On the Use of Empirical Bayes Estimates as Measures of Individual Traits

Liu, Siwei; Kuppens, Peter; Bringmann, Laura

Published in:
Assessment

DOI:
10.1177/1073191119885019

Publication date:
2021

Article

Assessment
2021, Vol. 28(3) 845–857
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1073191119885019
journals.sagepub.com/home/asm
SAGE

On the Use of Empirical Bayes Estimates as Measures of Individual Traits

Siwei Liu[1], Peter Kuppens[2], and Laura Bringmann[3]

Abstract

Empirical Bayes (EB) estimates of the random effects in multilevel models represent how individuals deviate from the population averages and are often extracted to detect outliers or used as predictors in follow-up analysis. However, little research has examined whether EB estimates are indeed reliable and valid measures of individual traits. In this article, we use statistical theory and simulated data to show that EB estimates are biased toward zero, a phenomenon known as "shrinkage." The degree of shrinkage and reliability of EB estimates depend on a number of factors, including Level-1 residual variance, Level-1 predictor variance, Level-2 random effects variance, and number of within-person observations. As a result, EB estimates may not be ideal for detecting outliers, and they produce biased regression coefficients when used as predictors. We illustrate these issues using an empirical data set on emotion regulation and neuroticism.

Keywords

multilevel modeling, empirical Bayes estimates, longitudinal data, stress reactivity, emotional inertia

Multilevel modeling is a common statistical technique for analyzing longitudinal data. When estimating a multilevel model, researchers typically distinguish between two types of effects: the *fixed effects*, which represent the average effects at the population level; and the *random effects*, which represent individuals' deviations from the fixed effects. Whereas sometimes researchers are interested in estimating and testing for the fixed effects parameters and the variances and covariances of the random effects, often such models are used to obtain the individual-specific random effects estimates when modeling as measures of individual traits (Asendorpf, 2006; Bai & Repetti, 2018; Bringmann et al., 2016; Mohr et al., 2013; van Eck, Berkhof, Nicolson, & Sulon, 1996). Random intercepts are used as indicators of individual differences in how people feel, think, and behave on average, and random slopes as indicators of individual differences in the dynamic processes underlying the psychological phenomenon of interest.

Also known as empirical Bayes (EB) estimates, these individual-specific random effects estimates have been used mainly for two purposes. One is to detect outliers. For example, Morrell and Brant (1991) fit a multilevel model to longitudinal data and used the EB estimates to detect outliers among older adults in the change of hearing perception. Cummings, Stoolmiller, Baker, Fien, and Kame'enui (2015) analyzed student achievement data from a sample of schools and used the EB estimates to evaluate whether a specific school lags behind other schools. The other common use is to extract the EB estimates and treat them as variables in other analyses such as predictors of an outcome variable in follow-up analyses. For example, health psychologists have used EB estimates to quantify individuals' emotional reactivity to stress and found consistent patterns linking stronger stress reactivity to negative health outcomes (Cohen, Gunthert, Butler, O'Neill, & Tolpin, 2005; Mroczek et al., 2015; Ong et al., 2013; Sin, Graham-Engeland, Ong, & Almeida, 2015). Similarly, personality and clinical psychologists have examined emotional inertia using multilevel autoregressive models and found that individual emotional inertia, as represented by the EB estimates of the autoregressive parameters, significantly predict mental health problems such as depression (Brose, Schmiedek, Koval, & Kuppens, 2015; Hamaker & Grasman, 2015; Koval, Kuppens, Allen, & Sheeber, 2012; Kuppens et al., 2012).

Despite the popularity of these approaches, little research has examined whether EB estimates are indeed reliable and valid measures of individual traits. Liu (2017, 2018) compared EB estimates to regression estimates from person-specific autoregressive models with time series data and

[1]University of California at Davis, CA, USA
[2]University of Leuven, Leuven, Belgium
[3]University of Groningen, Groningen, Netherlands

Corresponding Author:
Siwei Liu, Human Development and Family Studies, Department of Human Ecology, University of California, One Shields Ave., Davis, CA 95616, USA.
Email: sweliu@ucdavis.edu

found that EB estimates generally have better accuracy, and similar reliability, when all individuals in the sample have homogeneous dynamic patterns, and when the model used to analyze the data is correctly specified. In addition, accuracy and reliability are affected by the number of within-person observations and the random effects variance, but not by sample size or distribution of the random effects. Du and Wang (2018) compared the reliability of various intra-individual variability indicators based on time series data and found that EB estimates of the autoregressive parameters generally have lower reliability than other indicators, such as the intraindividual standard deviation. In terms of factors affecting reliability, they found similar results as Liu (2018). In addition, they also found substantial influences of the measurement scale reliability and a number of other factors, such as the intraindividual variance of the autoregressive process. Although these studies provide important insights to the properties of EB estimates, all are simulation studies that focus on a limited set of factors. As such, it is unclear whether additional factors not considered in the simulations may affect the results. In addition, these studies are based on a very specific model, the autoregressive model, where one variable serves as both the predictor and the outcome. It is thus not clear how well these results generalize to other multilevel models. In the current study, we aim to provide a more thorough discussion of these issues. First, based on statistical theory, we will provide an overview of the factors influencing EB estimates. Next, we will demonstrate our findings using simulated data and an empirical data set on emotional inertia. Throughout the article, we will use multilevel models for longitudinal data as examples. However, the issues discussed and the findings apply generally to all multilevel models, including those used to model other types of nested data, such as students nested within schools.

Empirical Bayes Estimates in Multilevel Models

In the multilevel modeling framework, longitudinal data from an independent sample of individuals can be represented by a two-level model:

$$\begin{aligned}\text{Level 1: } y_{it} &= \sum_{j=0}^{J} \beta_{ji} x_{jit} + \varepsilon_{it} \\ \text{Level 2: } \beta_{ji} &= \sum_{k=0}^{K} \gamma_{jk} w_{ki} + \theta_{ji}\end{aligned} \tag{1}$$

Level 1 is the within-person level, where y_{it} is the value of the outcome variable from individual i at time t, which is predicted by an intercept β_{0i} and J time-varying predictors x_{jit} with individual-specific coefficients $\beta_{ji}, j = 1, \ldots, J$. The residuals ε_{it} is typically assumed to be independent and normally distributed with mean zero and variance σ_{ε}^2. Level 2 is the between-person level, where the Level-1 intercept and coefficients are predicted by intercepts γ_{j0} and K time-invariant predictors w_{ki} with regression coefficients $\gamma_{jk}, k = 1, \ldots, K$. The Level-2 intercepts and regression coefficients are referred to as fixed effects because they are identical for all individuals and represent the average effects in the population. The Level-2 "residuals" θ_{ji} are referred to as random effects because they vary across individuals and represent how individual i deviates from his/her expected values based on the fixed effects. Random effects are typically assumed to follow a multivariate normal distribution with mean zero and covariance matrix Σ_θ. The parameters of the model thus consist of the fixed effects parameters γ_{jk}, the variance and covariance components in Σ_θ, and the pooled within-person residual variance σ_{ε}^2. These parameters are typically estimated using maximum likelihood or Bayesian methods (Hox, 2002).

Given the estimated parameters in the model, the random effects θ_{ji} can be predicted based on the Bayes theorem:

$$f\left(\theta_i \middle| y_i\right) = \frac{f\left(y_i \middle| \theta_i\right) f\left(\theta_i\right)}{f\left(y_i\right)} \tag{2}$$

Here, $f\left(\theta_i\right)$ is the prior distribution of the random effects, $f\left(y_i \middle| \theta_i\right)$ is the conditional distribution, that is, the distribution of the data conditional on the random effects, $f\left(y_i\right)$ is the marginal distribution of the data, and $f\left(\theta_i \middle| y_i\right)$ is the posterior distribution, that is, the distribution of the random effects conditional on the data. An estimate of θ_i can then be obtained by calculating the mean of the posterior distribution. In practice, this process utilizes parameters that are unknown and thus need to be estimated from the data, such as the fixed effects parameters and the variances and covariances of the random effects. Hence, the estimate of θ_i is an EB estimate (Candel & Winkens, 2003). As can be seen in Equation 2, the EB estimates are obtained utilizing information from both the data and the prior distribution. Because the prior distribution is usually a normal distribution with mean zero, EB estimates are biased toward zero, a phenomenon well known in the multilevel modeling literature as the "shrinkage effect" (Raudenbush & Bryk, 2002; Snijders & Bosker, 1999). In other words, the EB estimates obtained from a multilevel model will show a slightly narrower distribution compared to the intercept or slope values obtained from when one would fit a regression model per person, because they are fitted to follow a normal distribution.

Below we illustrate the "shrinkage effect" using a multilevel random intercept only model:

$$\begin{aligned}\text{Level 1: } y_{it} &= \beta_{0i} + \varepsilon_{it} \\ \text{Level 2: } \beta_{0i} &= \gamma_{00} + \theta_{0i}\end{aligned} \tag{3}$$

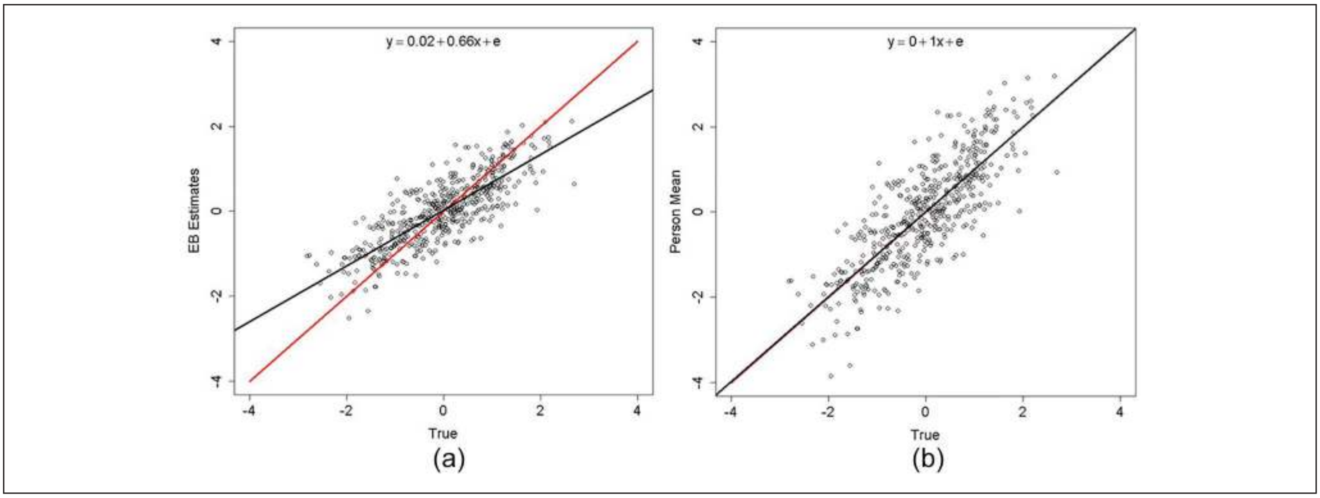

Figure 1. Scatter plots of the true random intercepts and (a) the EB estimates; (b) the person means. Red line represents *y* = *x*. Black line is the regression line of the data. Data are simulated based on a random intercept only model with $\sigma^2_{\theta_0} = 1$, $\sigma^2_\varepsilon = 3$, and $t_i = 6$ for all *i* from 1 to *N* = 500.

Because there is no predictor in this model, the EB estimates are simply weighted averages of the individual mean $\bar{y}_i$ and zero. For any individual in the sample, it can be shown that the weight given to the individual mean is

$$\lambda_i = \frac{\sigma^2_{\theta_0}}{\sigma^2_{\theta_0} + \frac{\sigma^2_\varepsilon}{t_i}} \tag{4}$$

where $\sigma^2_{\theta_0}$ is the Level-2 random effects variance, σ^2_ε is the Level-1 residual variance, and t_i is the number of observations from individual *i* (see McCulloch & Neuhaus, 2011, for derivation of this equation). λ_i is a ratio that ranges from zero to one. It is also known as the "shrinkage factor" because the EB estimates would be expected to be λ_i times the individual mean. For example, if we estimate a random intercept only model based on a longitudinal data set where all individuals are measured at 6 time points with $\sigma^2_{\theta_0} = 1$ and $\sigma^2_\varepsilon = 3$, all individuals will have a shrinkage factor of 2/3. This is shown in Figure 1a, which is a scatter plot of the true values and the EB estimates from a simulated data set with *N* = 500 individuals. In this plot, the black line represents the regression line when the EB estimates (*y*-axis) are regressed on the true values (*x*-axis). The red line represents *y* = *x*. Therefore, if the EB estimates are unbiased, the black line and red line would overlap. As shown in the figure, however, the intercept of the black line is close to zero, and the slope is close to 2/3, indicating that the EB estimates are expected to be 2/3 of the true values. On the other hand, simply taking the mean of each individual's repeated measures does not yield biased estimates, as shown in Figure 1b.

Factors Influencing the Empirical Bayes Estimates

In a more general multilevel model with predictors, EB estimates can be obtained in a similar fashion by weighting the information from the data and the information from the prior distribution. Information from the data can be represented by estimates obtained using ordinary least square (OLS) methods, that is, by fitting a regression model to each individual's data separately. Specifically, suppose we now represent the individual-level regression equation in matrix form:

$$\mathbf{Y_i} = \mathbf{X_i}\boldsymbol{\beta}_\mathbf{i} + \mathbf{E_i} \tag{5}$$

where $\mathbf{Y_i}$ is a $t_i \times 1$ vector of the outcome variable, $\mathbf{X_i}$ is a $t_i \times (J+1)$ matrix of predictors (including a column of 1 to estimate the intercept), $\boldsymbol{\beta}_\mathbf{i}$ is a $(J+1)\times 1$ vector of regression coefficients, and $\mathbf{E_i}$ is a $t_i \times 1$ vector of residuals. The OLS estimates are

$$\hat{\boldsymbol{\beta}}_\mathbf{i} = (\mathbf{X_i^T X_i})^{-1}\mathbf{X_i^T Y_i} \tag{6}$$

Importantly, in multilevel models, the weight assigned to the OLS estimates is associated with the uncertainty in estimating $\boldsymbol{\beta}_\mathbf{i}$, that is, the standard errors of $\hat{\boldsymbol{\beta}}_\mathbf{i}$. Intuitively, with lower uncertainty (i.e., a smaller standard error), more weight is assigned to the OLS estimates and less weight is assigned to the prior distribution, and vice versa. Hence, to calculate the weight, we will need to first estimate the standard error of $\hat{\boldsymbol{\beta}}_\mathbf{i}$. This can be done by taking the square root of the diagonal components of the parameter dispersion matrix:

$$\mathbf{V_i} = \sigma_{ei}^2(\mathbf{X_i^T X_i})^{-1} \tag{7}$$

where σ_{ei}^2 is the variance of $\mathbf{E_i}$ and can be estimated as

$$\hat{\sigma}_{ei}^2 = \frac{1}{t_i - J - 1}\sum_{m=1}^{t_i}(y_{mi} - \hat{y}_{mi})^2 \tag{8}$$

The $t_i - J - 1$ rather than t_i in the denominator makes $\hat{\sigma}_{ei}^2$ an unbiased estimate of σ_{ei}^2.

Next, the EB estimates of $\boldsymbol{\beta_i}$ for individual *i* in a multilevel model can be obtained with the following equation:

$$\hat{\boldsymbol{\beta}}_{\mathbf{i}}^{\mathbf{EB}} = \boldsymbol{\Lambda_i}\hat{\boldsymbol{\beta}}_{\mathbf{i}} + (\mathbf{I} - \boldsymbol{\Lambda_i})\mathbf{W_i}\hat{\boldsymbol{\gamma}} \tag{9}$$

where $\mathbf{I}$ is an identity matrix, $\mathbf{W}_i$ is a matrix containing the Level-2 predictors, $\hat{\boldsymbol{\gamma}}_{\mathbf{i}}$ is the fixed effects parameter estimates, and

$$\boldsymbol{\Lambda_i} = \boldsymbol{\Sigma_\theta}(\boldsymbol{\Sigma_\theta} + \mathbf{V_i})^{-1} \tag{10}$$

(Raudenbush & Bryk, 2002). In other words, the EB estimates are weighted averages of the OLS estimates and the fixed effects estimates, with $\boldsymbol{\Lambda_i}$ being the weighting matrix for the OLS estimates $\hat{\boldsymbol{\beta}}_{\mathbf{i}}$, and $(\mathbf{I} - \boldsymbol{\Lambda_i})$ being the weighting matrix for the fixed effects estimates. The degree to which the EB estimates are shrunk toward the fixed effects is thus determined by $\boldsymbol{\Lambda_i}$. By examining $\boldsymbol{\Lambda_i}$, we can identify factors that affect the degree of shrinkage.

According to Equation 10, more weight is assigned to the OLS estimates (and hence less shrinkage) when there is a larger Level-2 variance and a lower degree of uncertainty in the OLS estimates. By expanding Equation 7, we see that the uncertainty in the OLS estimates is in turn affected by the within-person residual variance σ_{ei}^2, the within-person variance of the predictors, and the number of observations per person. For example, consider the scenario where there is only one time-varying predictor x_{it}, and hence $\mathbf{X_i}$ is an $t_i \times 2$ matrix with 1s on the first column to represent the intercept and x_{it} on the second column. In this case,

$$\mathbf{V_i} = \sigma_{ei}^2(\mathbf{X_i^T X_i})^{-1} = \sigma_{ei}^2\begin{bmatrix} t_i & \sum_{t=1}^{t=t_i} x_{it} \\ \sum_{t=1}^{t=t_i} x_{it} & \sum_{t=1}^{t=t_i} x_{it}^2 \end{bmatrix}^{-1} \tag{11}$$

The standard error of the regression coefficient associated with x_{it} is the square root of the second diagonal component in this matrix, which is a function of σ_{ei}^2 and the within-person sum of square of x_{it}. The latter is in turn a function of the within-person variance of x_{it}, and t_i when x_{it} is centered within person (i.e., has a mean of zero):

$$\sum_{t=1}^{t=t_i} x_{it}^2 = \sum_{t=1}^{t=t_i}(x_{it} - 0)^2 = (t_i - 1)\sigma_{xi}^2 \tag{12}$$

Combining Equations 11 and 12, it can be seen that larger within-person residual variance is associated with higher uncertainty in the OLS estimates, and hence lower weights assigned to the OLS estimates. In contrast, larger within-person variance in Level-1 predictors and more observations per person are associated with lower uncertainty, and hence higher weights assigned to the OLS estimates.

To further illustrate the different factors affecting EB estimates, we simulated data based on a hypothetical daily diary study on emotional reactivity to stress, where individuals report on positive affect (PA) and stress daily for *T* days. The multilevel model below describes the relation between the two variables:

$$\begin{aligned} &\text{Level 1: } PA_{it} = \beta_{0i} + \beta_{1i}Stress_{it} + \varepsilon_{it} \\ &\text{Level 2: } \beta_{0i} = \gamma_{00} + \theta_{0i} \\ &\qquad\qquad \beta_{1i} = \gamma_{10} + \theta_{1i} \end{aligned} \tag{13}$$

In this model, an individual's emotional reactivity to stress is represented by β_{1i}, which has been shown to predict long-term health outcomes in previous research (Mroczek et al., 2015; Sin et al., 2015). Therefore, we focus on examining factors affecting the shrinkage of this parameter. Specifically, we first simulated data for $N = 500$ individuals based on a set of parameter values: $\gamma_{00} = 5$, $\gamma_{10} = 0.2$, $\sigma_\varepsilon^2 = 1$, $\sigma_{Stress}^2 = 1$, $\sigma_{\theta_0}^2 = 1$, $\sigma_{\theta_1}^2 = 1$, $\sigma_{\theta_0\theta_1} = 0$, and $T = 8$.[1] Because all individuals shared the same simulated values, they would have the same shrinkage factor. Figure 2a shows the scatter plot of the true values versus the EB estimates of stress reactivity, which can be interpreted in the same way as Figure 1. The empirical shrinkage factor, as represented by the slope of the regression line, was $\lambda = 0.83$.

Next, we simulated four other data sets doubling the size of σ_ε^2, σ_{Stress}^2, $\sigma_{\theta_1}^2$, or *T*, respectively, while keeping all other values unchanged. The corresponding scatter plots are shown in Figure 2b to e. For these four conditions, we obtained an empirical shrinkage factor of $\lambda = 0.79$, 0.92, 0.85, 0.94, respectively. These values changed in the direction we expect. More severe shrinkage (i.e., a smaller shrinkage factor) was associated with a larger σ_ε^2, whereas less severe shrinkage was associated with a larger σ_{Stress}^2, a larger variance of θ_{1i}, and a larger value of *T*.

Because the reliability of the EB estimates is simply the square of the correlation between themselves and the true values, and that correlation is the same as the standardized slope of the regression line, EB reliability is influenced by the same factors, namely, σ_ε^2, σ_{Stress}^2, $\sigma_{\theta_1}^2$, and *T*. The empirical reliability in the five simulated data sets above was 0.85, 0.77, 0.92, 0.86, and 0.94, respectively. This indicated that high reliability (≥0.80) could be achieved with large enough

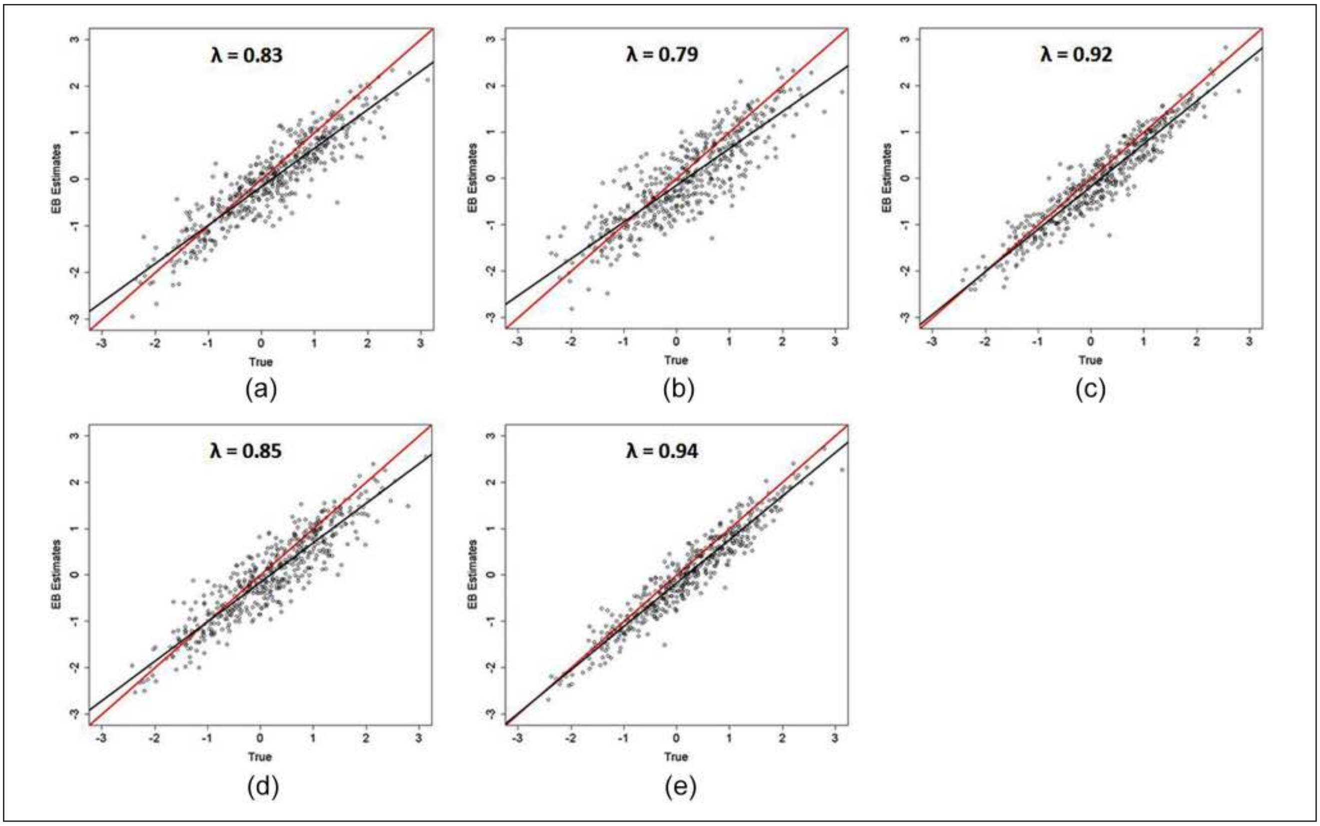

Figure 2. Scatter plots of the true values and the EB estimates of stress reactivity based on (a) the original set up; (b) doubled σ^2_ε; (c) doubled σ^2_{Stress}; (d) doubled $\sigma^2_{\theta_1}$; and (e) doubled T. Red line represents $y = x$. Black line is the regression line of the data with slope equal to λ.

Level-1 predictor variance, random effects variance, number of within-person observations, and a small enough Level-1 residual variance. However, whether or not a specific data set would yield high enough reliability in the EB estimates can be difficult to determine, because in reality, some of these values are likely to vary across individuals. For instance, some individuals experience large day-to-day variability in their stress levels while others hardly fluctuate, leading to different levels of σ^2_{Stress}. Similarly, even if researchers aim to have a balanced data set, participants in the study often have vastly different amount of missing values, leading to different levels of T. As a result, not all individuals in the sample have the same shrinkage factor. The EB estimates for those with more missing data and/or little variation in the predictors would be pulled toward the fixed effects estimates more strongly. This introduces additional noise to the rank ordering of EB estimates.

Nevertheless, because EB estimates are closely related to the OLS estimates, their reliability can be approximated by reliability of the OLS estimates, which is easy to obtain. In a recent simulation study, Neubauer, Voelkle, Voss, and Mertens (2019) showed that the two reliability coefficients are highly correlated ($r > 0.95$) and similar in size, especially when they are large. Therefore, one may calculate the OLS reliability to approximate the EB reliability. The OLS reliability is

$$reliability(\hat{\beta}_q) = \frac{1}{N}\sum_{i=1}^{N} \sigma_{qq} / (\sigma_{qq} + v_{qqi}) \qquad (14)$$

where σ_{qq} and ν_{qqi} are the qth diagonal elements of $\mathbf{\Sigma_\theta}$ and $\mathbf{V_i}$, respectively (Raudenbush & Bryk, 2002). For example, the OLS reliability for the stress reactivity coefficients in the first simulated data set was

$$\begin{aligned} reliability(\hat{\beta}_1) &= \frac{1}{N}\sum_{i=1}^{N} \frac{\sigma_{22}}{\sigma_{22} + \dfrac{\sigma^2_{\varepsilon i}}{\sigma^2_{xi} * (t_i - 1)}} \\ &= \frac{1}{500}\sum_{i=1}^{500} \frac{1}{1 + \dfrac{1}{1 * (8-1)}} = 0.875 \end{aligned} \qquad (15)$$

The empirical EB reliability, 0.85, was very similar to this value. For the other four simulated data sets, the OLS reliability was 0.78, 0.93, 0.93, and 0.94, respectively. These values, again, were similar to the empirical EB reliability obtained earlier.

To summarize, EB estimates are not perfect measures of individual traits. Their accuracy and reliability depend on a variety of factors, including Level-1 residual variance, Level-1 within-person predictor variance, Level-2 random effects variance, and number of within-person observations. Specifically, accuracy and reliability increases with a larger Level-1 predictor variance, Level-2 random effects variance, and/or more observations per person, and decreases if the Level-1 residuals variance becomes large. In addition, some of these characteristics may vary across individuals, resulting in different levels of shrinkage and introducing additional noise to the rank ordering of individuals. For instance, the EB estimate of an individual with an extremely high OLS value (i.e., an outlier) may be shrunk strongly toward the fixed effects because there are many missing data from this person. As a result, EB estimates may not be ideal for identifying outliers. Moreover, it is well known that measurement errors in a predictor variable will lead to a downward bias in the regression coefficient in regression models (Hutcheon, Chiolero, & Hanley, 2010). Hence, low reliability in the EB estimates may be a concern when they are used as predictors in follow-up analyses.

To further examine these issues, in particular issues associated with using EB estimates as predictors, below we present a simulation study in which we compared this two-step approach (i.e., *first* extract the EB estimates, *then* predict another person-level variable) to an alternative one-step approach (i.e., multilevel structural equation modeling), where the random effects are treated as latent variables in a general latent variable modeling framework (B. O. Muthen, 2002). Specifically, we aim to use simulated data to examine (1) whether OLS reliability calculated based on Equation 14 is a good approximation of EB reliability; (2) bias in the two-step approach when EB reliability is low; and (3) whether the one-step approach produces unbiased results. We expect this to be the case because the one-step approach takes into account reliability in the random effects when evaluating its relation with a distal outcome.

Method

Previous studies on EB reliability (Du & Wang, 2018; Liu, 2018) have mostly focused on multilevel autoregressive models. We chose here to simulate data based on the stress reactivity model represented in Equation 13. This model is more general than autoregressive models in that the predictor and outcome variables are different. Hence, results in our simulation study can more readily be generalized to other settings involving multilevel models. In the literature, the study by Mroczek et al. (2015) reported all four parameters necessary for estimating OLS reliability. In that study, 181 older adults reported negative affect, positive affect, and stress for 8 consecutive days. Greater reactivity in positive affect (but not negative affect) to stress was found to increase mortality risk. The authors reported a Level-1 residual variance of 18.60 and a Level-2 random effects variance of 1.32. Although the Level-1 predictor variance was not reported, the predictor (stress) was a dichotomous variable (yes or no) whose variance was bound between 0 and 0.25. Therefore, we simulated data following a factorial design that involved the following four factors which were fixed across individuals within each replication: (1) Level-1 residual variance σ_{ε}^{2} = 18.60 or 37.20; (2) Level-1 predictor variance σ_{Stress}^{2} = .25 or .50; (3) Level-2 random effects variance $\sigma_{\theta_1}^{2}$ = 1.32 or 2.64; and (4) Number of within-person observations T = 8, 40, or 80. To focus on the random slope, the intercept was fixed to zero for all individuals. A distal outcome variable was generated to have a correlation of .50 with θ_1. We replicated each condition 1,000 times and sample size was set to 181 across all conditions.

Each data set was analyzed using both the two-step approach (where EB estimates are obtained) and the one-step approach (multilevel structural equation modeling). The two-step analyses were conducted in R with restricted maximum likelihood estimation using the *nlme* package (Pinheiro, Bates, DebRoy, Sarkar, & R Core Team, 2016). The one-step approach was carried out in R with Bayesian estimation method using the *MplusAutomation* package (Hallquist & Wiley, 2018) which utilizes Mplus 8.1 (L. K. Muthen & Muthen, 1998-2017). The Bayesian method is the only estimation method in Mplus that produces a standardized coefficient estimate for predicting a distal outcome variable.

Results

Comparing Estimated Reliability to Empirical Reliability

With the two-step approach, there were convergence problems in a small number of replications (0% when T = 8; < 0.1% when T = 40; 0.71% when T = 80). For models that successfully converged, we calculated empirical EB reliability by taking the square of the correlation between the EB estimates and the true values. We compared this to the reliability calculated using Equation 14 based on estimates of σ_{ε}^{2} and $\sigma_{\theta_1}^{2}$ from the multilevel models (hereinafter referred to as *estimated reliability*).[2] As shown in Table 1, our simulation yielded a wide range of empirical reliability values (*Mean* = .18 when T = 8; *Mean* = .49 when T = 40; *Mean* = .65 when T = 80). Within each simulation condition, the average empirical reliability tended to be slightly larger than the average estimated reliability. Overall, however, the two indices were highly correlated (r = .97) and similar in size, especially when the values were high (Figure 3). This pattern was consistent with

Table 1. Means and Standard Deviations (in Parentheses) of Estimated Reliability, Empirical Reliability, and Standardized Regression Coefficients From the Two-Step and One-Step Approaches Across 1,000 Replications Under Various Simulation Conditions.

			Estimated reliability	Empirical reliability	Two-step coefficient	One-step coefficient
T = 8						
$\sigma^2_\varepsilon = 18.60$	$\sigma^2_{Stress} = .25$	$\sigma^2_{\theta_i} = 1.32$	.10 (.07)	.12 (.04)	.17 (.07)	.52 (.19)
		$\sigma^2_{\theta_i} = 2.64$	.20 (.08)	.22 (.05)	.23 (.07)	.55 (.16)
	$\sigma^2_{Stress} = .50$	$\sigma^2_{\theta_i} = 1.32$	.19 (.08)	.21 (.05)	.23 (.07)	.55 (.17)
		$\sigma^2_{\theta_i} = 2.64$	.33 (.07)	.35 (.06)	.29 (.07)	.52 (.12)
$\sigma^2_\varepsilon = 37.20$	$\sigma^2_{Stress} = .25$	$\sigma^2_{\theta_i} = 1.32$	.07 (.06)	.07 (.04)	.12 (.07)	.43 (.25)
		$\sigma^2_{\theta_i} = 2.64$	.11 (.08)	.12 (.05)	.17 (.07)	.53 (.19)
	$\sigma^2_{Stress} = .50$	$\sigma^2_{\theta_i} = 1.32$	.11 (.07)	.12 (.05)	.17 (.07)	.52 (.19)
		$\sigma^2_{\theta_i} = 2.64$	.19 (.08)	.21 (.05)	.22 (.07)	.54 (.16)
T = 40						
$\sigma^2_\varepsilon = 18.60$	$\sigma^2_{Stress} = .25$	$\sigma^2_{\theta_i} = 1.32$	.40 (.07)	.40 (.06)	.32 (.07)	.51 (.11)
		$\sigma^2_{\theta_i} = 2.64$	.58 (.04)	.58 (.05)	.38 (.06)	.50 (.08)
	$\sigma^2_{Stress} = .50$	$\sigma^2_{\theta_i} = 1.32$	.57 (.04)	.58 (.05)	.38 (.06)	.51 (.08)
		$\sigma^2_{\theta_i} = 2.64$	.73 (.03)	.73 (.04)	.43 (.06)	.50 (.07)
$\sigma^2_\varepsilon = 37.20$	$\sigma^2_{Stress} = .25$	$\sigma^2_{\theta_i} = 1.32$	.25 (.08)	.26 (.06)	.25 (.07)	.53 (.16)
		$\sigma^2_{\theta_i} = 2.64$	.40 (.06)	.41 (.05)	.32 (.06)	.51 (.10)
	$\sigma^2_{Stress} = .50$	$\sigma^2_{\theta_i} = 1.32$	.40 (.06)	.41 (.06)	.32 (.06)	.51 (.11)
		$\sigma^2_{\theta_i} = 2.64$	.57 (.05)	.58 (.05)	.38 (.06)	.50 (.08)
T = 80						
$\sigma^2_\varepsilon = 18.60$	$\sigma^2_{Stress} = .25$	$\sigma^2_{\theta_i} = 1.32$	.58 (.05)	.58 (.05)	.38 (.06)	.50 (.08)
		$\sigma^2_{\theta_i} = 2.64$	.74 (.03)	.74 (.03)	.43 (.06)	.50 (.07)
	$\sigma^2_{Stress} = .50$	$\sigma^2_{\theta_i} = 1.32$	.73 (.03)	.73 (.03)	.43 (.06)	.50 (.07)
		$\sigma^2_{\theta_i} = 2.64$	.85 (.02)	.85 (.02)	.46 (.05)	.50 (.06)
$\sigma^2_\varepsilon = 37.20$	$\sigma^2_{Stress} = .25$	$\sigma^2_{\theta_i} = 1.32$	.40 (.06)	.41 (.06)	.32 (.06)	.50 (.10)
		$\sigma^2_{\theta_i} = 2.64$	.58 (.05)	.58 (.05)	.38 (.06)	.50 (.08)
	$\sigma^2_{Stress} = .50$	$\sigma^2_{\theta_i} = 1.32$	.58 (.04)	.58 (.05)	.38 (.06)	.50 (.08)
		$\sigma^2_{\theta_i} = 2.64$	.73 (.03)	.73 (.03)	.43 (.05)	.49 (.07)

Note. Shaded cells represent conditions where relative bias is larger than 10%.

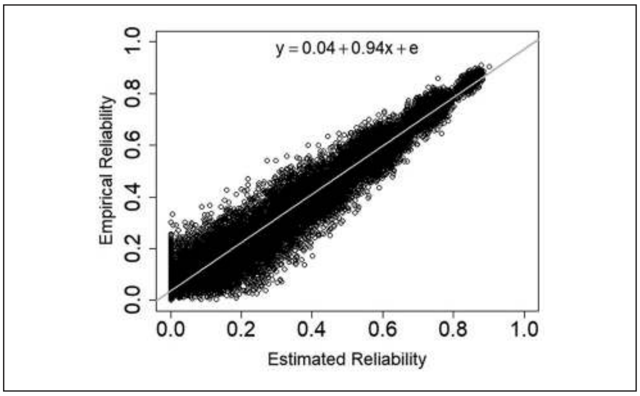

Figure 3. Scatter plot of estimated and empirical reliability of EB estimates across all simulation conditions.

findings in Neubauer et al. (2019). As pointed out in Neubauer et al. (2019), the larger inconsistency at the lower end of the spectrum is less a problem because researchers are typically interested in evaluating whether or not there is sufficient reliability (≥.80), rather than the exact reliability coefficient. For instance, in practice it does not matter whether reliability is .20 or .40, because both scenarios suggest reliability is too low for the measure to be used. Hence, consistency at the higher end of the spectrum is more important.

Bias in Regression Coefficients in the Two-Step and One-Step Approaches

Table 1 also shows the means and standard deviations of the standardized regression coefficient from the two-step and one-step approaches under various simulation conditions. Shaded cells represent conditions where relative bias (i.e., proportion of bias to the true value) is larger than 10%, which is often considered the threshold of acceptable bias. Consistent with our conjecture, the one-step approach generally yielded estimates that were very close to the true value (.50), whereas the two-step approach yielded estimates that were negatively biased. Not surprisingly, the

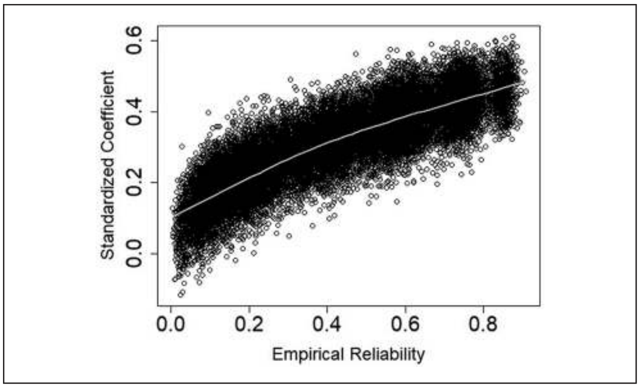

Figure 4. Scatter plot of empirical reliability and standardized regression coefficient (True value = .50) in the two-step approach. Grey line is a smooth curve computed by the *loess* method.

amount of bias was related to empirical reliability, with lower reliability yielding more bias. This can be seen in Figure 4, where the grey line represents a smooth curve fitted to the scatter plot of the two using the loess method. Ignoring the nonlinearity in the relation, we obtained a correlation of .85 between the two.

We also examined how often the 95% confidence intervals were able to cover the true value using the two approaches. Averaging across simulation conditions, the coverage rate of the one-step approach was 95% (Range = [92%, 97%] across conditions; not shown in Table 1). In contrast, the two-step approach yielded an average coverage rate of 2% when $T = 8$, 35% when $T = 40$, and 49% when $T = 80$ (Range = [0%, 89%] across conditions; not shown in Table 1). Clearly, the one-step approach is preferable to the two-step approach when a distal outcome variable is involved, especially when reliability is low.

In sum, results from our simulation study with simulation parameters similar to those obtained from Mroczek et al. (2015), suggest that (1) OLS reliability calculated based on Equation 14 is indeed a good approximation of EB reliability, especially when reliability is high; (2) the two-step approach yields biased estimates and the degree of bias is negatively associated with EB reliability; and (3) the one-step approach produces unbiased regression coefficients.

Empirical Example

In the following, we will use an empirical example on emotional inertia to illustrate issues with using EB estimates as measures of individual traits. The data we use has been published elsewhere (Bringmann et al., 2013; Bringmann et al., 2016; Koval et al., 2012; Pe, Koval, & Kuppens, 2013; Pe, Raes, & Kuppens, 2013). Ninety-five undergraduate students from KU Leuven in Belgium (age: *Mean* = 19 years, *SD* = 1; 62% female) participated in an experience sampling methods (ESM) study. Over the course of 7 days, participants carried a palmtop computer on which they were asked to fill out questions about mood and social context in their daily lives 10 times a day. Participants were beeped to fill out the ESM questionnaires at random times within 90-minute windows. They were asked to rate, among other things, their current feelings of negative and positive emotions on a continuous slider scale, ranging from 1 (*not at all, e.g., angry*) to 100 (*very, e.g., angry*). For the current analyses, we focus on one emotion variable, *sad*. On average, the response rate is 78% (*SD* = 7%) on this item. However, there is substantial variability in response rate across individuals. Figure 5 shows how the number of within-person observations is distributed in the sample. Whereas many participants responded to more than 50 beeps, some provided fewer than 40 observations. This type of skewed distribution is typical in ESM studies.

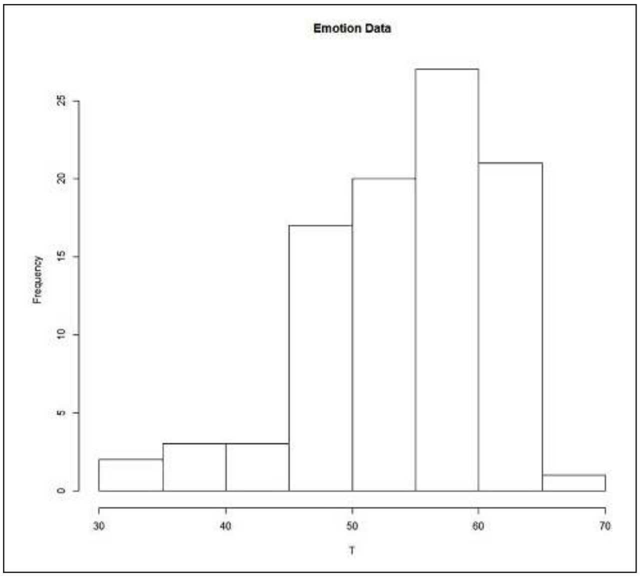

Figure 5. Distribution of number of within-person observations on the variable *sad*.

Following the literature (Hamaker & Grasman, 2015; Kuppens et al., 2012), we estimate a multilevel first-order autoregressive (AR[1]) model to investigate emotional inertia in sadness:

$$\begin{aligned} &\text{Level 1: } Sad_{it} = \beta_{0i} + \beta_{1i} Sad_{i(t-1)} + \varepsilon_{it} \\ &\text{Level 2: } \beta_{0i} \quad = \gamma_{00} + \theta_{0i} \\ &\qquad\qquad\ \beta_{1i} = \gamma_{10} + \theta_{1i} \end{aligned} \tag{16}$$

In this model, Sad_{it} is the score of sadness for individual i at time t, $Sad_{i(t-1)}$ is the same variable measured at the previous time point, centered within person. β_{0i} is the intercept that represents the average level of sadness for

Table 2. Results of the Autoregressive Model Using Multilevel Modeling and Ordinary Least Square (OLS) Methods.

	Fixed effects		Random effects			
	Intercept	AR(1)	Intercept variance	AR(1) variance	Intercept-AR(1) correlation	Residuals variance
Multilevel	18.05* (1.30)	0.26* (0.02)	154.50	0.03	0.60	224.10
OLS	18.05* (1.30)	0.25* (0.02)	158.26	0.05	0.51	185.78

*$p < .05$.
Note. Standard errors are in parentheses.

individual i at average levels of sadness at the previous time point, and β_{1i} is the AR(1) coefficient that typically ranges from −1 to 1. A large, positive β_{1i} (e.g., 0.6) suggests that sadness at a previous time point strongly predicts sadness at the current time point, which indicates a high level of emotional inertia. Previous research has showed that this individual trait is associated with personality and mental health measures such as neuroticism and depression (Kuppens et al., 2012; Suls, Green, & Hillis, 1998).

We will evaluate two different usages of the EB estimates. One is to detect individuals with exceptionally strong emotional inertia as this may indicate potential mental health problems. Since the true values of emotional inertia are unknown, we will compare the EB estimates to the OLS estimates. In an AR(1) model, the latter has a downward but negligible bias with at least 20 observations (Hamaker & Grasman, 2015; Liu, 2017). Therefore, it provides a good representation of the true values assuming data are missing at random. Next, we will compare the two-step approach and one-step approach when emotional inertia is used to predict neuroticism. Neuroticism was assessed during the introductory session before ESM with the Dutch version of the Ten Item Personality Inventory (Gosling, Rentfrow, & Swann, 2003; Hofmans, Kuppens, & Allik, 2008), resulting in a score ranging from 1 to 7 ($M = 3.4$; $SD = 1.5$). The code for the one-step approach is included in the appendix.

Estimates of Fixed Effects and Variances/ Covariances of Random Effects

Table 2 shows estimates of the fixed effects and variance components associated with the random effects from the multilevel and OLS AR(1) models. Although these statistics are not the focus of the current study, we present them here because they are typically of interest to empirical researchers. For the OLS approach, the fixed effects are computed by averaging the individual OLS estimates, and their standard errors are computed by dividing the standard deviations of the OLS estimates by the square root of N. Furthermore, the residuals variance represents the average residuals variance across all individuals.

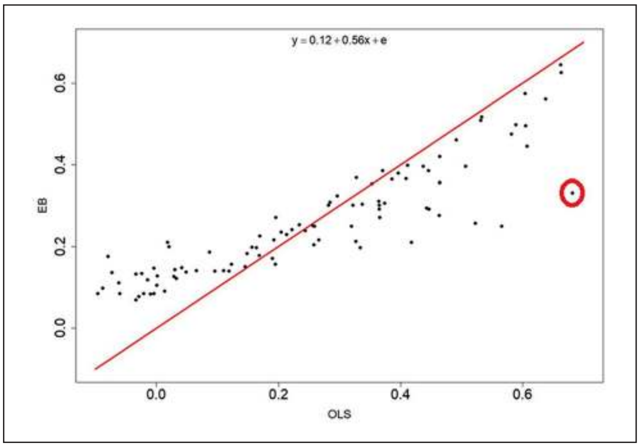

Figure 6. Scatter plot of OLS and EB estimates of the AR(1) parameters. Red line represents $y = x$.

Comparing the two approaches, the fixed effects estimates and their standard errors are almost identical. The average intercept is 18.05, and the average AR(1) coefficient is about 0.26. Hence, the level of sadness is generally low in this sample and emotional inertia is weak, but significantly different from zero. The random effects variances are also similar, but they are smaller for the multilevel modeling approach due to shrinkage. The correlation between the intercepts and AR(1) coefficients is positive and strong in both approaches, indicating that individuals with higher levels of sadness tend to show stronger emotional inertia. Finally, the residual variance is similar but slightly smaller in the OLS approach.

Identifying Individuals With High Emotional Inertia

Figure 6 shows a scatter plot of the EB estimates (y-axis) and the OLS estimates (x-axis) of the AR(1) parameters. As expected, the EB estimates have a narrower spread than the OLS estimates due to the shrinkage effect. Although the correlation between the two types of estimates is very high ($r = 0.90$), there is also substantial discrepancy, especially at the two ends of the spectrum. A particularly problematic case is circled in red. Specifically, this individual has the highest OLS estimate of the sample, with $\breve{\beta}_{1i} = 0.68$. The EB estimate, however, is only 0.33, which is at the 73rd percentile of the sample. In other words, the EB estimate for this individual is severely shrunk toward the sample mean. This is likely due to missing data, as this individual only provided 38 observations, a number much lower than the sample average. As a result, even though this individual appears to have the strongest emotional inertia, he/she could not be identified using the EB estimates.

Using Emotional Inertia to Predict Neuroticism

The estimated reliability of the AR(1) coefficients is 0.56, which is not acceptable. In practice, this suggests that the two-step approach should *not* be used. However, we conduct both the two-step and one-step approaches here for the purpose of comparison. Results show that both approaches produce standardized regression coefficients that are positive and significant. Specifically, using the EB estimates to predict neuroticism, we obtain a standardized regression coefficient of 0.44 ($t = 4.70, p < .001$). In comparison, the one-step approach yields a standardized coefficient of 0.53 ($t = 5.25, p < .001$). Therefore, results from both approaches suggest that higher emotional inertia is associated with higher levels of neuroticism, which is consistent with the literature. However, the two-step approach produces a slightly smaller standardized coefficient, potentially due to low reliability of the EB estimates.

Discussion

In this article, we examine whether the EB estimates from multilevel models are reliable and valid measures of individual traits. Based on statistical theory and simulated data, we show that the accuracy and reliability of EB estimates depend on a number of factors, including Level-1 residual variance, Level-1 within-person predictor variance, Level-2 random effects variance, and number of within-person observations.

In general, we find that EB estimates are not appropriate for detecting outliers because they are known to shrink toward the sample mean and the degree of shrinkage may vary across individuals. This variability in degree of shrinkage may be due to different factors, such as variability in the Level-1 predictor variance and number of within-person observations (or reversely, number of missing value) across individuals. For instance, if we assume missing at random, we could see in our empirical example that using the EB estimates results in failure in identifying at least one individual with exceptionally high emotional inertia due to a large amount of missing data from that individual. On the other hand, OLS estimates provide less biased results and are more appropriate. However, the OLS estimates are obtained by fitting person-specific regression models to each individual's data. Therefore, researchers should be cautious of using them if the number of within-person observations is small. In the person-specific modeling literature, a minimum of 50 is generally recommended (Liu, 2017). It is also important to note that data could be missing not at random. For example, in a daily diary study on stress reactivity, a participant may only complete surveys on days when he/she is most reactive to stress. In this case, outliers identified by OLS methods may be "false alarms" whereas EB estimates are somewhat protected because missing data are compensated by group-level information. Therefore, missing data mechanisms should also be considered in outlier detection.

Because EB estimates are not perfect measures of the true individual traits, they tend to produce biased regression coefficients when used as predictors. In general, this makes the two-step approach that involves the extraction of EB estimates less preferable than the one-step approach where they are treated as latent scores. The one-step approach, however, may be difficult to carry out in empirical research because of model convergence problems (e.g., Mroczek et al., 2015). If researchers are restricted to use the two-step approach, we recommend them to first evaluate reliability of the EB estimates. This could be done by calculating OLS reliability using Equation 14, which provides a good approximation especially when reliability is high. A set of R code for doing this is available in Neubauer et al. (2019). Alternatively, a third approach may be used if researchers simply aim to evaluate whether the random effects are associated with a covariate, without strong theoretical reasons to denote the random effects as predictor. This third approach involves using the covariate to predict the random effects, which can be easily implemented in the multilevel modeling framework. For instance, if researchers simply want to examine whether emotional inertia is related to neuroticism, neuroticism can be added as a Level-2 predictor of the AR(1) random effects (e.g., Kuppens, Allen, & Sheeber, 2010). Statistically, this approach would yield similar results as the one-step approach if there is no other variable in the model.

We now mention the limitations of the current study and provide notes of caution in interpreting the results. First, the finding that EB estimates shrink toward zero does *not* indicate that all Bayesian estimates tend to be biased. Rather, the shrinkage effect shown here is a result of having prior distributions with zero means, which is a logical choice given that random effects by definition are deviations from the fixed effects and hence always have means of zero. Non-Bayesian methods for estimating random effects in multilevel models, such as those described by Henderson (1984, 1990), would produce equivalent results and lead to the same shrinkage effects (Robinson, 1991). From a machine learning perspective, such shrinkage effects exist for a reason—trading non-biasedness for less variance makes the estimator less sensitive to outliers and would lead to better predictions in the long run (Yarkoni & Westfall, 2017).

Second, this study is based on the assumption that the variables used in the multilevel model are without measurement errors, which is unlikely to be true in reality. Du and Wang (2018) examined the influence of measurement scale reliability and found that measurement error in an autoregressive process could significantly hamper the reliability of the EB estimates of AR(1) parameters. For example, when measurement scale reliability was .50, the EB reliability was generally lower than .20 across the simulation conditions they examined. However, with a perfect measurement scale the EB reliability was higher than .80 with at least 100 observations. Hence, when evaluating EB reliability, measurement scale reliability should be taken into account. Importantly, a new method has recently been developed that includes measurement error in multilevel autoregressive modeling (Schuurman & Hamaker, 2019). This method, termed *multilevel measurement error vector autoregressive model* (MEVAR), performs particularly well when the autoregressive effects are large, and hence could be considered when strong inertia is expected.

Third, in this study we examine shrinkage based on the standard assumptions of multilevel models, which may not be true. For example, we assume that the random effects follow a (multivariate) normal distribution. In practice, this assumption can be relaxed, although it is rarely done. McCulloch and Neuhaus (2011) found that specifying different prior distributions for the random effects would change the distributions of the EB estimates, but their average accuracy, as measured by mean square error, is little affected. In our simulation and empirical example, we also assume that the Level-1 residual variance is constant across individuals. This again can be relaxed. Jongerling, Laurenceau, and Hamaker (2015) found that failing to allow for individual differences in the Level-1 residual variance would lead to biased estimates in the multilevel AR(1) model. However, it is unknown how these factors (i.e., assuming different prior distributions, allowing individual differences in Level-1 residual variance) may influence reliability of the EB estimates, which will need to be examined in future research.

Finally, the validity of using EB estimates also relies on the correct specification of the multilevel model from which they are obtained. In studies involving intensive longitudinal data (e.g., ESM data), it is often challenging to specify one single model to sufficiently describe the data from all individuals due to the large amount of between-person heterogeneity in their dynamic processes (e.g., Wright, Beltz, Gates, Molenaar, & Simms, 2015). In this case, multilevel modeling and the corresponding EB estimates may no longer be appropriate. For instance, Liu (2017) found that when individuals in the sample are characterized by heterogeneous dynamic processes, multilevel models produce biased results at the population level and larger prediction errors at the individual level than OLS methods when $T \geq 50$. Hence, individual-level processes may be better captured by person-specific analysis or clustering methods specifically developed to model heterogeneous processes (e.g., Lane, Gates, Pike, Beltz, & Wright, 2019).

To summarize, this study demonstrates that EB estimates are not ideal measures of individual traits because they are biased toward zero and do not always reliably represent individual differences. Given sufficient within-person measurements, we recommend researchers to at least consider alternative approaches such as OLS methods for detecting outliers. For predicting a person-level outcome variable, we generally recommend the one-step approach, which could be conveniently carried out in Mplus (L. K. Muthen & Muthen, 1998-2017). If the one-step approach leads to convergence problems and the two-step approach has to be used, researchers should evaluate EB reliability to determine the potential bias in the target regression coefficient. With large enough Level-1 predictor variance, random effects variance, number of within-person observations, and a small enough Level-1 residual variance, sufficient EB reliability may be achieved.

Appendix

Mplus Code for the One-Step Approach

```
TITLE:          Random slope predicting neuro
DATA: FILE = Data95_mplus.dat;
VARIABLE: NAMES = ID sad sad1 neuro;
  WITHIN = sad1;
     BETWEEN = neuro;
     CLUSTER = ID;
  MISSING = ALL (9999 9998);
DEFINE:   CENTER sad1 (GROUPMEAN);
ANALYSIS:   TYPE = TWOLEVEL RANDOM;
       ESTIMATOR = BAYES
MODEL:
       %WITHIN%
       ar | sad ON sad1;
       %BETWEEN%
       sad WITH ar;
       neuro ON ar;
OUTPUT:   STANDARDIZED;
```

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Siwei Liu https://orcid.org/0000-0002-2972-426X

Notes

1. These values are selected for convenience of calculation. In the more comprehensive simulation study below, we use more realistic values that are obtained from real data.
2. We also estimated reliability based on results from OLS regressions. We found that this method tended to yield positively biased values due to a positively biased individual slope variance and overfitting (i.e., the estimated residual variance tended to be smaller than the true residual variance). Therefore, we decided to use estimates from the multilevel models. This approach was also adopted in Neubauer et al. (2019).

References

Asendorpf, J. B. (2006). Typeness of personality profiles: A continuous person-centred approach to personality data. *European Journal of Personality*, *20*, 83-106. doi:10.1002/per.575

Bai, S., & Repetti, R. L. (2018). Negative and positive emotion responses to daily school problems: Links to internalizing and externalizing symptoms. *Journal of Abnormal Child Psychology*, *46*, 423-435. doi:10.1007/s10802-017-0311-8

Bringmann, L. F., Pe, M. L., Vissers, N., Ceulemans, E., Borsboom, D., Vanpaemel, W., . . .Kuppens, P. (2016). Assessing temporal emotion dynamics using networks. *Assessment*, *23*, 425-435. doi:10.1177/1073191116645909

Bringmann, L. F., Vissers, N., Wichers, M., Geschwind, N., Kuppens, P., Peeters, F., . . .Tuerlinckx, F. (2013). A network approach to psychopathology: New insights into clinical longitudinal data. *PLoS One*, *8*, e60188. doi:10.1371/journal.pone.0060188

Brose, A., Schmiedek, F., Koval, P., & Kuppens, P. (2015). Emotional inertia contributes to depressive symptoms beyond perseverative thinking. *Cognition and Emotion*, *29*, 527-538. doi:10.1080/02699931.2014.916252

Candel, M. J. J. M., & Winkens, B. (2003). Performance of empirical Bayes estimators of level-2 random parameters in multilevel analysis: A Monte Carlo study for longitudinal designs. *Journal of Educational and Behavioral Statistics*, *28*, 169-194.

Cohen, L. H., Gunthert, K. C., Butler, A. C., O'Neill, S. C., & Tolpin, L. H. (2005). Daily affective reactivity as a prospective predictor of depressive symptoms. *Journal of Personality*, *73*, 1687-1714. doi:doi:10.1111/j.0022-3506.2005.00363.x

Cummings, K. D., Stoolmiller, M. L., Baker, S. K., Fien, H., & Kame'enui, E. J. (2015). Using school-level student achievement to engage in formative evaluation: Comparative school-level rates of oral reading fluency growth conditioned by initial skill for second grade students. *Reading and Writing*, *28*, 105-130. doi:10.1007/s11145-014-9512-5

Du, H., & Wang, L. (2018). Reliabilities of intraindividual variability indicators with autocorrelated longitudinal data: Implications for longitudinal study designs. *Multivariate Behavioral Research*, *53*, 502-520. doi:10.1080/00273171.2018.1457939

Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, *37*, 504-528. doi:10.1016/S0092-6566(03)00046-1

Hallquist, M. N., & Wiley, J. F. (2018). MplusAutomation: An R package for facilitating large-scale latent variable analyses in Mplus. *Structural Equation Modeling*, *25*, 621-638. doi:10.1080/10705511.2017.1402334

Hamaker, E. L., & Grasman, R. P. P. P. (2015). To center or not to center? Investigating inertia with a multilevel autoregressive model. *Frontiers in Psychology*, *5*, 1492. doi:10.3389/fpsyg.2014.01492

Henderson, C. R. (1984). *Applications of linear models in animal breeding*. Ontario, Canada: University of Guelph.

Henderson, C. R. (1990). Statistical methods in animal improvement: Historical overview. In D. Gianola & K. Hammond (Eds.), *Statistical methods for genetic improvement of livestock* (pp. 2-14). Berlin, Germany: Springer-Verlag.

Hofmans, J., Kuppens, P., & Allik, J. (2008). Is short in length short in content? An examination of the domain representation of the ten item Personality Inventory scales in Dutch language. *Personality and Individual Differences*, *45*, 750-755. doi:10.1016/j.paid.2008.08.004

Hox, J. (2002). *Multilevel analysis: Techniques and applications*. Mahwah, NJ: Erlbaum.

Hutcheon, J. A., Chiolero, A., & Hanley, J. A. (2010). Random measurement error and regression dilution bias. *British Medical Journal*, *340*, c2289. doi:10.1136/bmj.c2289

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*, 334-349. doi:10.1080/00273171.2014.1003772

Koval, P., Kuppens, P., Allen, N. B., & Sheeber, L. (2012). Getting stuck in depression: The roles of rumination and emotional inertia. *Cognition and Emotion*, *26*, 1412-1427. doi:10.1080/02699931.2012.667392

Kuppens, P., Allen, N. B., & Sheeber, L. B. (2010). Emotional inertia and psychological maladjustment. *Psychological Science*, *21*, 984-991. doi:10.1177/0956797610372634

Kuppens, P., Sheeber, L. B., Yap, M. B. H., Whittle, S., Simmons, J. G., & Allen, N. B. (2012). Emotional inertia prospectively predicts the onset of depressive disorder in adolescence. *Emotion*, *12*, 283-289. doi:10.1037/a0025046

Lane, S. T., Gates, K. M., Pike, H. K., Beltz, A. M., & Wright, A. G. C. (2019). Uncovering general, shared, and unique temporal patterns in ambulatory assessment data. *Psychological Methods*, *24*(1), 54. doi:10.1037/met0000192

Liu, S. (2017). Person-specific versus multilevel autoregressive models: Accuracy in parameter estimates at the population and individual levels. *British Journal of Mathematical and Statistical Psychology*, *70*, 480-498. doi:10.1111/bmsp.12096

Liu, S. (2018). Accuracy and reliability of autoregressive parameter estimates: A comparison between person-specific and multilevel modeling approaches. In M. Wiberg, S. Culpepper, R. Janssen, J. Gonzalez & D. Molenaar (Eds.), *Quantitative psychology* (pp. 385-394). Cham, Switzerland: Springer.

McCulloch, C. E., & Neuhaus, J. M. (2011). Prediction of random effects in linear and generalized linear models under model misspecification. *Biometrics*, *67*, 270-279. doi:10.1111/j.1541-0420.2010.01435.x

Mohr, C. D., Brannan, D., Wendt, S., Jacobs, L., Wright, R., & Wang, M. (2013). Daily mood-drinking slopes as predictors: A new take on drinking motives and related outcomes. *Psychology of Addictive Behaviors. Journal of the Society of Psychologists in Addictive Behaviors*, *27*, 944-955. doi:10.1037/a0032633

Morrell, C. H., & Brant, L. J. (1991). Modelling hearing thresholds in the elderly. *Statistics in Medicine*, *10*, 1453-1464. doi:10.1002/sim.4780100912

Mroczek, D. K., Stawski, R. S., Turiano, N. A., Chan, W., Almeida, D. M., Neupert, S. D., & Spiro, I. I. I. A. (2015). Emotional reactivity and mortality: Longitudinal findings from the VA normative aging study. *Journals of Gerontology: Series B*, *70*, 398-406. doi:10.1093/geronb/gbt107

Muthen, B. O. (2002). Beyond SEM: General latent variable modeling *Behaviormetrika*, *29*(1), 81-117. doi:10.2333/bhmk.29.81

Muthen, L. K., & Muthen, B. O. (1998-2017). *Mplus user's guide* (8th ed.). Los Angeles, CA: Author.

Neubauer, A., Voelkle, M. C., Voss, A., & Mertens, U. K. (2019). Estimating reliability of within-person couplings in a multilevel framework. *Journal of Personality Assessment*. Advance online publication. doi:10.1080/00223891.2018.1521418

Ong, A. D., Exner-Cortens, D., Riffin, C., Steptoe, A., Zautra, A., & Almeida, D. M. (2013). Linking stable and dynamic features of positive affect to sleep. *Annals of Behavioral Medicine*, *46*(1), 52-61. doi:10.1007/s12160-013-9484-8

Pe, M. L., Koval, P., & Kuppens, P. (2013). Executive well-being: Updating of positive stimuli in working memory is associated with subjective well-being. *Cognition*, *126*, 335-340. doi:10.1016/j.cognition.2012.10.002

Pe, M. L., Raes, F., & Kuppens, P. (2013). The cognitive building blocks of emotion regulation: Ability to update working memory moderates the efficacy of rumination and reappraisal on emotion. *PLoS One*, *8*, e69071. doi:10.1371/journal.pone.0069071

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Core Team. (2016). nlme: Linear and nonlinear mixed effects models (V. 3.1-126). Retrieved from http://CRAN.R-project.org/package=nlme

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

Robinson, G. K. (1991). That BLUP is a good thing: The estimation of random effects. *Statistical Science*, *6*(1), 15-51.

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

Sin, N. L., Graham-Engeland, J. E., Ong, A. D., & Almeida, D. M. (2015). Affective reactivity to daily stressors is associated with elevated inflammation. *Health Psychology*, *34*, 1154-1165. doi:10.1037/hea0000240

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Newbury Park, CA: Sage.

Suls, J., Green, P., & Hillis, S. (1998). Emotional reactivity to everyday problems, affective inertia, and neuroticism. *Personality and Social Psychology Bulletin*, *24*, 127-136. doi:10.1177/0146167298242002

van Eck, M., Berkhof, H., Nicolson, N., & Sulon, J. (1996). The effects of perceived stress, traits, mood states, and stressful daily events on salivary cortisol. *Psychosomatic Medicine*, *58*, 447-458.

Wright, A. G. C., Beltz, A. M., Gates, K. M., Molenaar, P. C. M., & Simms, L. J. (2015). Examining the dynamic structure of daily internalizing and externalizing behavior at multiple levels of analysis. *Frontiers in Psychology*, *6*, 1914. doi:10.3389/fpsyg.2015.01914

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*, 1100-1122. doi:10.1177/1745691617693393