Behavior Research Methods (2023) 55:3872–3891
https://doi.org/10.3758/s13428-022-01995-1

Estimating reliabilities and correcting for sampling error in indices of within-person dynamics derived from intensive longitudinal data

Stefan Schneider[1,2,3] **· Doerte U. Junghaenel**[1,2,3]

Accepted: 28 September 2022 / Published online: 19 October 2022

Abstract
Psychology has witnessed a dramatic increase in the use of intensive longitudinal data (ILD) to study within-person processes, accompanied by a growing number of indices used to capture individual differences in within-person dynamics (WPD). The reliability of WPD indices is rarely investigated and reported in empirical studies. Unreliability in these indices can bias parameter estimates and yield erroneous conclusions. We propose an approach to (a) estimate the reliability and (b) correct for sampling error of WPD indices using "Level-1 variance-known" (V-known) multilevel models (Raudenbush & Bryk, 2002). When WPD indices are calculated for each individual, the sampling variance of the observed WPD scores is typically falsely assumed to be zero. V-known models replace this "zero" with an approximate sampling variance fixed at Level 1 to estimate the true variance of the index at Level 2, following random effects meta-analysis principles. We demonstrate how V-known models can be applied to a broad range of emotion dynamics commonly derived from ILD, including indices of the average level (mean), variability (intraindividual standard deviation), instability (probability of acute change), bipolarity (correlation), differentiation (intraclass correlation), inertia (autocorrelation), and relative variability (relative standard deviation) of emotions. A simulation study shows the usefulness of V-known models to recover the true reliability of these indices. Using a 21-day diary study, we illustrate the implementation of the proposed approach to obtain reliability estimates and to correct for unreliability of WPD indices in real data. The techniques may facilitate psychometrically sound inferences from WPD indices in this burgeoning research area.

Keywords Reliability · Intensive longitudinal data · Variability · Within-person dynamics · Emotion dynamics · Meta-analysis · Level-1 variance-known multilevel models

Introduction

Psychology and other social sciences have witnessed a dramatic increase in attention on the measurement of within-person, dynamic processes (Hamaker & Wichers, 2017; Molenaar & Campbell, 2009). In part, this is facilitated by assessment methods that enable the collection of intensive longitudinal data (ILD), including ecological momentary assessments, daily diaries, and ambulatory assessments of physiological and behavioral data. It has become clear that patterns of intraindividual variability contain essential information for understanding how individuals differ from each other, and researchers have developed a steadily growing number of indices for the purpose of capturing within-person dynamics (WPD).

The field of emotion dynamics research provides a paradigm example to illustrate this. In addition to mean levels of affect, multiple indices have been developed to capture the magnitude of emotion fluctuations [e.g., intraindividual standard deviation (ISD), coefficient of variation (Mestdagh et al., 2018; Ram & Gerstorf, 2009; Wang & Grimm, 2012)], to capture temporal dependencies in affective states [e.g., autocorrelation (affect inertia; Kuppens et al., 2010), frequency and damping of oscillating patterns (Chow et al., 2005)], and to capture the interplay of

✉ Stefan Schneider
schneids@usc.edu

[1] Dornsife Center for Self-Report Science & Center for Economic and Social Research, University of Southern California, 635 Downey Way, Los Angeles, CA 90089-3332, USA

[2] Department of Psychology, University of Southern California, Los Angeles, CA, USA

[3] Leonard Davis School of Gerontology, University of Southern California, Los Angeles, CA, USA

several emotions [e.g., emotional bipolarity (Dejonckheere et al., 2018), mixed emotions (Schneider & Stone, 2015), and emotion differentiation (Kashdan et al., 2015)].

In view of the increasing adoption of new indices of WPD, it is paramount that they provide psychometrically sound measurements (Brose et al., 2020). The reliability of WPD indices is a basic psychometric property that is not commonly investigated and rarely reported in empirical studies. Strategies for reliability examination that are familiar to most applied researchers (e.g., internal consistency reliability) are not applicable to indices of WPD leaving researchers with few statistical tools to quantify the reliability and to mitigate the impact of unreliable measurement of these indices in their studies.

In this article, we present an approach to estimate the reliability and to correct for unreliability in indices of WPD using "Level-1 variance-known" (V-known) multilevel models (a term introduced by Raudenbush & Bryk, 2002). These models provide a very flexible framework to partition the variance of observed scores into true (or systematic) between-person (Level 2) variance and variance due to sampling error (Level 1). Unlike traditional multilelvel models that use raw within-person data, V-known models are applicable if summary statistics, such as indices of emotion dynamics, are created from within-person data.

Our article is organized as follows. We start by outlining challenges with estimating the reliability and correcting for unreliability of WPD indices. We then describe how V-known models can be used for reliability estimation and unreliability correction for a broad range of WPD indices. Next, we present results of a Monte Carlo simulation to evaluate whether V-known models can adequately recover the true reliability of these indices, followed by a real data illustration of the proposed methods. We end with a discussion emphasizing future directions. Even though we illustrate the methods using examples from the literature on emotion indices, we emphasize that they are equally applicable to WPD indices in other research domains.

Reliability of individual differences in the context of intensive longitudinal data

Challenges associated with estimating the reliability of indices of within-person dynamics

Reliability is most commonly defined within the framework of classical test theory (CTT). In CTT, the variance of observed scores between individuals σ^2_{TOT} is the sum of the variance in true scores τ^2 and the variance of random errors σ^2_ε, such that

$$\sigma^2_{TOT} = \tau^2 + \sigma^2_\varepsilon. \quad (1)$$

Reliability is defined as the proportion of variance in the observed scores between individuals that can be attributed to variance in true scores rather than error variance:

$$rel = \frac{\tau^2}{\sigma^2_{TOT}} = \frac{\tau^2}{\tau^2 + \sigma^2_\varepsilon}. \quad (2)$$

For individual difference measures derived from ILD, the reliability of a measure of intraindividual means $\bar{y}$ can be readily estimated from empirical data. The proportion of true variance in means is captured by the intraclass correlation (ICC), which can be calculated based on a one-way ANOVA with random effects (Raudenbush & Bryk, 2002). Applying the Spearman–Brown prediction formula to the ICC, the reliability of intraindividual means computed from T measurement occasions is

$$rel_{\bar{y}} = \frac{(I\hat{C}C)T}{1 + I\hat{C}C(T-1)} = \frac{\hat{\tau}^2_\mu}{\hat{\tau}^2_\mu + (\hat{\sigma}^2/T)} = \frac{\hat{\tau}^2_\mu}{\hat{\tau}^2_\mu + \hat{\sigma}^2_\varepsilon}. \quad (3)$$

Importantly, this strategy is limited to intraindividual means and does not extend to other indices of WPD. For example, when a measure of within-person variability is created by calculating the intraindividual standard deviation (ISD), the ICC cannot be used to estimate the true score variance. In this case, the variation of observations around a person's mean, which comprises the error variance component in the ICC, is itself used to create the individual difference measure. Consequently, conventional analytic methods to separating true score and error variance components from ILD do not apply.

Some methods of reliability estimation can be used for any index of WPD, but each of these methods has some practical limitations. For example, a coefficient of test–retest reliability can be estimated for any WPD index that is computed at two different time periods (Eid & Diener, 1999) or when deriving parallel WPD indices from odd and even sampling occasions (Wendt et al., 2020). However, these approaches assume that people's true scores on the WPD of interest do not change across the time periods or sampling occasions (Fleeson, 2001). In addition, parallel-test reliability of WPD indices could be calculated based on test forms that are constructed to be equivalent (Borsboom, 2003; Lord & Novick, 1968), but parallel forms are not often available for ILD (Hu et al., 2016).

What is known about the reliability of indices of WPD is largely based on Monte Carlo simulation studies (Du & Wang, 2018; Estabrook et al., 2012; Wang & Grimm, 2012). These simulations suggest that WPD indices addressing ISDs, mean squared successive differences, or autocorrelations may be much less reliably measured than

intraindividual means, especially when few measurement occasions are available per person. However, simulation studies operate under idealized assumptions and they require that the true population variance of a measure is known *a priori*, which is not the case when we are interested in estimating reliabilities in empirical studies. As emphasized by the APA Task Force of Statistical Inference (1999), reliability depends on the specific study and population investigated. Furthermore, as will be discussed next, we may wish to not only obtain an estimate of the reliability of our WPD measures, but to also correct parameter estimates for error variance in these measures.

Challenges associated with correcting for sampling error in indices of within-person dynamics

Analyses of WPD indices commonly follow a two-step procedure: the index is calculated separately for each individual in the sample in a first step, and the second step uses the resulting scores as a manifest (i.e., observed) independent, dependent, or mediator variable in the analysis model. The problem associated with this approach is that the manifest scores are treated as proxies for their corresponding true scores, whereby it is falsely assumed that there is no error variance and no uncertainty about the true values, which can result in biased parameter estimates and incorrect inference (Bollen, 1989; Fritz et al., 2016; Kenny, 1979).

Latent variable models correct for such biases resulting from unreliable measurement (Bollen, 1989; Kenny, 1979). Statistical and software options to model individual differences in WPD as latent variables have expanded considerably in recent years, especially due to rapid advances in multilevel modeling techniques. For instance, the mixed effects location scale model by Hedeker et al. (2012) enables the modeling of individual differences in intraindividual means and (log) variances jointly as dependent variables, using standard software (i.e., SAS PROC NLMIXED; Cary, NC, USA). Expanding beyond this, multilevel time-series models (or dynamic structural equation models) allow for random effects in any parameter of a within-person time series model (e.g., autoregressive and cross-lagged parameters, residual variances), and these have been implemented in WinBUGS (Jongerling et al., 2015; Wang et al., 2012) and M*plus* software (Asparouhov et al., 2018; Hamaker et al., 2018; McNeish & Hamaker, 2020).

Despite the increasing versatility of these multilevel modeling strategies, the use of manifest variable (two-step) approaches for analyzing WPD still predominates in applied research. Computing indices of WPD as manifest variables is often easily accomplished using standard software, such that researchers may gravitate to using manifest variables for pragmatic reasons. Indices of WPD that involve nonlinear combinations of parameters (e.g., coefficient of variation, computed as the ratio of the standard deviation to the mean) are easily computed but difficult to estimate directly in a multilevel model. Moreover, using a two-step approach, procedures used to calculate indices of WPD can be tailored to the individual, for example, when choosing between different functions to remove temporal trends from the data before calculating an index (Wang et al., 2012). As illustrated below, V-known multilevel models may serve as one strategy to overcome some of these challenges.

Applying "Level-1 Variance-Known" multilevel models to indices of within-person dynamics

The general idea behind the use of V-known multilevel models is to capitalize on the flexibility and versatility of the familiar two-step method to examining WPD indices, while addressing problems associated with using these indices as manifest variables. As explained above, in the traditional two-step approach, an index is first calculated for each individual and its sampling error variance is subsequently assumed to be zero. In V-known multilevel models, the goal is to replace this "zero" sampling variance with a much more reasonable value. As highlighted by Raudenbush and Bryk (1985, 1987, 2002), even though the index is represented by a summary statistic (i.e., a single value per individual), an estimate of the sampling variance of most summary statistics (i.e., the square of its standard error) can be derived alongside the statistic. Treating this sampling variance as "known" at Level 1 (the within-person level) in multilevel models makes it possible to treat any given index of WPD as an unobserved, latent variable that is measured with imprecision by the observed values of the index.

In the V-known multilevel model, the within-person model refers to the variation of the index due to sampling variation:

$$\text{Level 1}: d_i = \delta_i + \varepsilon_i, \text{where } \varepsilon_i \sim N\left(0, \sigma^2{}_{\varepsilon i}\right). \tag{4a}$$

In this equation, d_i represents the observed WPD index d derived from the ILD of person i, δ_i refers to the true value of the index for person i, and ε_i is a normally distributed random residual with mean 0 and variance $\sigma^2{}_{\varepsilon i}$. The latter is the sampling variance that is assumed to be known and it is fixed at the value derived alongside the index for each individual.

The between-person model refers to the variation of the true values of the index across individuals.

$$\text{Level 2}: \delta_i = \gamma_0 + u_i, \text{where } u_i\, N\left(0, \tau^2\right). \tag{4b}$$

Here, the true unknown (i.e., latent) values for each person δ_i are expressed as the sum of the overall population mean of the index γ_0 and the person's random deviation from this

grand mean u_i. The random deviations from the grand mean are assumed to follow a normal distribution with mean 0 and variance τ^2, which represents the true between-person variance of the index.

Substituting (4b) into (4a) yields a combined model for the observed values of the index:

$$d_i = \gamma_0 + u_i + \varepsilon_i, \quad \text{where } u_i \sim \mathrm{N}\left(0, \tau^2\right) \text{ and } \varepsilon_i \sim N\left(0, \sigma^2_{\varepsilon i}\right). \tag{5}$$

It is now easy to see that the observed variable is decomposed into true and sampling error variance components. Note that the sampling variance $\sigma^2_{\varepsilon i}$ has a subscript *i*, indicating that the sampling variance does not need to be the same across individuals. In practice, WPD indices are likely calculated from differing numbers of observations (e.g., due to missed assessments), such that the index will differ in precision across individuals. The V-known multilevel model accommodates this by allowing the sampling variance to be individual-specific.

Relationship between the proposed approach and random-effects meta-analysis

The principles of V-known multilevel models are not new. In fact, their most prominent application can be found in research using (random effects) meta-analysis (Raudenbush & Bryk, 1985, 2002). A key concern in meta-analysis is how to appropriately aggregate summary statistics represented by effect sizes across studies. To appropriately take the sampling error in study-specific effect size estimates into account, and to detect the true heterogeneity between studies, the sampling variance of the effect sizes is derived from the research reports and treated as "known" (Fernández-Castilla et al., 2020; Pastor & Lazowski, 2018). Whereas meta-analysis applications presuppose a multilevel structure where participants are nested within studies, the multilevel structure of dynamic within-person processes involves measurement occasions nested within individuals. Thus, the proposed application of V-known models can be viewed as an application of meta-analysis to the synthesis of WPD scores across individuals. Correspondingly, WPD indices can be analyzed with V-known models using a broad range of statistical packages that are commonly used for conducting meta-analysis models, including multilevel modeling software (Mplus; Cheung, 2015a; SAS PROC MIXED; van Houwelingen et al., 2002) and software dedicated to meta-analysis (the R package metaSEM; Cheung, 2015b; the R package metafor; Viechtbauer, 2010).

Testing the reliability of indices of within-person dynamics with V-known models

In the meta-analysis literature, the I^2 statistic is a popular measure assessing "the percentage of total variation across studies that is due to heterogeneity rather than chance" (Higgins et al., 2003; p. 558). Despite its labeling as a measure to quantify heterogeneity rather than the reliability of summary statistics, the definition of I^2 is identical with the classical test theory definition of reliability in Eq. (2) above:

$$I^2 = \frac{\tau^2}{\tau^2 + \sigma^2_{\varepsilon}}. \tag{6}$$

This means that in the context of research using ILD, the I^2 statistic can be used to estimate the reliability of any WPD index.

Notably, like Eq. (2), Eq. (6) shows a common sampling variance term σ^2_{ε} for all individuals, whereas in reality the sampling variance will often be individual specific, that is, $\sigma^2_{\varepsilon i}$. To accommodate the heterogeneity of sampling variances in meta-analyses, the I^2 statistic defines σ^2_{ε} as the *typical* value of the individual sampling variances $\sigma^2_{\varepsilon i}$, whereby multiple variants of I^2 have been proposed that differ in how the typical sampling variance is calculated.

In a widely influential article, Higgins and Thompson (2002) defined the typical value of the sampling variance as

$$\sigma^2_{\varepsilon(I)} = \frac{(n-1)\sum_{i=1}^{n} 1/\sigma^2_{\varepsilon i}}{\left(\sum_{i=1}^{n} 1/\sigma^2_{\varepsilon i}\right)^2 - \sum_{i=1}^{n} 1/\sigma^2_{\varepsilon i}}, \tag{7}$$

and this version is implemented in most meta-analysis programs.

Xiong et al. (2008, 2010) proposed that the arithmetic mean of the sampling variances across individuals can be used:

$$\sigma^2_{\varepsilon(II)} = \frac{\sum_{i=1}^{n} \sigma^2_{\varepsilon i}}{n}. \tag{8}$$

Both versions of I^2 (with likelihood based 95% CIs) can be estimated in the R package metaSEM (Cheung, 2015b). As noted by Xiong et al. (2010), the different versions give similar estimates unless the individual differences in the sampling variance are pronounced. In the context of reliability estimation, a potential advantage of using the arithmetic mean (Equation 8) is that the resulting values will be consistent with the classical test theory definition of the *reliability index,* that is, the squared correlation between the true and observed scores (Estabrook et al., 2012; Wang & Grimm, 2012).

Apart from obtaining reliability coefficients, we can also test the null hypothesis that there is no variance in the true scores of a given WPD index between individuals. One way to do this is to conduct a likelihood ratio test comparing the deviance (−2 times the log-likelihood) statistics between the model in Equation 5, in which the true variance is freely estimated as a random effect, with the fit of a model that constrains the between-person variance parameter u_i to zero

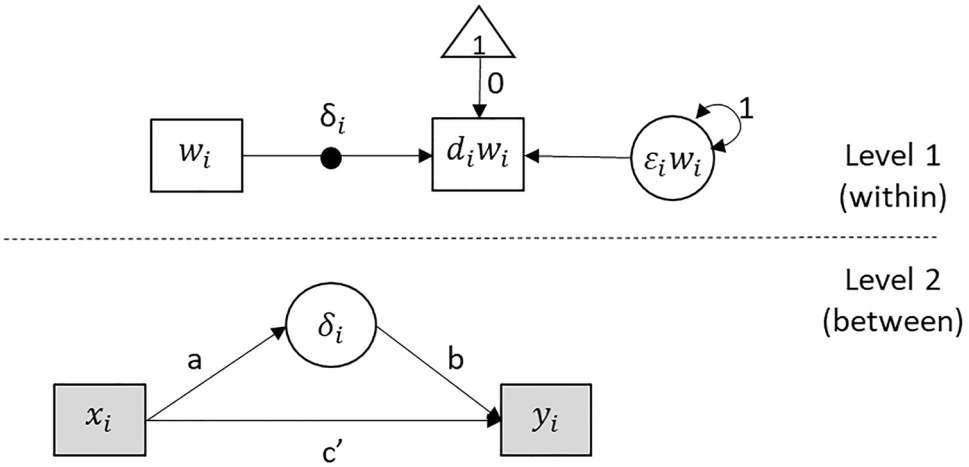


Fig. 1 V-known multilevel structural equation model. Intercepts and residual variances at Level 2 are omitted for simplicity

(referred to as a fixed-effects model in the meta-analysis literature) (Raudenbush & Bryk, 2002). Alternatively, the Q-statistic (Cochran, 1954) can be used for this purpose and is available from virtually all meta-analysis programs (Pastor & Lazowski, 2018).

Relating indices of within-person dynamics to other variables in V-known analysis models

The V-known model makes it possible to estimate any given index of WPD as an unobserved, latent variable. Relating this latent variable to other variables (e.g., other between-person characteristics) yields coefficients that are corrected for sampling error in the index. For brevity, we limit our presentation to basic examples. Interested readers are referred to the extensive literature on V-known multilevel, mixed effects meta-analysis, and meta-analytic structural equation modeling (meta-SEM) (e.g., Card, 2012; Cheung, 2015a; Demidenko, 2013; Lipsey & Wilson, 2001; Raudenbush & Bryk, 2002).

Oftentimes, a WPD index will be a dependent variable in the model. For example, many research questions evolve around predictors of individual differences in emotion dynamics, such as whether they differ by age or other person characteristics. The unconditional (random effects) V-known model shown in Eq. (4a) and (b) can be readily extended into conditional (mixed effects) models by including between-person predictor variables $x_1, \ldots, x_S$ at Level 2:

$$\begin{aligned} &\text{Level 1}: d_i = \delta_i + \varepsilon_i, \\ &\text{Level 2}: \delta_i = \gamma_0 + \gamma_1 x_{1i} + \gamma_2 x_{2i} + \cdots + \gamma_S x_{Si} + u_i, \\ &\qquad \text{where } \varepsilon_i \sim N\left(0, \sigma^2_{\varepsilon i}\right) \text{ and } u_i \sim N\left(0, \tau^2\right). \end{aligned} \tag{9}$$

This model is known as meta-regression model in the meta-analysis literature (Lipsey & Wilson, 2001). The Level 1 equation remains the same as in Eq. (4a) as its role is to account for the variation of the index of WPD due to sampling variance, which is fixed at $\sigma^2_{\varepsilon i}$ for each individual. At Level 2, γ_0 is now the intercept of the latent variable representing the index, $\gamma_1, \ldots, \gamma_S$ are the regression coefficients for the set of predictor variables, and τ^2 is the residual true between-person variance of the index.

Other research questions will require more flexible modeling of WPD indices as independent variables, covariates, moderators, or mediator variables. V-known applications to multilevel SEM can accommodate the modeling of WPD indices as latent variables in these situations. As demonstrated by Cheung (2008), multilevel SEM models can readily accommodate V-known problems using a specific parameterization of a weighted least squares approach. To implement a V-known model in multilevel SEM programs such as M*plus*, all terms at Level 1 (including the intercept) need to be weighted by the inverse standard error of the observed index (i.e., one over the square root of the known sampling variance: $w_i = 1/\sqrt{\sigma^2_{\varepsilon i}}$). A diagram of the V-known multilevel SEM model is shown in Fig. 1. Specifically, the observed index of WPD d_i is weighted by (i.e., multiplied with) w_i, and this variable $d_i w_i$ is regressed on the weight variable w_i in a Level 1 model without intercept (i.e., the intercept of $d_i w_i$ is fixed at 0). The resulting regression parameter implicitly weights the individual-specific sampling variance of the index such that the error term $\varepsilon_i w_i$ is a normally distributed random residual with mean = 0 and variance = 1 for each individual; accordingly, the residual variance of $\varepsilon_i w_i$ needs to be fixed at 1 in the model (see Fig. 1).

Table 1 Formulas to derive approximate sampling variances of common indices of within-person emotion dynamics

Emotion dynamic and example index	Transformation to normalize the sampling distribution of the index	Sample estimator	Approximate sampling variance
Average emotion level: Mean of a person's emotion ratings across measurement occasions	None	$\bar{y}$	$\frac{\sigma^2}{T_i}$
Emotion variability: Intraindividual standard deviation of emotion ratings across measurement occasions	Natural logarithm of standard deviation, with correction for small sample bias (Raudenbush & Bryk, 1987)	$ISD_{RB,i} = \log\left(ISD_i\right) + \frac{1}{2(T_i-1)}$	$\frac{1}{2(T_i-1)}$
Probability of acute change (emotional instability): Proportion of successive absolute changes in emotion ratings that exceed a researcher-selected threshold	Freeman-Tukey double arcsine transformation of proportion (Freeman & Tukey, 1950)	$\hat{p}_{FT,i} = \arcsin\sqrt{\frac{\hat{p}_i T_i}{T_i+1}} + \arcsin\sqrt{\frac{\hat{p}_i T_i+1}{T_i+1}}$	$\frac{1}{T_i+0.5}$
Emotional bipolarity: Correlation of positive and negative emotion ratings across measurement occasions	Fisher z-transformation of Pearson correlation coefficient (Fisher, 1915)	$z\prime_i = 0.5\log\frac{(1+r_i)}{(1-r_i)}$	$\frac{1}{(T_i-3)}$.
Emotion differentiation: Intraclass correlation of ratings on multiple emotion items of the same valence across measurement occasions	Fisher z-transformation of intraclass correlation coefficient (Fisher, 1938)	$zICC_i = 0.5\log\frac{1+(K_i-1)ICC_i}{1-ICC_i}$,	$\frac{K_i}{2(T_i-2)(K_i-1)}$.

ISD intraindividual standard deviation; $\hat{p}$ = proportion; r = Pearson correlation coefficient, *ICC* intraclass correlation coefficient; T_i = number of observations for individual i; K_i = average number of emotion items per measurement occasion

In a random-effects V-known multilevel SEM model, the regression of $d_i w_i$ on w_i has a random slope (shown as a filled circle at Level 1). At Level 2, these random regression slopes are modeled as a latent variable δ_i capturing the true variance of the WPD index. As is typical for SEM, the latent index δ_i can now be flexibly incorporated into any structural model; for example, as a latent predictor variable, latent mediator variable, or as a moderator variable using latent variable interactions (Cheung, 2015a). Figure 1 shows an example in which the latent index δ_i is a mediator variable of the relationship between a Level 2 predictor variable x_i and a Level 2 dependent variable y_i. In this example, the path coefficients are corrected for the unreliability of the index used as a mediator variable. The necessary procedures are implemented in the specialized R package metaSEM (Cheung, 2015b), but they can also be estimated with any multilevel SEM software, such as the general latent variable modeling program M*plus*, or using the gsem command in STATA (Palmer & Sterne, 2015).

Obtaining approximate sampling variances for common indices of within-person dynamics

In order to fit V-known multilevel models to data involving WPD indices, it is necessary to find the appropriate sampling variance estimate for the scores on a given WPD index. Even though this is not a trivial task, we propose three strategies that provide the building blocks for estimating the sampling variance of virtually any WPD index. First, established formulas are available to obtain approximate sampling variances and these can be readily applied to common WPD indices. Second, for WPD indices that are derived from individual-specific analysis (e.g., per-subject time series or differential equation) models, standard errors of the parameter estimates can be used. A third and most flexible strategy is to use resampling methods to approximate the sampling variance of an index.

Strategy 1: Computational formulas to derive the sampling variance

Table 1 shows formulas to approximate the sampling variance for several indices commonly used in emotion research, including the mean (average emotion level), ISD (emotion variability), probability of acute change (emotional instability), intraindividual correlation (emotional bipolarity), and intraclass correlation (emotion differentiation). As can be seen in the table, for most indices to be used in V-known models, it is advantageous to first transform the statistic into a measure for which the sampling distribution is closer to a normal distribution. This also has the added benefit that the sampling variance is well approximated purely as a function of the number of observations per person. For example, Raudenbush and Bryk (1987) described a transformation of the standard deviation using the natural logarithm (with correction for bias in small samples), and we propose to apply this transformation when using the ISD as an index of WPD. Whereas the sampling variance of the untransformed

ISD depends on the true (unknown) value of the ISD, the sampling variance of this transformed index of variability is well approximated as a function of the number of measurement occasions (see Table 1).

Strategy 2: Sampling variances for parameters in individual-specific analysis models

The second strategy applies to indices of WPD that are derived from linear or nonlinear regression models, time series models, or differential equation models.[1] Because the standard error of a parameter estimated in these statistical models is the standard deviation of its sampling distribution, the squared standard error of the parameters can be used to approximate the sampling variance of these indices. This strategy is rarely implemented in traditional meta-analyses that aim to synthesize effect sizes found in research reports because the parameters and associated standard errors from more complex models are often not comparable across studies (Becker & Wu, 2007). However, the strategy is feasible when applying a two-step approach to the analysis of WPD because the indices and their standard errors are estimated from the raw data using the same statistical model for each individual in the sample.

To provide an example from the emotion literature, *emotional inertia* is conceptualized as the degree to which prior emotions affect current emotions (Kuppens et al., 2010), and it can be estimated from a first-order autoregressive [AR(1)] model. In this model, the value at time t is linearly related to the value at the immediately preceding time point,

$$y_{it} = c_i + \phi_1 y_{it-1} + \varepsilon_{it}, \tag{10}$$

where c is the individual's intercept (the expected score when $y_{t-1} = 0$), ϕ_1 is the AR(1)-parameter, and ε_t is the residual. Autoregressive parameters can be estimated using various methods, including the Yule–Walker method, ordinary least squares estimation, and maximum likelihood estimation (for comparison of these estimators in short time series, see Krone et al., 2017), and the squared standard errors of the parameters can be used as approximate sampling variance for use in V-known models.

[1] Examples from the emotion literature include autoregressive parameters from $N = 1$ time series models to capture emotional inertia (Kuppens et al., 2010), transition probabilities from Markov-switching models capturing vacillations between emotional states (Hamaker et al., 2016), or parameters from the damped linear oscillator model to measure the periodicity and damping of oscillations in emotion levels (Chow et al., 2005).

Strategy 3: Using resampling methods to derive the sampling variances

When standard methods for computing the sampling variance cannot be applied or are difficult to apply, resampling methods such as bootstrapping or jackknifing provide a general purpose approach to estimating the variance of an estimator (Wolter, 2007). To use an example from research on emotion dynamics, Mestdagh et al. (2018) proposed *relative intraindividual variability* measures that are calculated by dividing the observed variability by the maximum possible variability given the observed mean. Assuming measurements of a continuous variable on a scale with a lower bound LB and an upper bound UB, the relative intraindividual standard deviation ISD* is given by

$$\mathrm{ISD}_i^* = \log\left(\frac{ISD_i}{\sqrt{(\bar{y}_i - LB)(UB - \bar{y}_i)}}\right), \tag{11a}$$

where the natural log is used to normalize the distribution of the index.

We can use the jackknife (or leave one out) method (Efron & Stein, 1981) to estimate the sampling variance of the index. The method is based upon sequentially leaving out one observation from the dataset and re-computing the index for the remaining observations. That is, for an individual with T measurement occasions, sequentially omit the t^{th} observation from the dataset and compute the index $\widehat{ISD}^*_{(t)}$ on the remaining $T - 1$ observations. A total of T jackknife replicates of the index are generated this way. The sampling variance is calculated using the sum of squared deviations of the jackknife replicates from the mean of the replicates $\widehat{ISD}^*_{(\bullet)}$:

$$\sigma^2_{\varepsilon i, ISD^*} = \frac{T_i - 1}{T_i}\sum_{t=1}^{T}\left(\widehat{ISD}^*_{i(t)} - \widehat{ISD}^*_{i(\bullet)}\right)^2. \tag{11b}$$

In sum, approximate sampling variances can be obtained for a large variety of statistics. This makes V-known models attractive as a general method for reliability estimation and unreliability correction of measures from the steadily expanding universe of WPD indices.

Monte Carlo study of reliabilities estimated from V-known multilevel models

Before we apply the V-known multilevel model to indices of emotion dynamics in empirical data, we present results from a Monte Carlo simulation study to investigate how well the V-known approach can recover the true population

Table 2 Model parameters for Monte Carlo simulation

Parameter type	Parameter values
Simulation conditions: measurement occasions per individual	$T = 5, 10, 15, 20, 25, 50$
Sample size N	Constant at 100
True population between-person variance τ^2	Adjusted per index and condition such that true reliability = 0.5
True population mean γ_0 and additional index specific parameters	
Intraindividual mean $\bar{y}$	$\gamma_0 = 0$; within-person variance $\sigma^2 = 1$
Raudenbush–Bryk transformed standard deviation ISD_{RB}	$\gamma_0 = 0$ (for an untransformed mean intraindividual standard deviation of approximately $\gamma_0 = 1.0$)
Freeman–Tukey transformed proportion $\hat{p}_{FT}$	γ_0 was adjusted so that mean untransformed probability of acute changes was $\gamma_0 = .10$
Fisher z-transformed correlation z'	$\gamma_0 = -.55$ (for an untransformed correlation of $\gamma_0 = -.50$)
Fisher z-transformed intraclass correlation zICC	$\gamma_0 = 1.07$ (for an untransformed population mean ICC of $\gamma_0 = .60$); $K = 5$ items per measurement occasion
First-order autocorrelation AR(1)	$\gamma_0 = .20$
Log-transformed relative standard deviation ISD*	$\gamma_0 = -1.204$ (for an untransformed ISD* $\gamma_0 = 0.30$); individual means $\bar{y}$ used to compute the ISD* were distributed with a population mean of $\gamma_0 = 40$ and between-person variance of $\tau^2 = 64$, and were uncorrelated with ISD*; minimum and maximum of the rating scale were assumed to be 0 and 100

reliability of various indices under different sampling conditions. The seven indices selected for this simulation were described in the previous section: the mean $\bar{y}$ (average emotion level), Raudenbush–Bryk transformed standard deviation ISD_{RB} (emotion variability), Freeman–Tukey transformed proportion $\hat{p}_{FT}$ (emotional instability), Fisher z-transformed Pearson correlation z' (emotional bipolarity), Fisher z-transformed intraclass correlation zICC (emotion differentiation), first-order autocorrelation AR(1) (emotional inertia), and relative standard deviation ISD* (relative emotion variability).

Model parameters

For each index, six conditions were simulated in which the number of measurement occasions per individual was manipulated as $T = 5, 10, 15, 20, 25$, and 50. We focused on these conditions because the sampling variance approximations to be used in the V-known models will be increasingly more fallible the fewer observations are available per person (Lin, 2018; Raudenbush & Bryk, 1987, 2002). Five measurement occasions probably approaches the lower bound for any ILD study, whereas 50 measurements represents a moderate to high number in emotion research using ILD (Houben et al., 2015). The number of individuals in each sample was held constant at $N = 100$, a sample size that is typical for ILD studies using momentary assessments and daily diaries. We did not vary the true reliability but instead held it constant at $rel = 0.5$ for each index and number of measurement occasions. A reliability of 0.5 is halfway between the possible values of 0.0 and 1.0 and this was selected to ensure that any bias in the model estimated reliabilities would be equally detectable in both negative (model estimated values ranging from 0 to < 0.5) and positive (values ranging from > 0.5 to 1.0) directions.

True scores for each index were generated from a normally distributed population with mean γ_0 and between-person variance τ^2. Given that the sampling variance varies across indices and conditions, the true between-person variance τ^2 was adjusted accordingly to hold the true reliability constant at 0.5. To do this, we generated data from large samples of 10^5 individuals, determined the reliability based on the squared correlation between the true and observed values, and iteratively refined the value of τ^2 to arrive at the desired reliability, separately for each index and condition. The population means γ_0 of each index (as well as additional parameters required for the calculation of some of the indices) are shown in Table 2 and they were selected to reflect values that would be reasonable to expect in research on emotion dynamics.

Data generation and reliability estimation

In each simulation, true scores of a given index were generated for 100 individuals, observed raw scores for T measurement occasions were generated for each individual based on the true score, and observed scores for the index were then calculated from these raw scores and transformed as necessary (i.e., as shown in Table 1 and Eq. 11a). The sampling variances for each index were estimated as follows: For $\bar{y}$, ISD_{RB}, $\hat{p}_{FT}$, z', and zICC, we used the formulas shown in Table 1. The AR(1) parameter and its standard error were estimated using ordinary least squares regression. The sampling variance of ISD* was estimated using the jackknife

procedure (Eq. 11b). V-known models were then fitted using maximum likelihood parameter estimation in M*plus* version 8.3 (Muthén & Muthén, 2017) to estimate the reliability of each index via the I^2 statistic (using the arithmetic mean definition in Eq. 8) from the observed scores in each sample.

A total of 1000 replications were used for the Monte Carlo study, which means that 1000 datasets were generated and analyzed for each of the seven indices and each of the six conditions (5, 10, 15, 20, 25, and 50 measurement occasions). The exception was that we did not simulate the AR(1) index for the condition using five measurement occasions because the between-person variance that would have been required to obtain a true reliability of 0.5 was very large for this index, which would have resulted in non-stationary time series [i.e., true AR(1) parameters exceeding an absolute value of 1.0] for a number of simulated individuals. The simulated data were generated in SAS version 9.4 (Cary, NC) and analyzed using the external Monte Carlo facilities in M*plus*.

Evaluation criteria

The performance of the V-known models was evaluated with two criteria: (a) relative percent parameter bias, and (b) 95% coverage rates.

Parameter bias refers to the extent to which the reliability coefficient estimated from the V-known models deviates from (i.e., is higher or lower than) the true population parameter, on average across the 1000 replications. Relative percent parameter bias was calculated as $100*[(\hat{rel} - 0.5) / 0.5]$, where $\hat{rel}$ is the average V-known model estimated reliability and 0.5 is the true reliability in the population. In line with Muthén and Muthén (2002), parameter bias was considered acceptable if the average model estimated reliability was within ± 10% of the true population reliability (i.e., the average estimated reliability was between .45 and .55).

The 95% coverage represents the proportion of replications for which the 95% confidence interval includes the population value. Accurate coverage translates directly to an accurate type I error rate. Because the upper and lower confidence limits of a reliability coefficient tend to be asymmetric, we estimated coverage rates using a transformation of the reliability coefficients suggested by Bonett (2002; Kelley & Pornprasertmanit, 2016). Coverage should ideally be near 0.95 if a procedure is working well, and values between 0.91 and 0.98 can be considered good coverage (Muthén & Muthén, 2002).

Simulation study results

The V-known models converged in all 41,000 simulation data sets. The simulation results are summarized in Fig. 2, where panel A shows the relative percent parameter bias and panel B shows the coverage rates for the estimated reliability of each index. The number of measurement occasions used to calculate each index are shown on the *x*-axis in the figure. As can be seen, the performance of the V-known model improved both in terms of bias and coverage rates as the number of measurement occasions increased, with estimated reliability coefficients asymptotically approaching the true reliability of 0.5 and coverage approaching the nominal rate of .95. This reflects the established finding that estimators of the sampling variance are based on approximations that are asymptotically unbiased.

When only five measurement occasions per person were used to calculate the indices, bias in the reliability estimates exceeded an absolute value of 10% for every index. Reliabilities were overestimated for indices of $\bar{y}$ (+12% bias) and ISD_{RB} (+21% bias), whereas they were underestimated for $\hat{p}_{FT}$ (– 24% bias), *z'* (– 16% bias), *z*ICC (– 14% bias), and iSD* (– 17% bias). Coverage rates ranged between .58 and .85 across the different indices in this condition.

When ten or more measurement occasions per person were generated, bias in the reliability estimates was below ± 10% of the true population reliability for each simulated index. With as few as ten measurement occasions, bias was within acceptable limits and ranged from – 8% to + 9% across indices; coverage rates were within acceptable limits for reliability estimates of $\hat{p}_{FT}$, *z'*, and *z*ICC, but were below .90 for the remaining indices. With 15 measurement occasions, coverage rates fell within acceptable ranges of .91 to .95 for all indices except for the AR(1) index, which required 20 measurement occasions for adequate coverage rates. Together, this suggests that the V-known model estimated reliabilities may not replicate the true reliability well in studies that collect very few measurements per person, such as studies using 1 week (7 days) of daily diaries. However, for all indices, reasonably unbiased reliability point estimates were obtained with ten measurement occasions, and accurate reliability estimates with confidence intervals with 15 to 20 measurement occasions under the conditions of this simulation.

Real-data application with indices of emotion dynamics

We next present an analysis with real data to illustrate the information obtained when applying V-known models to emotion WPD indices. The specific indices we investigate are the same as those in the simulation study; in the current application, we examine these indices for both positive affect (PA) and negative affect (NA). First, we apply the models to estimate reliability coefficients of the various indices in an actual data set. Second, we illustrate how the V-known

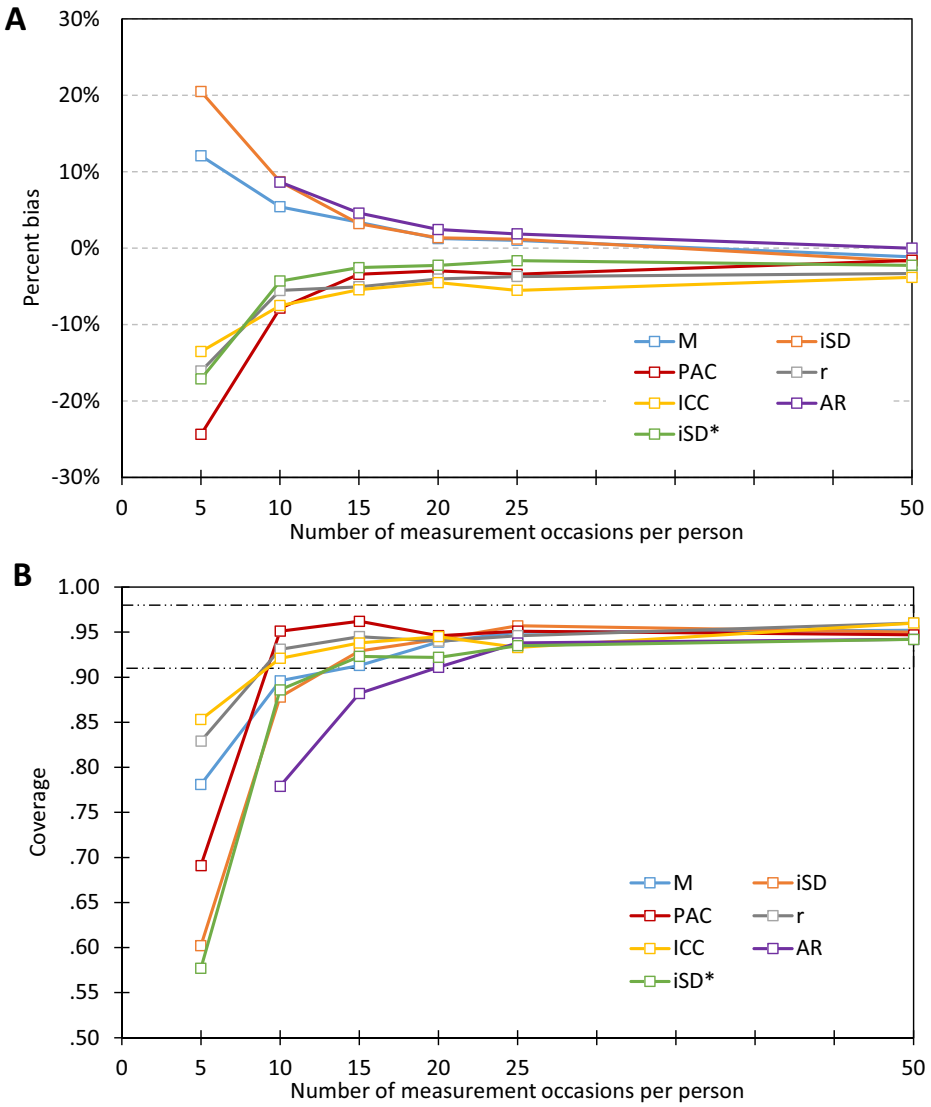


Fig. 2 Relative percentage bias (**a**) and coverage rates (**b**) of indices of within-person dynamics for different numbers of measurement occasions in the simulation study. ISD_{RB} = Raudenbush–Bryk transformed standard deviation, $\hat{p}_{FT}$ = Freeman–Tukey transformed proportion, ISD* = relative standard deviation, AR(1) = first-order autocorrelation, *z*ICC = Fisher *z*-transformed intraclass correlation, *z'* = Fisher *z*-transformed Pearson correlation

approach corrects for unreliability due to sampling error in the indices.

For this illustration, we used a dataset from a daily diary study in which participants (aged 21 to 91 years) rated their emotions over 21 days (Junghaenel et al., 2021). Daily emotions were assessed using 6 PA and 6 NA items, each rated "since waking up today" on a 7-point (not at all – extremely) scale. Analyses included 495 participants who completed the assessments on average on 20.0 (SD = 1.77, range = 6 to 21) out of the 21 days.

Calculation of emotion indices and sampling variances

Indices of $\bar{y}$ (average emotion levels), ISD_{RB} (emotion variability), $\hat{p}_{FT}$ (emotional instability), AR(1) (emotional inertia), and ISD* (relative emotion variability), and z'

(emotional bipolarity) were calculated based on the composite scores for PA and NA. The zICC (emotion differentiation) index, which is based on responses on multiple emotion items of the same valence, was calculated from the raw responses on the six PA and six NA items. All indices were transformed following Table 1 and Eq. (11a).

Several additional considerations regarding the calculation of the emotion indices are noteworthy. First, several indices [ISD_{RB}, AR(1)] are sensitive to temporal trends in the data. Based on visual inspection and linear regression models of linear temporal tends for each participant, linear trends were removed from the data for 57 (PA) and 34 (NA) individuals before calculating ISD_{RB} and AR(1) indices (Wang et al., 2012). Second, the "probability of acute change" index $\hat{p}_{FT}$ is calculated as the proportion of successive change scores that exceed an *a priori* selected threshold; following prior research (Jahng et al., 2008), we defined acute changes as successive increases or decreases exceeding the 90th percentile of observed successive changes (which translated to successive changes exceeding ±1.167 points for both PA and NA). Third, indices that take the temporal ordering of observations into account [$\hat{p}_{FT}$ and AR(1)] are sensitive to unequally spaced time intervals (Hamaker et al., 2018), which occurred if a day was missed by a participant. To address this issue, only successive changes between consecutive nonmissing days were included in the computation of $\hat{p}_{FT}$. To estimate the AR(1) parameter and sampling variance from unequally spaced data, we took advantage of a continuous time model using a "spatial power" covariance function for the estimation of autocorrelated residuals, implemented in SAS PROC MIXED (Schwartz & Stone, 2007). Note that whereas PROC MIXED is typically used for multilevel analyses, we applied it separately to each individual's time series to obtain one AR(1) parameter per individual.

With few exceptions, observed scores of each index could be calculated for all 472 participants. For one participant, the *z*ICC index for NA was not estimable because the NA items showed no within-day variance. In addition, the AR(1) model did not converge for $n = 5$ (PA) and $n = 2$ (NA) participants.

The sampling variances of $\bar{y}$, ISD_{RB}, $\hat{p}_{FT}$, *z'*, and *z*ICC were calculated using the formulas in Table 1.[2] The AR(1) parameter was estimated using restricted maximum likelihood estimation and its squared standard error was used as approximate sampling variance. The jackknife procedure (Eq. 11b) was used to obtain the sampling variance of ISD*.

Estimated reliabilities of indices of emotion dynamics

To estimate the reliability of each index, unconditional V-known models were applied to the observed values and approximate sampling variances. We used the R package metaSEM for this purpose (Cheung, 2015b). Appendix 1 shows metaSEM code used to obtain the following basic statistics: (a) the estimated true variance τ^2, (b) a test of significant between-person variance in the index via the Q-statistic, and (c) the estimated reliability based on the I^2 statistic, together with likelihood-based 95% confidence intervals. The metaSEM code and data can also be accessed at https://osf.io/nk6zv/.

Results are summarized in Table 3. The Q-statistic rejected the null-hypothesis of no true between-person variance for all indices. However, the estimated reliabilities I^2 varied substantially between different indices, with reliability coefficients exceeding .95 for $\bar{y}$; coefficients approaching or exceeding a level of .90 for ISD_{RB}; coefficients ranging between approximately .70 and .85 for ISD*, *z*ICC, and *z'*; coefficients in the range of .60 to .70 for $\hat{p}_{FT}$; and coefficients between .35 and .40 for the AR(1) index.

A side-by-side comparison of distributional characteristics of the V-known estimated true scores and observed (manifest variables) scores is also shown in Table 3. The means of the estimated latent variables are very close to the means of observed scores, whereas the model estimated variances τ^2 are smaller than the variances of the manifest variables by magnitudes corresponding with the (un-)reliability of each index. The caterpillar plots in Fig. 3 illustrate the difference between observed scores and empirical Bayes estimates of the true scores (with 95% confidence intervals) derived from the V-known models. For indices with moderate to low reliability [e.g., $\hat{p}_{FT}$, AR(1)], the empirical Bayes estimates are substantially more concentrated around the mean than the observed scores.

Applying the V-known model to correct for unreliability in indices of emotion dynamics

In additional real-data analyses, we used the V-known approach to examine an applied research question from the psychology of aging and emotions, using the same dataset and emotion indices as above. A well-established finding is that older age is associated with lower depression levels, and it has been hypothesized that age-related improvements in the dynamics of emotion regulation account for this effect (Carstensen et al., 2000). With this hypothesis in mind, we

[2] Because ISD_{RB} was calculated from detrended PA and NA data for some participants, the calculation of the sampling variance shown in Table 1 was adjusted accordingly. Linear detrending reduces the degrees of freedom associated with the standard deviation by one, such that the approximate sampling variance of ISD_{RB} becomes 1/[2*(*T*-2)] instead of 1/[2*(*T*-1)] for these participants (Raudenbush & Bryk, 2002). This adjustment of the sampling variance had minimal impact on the results.

Table 3 Estimated reliabilities and descriptive statistics of indices of emotion dynamics from empirical data

Concept	Index	V-known multilevel approach						Manifest variable approach	
		Reliability [a]		Heterogeneity test		Latent variable distribution		Observed variable distribution	
		Estimate	95% CI	*df*	Q[b]	Mean	Variance τ^2	Mean	Variance
Average emotion level	PA $\bar{y}$	.98	(.97; .98)	471	74260.99	4.34	1.46	4.33	1.48
	NA $\bar{y}$	.97	(.96; .97)	471	40845.90	2.07	0.87	2.09	0.92
Emotion variability	PA ISD_{RB}	.90	(.88; .91)	471	4583.17	– 0.50	0.23	– 0.48	0.26
	NA ISD_{RB}	.95	(.95; .96)	471	10300.94	– 0.65	0.55	– 0.64	0.58
Emotional instability	PA $\hat{p}_{FT}$	.62	(.57; .66)	471	1266.20	0.82	0.09	0.83	0.14
	NA $\hat{p}_{FT}$	.67	(.62; .71)	471	1458.15	0.77	0.11	0.77	0.16
Relative emotion variability	PA ISD*	.80	(.77; .82)	471	4976.75	– 1.42	0.19	– 1.44	0.24
	NA ISD*	.71	(.68; .74)	471	5520.97	– 1.20	0.14	– 1.19	0.19
Emotional inertia	PA AR(1)	.35	(.28; .42)	466	730.53	0.21	0.03	0.20	0.08
	NA AR(1)	.39	(.32; .45)	469	791.64	0.19	0.04	0.18	0.08
Emotion differentiation	PA zICC	.85	(.83; .87)	471	3164.50	0.63	0.19	0.63	0.23
	NA zICC	.79	(.76; .81)	470	2252.59	0.59	0.13	0.59	0.16
Emotional bipolarity	z'	.72	(.68; .75)	471	1690.12	– 0.56	0.15	– 0.56	0.21

PA positive affect, *NA* negative affect, *CI* confidence interval, $\bar{y}$ =mean, ISD_{RB} Raudenbush-Bryk transformed standard deviation, $\hat{p}_{FT}$ Freeman–Tukey transformed proportion, *ISD** relative standard deviation, *AR(1)* first-order autocorrelation, *z*ICC Fisher *z*-transformed intraclass correlation, *z'* Fisher *z*-transformed Pearson correlation

[a]Reliabilities use the arithmetic mean of the sampling variances across individuals. [b]$p < 0.001$ for all Q values

chose to examine a set of mediator models addressing the question whether any given index of emotion dynamics would account for (i.e., mediate) age differences in depressive symptoms. We selected a mediation hypothesis to illustrate how the V-known approach can be flexibly incorporated into analysis models that are common in applied research. However, the reader should bear in mind that the analyses are based on cross-sectional age differences and cannot address the problem of causal processes that unfold over time (Maxwell & Cole, 2007).

Age and self-reported depression were assessed at the beginning of the daily diary study. Age (mean = 50.87, SD = 16.12, range = 21 to 91 years) was scaled in decades for analysis. Depression was assessed via computerized adaptive testing using the NIH Patient-Reported Outcomes Measurement Information System (PROMIS; Pilkonis et al., 2011), and was scaled on a T-score metric (mean = 50, SD = 10 in the general population); the mean depression T-score in the present sample was 49.88 (SD = 8.63). Older age was associated with significantly lower depression scores ($b = -1.35$, SE = .24, $p < 0.001$).

All mediation models were estimated in M*plus*. In one set of models, each index of emotion dynamics was treated as a manifest intermediate variable of the age-depression relationship in a single-level mediation analysis. A corresponding set of V-known multilevel mediation analysis models treated each index as a latent intermediate variable. The V-known mediation model is illustrated in Fig. 1 (discussed above), where x_i represents age, y_i represents depression, and δ_i is the latent variable representation of a given index on the between-person level. Annotated M*plus* code used to estimate the V-known mediation model is provided in Appendix 2. All mediation models were implemented using Bayesian estimation (with default diffuse priors and a minimum of 10,000 iterations) to accommodate the asymmetric nature of the sampling distribution of indirect effects. The data and M*plus* code (together with OpenMx code to estimate the same model in R) are also available at https://osf.io/nk6zv/.

Results for the manifest and latent mediator variable versions of each index are shown in Table 4. The effects of age on the emotion indices (a-path in each model) were virtually identical between the manifest and latent variable versions of each index. This reflects the finding that measurement error in a dependent variable does not bias its (unstandardized) regression coefficient (Cole & Preacher, 2014). Controlling for age, the effects of the emotion indices on depression (b-path in each model) were consistently larger for latent compared to manifest versions of the indices. Conversely, the direct effects of age on depression (c'-path in each model) were consistently smaller when latent rather than manifest versions of the indices were used. With few exceptions [ISD* for NA, AR(1) for NA and PA, z'], indirect effects of age on depression via the emotion

indices were all significant, indicating that the indices of emotion dynamics (entered in separate models) partially accounted for the relationship between age and depression. Although the pattern of significant parameters did not differ between manifest and latent mediator models, the indirect effects (as well as the proportions mediated) were consistently larger for latent compared to manifest versions, suggesting that biases from ignoring sampling error were reduced or eliminated in the V-known multilevel mediation models.

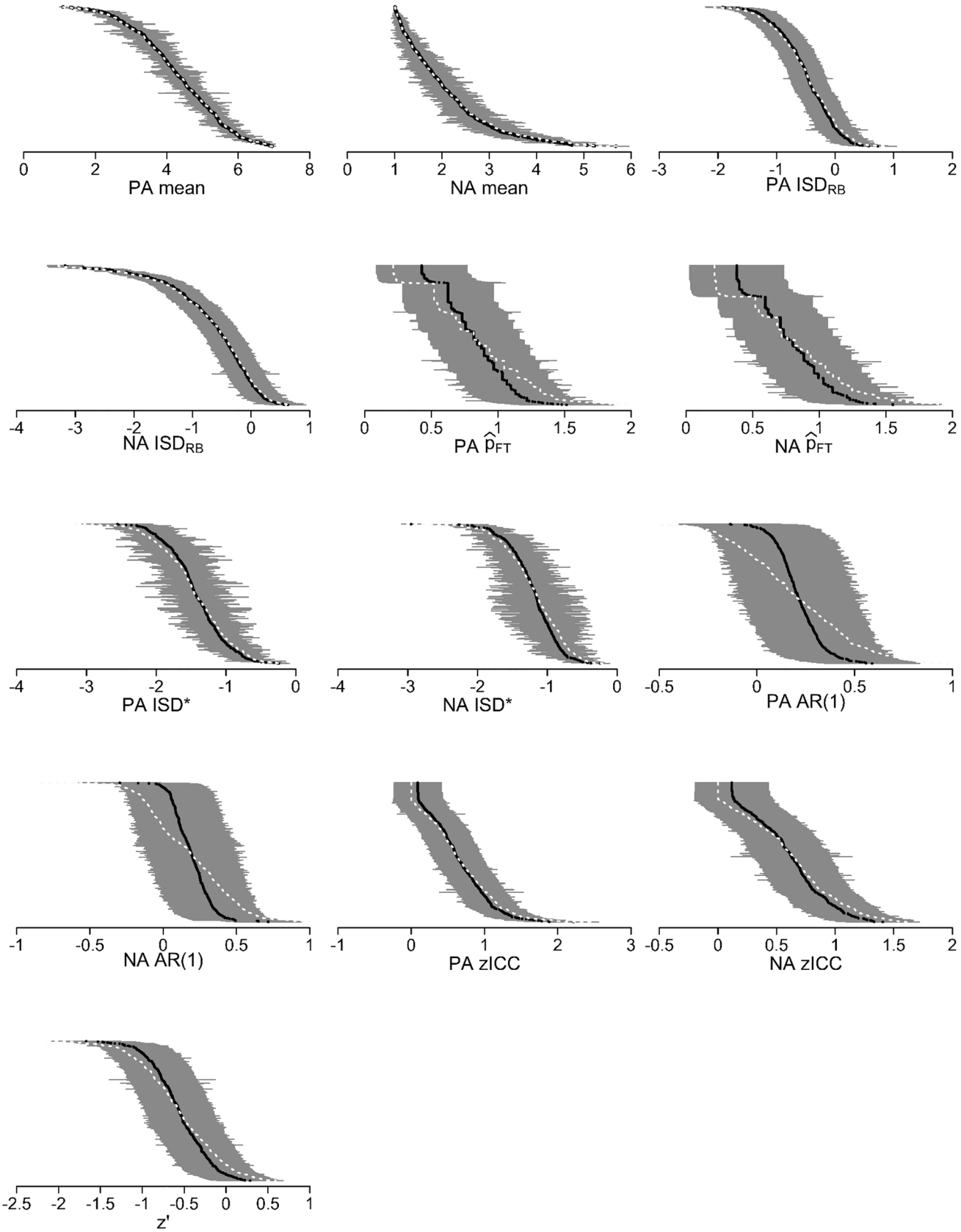
Discussion

ILD provide many novel opportunities for characterizing individual differences in the ebb and flow of behaviors and experiences, and social scientists have enthusiastically embraced the idea that the measurement of WPD with ILD may provide many new insights into individuals' health, well-being, and standing in society. To be able to test whether this is in fact the case, however, it is necessary that the measures themselves are of high quality. For example, a provocative study by Dejonckheere et al. (2019) called into question the incremental validity of many indices of emotion dynamics in predicting psychological wellbeing above and beyond simple measures of mean affect, suggesting the possibility that attempts to derive increasingly more granular indices of WPD from ILD may merely create a false sense of scientific progress. As we have shown, disregarding unreliability in indices of WPD can distort parameter estimates and yield erroneous conclusions about the validity of the indices. Given the proliferation of new indices of WPD, it is paramount that researchers have the tools to evaluate which indices ensure consistent and reproducible measurement and under which circumstances.

The V-known multilevel modeling approach can be used to estimate the reliability of a wide range of indices of WPD applied in empirical research. Our Monte Carlo simulation study indicated that the V-known model can recover the reliability of these indices well as long as the number of measurement occasions per person is not very small. When less than ten measurement occasions were used, the reliabilities were underestimated for some indices and overestimated for others, and the estimated latent variances of the indices were over- and underestimated accordingly. This parallels well-documented issues encountered when meta-analyses are conducted based on studies with small sample sizes (i.e., with few observations available to calculate sampling variances at Level 1), given that the estimated sampling variances do not approximate the true sampling variances well in this case (Lin, 2018). One possible way to improve the estimated sampling variance for many different types of indices might be to use Bayesian estimation with informative priors rather than relying on asymptotic distributions of their sampling variances (McNeish, 2016). Alternatively, it might be possible to develop models that account for the uncertainty about the sampling variance in an index by treating the sampling variances as estimates rather than assuming that they are known (Malzahn et al., 2000). More extensive simulations are needed to determine the performance of the V-known application across a range of indices, parameter estimators, and distributional characteristics.

The results of our empirical analysis of the reliability of several indices of emotion dynamics demonstrated the utility of the V-known approach in a 21-day diary study. Out of the tested indices, the measurement of average emotion levels ($\bar{y}$) showed the highest reliabilities and measures of emotional inertia [AR(1)] showed by far the lowest reliability. This finding corresponds with earlier results from Monte Carlo simulations that found very low reliabilities for the AR(1) parameter under many simulated conditions (Du & Wang, 2018). Emotion variability (ISD_{RB}) was very reliably measured in this sample, contrary to claims that ISDs have generally poor reliability based on simulated conditions (Estabrook et al., 2012). Other indices that were tested here (probability of acute change, relative emotion variability, and emotion bipolarity and differentiation) have repeatedly been used in ILD research on emotion dynamics with very little to no knowledge about their reliability. In our sample, these indices showed moderate-to-good reliabilities ranging from .62 to .85. However, we emphasize that reliabilities depend on the characteristics of the sample and our empirical results may not generalize to other samples. We encourage researchers to apply the techniques presented here to their own samples in future ILD studies and to report the resulting reliabilities in their publications. Ultimately, this would provide a cumulative evidence base that could be used to evaluate the reliabilities of commonly used indices of WPD—across different samples, measurement instruments, and research areas—in reliability generalization studies.

As we have shown, the V-known approach also makes it possible to treat indices of WPD as latent variables in analysis models in order to correct the parameter estimates for sampling error in the indices. Apart from the mediation model examples in our empirical demonstration, the V-known approach can be flexibly integrated in a general latent variable modeling framework (e.g., as provided in M*plus*), for example, to examine latent variable interactions or to identify latent subgroups of individuals with different underlying dynamics using finite mixture models (Cheung, 2008, 2015a). Multiple indices can also be simultaneously handled in multivariate V-known models (Cheung, 2013; Jackson et al., 2011). When doing this, it is necessary to carefully consider potential non-independencies of multiple indices within individuals. Whereas the covariance between


PA mean
NA mean
PA ISD$_{RB}$
NA ISD$_{RB}$
PA $\widehat{p}_{FT}$
NA $\widehat{p}_{FT}$
PA ISD*
NA ISD*
PA AR(1)
NA AR(1)
PA zICC
NA zICC
z'

◂**Fig. 3** Caterpillar plots of indices of emotion dynamics. *Black dotted lines* represent the empirical Bayes estimated values for each index, sorted in ascending order. *Gray shaded areas* represent 95% confidence intervals around the empirical Bayes estimates. *Light gray dotted lines* represent the observed values for each index. PA = positive affect, NA = negative affect. ISD_{RB} = Raudenbush–Bryk transformed standard deviation (*emotion variability*), $\hat{p}_{FT}$ = Freeman–Tukey transformed proportion (*emotional instability*), ISD* = relative standard deviation (*relative emotion variability*), AR(1) = first-order autocorrelation (*emotional inertia*), *z*ICC = Fisher *z*-transformed intraclass correlation *(emotion differentiation)*, *z'* = Fisher *z*-transformed Pearson correlation (*emotional bipolarity*)

indices can be freely estimated at Level 2 (between-person), covariances of observations from which the indices are derived need to be specified as known at Level 1 (within-person) in a multivariate V-known model. For example, when deriving indices of ISD_{RB} based on both PA and NA ratings of the same individuals in ILD, the covariance of the two indices can be approximated as $cov_i = r^2_i / (2\ T_i)$, where r is the within-person correlation of PA and NA ratings of individual i (Raudenbush & Bryk, 2002). Alternatively, robust variance estimation has been proposed to account for non-independent summary statistics in multivariate meta-analyses (Hedges et al., 2010; Tipton, 2015), although further research is needed for this approach.

When considering the utility of V-known models for the purpose of correcting for sampling error in indices of WPD, it needs to be kept in mind that V-known models may not perform as well as multilevel models in which WPD indices are estimated directly as parameters from the raw data (e.g., via multilevel time-series analysis in M*plus*), especially when the number of measurement occasions per person is low. For multilevel time-series parameters, Schultzberg and Muthén (2018) have shown that models with many individuals and few measurement occasions recovered the true WPD parameters [individual means, log intraindividual variability, and AR(1)] substantially better than models with few individuals and many measurement occasions. It would be worthwhile to formally investigate to what extent and under which conditions multilevel time-series models outperforms the V-known approach for indices that can be estimated equally using both approaches. In addition, an intriguing possibility is to couple the two approaches in a joint multilevel model. That is, WPD indices that can be estimated directly as multilevel time series parameters (Hamaker et al., 2018; McNeish & Hamaker, 2020) could be combined with indices for which the use of a two-step approach is the only option in the same model, using the V-known approach for the latter indices.

A strength of the V-known multilevel model is that it takes individual differences in the amount of sampling variance into account. Thus, estimated reliabilities and latent variables underlying the observed WPD indices are explicitly adjusted for unbalanced numbers of measurement occasions that frequently occur as part of ILD study designs or due to participant noncompliance with the study protocol. In extreme cases, however, participants who drop out of an ILD study or who show low compliance will only contribute very few measurements to the derived WPD indices, which, in view of our finding that the sampling variances may not be approximated well for these cases, may potentially bias parameter estimates (e.g., reliabilities) (Lin, 2018). The extent of the resulting biases may depend on the amount of imbalance in sampling variances (e.g., the number of participants with few measurements relative to the overall sample size), the amount of true variance between individuals, and the specific WPD index, and these various influences may be addressed in additional simulation studies. It is also important to note that, unless missed measurement occasions are missing completely at random, the observed values of WPD indices may themselves be biased and V-known models do not account for this. Graphical tools available for meta-analysis (funnel plots, residual diagnostic plots, outlier plots) may be helpful to diagnose such biases in WPD indices (Viechtbauer & Cheung, 2010). For example, in funnel plots, which plot the observed point estimates of an index against their sampling variances, an asymmetric funnel may indicate that lower compliance (higher sampling variances) is associated with the values on a WPD index, suggesting the possibility that assessments were not missed completely at random.

It is also important to note that the methods presented in this article exclusively focused on unreliability due to sampling error. Another common source of unreliability in observed scores of indices of WPD is measurement error at each sampling occasion that is due to unreliability in the measurement instrument (Lüdtke et al., 2011). Prior simulation studies suggest that moderate-to-high levels of error variance in measurement instruments can substantially contribute to the overall unreliability of intraindividual variability indices (Du & Wang, 2018; Estabrook et al., 2012; Wang & Grimm, 2012). In future research, possible extensions to the strategies presented here could be considered that attempt to incorporate the role of measurement error in V-known models. For example, a key element in the meta-analysis methods proposed by Hunter and Schmidt (2004) is the application of various "artifact corrections", including a correction of the observed values of a summary statistic (and of its corresponding sampling variance) for attenuation due to measurement error. More research on the feasibility of incorporating measurement errors in the estimation of the reliability of WPD indices is warranted.

Table 4 Results from mediation models with indices of emotion dynamics as manifest versus latent intermediate variables

Emotion index	Age ➔ emotion index (a-path)		Emotion index ➔ depression (b-path)		Age ➔ Depression (c'-path)		Indirect effect (a*b-paths)		Proportion mediated
	Est.	95% CI	Est.	95% CI	Est.	95% CI	Est.	95% CI	Est.
PA $\bar{y}$									
Manifest	0.11	(0.04; 0.18)	– 3.73	(– 4.26; – 3.20)	– 0.95	(– 1.34; – 0.52)	– 0.41	(– 0.68; – 0.16)	.30
Latent	0.11	(0.04; 0.18)	– 3.83	(– 4.37; – 3.30)	– 0.92	(– 1.32; – 0.52)	– 0.42	(– 0.70; – 0.15)	.32
NA $\bar{y}$									
Manifest	– 0.19	(– 0.24; – 0.14)	5.82	(5.18; 6.47)	– 0.25	(– 0.63; 0.15)	– 1.10	(– 1.43; – 0.79)	.81
Latent	– 0.19	(– 0.24; – 0.14)	6.15	(5.48; 6.82)	– 0.18	(– 0.57; 0.20)	– 1.16	(– 1.53; – 0.83)	.87
PA ISD_{RB}									
Manifest	– 0.10	(– 0.13; – 0.07)	2.45	(0.87; 4.01)	– 1.10	(– 1.62; – 0.61)	– 0.25	(– 0.43; – 0.09)	.18
Latent	– 0.10	(– 0.13; – 0.07)	2.72	(0.94; 4.44)	– 1.08	(– 1.59; – 0.56)	– 0.27	(– 0.48; – 0.09)	.20
NA ISD_{RB}									
Manifest	– 0.15	(– 0.19; – 0.11)	4.24	(3.25; 5.22)	– 0.73	(– 1.21; – 0.27)	– 0.62	(– 0.85; – 0.41)	.46
Latent	– 0.15	(– 0.19; – 0.11)	4.45	(3.42; 5.45)	– 0.70	(– 1.18; – 0.23)	– 0.65	(– 0.91; – 0.43)	.48
PA $\hat{p}_{FT}$									
Manifest	– 0.06	(– 0.09; – 0.04)	2.19	(0.12; 4.26)	– 1.21	(– 1.70; – 0.71)	– 0.14	(– 0.29; – 0.01)	.10
Latent	– 0.06	(– 0.08; – 0.04)	3.52	(– 0.08; 7.05)	– 1.13	(– 1.65; – 0.61)	– 0.21	(– 0.47; – 0.002)	.16
NA $\hat{p}_{FT}$									
Manifest	– 0.07	(– 0.09; – 0.05)	6.35	(4.46; 8.19)	– 0.90	(– 1.38; – 0.44)	– 0.44	(– 0.66; – 0.27)	.33
Latent	– 0.07	(– 0.09; – 0.05)	9.77	(6.90; 12.75)	– 0.70	(– 1.20; – 0.19)	– 0.65	(– 0.98; – 0.39)	.48
PA ISD*									
Manifest	– 0.09	(– 0.11; – 0.06)	2.05	(0.42; 3.70)	– 1.17	(– 1.65; – 0.68)	– 0.18	(– 0.34; – 0.04)	.13
Latent	– 0.09	(– 0.11; – 0.06)	2.26	(0.30; 4.28)	– 1.16	(– 1.66; – 0.64)	– 0.19	(– 0.39; – 0.03)	.14
NA ISD*									
Manifest	– 0.03	(– 0.05; – 0.004)	0.06	(– 1.78; 1.90)	– 1.35	(– 1.82; – 0.87)	– 0.001	(– 0.06; 0.06)	.001
Latent	– 0.04	(– 0.06; – 0.01)	1.45	(– 0.79; 3.73)	– 1.30	(– 1.78; – 0.80)	– 0.05	(– 0.16; 0.03)	.04
PA AR(1)									
Manifest	0.01	(– 0.01; 0.02)	3.10	(0.37; 5.77)	– 1.36	(– 1.84; – 0.89)	0.01	(– 0.04; 0.08)	– .01
Latent	0.01	(– 0.01; 0.02)	6.72	(0.03; 13.73)	– 1.40	(– 1.89; – 0.91)	0.03	(– 0.09; 0.20)	– .02
NA AR(1)									
Manifest	– 0.01	(– 0.03; 0.01)	5.39	(2.73; 8.08)	– 1.30	(– 1.77; – 0.83)	– 0.05	(– 0.15; 0.04)	.04
Latent	– 0.01	(– 0.03; 0.01)	11.66	(5.59; 17.98)	– 1.22	(– 1.73; – 0.72)	– 0.13	(– 0.39; 0.06)	.10
PA *z*ICC									
Manifest	– 0.08	(– 0.10; – 0.05)	2.20	(0.59; 3.84)	– 1.18	(– 1.68; – 0.71)	– 0.17	(– 0.32; – 0.04)	.12
Latent	– 0.08	(– 0.10; – 0.05)	2.57	(0.60; 4.53)	– 1.15	(– 1.65; – 0.65)	– 0.19	(– 0.38; – 0.05)	.14
NA *z*ICC									
Manifest	– 0.03	(– 0.06; – 0.01)	2.00	(0.15; 3.86)	– 1.29	(– 1.76; – 0.81)	– 0.06	(– 0.16; – 0.002)	.04
Latent	– 0.03	(– 0.05; – 0.01)	2.58	(0.15; 5.06)	– 1.27	(– 1.74; – 0.79)	– 0.08	(– 0.20; – 0.001)	.06
z'									
Manifest	0.007	(– 0.02; 0.03)	– 2.50	(– 4.13; – 0.85)	– 1.33	(– 1.81; – 0.86)	– 0.02	(– 0.10; 0.05)	.01
Latent	0.007	(– 0.02; 0.03)	– 3.40	(– 5.74; – 1.11)	– 1.33	(– 1.80; – 0.84)	– 0.02	(– 0.13; 0.07)	.02

PA positive affect, *NA* negative affect, *CI* credible interval, $\bar{y}$ =mean, ISD_{RB} Raudenbush–Bryk transformed standard deviation, $\hat{p}_{FT}$ Freeman–Tukey transformed proportion, *ISD** relative standard deviation, *AR(1)* first-order autocorrelation, *z*ICC Fisher *z*-transformed intraclass correlation, *z'* Fisher *z*-transformed Pearson correlation.

In conclusion, we have demonstrated how V-known multilevel models can be used for the purposes of estimating reliabilities and correcting for sampling error in indices of WPD. The proposed approach joins other applications that have increasingly recognized the versatility of V-known or meta-analysis models for use with primary data (coordinated multi-study analysis, Hofer & Piccinin, 2009; individual participant data meta-analysis, Riley et al., 2010). We hope that the use of V-known multilevel models in empirical studies may contribute to the toolkit available for research on indices of WPD and may add to psychometrically sound inferences in this burgeoning research area.

Appendix 1

The syntax below can be used to estimate the reliability of any index of within-person dynamics via the I^2 statistic in the R package metaSEM. After loading the package, the function meta() is used to conduct a univariate unconditional meta-analysis. The arguments y and v are used to specify the observed values and sampling variances of the index, respectively. The argument I2="I2am" is used to specify that the reliability is estimated based on the arithmetic mean of the subject-specific sampling variances (arguments I2="I2q" and I2="I2hm" can be specified for alternative definitions of the typical sampling variance based on the Q-statistic or harmonic mean, respectively). The argument intervals.type="LB" requests a likelihood based 95% confidence interval for the reliability coefficient. Summary() is used to extract the results. A random-effects meta-analysis is fitted and the Q statistic testing the null-hypothesis of no true individual differences in the index is given alongside the I^2 statistic by default.

```
library(metaSEM)

summary(meta(data=emo_indices, y=index, v=sampvar, I2="I2am", intervals.type="LB"))
```

Appendix 2

The M*plus* code below can be used to estimate a V-known multilevel mediation model. The observed index and sampling variance are transformed in the define command as necessary to implement the V-known model. These transformed variables are specified as within variables in the variable command. In the within part of the model, the transformed index is regressed on the transformed sampling variance with a random slope *delta*. The mean and residual variance of the transformed index variable must be constrained at 0 and 1, respectively. In the between part of the model, *delta* is specified as a latent mediator of the relationship between age and dep.

```
TITLE:
V-known multilevel mediation;

DATA:
FILE IS emo_indices.dat;

DEFINE:
w = sqrt(sampvar**(-1));  !transform the sampling variance sampvar into w
indexw = index * w;       !transform the index by multiplying it with w

VARIABLE:
NAMES ARE
id                        !ID variable
age                       !between-person predictor variable
dep                       !between-person outcome variable
index                     !index of within-person dynamics
sampvar;                  !sampling variance for index of within-person dynamics

USEVARIABLES IS
age dep w indexw;         !use variables that were transformed in the DEFINE command

MISSING = .;              !missing value flag
WITHIN IS w indexw;       !declare transformed sampling variance and index as within
BETWEEN IS age dep;       !declare observed age and depression scores as between
CLUSTER IS id;            !declare ID as clustering variable

ANALYSIS:
TYPE IS TWOLEVEL RANDOM;  !two-level random effects model;
ESTIMATOR = BAYES;
BITER = (10000);          !minimal number of iterations set at 10000;

MODEL:
%WITHIN%                  !specify the Level-1 V-known model on the within level
delta | indexw on w;      !regress the transformed index on w, with random slope delta
[indexw@0];               !fix the intercept of the transformed index at 0
indexw@1;                 !fix the residual variance of the transformed index at 1

%BETWEEN%                 !specify the mediation model on the between level
delta on age (a);         !regress the latent index on age (a-path)
dep on delta (b);         !regress depression on the latent index (b-path)
dep on age;               !regress depression on age (c'-path)

MODEL CONSTRAINT:
new (ind);
ind = a*b;                !estimate the indirect effect via the latent index;

OUTPUT:
CINTERVAL;                !output credible intervals
```

Funding This work was supported by grants from the National Institute on Aging (R37AG057685, R21AG061364, R01AG042407, and R01AG068190).

Declarations

Financial interests The authors have no competing interests to declare that are relevant to the content of this article.

Ethics approval Approval was obtained from the Institutional Review Board of the University of Southern California. The procedures used in this study adhere to the tenets of the Declaration of Helsinki.

Consent to participate Informed consent was obtained from all individual participants included in the study.

References

Asparouhov, T., Hamaker, E. L., & Muthén, B. (2018). Dynamic structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal, 25*(3), 359–388.

Becker, B. J., & Wu, M.-J. (2007). The synthesis of regression slopes in meta-analysis. *Statistical Science, 22*(3), 414–429.

Bollen, K. A. (1989). *Structural equations with latent variables*. Wiley.

Bonett, D. G. (2002). Sample size requirements for testing and estimating coefficient alpha. *Journal of Educational and Behavioral Statistics, 27*(4), 335–340.

Borsboom, D. (2003). *Conceptual issues in psychological measurement*. Ipskamp.

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

Card, N. A. (2012). *Applied meta-analysis for social science research*. Guilford Press.

Carstensen, L. L., Pasupathi, M., Mayr, U., & Nesselroade, J. R. (2000). Emotional experience in everyday life across the adult life span. *Journal of Personality and Social Psychology, 79*(4), 644–655.

Cheung, M. W.-L. (2008). A model for integrating fixed-, random-, and mixed-effects meta-analyses into structural equation modeling. *Psychological Methods, 13*(3), 182–202.

Cheung, M.W.-L. (2013). Multivariate meta-analysis as structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal, 20*(3), 429–454.

Cheung, M. W.-L. (2015a). *Meta-analysis: A structural equation modeling approach*. John Wiley & Sons.

Cheung, M. W.-L. (2015b). metaSEM: an R package for meta-analysis using structural equation modeling. *Frontiers in Psychology, 5*, 1521.

Chow, S.-M., Ram, N., Boker, S. M., Fujita, F., & Clore, G. (2005). Emotion as a thermostat: Representing emotion regulation using a damped oscillator model. *Emotion, 5*(2), 208–225.

Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics, 10*(1), 101–129.

Cole, D. A., & Preacher, K. J. (2014). Manifest variable path analysis: Potentially serious and misleading consequences due to uncorrected measurement error. *Psychological Methods, 19*(2), 300–315.

Dejonckheere, E., Mestdagh, M., Houben, M., Erbas, Y., Pe, M., Koval, P., & Kuppens, P. (2018). The bipolarity of affect and depressive symptoms. *Journal of Personality and Social Psychology, 114*(2), 323–341.

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

Demidenko, E. (2013). *Mixed models: Theory and applications with R* (2nd ed.), Wiley-Interscience.

Du, H., & Wang, L. (2018). Reliabilities of intraindividual variability indicators with autocorrelated longitudinal data: Implications for longitudinal study designs. *Multivariate Behavioral Research, 53*(4), 502–520.

Efron, B., & Stein, C. (1981). The jackknife estimate of variance. *The Annals of Statistics, 9*, 586–596.

Eid, M., & Diener, E. (1999). Intraindividual variability in affect: Reliability, validity, and personality correlates. *Journal of Personality and Social Psychology, 76*(4), 662–676.

Estabrook, R., Grimm, K. J., & Bowles, R. P. (2012). A Monte Carlo simulation study of the reliability of intraindividual variability. *Psychology and Aging, 27*(3), 560–576.

Fernández-Castilla, B., Jamshidi, L., Declercq, L., Beretvas, S. N., Onghena, P., & Van den Noortgate, W. (2020). The application of meta-analytic (multi-level) models with multiple random effects: A systematic review. *Behavior Research Methods, 52*, 2031–2052.

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika, 10*(4), 507–521.

Fisher, R. A. (1938). *Statistical methods for research workers* (7th ed.), Oliver and Boyd.

Fleeson, W. (2001). Toward a structure-and process-integrated view of personality: Traits as density distributions of states. *Journal of Personality and Social Psychology, 80*(6), 1011–1027.

Freeman, M. F., & Tukey, J. W. (1950). Transformations related to the angular and the square root. *The Annals of Mathematical Statistics, 21*(4), 607–611.

Fritz, M. S., Kenny, D. A., & MacKinnon, D. P. (2016). The combined effects of measurement error and omitting confounders in the single-mediator model. *Multivariate Behavioral Research, 51*(5), 681–697.

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

Hamaker, E. L., Grasman, R. P., & Kamphuis, J. H. (2016). Modeling BAS dysregulation in bipolar disorder: Illustrating the potential of time series analysis. *Assessment, 23*(4), 436–446.

Hamaker, E. L., Asparouhov, T., Brose, A., Schmiedek, F., & Muthén, B. (2018). At the frontiers of modeling intensive longitudinal data: Dynamic structural equation models for the affective measurements from the COGITO study. *Multivariate Behavioral Research, 53*(6), 820–841.

Hedeker, D., Mermelstein, R. J., & Demirtas, H. (2012). Modeling between-subject and within-subject variances in ecological momentary assessment data using mixed-effects location scale models. *Statistics in Medicine, 31*(27), 3328–3336.

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods, 1*(1), 39–65.

Higgins, J. P., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine, 21*(11), 1539–1558.

Higgins, J. P., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal, 327*(7414), 557–560.

Hofer, S. M., & Piccinin, A. M. (2009). Integrative data analysis through coordination of measurement and analysis protocol across independent longitudinal studies. *Psychological Methods, 14*(2), 150.

Houben, M., Van Den Noortgate, W., & Kuppens, P. (2015). The relation between short-term emotion dynamics and psychological

well-being: A meta-analysis. *Psychological Bulletin, 141*(4), 901–930.

Hu, Y., Nesselroade, J. R., Erbacher, M. K., Boker, S. M., Burt, S. A., Keel, P. K., ... Klump, K. (2016). Test reliability at the individual level. *Structural Equation Modeling: A Multidisciplinary Journal, 23*(4), 532–543.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. Sage.

Jackson, D., Riley, R., & White, I. R. (2011). Multivariate meta-analysis: Potential and promise. *Statistics in Medicine, 30*(20), 2481–2498.

Jahng, S., Wood, P. K., & Trull, T. J. (2008). Analysis of affective instability in ecological momentary assessment: Indices using successive difference and group comparison via multilevel modeling. *Psychological Methods, 13*(4), 354–375.

Jongerling, J., Laurenceau, J.-P., & Hamaker, E. L. (2015). A multilevel AR (1) model: Allowing for inter-individual differences in trait-scores, inertia, and innovation variance. *Multivariate Behavioral Research, 50*(3), 334–349.

Junghaenel, D. U., Broderick, J. E., Schneider, S., Wen, C. K. F., Mak, H. W., Goldstein, S., ... Stone, A. A. (2021). Explaining age differences in the memory-experience gap. *Psychology and Aging, 36*(6), 679–693.

Kashdan, T. B., Barrett, L. F., & McKnight, P. E. (2015). Unpacking emotion differentiation: Transforming unpleasant experience by perceiving distinctions in negativity. *Current Directions in Psychological Science, 24*(1), 10–16.

Kelley, K., & Pornprasertmanit, S. (2016). Confidence intervals for population reliability coefficients: Evaluation of methods, recommendations, and software for composite measures. *Psychological Methods, 21*(1), 69–92.

Kenny, D. A. (1979). *Correlation and causality.* Wiley.

Krone, T., Albers, C. J., & Timmerman, M. E. (2017). A comparative simulation study of AR (1) estimators in short time series. *Quality & Quantity, 51*(1), 1–21.

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

Lin, L. (2018). Bias caused by sampling error in meta-analysis with small sample sizes. *PloS One, 13*(9), e0204056.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical Meta-Analysis*. Sage Publications.

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

Lüdtke, O., Marsh, H. W., Robitzsch, A., & Trautwein, U. (2011). A 2×2 taxonomy of multilevel latent contextual models: Accuracy–bias trade-offs in full and partial error correction models. *Psychological Methods, 16*(4), 444–467.

Malzahn, U., Böhning, D., & Holling, H. (2000). Nonparametric estimation of heterogeneity variance for the standardised difference used in meta-analysis. *Biometrika, 87*(3), 619–632.

Maxwell, S. E., & Cole, D. A. (2007). Bias in cross-sectional analyses of longitudinal mediation. *Psychological Methods, 12*(1), 23–44.

McNeish, D. (2016). On using Bayesian methods to address small sample problems. *Structural Equation Modeling: A Multidisciplinary Journal, 23*(5), 750–773.

McNeish, D., & Hamaker, E. (2020). A primer on two-level dynamic structural equation models for intensive longitudinal data in Mplus. *Psychological Methods, 25*, 610–635.

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

Molenaar, P. C. M., & Campbell, C. G. (2009). The New Person-Specific Paradigm in Psychology. *Current Directions in Psychological Science, 18*(2), 112–117.

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling, 9*, 599–620.

Muthén, L. K., & Muthén, B. O. (2017). *Mplus: Statistical Analysis with Latent Variables: User's Guide (Version 8)*. Muthén & Muthén.

Palmer, T. M., & Sterne, J. A. (2015). Fitting fixed-and random-effects meta-analysis models using structural equation modeling with the sem and gsem commands. *The Stata Journal, 15*(3), 645–671.

Pastor, D. A., & Lazowski, R. A. (2018). On the multilevel nature of meta-analysis: a tutorial, comparison of software programs, and discussion of analytic choices. *Multivariate Behavioral Research, 53*(1), 74–89.

Pilkonis, P. A., Choi, S. W., Reise, S. P., Stover, A. M., Riley, W. T., Cella, D., & Grp, P. C. (2011). Item Banks for Measuring Emotional Distress From the Patient-Reported Outcomes Measurement Information System (PROMIS (R)): Depression, Anxiety, and Anger. *Assessment, 18*(3), 263–283.

Ram, N., & Gerstorf, D. (2009). Time-Structured and Net Intraindividual Variability: Tools for Examining the Development of Dynamic Characteristics and Processes. *Psychology and Aging, 24*(4), 778–791.

Raudenbush, S. W., & Bryk, A. S. (1985). Empirical Bayes meta-analysis. *Journal of Educational Statistics, 10*(2), 75–98.

Raudenbush, S. W., & Bryk, A. S. (1987). Examining correlates of diversity. *Journal of Educational and Behavioral Statistics, 12*(3), 241–269.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Sage.

Riley, R. D., Lambert, P. C., & Abo-Zaid, G. (2010). Meta-analysis of individual participant data: rationale, conduct, and reporting. *British Medical Journal, 340*, c221.

Schneider, S., & Stone, A. A. (2015). Mixed emotions across the adult life span in the United States. *Psychology and Aging, 30*(2), 369–382.

Schultzberg, M., & Muthén, B. (2018). Number of subjects and time points needed for multilevel time-series analysis: A simulation study of dynamic structural equation modeling. *Structural Equation Modeling: a Multidisciplinary Journal, 25*(4), 495–515.

Schwartz, J. E., & Stone, A. A. (2007). The analysis of real-time momentary data: A practical guide. In A. A. Stone, S. Shiffman, A. Atienza, & L. Nebeling (Eds.), *The Science of Real-Time Data Capture: Self-Report in Health Research* (pp. 76–113). Oxford University Press.

Tipton, E. (2015). Small sample adjustments for robust variance estimation with meta-regression. *Psychological Methods, 20*(3), 375–393.

van Houwelingen, H. C., Arends, L. R., & Stijnen, T. (2002). Advanced methods in meta-analysis: Multivariate approach and meta-regression. *Statistics in Medicine, 21*(4), 589–624.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3), 1–48.

Viechtbauer, W., & Cheung, M. W. L. (2010). Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods, 1*(2), 112–125.

Wang, L., & Grimm, K. J. (2012). Investigating Reliabilities of Intraindividual Variability Indicators. *Multivariate Behavioral Research, 47*(5), 771–802. https://doi.org/10.1080/00273171.2012.715842

Wang, L. P., Hamaker, E., & Bergeman, C. (2012). Investigating inter-individual differences in short-term intra-individual variability. *Psychological Methods, 17*(4), 567–581.

Wendt, L. P., Wright, A. G., Pilkonis, P. A., Woods, W. C., Denissen, J., Kühnel, A., & Zimmermann, J. (2020). Indicators of affect dynamics: structure, reliability, and personality correlates. *European Journal of Personality.*

Wilkinson, L., & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594–604.

Wolter, K. (2007). *Introduction to variance estimation*. Springer Science & Business Media.

Xiong, C., Gao, F., Yan, Y., Luo, J., Sung, Y., & Shi, G. (2008). Measuring Overall Heterogeneity in Meta-Analyses: Application to CSF Biomarker Studies in Alzheimer's Disease. *Journal of Modern Applied Statistical Methods, 7*(1), 24.

Xiong, C., Miller, J. P., & Morris, J. C. (2010). Measuring study-specific heterogeneity in meta-analysis: Application to an antecedent biomarker study of Alzheimer's disease. *Statistics in Biopharmaceutical Research, 2*(3), 300–309.

Open practices statement The datasets analyzed during the current study and software code are available at https://osf.io/nk6zv/. The study was not preregistered.