloaded from https://www.cambridge.org/core. 13 Nov 2026 at 03:02:29, subject to the Cambridge Core terms of use.



(a) Between-person level (b) Within-person level

Figure 1. Between- (Panel (a)) and within-person (Panel (b)) reliability estimates for the single-item negative affect assessment in the Exam Results Study. In each panel, there are three (or two) gray vertical strips, corresponding to the instantaneity ($\rho_0 = R_{TM}$), moment-to-moment long-term reliability (R_{TD}), and stability ($\rho_\infty = R_{TP}$, only for between-person level). In between, we find the moment-to-moment reliabilities ρ_M (for time lag $\Delta_{kk'} = 1, \ldots, 9$) and day-to-day reliabilities ρ_D (for time lag $\Delta_{jj'} = 1, \ldots, 8$). All reliabilities are for the single negative affect item, but for comparison, the reliability of the average day score is added as well.

both parameters, the posteriors are substantially larger than zero. However, the values of the day and moment autoregressive parameters are not directly comparable because they depend on time intervals with different units between assessments (1 day at the day level and around 1.2 hours at the moment level). This suggests considerable variability in negative affect within individuals over time.

Based on the estimated parameters and the formulas presented in this article, we can calculate all the reliability coefficients. In Figure 1, both the between-person (left panel) and within-person (right panel) reliability estimates are presented. In each plot, the x-axis represents the time lag between two assessments. The time lag ranges from 0 (for the instantaneity) across 10 moments ($\Delta_{kk'} = 1, \ldots, 9$) and then across 9 days ($\Delta_{jj'} = 1, \ldots, 8$), ending with a very large time lag (for the stability). The reliability estimates lie on the y-axis. In the plot, each dot represents the posterior mean of the reliability estimates, and the thick and thin vertical bars represent the 50% and 95% HDIs, respectively. These results are also given in Table B1 in Appendix B.

We can now address our three research questions using the reliability estimates presented in Figure 1. In short, these reliability coefficients indicate how well the single-item instrument can distinguish measurements across different time intervals: from very short intervals (nearly instantaneous), to short intervals (hours), to longer intervals (days), and ultimately very long intervals (reflecting stability).

To answer question (1), we quantify how well the single-item measure reflects a trait-like tendency that allows us to distinguish between persons. We look at the long-term between-person reliability, or stability R^B_{TP}. As shown in panel (a) of Figure 1, this value is 0.23 (95% HDI [0.14, 0.31]). This low coefficient indicates that only 23% of the variance in a single measurement can be attributed to stable, trait-like differences between individuals. Consequently, the current design combined with a single-item measure has very limited utility for distinguishing people based on a stable negative affect level.

To answer question (2), we need to determine the extent to which repeating the measurement under the exact same conditions at a given moment would yield (dis)similar results. We refer to the instantaneous reliability coefficient. Between-person instantaneous reliability $R^B_{TM} = 0.76$ (95% HDI [0.69,0.84]), suggesting that the measurements can effectively differentiate individuals' current states at a specific moment. Concurrently, within-person instantaneous reliability $R^W_{TM} = 0.68$ (95% HDI [0.60,0.79]), indicating that it reasonably captures a person's momentary state.

Downloaded from https://www.cambridge.org/core. 13 Nov 2026 at 03:02:29, subject to the Cambridge Core terms of use.

Finally, to answer question (3), we quantify the degree of agreement of single-item measurements across time. This question can be separated into two parts: between-person reliability (see Figure 1a) and within-person reliability (see Figure 1b). At the between-person level, the instantaneous reliability initially drops steeply but then stabilizes around $R^{B}_{TD} = 0.50$ (95% HDI [0.46,0.54]) after three time units ($\Delta_{kk'} = 3$), corresponding approximately to $3.6 = 12/10 * (4-1)$ hours. When considering longer time intervals, the between-person reliability is approximately 0.38 (95% HDI [0.33,0.43]) for one day and stabilizes after four days to 0.23 (the stability reliability). This shows that the negative affect scale is capable of distinguishing between persons only across time intervals as small as 1.2 hours. This is partially due to the low autocorrelation between the observations made at different moments and on different days, meaning that negative affects do not persist over time. Averaging observations over days increases reliability to allow the distinction between students when measurements are taken one day apart ($\rho^{B}_{D}(\Delta_{jj'} = 1)$=0.70, 95% HDI [0.65,0.74]) but decreases too much after two days. The primary trade-off of using average daily scores is the loss of the personal dynamic patterns within the day.

In contrast, at the within-person level, the instantaneous reliability quickly drops to stabilize around 0.35 (95% HDI [0.30,0.41]) after 3.6 hours within the same day. Day-to-day similarity is around 0.19 (95% HDI [0.14,0.17]) after one day lag, and rapidly decreases. This shows that, within a typical person, the negative affect does not persist in time. This pattern demonstrates that, for a typical individual, negative affect is highly dynamic and does not persist strongly over time, due to the combination of the small autoregressive effect and stationary day variance. The measurement is sensitive enough to capture momentary states, but these states are transient, highlighting the fluctuating nature of emotional experiences. Likewise, using average daily scores can increase reliability, but only to a limited extent.

6. Discussion

The main contribution of this article is to propose a general framework for single-item reliability for a common type of ILD (moments within days within persons). Starting from a three-level LMM with two AR(1) processes to account for the multilevel structure as well as the serial correlation structure, a family of reliability coefficients (both at the between-person and within-person level) can be derived and interpreted based on two different approaches: the correlation and variance partitioning approaches.

The correlation approach defines reliability at between- and within-person levels as a correlation (conditioned on a specific person at the within-person level) between two measurements taken at different time points, emphasizing the sheer degree of similarity between them. Specifically, we separate test–retest reliability into four types based on the time interval considered: instantaneity (extremely short-term, ρ_0), moment-to-moment (ρ_M), day-to-day (ρ_D), and stability (extremely long-term, ρ_∞), reflecting the three-level structure of the ILD. This is crucial in the presence of ILD, as ignoring the serial correlation may lead to biased estimates of the variance components, which in turn result in misleading reliability estimates.

In contrast, the variance partitioning approach defines reliability as the proportion of total variance to the systematic variance (e.g., variance of the person's true scores). It stresses the ability of the measurement instrument to distinguish the target of interest from measurement error. We further extended Laenen et al.'s (2007) reliability R_T into three subtypes: reliability of the momentary true scores (R_{TM}), daily true scores (R_{TD}), and person (trait) scores (R_{TP}). Each of them can be further defined at between- and within-person levels, with the exception of within-person reliability of the trait score (R^{W}_{TP}), which is undefined. We have elucidated the connections between the various coefficients. In addition, we showed how this family of coefficients is directly related to reliability coefficients arising from simpler models, most notably the CTT model.

Finally, our framework has been applied to the empirical Exam Results Study (Kalokerinos et al., 2019). As expected, the test–retest reliability (based on the correlation interpretation) gradually decreased as the time lag increased, and the between-person reliability (ranging from 0.23 to 0.78) was uniformly higher than the within-person reliability (ranging from 0.01 to 0.68). Based on these

Downloaded from https://www.cambridge.org/core. 13 Nov 2026 at 03:02:29, subject to the Cambridge Core terms of use.

results, we can conclude that after around 3.6 hours within a day, the between-subject agreement has reached its asymptote. In the same vein, the between-subject agreement of measurements taken about four days apart is close to that of the long-term stability. This gives some hints for researchers about the optimal time interval between measurements when designing an intensive longitudinal study to achieve a desired level of reliability. For instance, Adolf et al. (2021) provide a methodological framework for determining optimal sampling rates in first-order autoregressive models, which are commonly used in such studies. Furthermore, the reliability of averaged daily scores is higher than that of single scores. This improvement was more pronounced for the between-person reliability than the within-person reliability. This study underscores that reliability is not a static value but one that depends on the specific context.

A first implication of our approach is that it extends the modern psychometric viewpoint that reliability is a property of observed scores, and thus depends on the particular sample and context (such as sample characteristics, measurement conditions, and temporal factors), rather than an immutable characteristic of the measurement instrument (e.g., Streiner, 2003; Thompson, 2003). In the context of ILD, our approach demonstrates that the reliability of these scores is not solely dependent on sample and setting, but is deeply dynamic and sensitive to the temporal structure of the data and the chosen level of analysis. Our approach aligns with the current trends emphasizing reliability as a multifaceted property shaped by temporal variation and the analytic time span (Tisak & Tisak, 1996; Vangeneugden et al., 2004, 2005). Different types of reliability can further be distinguished based on the time interval considered (e.g., instantaneity and stability). In addition, reliability can be different depending on the chosen level of analysis, for instance, whether the focus is on the whole population of people or a specific person across different occasions (Lin et al., 2025; Schuurman & Hamaker, 2019). The choice of one or more reliability coefficients depends on the research question, study design, and the analysis level of interest.

A second implication of our work is that in the variance partitioning approach, an explicit decision is needed about the partitioning of the variance components. This means that, depending on the context, a variance component can be considered as unsystematic (error) in one context and as systematic in another. The same decision shows up in the correlation approach because the choice of the coefficient carries with it an assumption about the extent to which a score generalize across different time points or situations (e.g., when looking at day-to-day reliability, all within-day process are treated as error). This conceptual flexibility is not unique to our framework. For instance, in factor analysis, McDonald's (1999) omega is extended as the reliability of the general factor (ω_g) and the total reliability of the test (ω_t). In that context, some specific factor variances are treated as error when calculating ω_g but as systematic variance when calculating ω_t (Revelle & Condon, 2019; Zinbarg et al., 2005).

A major theme underlying the choice of coefficient or the variance partitioning is the status of the serial processes (i.e., $\tau_d^2\phi_d^{\Delta_{jj'}}$ and $\tau_m^2\phi_m^{\Delta_{kk'}}$): Are these error processes or should they be considered as systematic? When treated as a systematic source of variance, high serial correlation arising from carry-over effects such as practice, fatigue, or memory may artificially inflate reliability, as these effects do not reflect true changes in the underlying latent state. For such cases, we recommend using the stability reliability, ρ_∞ (analogous to the reliability of a person's trait score, R_{TP}), which treats the serial process as error or unsystematic. However, the decision of selecting a reliability coefficient depends on the research context. In applications where physiological indicators such as blood pressure or blood glucose are measured, serial correlation is an inherent and meaningful part of the construct. In these contexts, short-term test–retest reliability and instantaneous reliability, ρ_0 (analogous to the reliability of a momentary score, R_{TM}), may be the most appropriate and informative, depending on the specific research objectives. For the measurement of affect and mood, for which ILD designs are often used in psychology, researchers are often interested in the autocorrelation, and the serial processes may thus be considered as systematic.

A third implication of our work is that it shows very clearly that the two approaches to define reliability (correlation and variance partitioning) are closely related, also in complex designs.

Downloaded from https://www.cambridge.org/core. 13 Nov 2026 at 03:02:29, subject to the Cambridge Core terms of use.

Specifically, under a particular model, we analytically demonstrate that the correlation-based reliability coefficients, expressed as functions of the variance components, are equivalent or almost equivalent to the variance partitioning-based reliability coefficients (depending on the treatment of the time span). Our framework, therefore, helps to unify the two perspectives and provides a deeper understanding of the reliability concept in intensive longitudinal measurements. In addition, this article facilitates the comparison of the two approaches under different models and scenarios. The reliability coefficients proposed in this article can be seen as extensions of reliability coefficients based on CTT, generalizability coefficients and dependability coefficients in GT, and ICCs derived under the ANOVA framework when considering LMM without serial correlation (Lin et al., 2025; Vangeneugden et al., 2004, 2005).

A major advantage of the correlation approach over the variance partitioning approach is that it provides a principled way of defining a family of reliability coefficients. Moreover, the interpretation is immediately clear because the reliability coefficient quantifies to which extent the observed score generalizes to another moment, day, or person.

6.1. Limitation and future work

Some limitations of the present study should be noted, along with corresponding avenues for future research. First, any reliability coefficient is contingent on the statistical model it is derived from. This holds for coefficients defined in the context of CTT or GT, and our approach is no exception. In this article, we use the equivalent of an empty model as is common practice in the multilevel literature (see, e.g., Hox, 2010; Snijders & Bosker, 2012). In multilevel modeling, one often starts with fitting an empty model to study the correlation between responses and the partitioning of the variance across the different levels (Hox, 2010). The major difference is that we have added the two serial correlation processes to account for possible time dependencies. Obviously, the empty model could be further extended with adjustments to the serial correlation process (e.g., more complicated than an AR(1) process) or heterogeneity of variances (e.g., making them dependent on the day). Such extensions would, however, require careful consideration of model identifiability. In the second step, covariates can be included in the model to explain the variance.

Second, the present study focuses on theoretically defining the reliability of single-item measures in ILD under a three-level LMM with autocorrelation processes. However, other approaches for estimating single-item reliability in ILD have been proposed in literature. For instance, as highlighted in the previous paragraph, the ML-MEAR framework (Schuurman & Hamaker, 2019; Schuurman et al., 2015) utilizes a state-space representation that allows for person-specific AR(1) parameters and individual measurement error variances, enabling the computation of personalized within-person reliability coefficients. In addition, this framework can be extended to accommodate multiple items (i.e., ML-MEVAR, where "V" stands for vector), which partitions variance while accounting for cross-lagged effects between items. Nevertheless, unlike our proposed framework, the ML-ME(V)AR does not explicitly distinguish between reliability at different hierarchical time levels (e.g., momentary vs. daily reliability), nor does it evaluate how between- and within-person reliability dynamically changes as a function of the time lag—both of which are central conceptual contributions of our current article.

Alternatively, Dejonckheere et al. (2022) proposed a model-free (design-based) approach that leverages specialized data collection schemes (e.g., immediate and delayed retests) to empirically isolate measurement error variance. This model-free approach avoids the need to specify or correctly guess an underlying autocorrelation structure. However, because estimating instantaneous reliability through this method requires a programmed immediate retest design, which is extremely rare in existing or retrospective ILD datasets, we did not incorporate it into our generalized model-based analytical framework. For a comprehensive review and empirical comparison of these methods, we direct readers to recent work by Leertouwer et al. (2025).

Third, this article focuses exclusively on the reliability of single-item measures. A critical limitation of single-item measures is their heightened susceptibility to measurement error, compared to composite scores. Consequently, some intensive longitudinal studies employ two or more items (or multiple raters)

Downloaded from https://www.cambridge.org/core. 13 Nov 2026 at 03:02:29, subject to the Cambridge Core terms of use.

to measure a construct, allowing for the quantification of internal consistency or inter-rater reliability (Castro-Alvarez, Tendeiro, Meijer, et al., 2022; Cranford et al., 2006; Geldhof et al., 2014; Hu et al., 2016; Koch et al., 2025; Lai, 2021; Nacke & Mayer, 2025; Nezlek, 2017; Shrout & Lane, 2012a; Xiao et al., 2023) for a comprehensive review, see castroalvarez et al., 2025, leertouwer et al., 2025). Therefore, the extension of the current framework to accommodate multiple-item measures represents a valuable and necessary direction for future research. A promising step in this direction is the recent work by Hunter (2026), who proposed a coefficient κ under the multilevel state-space modeling framework to estimate dynamic reliability, which can encompass many of the aforementioned models as special cases.

Finally, the current framework is tailored for continuous outcomes with a normality assumption. However, data in intensive longitudinal studies are often of other types, such as binary, ordinal, or bounded (as with our negative affect data, which were bounded from 0 to 100). A valuable extension of this work would be to adapt the framework for such data types. This would likely involve a generalized linear mixed-effects model (Molenberghs et al., 2007) or another suitable approach (e.g., Vanbelle et al., 2025, for a comparison review on binary scales), and subsequently redefining the reliability coefficients within that new context.

Data availability statement. The preprocessed data, code for analysis, and supplementary materials can be found in Open Science Framework (OSF): https://osf.io/hnkuc.

Acknowledgements. The resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation – Flanders (FWO) and the Flemish Government. The authors acknowledge the use of Google Gemini 2.5 Pro for assistance with proofreading and correcting grammatical errors in this article. All intellectual content, data analysis, and the core text were originally conceived and written by the authors, who take full responsibility for the final manuscript's content and integrity.

Author contributions. Conceptualization: T.-Y.L., F.T., and S.V.; Formal analysis: T.-Y.L.; Funding acquisition: F.T. and S.V.; Methodology: T.-Y.L.; Software: T.-Y.L.; Supervision: F.T. and S.V.; Writing—original draft: T.-Y.L.; Writing—review and editing: T.-Y.L., F.T., and S.V. All authors approved the final submitted draft.

Funding statement. This research was supported by grants from the Research Council of the KU Leuven (GPMU/22/014) awarded to S.V. and F.T., and from (C14/23/062) awarded to F.T.

Competing interests. The authors declare none.

Ethical standards. The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

References

Adolf, J. K., Loossens, T., Tuerlinckx, F., & Ceulemans, E. (2021). Optimal sampling rates for reliable continuous-time first-order autoregressive and vector autoregressive modeling. *Psychological Methods*, *26*(6), 701–718. https://doi.org/10.1037/met0000398

Ariens, S., Ceulemans, E., & Adolf, J. K. (2020). Time series analysis of intensive longitudinal data in psychosomatic research: A methodological overview. *Journal of Psychosomatic Research*, *137*, 110191. https://doi.org/10.1016/j.jpsychores.2020.110191

Bartko, J. J. (1966). The intraclass correlation coefficient as a measure of reliability. *Psychological Reports*, *19*(1), 3–11. https://doi.org/10.2466/pr0.1966.19.1.3

Bolger, N., & Laurenceau, J.-P. (2013). *Intensive longitudinal methods: An introduction to diary and experience sampling research*. The Guilford Press.

Boyle, G. J., Stankov, L., & Cattell, R. B. (1995). In D. H. Saklofske, & M. Zeidner (Eds.). *Measurement and statistical models in the study of personality and intelligence*. Springer. https://doi.org/10.1007/978-1-4757-5571-8_20

Brennan, R. L., & Kane, M. T. (1977a). An index of dependability for mastery tests. *Journal of Educational Measurement*, *14*(3), 277–289. https://www.jstor.org/stable/1434319

Brennan, R. L., & Kane, M. T. (1977b). Signal/noise ratios for domain-referenced tests. *Psychometrika*, *42*(4), 609–625. https://doi.org/10.1007/BF02295983

Brose, A., Schmiedek, F., Gerstorf, D., & Voelkle, M. C. (2020). The measurement of within-person affect variation. *Emotion*, *20*(4), 677–699. https://doi.org/10.1037/emo0000583

Downloaded from https://www.cambridge.org/core. 13 Nov 2026 at 03:02:29, subject to the Cambridge Core terms of use.

Calamia, M. (2019). Practical considerations for evaluating reliability in ambulatory assessment studies. *Psychological Assessment*, *31*(3), 285–291. https://doi.org/10.1037/pas0000599

Castro-Alvarez, S., Bringmann, L. F., Back, J., & Liu, S. (2025). The many reliabilities of psychological dynamics: An overview of statistical approaches to estimate the internal consistency reliability of intensive longitudinal data. *Psychological Methods*. https://doi.org/10.1037/met0000778

Castro-Alvarez, S., Tendeiro, J. N., de Jonge, P., Meijer, R. R., & Bringmann, L. F. (2022). Mixed-effects trait-state-occasion model: Studying the psychometric properties and the person-situation interactions of psychological dynamics. *Structural Equation Modeling: A Multidisciplinary Journal*, *29*(3), 438–451. https://doi.org/10.1080/10705511.2021.1961587

Castro-Alvarez, S., Tendeiro, J. N., Meijer, R. R., & Bringmann, L. F. (2022). Using structural equation modeling to study traits and states in intensive longitudinal data. *Psychological Methods*, *27*, 17–43. https://doi.org/10.1037/met0000393

Cattell, R. B. (1964). Validity and reliability: A proposed more basic set of concepts. *Journal of Educational Psychology*, *55*(1), 1–22. https://doi.org/10.1037/h0046462

Cranford, J. A., Shrout, P. E., Iida, M., Rafaeli, E., Yip, T., & Bolger, N. (2006). A procedure for evaluating sensitivity to within-person change: Can mood measures in diary studies detect change reliably? *Personality & Social Psychology Bulletin*, *32*(7), 917–929. https://doi.org/10.1177/0146167206287721

Cronbach, L. J. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. John Wiley & Sons.

Dejonckheere, E., Demeyer, F., Geusens, B., Piot, M., Tuerlinckx, F., Verdonck, S., & Mestdagh, M. (2022). Assessing the reliability of single-item momentary affective measurements in experience sampling. *Psychological Assessment*, *34*(12), 1138–1154. https://doi.org/10.1037/pas0001178

Diggle, P. J. (1988). An approach to the analysis of repeated measurements. *Biometrics*, *44*(4), 959–971. https://doi.org/10.2307/2531727

Fuller-Tyszkiewicz, M., Hartley-Clark, L., Cummins, R. A., Tomyn, A. J., Weinberg, M. K., & Richardson, B. (2017). Using dynamic factor analysis to provide insights into data reliability in experience sampling studies. *Psychological Assessment*, *29*(9), 1120–1128. https://doi.org/10.1037/pas0000411

Gabry, J., Češnovar, R., Johnson, A., & Bronder, S. (2025). *CmdStanR: R interface to 'CmdStan'*. [R package version 0.9.0]. https://mc-stan.org/cmdstanr/

Geldhof, G. J., Preacher, K. J., & Zyphur, M. J. (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological Methods*, *19*, 72–91. https://doi.org/10.1037/a0032138

Ghiselli, E. E., Campbell, J. P., & Zedeck, S. (1981). *Measurement theory for the behavioral sciences*. W. H. Freeman.

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

Hamaker, E. L. (2025). Analysis of intensive longitudinal data: Putting psychological processes in perspective. *Annual Review of Clinical Psychology*, *21*(1), 379–405. https://doi.org/10.1146/annurev-clinpsy-081423-022947

Hamaker, E. L., Ceulemans, E., Grasman, R. P. P. P., & Tuerlinckx, F. (2015). Modeling affect dynamics: State of the art and future challenges. *Emotion Review*, *7*(4), 316–322. https://doi.org/10.1177/1754073915590619

Ho, M.-H. R., Shumway, R., & Ombao, H. (2006). The state-space approach to modeling dynamic processes. In T. A. Walls, & J. L. Schafer (Eds.), *Models for intensive longitudinal data* (pp. 148–175). Oxford University Press.

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications*. (2nd ed.) Routledge/Taylor & Francis Group.

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: A Multidisciplinary Journal*, *23*(4), 532–543. https://doi.org/10.1080/10705511.2016.1148605

Hunter, M. D. (2026). Extending reliability to intensive longitudinal data with the Kalman filter [advance online publication]. *British Journal of Mathematical and Statistical Psychology*. https://doi.org/10.1111/bmsp.70039

Kalokerinos, E. K., Erbas, Y., Ceulemans, E., & Kuppens, P. (2019). Differentiate to regulate: Low negative emotion differentiation is associated with ineffective use but not selection of emotion-regulation strategies. *Psychological Science*, *30*(6), 863–879. https://doi.org/10.1177/0956797619838763

Koch, T., Jaehne, M. F., Riediger, M., Rauers, A., & Holtmann, J. (2025). Idiographic interrater reliability measures for intensive longitudinal multirater data [advance online publication]. *British Journal of Mathematical and Statistical Psychology*. https://doi.org/10.1111/bmsp.70022

Laenen, A., Alonso, A., & Molenberghs, G. (2007). A measure for the reliability of a rating scale based on longitudinal clinical trial data. *Psychometrika*, *72*(3), 443. https://doi.org/10.1007/s11336-007-9002-7

Laenen, A., Alonso, A., Molenberghs, G., & Vangeneugden, T. (2009a). A family of measures to evaluate scale reliability in a longitudinal setting. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *172*(1), 237–253. https://doi.org/10.1111/j.1467-985X.2008.00554.x

Laenen, A., Alonso, A., Molenberghs, G., & Vangeneugden, T. (2009b). Reliability of a longitudinal sequence of scale ratings. *Psychometrika*, *74*(1), 49. https://doi.org/10.1007/s11336-008-9079-7

Lai, M. H. C. (2021). Composite reliability of multilevel data: It's about observed scores and construct meanings. *Psychological Methods*, *26*(1), 90–102. https://doi.org/10.1037/met0000287

Downloaded from https://www.cambridge.org/core. 13 Nov 2026 at 03:02:29, subject to the Cambridge Core terms of use.

Leertouwer, I., Keijsers, L., van Roekel, E., & Schuurman, N. K. (2025). *A practical guide to estimating reliability of intensive longitudinal data*. https://doi.org/10.31234/osf.io/uq4sk_v1

Lin, T.-Y., Tuerlinckx, F., & Vanbelle, S. (2025). Reliability for multilevel data: A correlation approach. *Psychological Methods*. https://doi.org/10.1037/met0000738

Lodewyckx, T., Tuerlinckx, F., Kuppens, P., Allen, N. B., & Sheeber, L. (2011). A hierarchical state space approach to affective dynamics. *Journal of Mathematical Psychology*, *55*(1), 68–83. https://doi.org/10.1016/j.jmp.2010.08.004

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

Ma, Q., Mermelstein, R., & Hedeker, D. (2020). A three-level mixed model to account for the correlation at both the between-day and the within-day level for ecological momentary assessments. *Health Services and Outcomes Research Methodology*, *20*(4), 247–264. https://doi.org/10.1007/s10742-020-00220-w

Marcoulides, K. M. (2019). Reliability estimation in longitudinal studies using latent growth curve modeling. *Measurement: Interdisciplinary Research and Perspectives*, *17*(2), 67–77. https://doi.org/10.1080/15366367.2018.1522169

McDonald, R. P. (1999). *Test theory: A unified treatment*. Lawrence Erlbaum Associates Publishers.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*(1), 30–46. https://doi.org/10.1037/1082-989X.1.1.30

Molenberghs, G., Laenen, A., & Vangeneugden, T. (2007). Estimating reliability and generalizability from hierarchical biomedical data. *Journal of Biopharmaceutical Statistics*, *17*(4), 595–627. https://doi.org/10.1080/10543400701329448

Nacke, L., & Mayer, A. (2025). Level-specific reliability coefficients from the perspective of latent state-trait theory [advance online publication]. *British Journal of Mathematical and Statistical Psychology*. https://doi.org/10.1111/bmsp.70027

Nezlek, J. B. (2007). A multilevel framework for understanding relationships among traits, states, situations and behaviours. *European Journal of Personality*, *21*(6), 789–810. https://doi.org/10.1002/per.640

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., & Gable, S. L. (2001). Depression as a moderator of relationships between positive daily events and day-to-day psychological adjustment. *Personality and Social Psychology Bulletin*, *27*(12), 1692–1704. https://doi.org/10.1177/01461672012712012

Piepho, H.-P., & Ogutu, J. O. (2007). Simple state-space models in a mixed model framework. *The American Statistician*, *61*(3), 224–232. https://www.jstor.org/stable/27643898

R Core Team. (2025). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Revelle, W., & Condon, D. M. (2018). *Reliability*. Wiley Blackwell. https://doi.org/10.1002/9781118489772.ch23

Revelle, W., & Condon, D. M. (2019). Reliability from α to ω: A tutorial. *Psychological Assessment*, *31*(12), 1395–1411. https://doi.org/10.1037/pas0000754

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/abstract/DC1F07F1E5D4B09PQ/1

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*. https://doi.org/10.3389/fpsyg.2015.01038

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Shrout, P. E., & Lane, S. P. (2012a). Psychometrics. In M. R. Mehl, & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 302–320). The Guilford Press. https://psycnet.apa.org/record/2012-05165-017

Shrout, P. E., & Lane, S. P. (2012b). Reliability. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), *APA handbook of research methods in psychology, vol 1: Foundations, planning, measures, and psychometrics* (pp. 643–660). American Psychological Association. https://doi.org/10.1037/13619-034

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling: An introduction to basic and advanced multilevel modeling*. (2nd ed.) [Hier auch später erschienene, unveränderte Nachdrucke. - Literaturverz. S. [332] - 347]). Sage Publishers.

Stan Development Team. (2025a). Stan reference manual (version 2.36). https://mc-stan.org

Stan Development Team. (2025b). Stan user's guide (version 2.36). https://mc-stan.org

Stone, A. A., & Shiffman, S. (1994). Ecological momentary assessment (EMA) in behavorial medicine. *Annals of Behavioral Medicine*, *16*(3), 199–202. https://doi.org/10.1093/abm/16.3.199

Streiner, D. L. (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment*, *80*(1), 99–103. https://doi.org/10.1207/S15327752JPA8001_18

ten Hove, D., Jorgensen, T. D., & van der Ark, L. A. (2022). Interrater reliability for multilevel data: A generalizability theory approach. *Psychological Methods*, *27*(4), 650–666. https://doi.org/10.1037/met0000391

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

Downloaded from https://www.cambridge.org/core. 13 Nov 2026 at 03:02:29, subject to the Cambridge Core terms of use.

Tisak, J., & Tisak, M. S. (1996). Longitudinal models of reliability and validity: A latent curve approach. *Applied Psychological Measurement*, *20*(3), 275–288. https://doi.org/10.1177/014662169602000307

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*, *129*(1), 56–63. https://doi.org/10.1037/abn0000473

Tsimikas, J. V., & Ledolter, J. (1997). Mixed model representation of state space models: New smoothing results and their application to REML estimation. *Statistica Sinica*, *7*(4), 973–991. https://www.jstor.org/stable/24306167

Tuerlinckx, F., Kuppens, P., Ariens, S., Cloos, L., Dejonckheere, E., Lafit, G., Niemeijer, K., Revol, J., Schat, E., Schreuder, M., Vanhasbroeck, N., & Ceulemans, E. (2025). New developments in experience sampling methodology. *British Journal of Mathematical and Statistical Psychology*. https://doi.org/10.1111/bmsp.12398

Vanbelle, S., Engelhart, C. H., & Blix, E. (2025). Measuring agreement in diagnostics: A practical guide for researchers. *Statistics in Medicine*, *44*(23–24), e70299. https://doi.org/10.1002/sim.70299

Vangeneugden, T., Laenen, A., Geys, H., Renard, D., & Molenberghs, G. (2004). Applying linear mixed models to estimate reliability in clinical trial data with repeated measurements. *Controlled Clinical Trials*, *25*(1), 13–30. https://doi.org/10.1016/j.cct.2003.08.009

Vangeneugden, T., Laenen, A., Geys, H., Renard, D., & Molenberghs, G. (2005). Applying concepts of generalizability theory on clinical trial data to investigate sources of variation and their impact on reliability. *Biometrics*, *61*(1), 295–304. https://doi.org/10.1111/j.0006-341X.2005.031040.x

Vansteelandt, K., & Verbeke, G. (2016). A mixed model to disentangle variance and serial autocorrelation in affective instability using ecological momentary assessment data. *Multivariate Behavioral Research*, *51*(4), 446–465. https://doi.org/10.1080/00273171.2016.1159177

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of MCMC. *Bayesian Analysis*, *16*(2). [arXiv:1903.08008 [stat]]. https://doi.org/10.1214/20-BA1221

Verbeke, G., & Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. Springer. https://doi.org/10.1007/978-1-4419-0300-6

Vogelsmeier, L. V. D. E., Jongerling, J., & Maassen, E. (2024). Assessing and accounting for measurement in intensive longitudinal studies: Current practices, considerations, and avenues for improvement. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation*, *33*(8), 2107–2118. https://doi.org/10.1007/s11136-024-03678-0

Walls, T. A., & Schafer, J. L. (2006). *Models for intensive longitudinal data*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195173444.001.0001

Wilhelm, P., & Schoebi, D. (2007). Assessing mood in daily life. *European Journal of Psychological Assessment*, *23*(4), 258–267. https://doi.org/10.1027/1015-5759.23.4.258

Xiao, Y., Wang, P., & Liu, H. (2023). Assessing intra- and inter-individual reliabilities in intensive longitudinal studies: A two-level random dynamic model-based approach. *Psychological Methods*. https://doi.org/10.1037/met0000608

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β, and Mcdonald's ω_h: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, *70*(1), 123–133. https://doi.org/10.1007/s11336-003-0974-7

A. Derivations of reliability coefficients using the correlation approach

In this appendix, we provide a detailed derivation of the between-person and within-person reliability coefficients utilized in the correlation approach (e.g., Lin et al., 2025; Vangeneugden et al., 2004, 2005). While the overarching logic follows the proofs outlined by Lin et al. (2025), we extend the derivations to accommodate the generalized three-level linear mixed-effects model featuring two autoregressive, AR(1), processes.

Let y_{ijk} be the observed score of an intensive longitudinal design for person $i \in \mathcal{S}$ on day $j \in \mathcal{D}$ at moment $k \in \mathcal{M}$, where $\mathcal{S}$, $\mathcal{D}$, and $\mathcal{M}$ are the sets of all persons, days, and moments, respectively. Theoretically, we assume that $\mathcal{S}$, $\mathcal{D}$, and $\mathcal{M}$ may contain a (countably) infinite or sufficiently large finite number of elements, meaning that finite-population corrections for sampling are negligible. The data-generating model is defined as

$$y_{ijk} = \beta_0 + s_i + d_{ij} + v_{ij} + \omega_{ijk} + \varepsilon_{ijk}, \tag{A.1}$$

where β_0 is the grand mean; $s_i \sim \mathcal{N}(0, \sigma_s^2)$ is the person random effect; $d_{ij} \sim \mathcal{N}(0, \sigma_d^2)$ is the day random effect; v_{ij} is the day-level AR(1) process with stationary variance τ_d^2 and autoregressive parameter ϕ_d, ω_{ijk} is the moment-level AR(1) process with stationary variance τ_m^2 and autoregressive parameter ϕ_m; and $\varepsilon_{ijk} \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ is the measurement error. All aforementioned random variables are assumed to be independent of each other and have finite variances.

Before deriving the correlation coefficients, some fundamental statistical concepts must be clarified.

Downloaded from https://www.cambridge.org/core. 13 Nov 2026 at 03:02:29, subject to the Cambridge Core terms of use.

A.1. Marginal versus conditional variance

To generalize the reliability results to the broader population of persons, days, and moments, we must consider the random mechanism that generates the observed scores. Formally, we can conceptually treat the observed score y_{IJK} as a random variable, where I, J, and K are random indices representing a randomly drawn person, day, and moment, respectively. This means that y_{ijk} is then a realization of y_{IJK} when $I = i$, $J = j$, and $K = k$. We assume the probability of drawing any specific person from the population is uniform, as is the probability of drawing any specific day or moment. Under this fully marginal perspective, the variance of the person effect s_I across the population is $\mathrm{Var}(s_I) = \sigma_s^2$.

Conversely, when we evaluate reliability strictly at the within-person level—meaning we consider the scores for a specific person i rather than the whole population—we condition on that specific person i. Under this condition, the variance of the person effect yields zero (i.e., $\mathrm{Var}(s_I \mid I = i) = 0$) because the person is fixed. However, the conditional variances of the other components (d_{IJ}, v_{IJ}, ω_{IJK}, and ε_{IJK} conditional on $I = i$) are not zero, because the days and/or moments have not been specified and thus remain random.

Note that for the simplicity of notation, in the main text and hereafter, we still use s_i to denote the person effect for a random person i from the population $\mathcal{S}$ and the same for y_{ijk} and the remaining random effects (i.e., d, v, ω, and ε). If we want to denote a specific person i, then we condition on that person i in the variance or covariance, $\mathrm{Var}(\cdot \mid i)$ or $\mathrm{Cov}(\cdot, \cdot \mid i)$, respectively.

A.2. Stationarity of the AR(1) process

An AR(1) process is *stationary* if its marginal distribution remains constant over time (specifically, it is weak stationary if its mean and variance are time-invariant). Taking the day-level AR(1) process v_{ij} as an example, it is formulated as

$$v_{ij} = \phi_d v_{i,\,j-1} + \eta_{ij}, \quad \eta_{ij} \sim \mathcal{N}(0, \tau_d^{*2}),$$

where τ_d^{*2} is the variance of the innovation (i.e., white noise) process η_{ij}. By the definition of stationarity, the variance should be constant over time, meaning $\mathrm{Var}(v_{ij}) = \mathrm{Var}(v_{ij'})$ for all $j, j' \in \mathcal{D}$. Therefore, we have

$$\begin{aligned} &\mathrm{Var}(v_{ij}) = \mathrm{Var}(v_{i,\,j+1}) = \mathrm{Var}(\phi_d v_{ij} + \eta_{i,\,j+1}) = \phi_d^2 \mathrm{Var}(v_{ij}) + \tau_d^{*2} \\ \Rightarrow\ &\mathrm{Var}(v_{ij}) = \frac{\tau_d^{*2}}{1 - \phi_d^2} \equiv \tau_d^2 \end{aligned}$$

for all $|\phi_d| < 1$. In the main text, we bypass the innovation variance τ_d^{*2} and simply directly parameterize this constant marginal variance as τ_d^2.

Consequently, the covariance between two observations with a time lag of $\Delta_{jj'} = |j - j'|$ is given by

$$\begin{aligned} \mathrm{Cov}(v_{i,\,j}, v_{i,\,j'}) &= \mathrm{Cov}(v_{i,\,j}, v_{i,\,j-\Delta_{jj'}}) = \mathrm{Cov}\left(\phi_d^{\Delta_{jj'}} v_{i,\,j-\Delta_{jj'}} + \sum_{l=0}^{\Delta_{jj'}-1} \phi_d^l \eta_{i,\,j-l}, v_{i,\,j-\Delta_{jj'}}\right) \\ &= \phi_d^{\Delta_{jj'}} \mathrm{Var}(v_{i,\,j-\Delta_{jj'}}) = \tau_d^2 \phi_d^{\Delta_{jj'}}, \end{aligned}$$

which is a geometric decay based on the autoregressive parameter ϕ_d. Note that if $\Delta_{jj'} = 0$ (i.e., on the same day), the covariance is exactly equal to the marginal stationary variance τ_d^2. By combining all days into a vector $\boldsymbol{v}_i = (v_{i1}, \ldots, v_{iD})^\top$, the constructed covariance matrix of the day-level AR(1) process perfectly corresponds to $\boldsymbol{H}_d$ presented in Section 3.1.

The identical stationarity principle applies to the variance and covariance of the moment-level AR(1) process, ω_{ijk}, when the autoregressive parameter $|\phi_m| < 1$. Specifically, $\mathrm{Var}(\omega_{ijk}) = \tau_m^2$, and $\mathrm{Cov}(\omega_{ijk}, \omega_{ij'k'}) = \tau_m^2 \phi_m^{\Delta_{kk'}}$ if measured on the same day ($j = j'$), and 0 if measured on different days ($j \neq j'$). Correspondingly, combining all moments as a vector $\boldsymbol{\omega}_{ij} = (\omega_{ij1}, \ldots, \omega_{ijK})^\top$ yields the covariance matrix $\boldsymbol{H}_m$ presented in Section 3.1.

A.3. The derivation of between-person reliability

Between-person test–retest reliability for single-item measures is conceptualized as the (marginal) correlation between two measurements for the same person taken at different time points (lagged by $\Delta_{jj'}$ days and $\Delta_{kk'}$ moments) across the population, which can be expressed as follows:

$$\rho^{\mathrm{B}}(\Delta_{jj'}, \Delta_{kk'}) = \mathrm{Cor}(y_{ijk}, y_{ij'k'}) = \frac{\mathrm{Cov}(y_{ijk}, y_{ij'k'})}{\sqrt{\mathrm{Var}(y_{ijk})\,\mathrm{Var}(y_{ij'k'})}}. \tag{A.2}$$

Note that the person i here is randomly selected from the people population $\mathcal{S}$, not fixed.

Downloaded from https://www.cambridge.org/core. 13 Nov 2026 at 03:02:29, subject to the Cambridge Core terms of use.

The numerator in Equation (A.2) depends on whether the two moments occur on the same day or not. Following the principles outlined in Appendices A.1 and A.2, distributing the covariance yields:

$$\begin{aligned}\mathrm{Cov}(y_{ijk}, y_{ij'k'}) &= \mathrm{Cov}(s_i + d_{ij} + v_{ij} + \omega_{ijk} + \varepsilon_{ijk}, s_i + d_{ij'} + v_{ij'} + \omega_{ij'k'} + \varepsilon_{ij'k'}) \\ &= \mathrm{Var}(s_i) + \mathrm{Cov}(d_{ij}, d_{ij'}) + \mathrm{Cov}(v_{ij}, v_{ij'}) + \mathrm{Cov}(\omega_{ijk}, \omega_{ij'k'}) + \mathrm{Cov}(\varepsilon_{ijk}, \varepsilon_{ij'k'}) \\ &= \begin{cases} \sigma_s^2 + \sigma_d^2 + \tau_d^2 + \tau_m^2 \phi_m^{\Delta_{kk'}}, & \text{if } j = j' \\ \sigma_s^2 + \tau_d^2 \phi_d^{\Delta_{jj'}}, & \text{if } j \neq j' \end{cases} \\ &= \sigma_s^2 + \delta_{jj'}\sigma_d^2 + \tau_d^2 \phi_d^{\Delta_{jj'}} + \delta_{jj'}\tau_m^2 \phi_m^{\Delta_{kk'}},\end{aligned}$$

where $\delta_{jj'}$ is the Kronecker delta, which equals 1 if $j = j'$ and 0 otherwise. The above equation shows that if the measurements are taken on the same day ($j = j'$), the covariance includes the person variance σ_s^2, the day variance σ_d^2, the day-level AR(1) variance τ_d^2 (i.e., $\Delta_{jj'} = 0$, $\phi_d^0 = 1$), and the decayed moment-level AR(1) covariance $\tau_m^2 \phi_m^{\Delta_{kk'}}$. If the measurements are taken on different days ($j \neq j'$), the day effects d_{ij} and $d_{ij'}$ are independent, and there is no covariance between moment processes because they occur on different days. Hence, the covariance merely includes the person variance σ_s^2 and the decayed day-level AR(1) covariance $\tau_d^2 \phi_d^{\Delta_{jj'}}$.

In contrast, based on the independence of the random components, the denominator in Equation (A.2) is the straightforward sum of all variance components:

$$\begin{aligned}\mathrm{Var}(y_{ijk}) &= \mathrm{Var}(s_i + d_{ij} + v_{ij} + \omega_{ijk} + \varepsilon_{ijk}) \\ &= \mathrm{Var}(s_i) + \mathrm{Var}(d_{ij}) + \mathrm{Var}(v_{ij}) + \mathrm{Var}(\omega_{ijk}) + \mathrm{Var}(\varepsilon_{ijk}) \\ &= \sigma_s^2 + \sigma_d^2 + \tau_d^2 + \tau_m^2 + \sigma_\varepsilon^2.\end{aligned}$$

Crucially, under the stationarity assumption, this variance is constant regardless of the specific time points (i.e., $\mathrm{Var}(y_{ij'k'}) = \mathrm{Var}(y_{ijk})$).

Combining the numerator and denominator, we have the between-person reliability coefficient as follows:

$$\rho^{\mathrm{B}}(\Delta_{jj'}, \Delta_{kk'}) = \frac{\sigma_s^2 + \delta_{jj'}\sigma_d^2 + \tau_d^2 \phi_d^{\Delta_{jj'}} + \delta_{jj'}\tau_m^2 \phi_m^{\Delta_{kk'}}}{\sigma_s^2 + \sigma_d^2 + \tau_d^2 + \tau_m^2 + \sigma_\varepsilon^2}.$$

Therefore, we get the general reliability coefficient under the correlation approach for the three-level LMM with two AR(1) processes. By setting $\Delta_{jj'}$ and $\Delta_{kk'}$ to specific values, we can get the specific reliability coefficients for different time intervals (e.g., instantaneity, moment-to-moment, day-to-day, and stability).

A.4. The derivation of within-person reliability

Within-person test–retest reliability for single-item measures can be defined as the correlation between two measurements for the same person taken at different time points (lagged by $\Delta_{jj'}$ days and $\Delta_{kk'}$ moments), conditioned on that specific person i, which can be expressed as follows:

$$\rho^{\mathrm{W}}(\Delta_{jj'}, \Delta_{kk'}) = \mathrm{Cor}(y_{ijk}, y_{ij'k'} \mid i) = \frac{\mathrm{Cov}(y_{ijk}, y_{ij'k'} \mid i)}{\sqrt{\mathrm{Var}(y_{ijk} \mid i)\,\mathrm{Var}(y_{ij'k'} \mid i)}}. \tag{A.3}$$

The formulation mirrors the between-person reliability, except that the person i is fixed all along. It should be noted that because our current model assumes homogeneity of variances and autoregressive parameters across all individuals, the resulting within-person reliability coefficient is a single metric that technically applies to any person sampled from the population. However, this generalizability stems entirely from the model's homogeneity constraints. If these assumptions were relaxed, for instance, by specifying person-specific innovation variances or person-specific AR parameters, the same derivation would naturally yield person-specific reliability coefficients.

The numerator in Equation (A.3) also depends on whether the two moments occur on the same day or not. Based on the principles in Appendices A.1 and A.2, distributing the conditional covariance yields:

$$\begin{aligned}\mathrm{Cov}(y_{ijk}, y_{ij'k'} \mid i) &= \mathrm{Var}(s_i \mid i) + \mathrm{Cov}(d_{ij}, d_{ij'} \mid i) + \mathrm{Cov}(v_{ij}, v_{ij'} \mid i) \\ &\quad + \mathrm{Cov}(\omega_{ijk}, \omega_{ij'k'} \mid i) + \mathrm{Cov}(\varepsilon_{ijk}, \varepsilon_{ij'k'} \mid i) \\ &= \begin{cases} \sigma_d^2 + \tau_d^2 + \tau_m^2 \phi_m^{\Delta_{kk'}}, & \text{if } j = j' \\ \tau_d^2 \phi_d^{\Delta_{jj'}}, & \text{if } j \neq j' \end{cases} \\ &= \delta_{jj'}\sigma_d^2 + \tau_d^2 \phi_d^{\Delta_{jj'}} + \delta_{jj'}\tau_m^2 \phi_m^{\Delta_{kk'}}.\end{aligned}$$

Downloaded from https://www.cambridge.org/core. 13 Nov 2026 at 03:02:29, subject to the Cambridge Core terms of use.

Note that the person variance σ_s^2 is not included in the covariance because we have conditioned on a specific person i. Similarly, the denominator in Equation (A.3) is a conditional variance:

$$\begin{aligned}\mathrm{Var}(y_{ijk} \mid i) &= \mathrm{Var}(s_i \mid i) + \mathrm{Var}(d_{ij} \mid i) + \mathrm{Var}(v_{ij} \mid i) + \mathrm{Var}(\omega_{ijk} \mid i) + \mathrm{Var}(\varepsilon_{ijk} \mid i) \\ &= \sigma_d^2 + \tau_d^2 + \tau_m^2 + \sigma_\varepsilon^2,\end{aligned}$$

which is the sum of all variance components except the person variance σ_s^2. Because of the stationarity assumption, this variance is also constant regardless of the specific time points (i.e., $\mathrm{Var}(y_{ij'k'} \mid i) = \mathrm{Var}(y_{ijk} \mid i)$).

Combining the above results, the within-person reliability coefficient under the correlation approach for the three-level LMM with two AR(1) processes is given by

$$\rho^{\mathrm{W}}(\Delta_{jj'}, \Delta_{kk'}) = \frac{\delta_{jj'}\sigma_d^2 + \tau_d^2\phi_d^{\Delta_{jj'}} + \delta_{jj'}\tau_m^2\phi_m^{\Delta_{kk'}}}{\sigma_d^2 + \tau_d^2 + \tau_m^2 + \sigma_\varepsilon^2}.$$

By setting $\Delta_{jj'}$ and $\Delta_{kk'}$ to specific values, we can get the specific reliability coefficients for different time intervals (e.g., instantaneity, moment-to-moment, and day-to-day, but not stability).

B. Reliability estimation of negative affect in Exam Results Study

Table B1. Estimates of reliability coefficients for the single-item negative affect assessment in the Exam Results Study

	Reliability	Between-person		Within-person	
		Posterior mean (posterior SD)	Posterior 95% HDI	Posterior mean (posterior SD)	Posterior 95% HDI
Single score	$\rho_0 = R_{\mathbf{TM}}$	0.76 (0.05)	[0.69, 0.84]	0.68 (0.06)	[0.60, 0.79]
	$\rho_{\mathbf{M}}(\Delta_{kk'} = 1)$	0.63 (0.03)	[0.58, 0.67]	0.52 (0.04)	[0.46, 0.58]
	$\rho_{\mathbf{M}}(\Delta_{kk'} = 2)$	0.56 (0.03)	[0.51, 0.60]	0.43 (0.04)	[0.37, 0.49]
	$\rho_{\mathbf{M}}(\Delta_{kk'} = 3)$	0.53 (0.03)	[0.48, 0.57]	0.38 (0.04)	[0.33, 0.45]
	$\rho_{\mathbf{M}}(\Delta_{kk'} = 4)$	0.51 (0.02)	[0.47, 0.55]	0.36 (0.03)	[0.32, 0.43]
	$\rho_{\mathbf{M}}(\Delta_{kk'} = 5)$	0.50 (0.02)	[0.46, 0.54]	0.35 (0.03)	[0.31, 0.41]
	$\rho_{\mathbf{M}}(\Delta_{kk'} = 6)$	0.50 (0.02)	[0.46, 0.54]	0.35 (0.03)	[0.30, 0.41]
	$\rho_{\mathbf{M}}(\Delta_{kk'} = 7)$	0.50 (0.02)	[0.46, 0.54]	0.35 (0.03)	[0.30, 0.41]
	$\rho_{\mathbf{M}}(\Delta_{kk'} = 8)$	0.50 (0.02)	[0.46, 0.54]	0.35 (0.03)	[0.30, 0.41]
	$\rho_{\mathbf{M}}(\Delta_{kk'} = 9)$	0.50 (0.02)	[0.46, 0.54]	0.35 (0.03)	[0.30, 0.41]
	$R_{\mathbf{TD}}$	0.50 (0.03)	[0.45, 0.54]	0.35 (0.04)	[0.30, 0.41]
	$\rho_{\mathbf{D}}(\Delta_{jj'} = 1)$	0.38 (0.03)	[0.33, 0.43]	0.19 (0.04)	[0.14, 0.27]
	$\rho_{\mathbf{D}}(\Delta_{jj'} = 2)$	0.32 (0.03)	[0.27, 0.37]	0.11 (0.04)	[0.06, 0.19]
	$\rho_{\mathbf{D}}(\Delta_{jj'} = 3)$	0.28 (0.03)	[0.23, 0.34]	0.07 (0.04)	[0.03, 0.13]
	$\rho_{\mathbf{D}}(\Delta_{jj'} = 4)$	0.26 (0.05)	[0.20, 0.32]	0.04 (0.03)	[0.01, 0.09]

(Continued)

Downloaded from https://www.cambridge.org/core. 13 Nov 2026 at 03:02:29, subject to the Cambridge Core terms of use.

Table B1. (Continued)

	Reliability	Between-person		Within-person	
		Posterior mean (posterior SD)	Posterior 95% HDI	Posterior mean (posterior SD)	Posterior 95% HDI
	$\rho_{\mathbf{D}}(\Delta_{jj'} = 5)$	0.25 (0.04)	[0.18, 0.32]	0.02 (0.02)	[0.01, 0.07]
	$\rho_{\mathbf{D}}(\Delta_{jj'} = 6)$	0.24 (0.04)	[0.17, 0.31]	0.02 (0.02)	[0.00, 0.05]
	$\rho_{\mathbf{D}}(\Delta_{jj'} = 7)$	0.24 (0.05)	[0.16, 0.31]	0.01 (0.01)	[0.00, 0.03]
	$\rho_{\mathbf{D}}(\Delta_{jj'} = 8)$	0.23 (0.05)	[0.15, 0.31]	0.01 (0.01)	[0.00, 0.02]
	$\rho_{\infty} = R_{\mathbf{TP}}$	0.23 (0.05)	[0.14, 0.31]	–	–
Daily average score	$\rho_{\overline{\mathbf{D}}}(\Delta_{jj'} = 1)$	0.70 (0.03)	[0.65, 0.74]	0.47 (0.07)	[0.37, 0.69]
	$\rho_{\overline{\mathbf{D}}}(\Delta_{jj'} = 2)$	0.58 (0.04)	[0.52, 0.65]	0.27 (0.08)	[0.17, 0.42]
	$\rho_{\overline{\mathbf{D}}}(\Delta_{jj'} = 3)$	0.52 (0.05)	[0.44, 0.59]	0.16 (0.07)	[0.08, 0.29]
	$\rho_{\overline{\mathbf{D}}}(\Delta_{jj'} = 4)$	0.48 (0.05)	[0.39, 0.56]	0.10 (0.06)	[0.03, 0.21]
	$\rho_{\overline{\mathbf{D}}}(\Delta_{jj'} = 5)$	0.45 (0.06)	[0.35, 0.55]	0.06 (0.05)	[0.02, 0.15]
	$\rho_{\overline{\mathbf{D}}}(\Delta_{jj'} = 6)$	0.44 (0.07)	[0.32, 0.54]	0.04 (0.04)	[0.01, 0.10]
	$\rho_{\overline{\mathbf{D}}}(\Delta_{jj'} = 7)$	0.43 (0.07)	[0.30, 0.54]	0.02 (0.03)	[0.00, 0.07]
	$\rho_{\overline{\mathbf{D}}}(\Delta_{jj'} = 8)$	0.43 (0.08)	[0.29, 0.54]	0.01 (0.02)	[0.00, 0.05]

Note: ρ = reliability based on the correlation approach; R = reliability based on the variance partitioning approach; subscripts 0 = instantaneity, M = moment-to-moment; D = day-to-day, $\overline{\text{D}}$ = day-to-day for average scores, ∞ = stability, TM = momentary scores, TD = daily score, TP = personal trait scores; $\Delta_{kk'}$ = time lag between moments, and $\Delta_{jj'}$ = time lag between days.

Cite this article: Lin, T.-Y., Tuerlinckx, F. and Vanbelle, S. (2026). Single-Item Reliability for Intensive Longitudinal Data: A Three-Level Linear Mixed-Effect Autoregressive Modeling Approach. *Psychometrika*, 1–32. https://doi.org/10.1017/psy.2026.10133

Downloaded from https://www.cambridge.org/core. 13 Nov 2026 at 03:02:29, subject to the Cambridge Core terms of use.