Multivariate Behavioral Research

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/hmbr20

Estimating the Reliability of Aggregated and Within-Person Centered Scores in Ecological Momentary Assessment

Po-Hsien Huang [a] & Li-Jen Weng [a]

[a] Department of Psychology, National Taiwan University

Version of record first published: 15 Jun 2012.

To cite this article: Po-Hsien Huang & Li-Jen Weng (2012): Estimating the Reliability of Aggregated and Within-Person Centered Scores in Ecological Momentary Assessment, Multivariate Behavioral Research, 47:3, 421-441

To link to this article: http://dx.doi.org/10.1080/00273171.2012.673924

Multivariate Behavioral Research, 47:421–441, 2012

ISSN: 0027-3171 print/1532-7906 online
DOI: 10.1080/00273171.2012.673924

Psychology Press
Taylor & Francis Group

Estimating the Reliability of Aggregated and Within-Person Centered Scores in Ecological Momentary Assessment

Po-Hsien Huang and Li-Jen Weng
Department of Psychology, National Taiwan University

A procedure for estimating the reliability of test scores in the context of ecological momentary assessment (EMA) was proposed to take into account the characteristics of EMA measures. Two commonly used test scores in EMA were considered: the aggregated score (AGGS) and the within-person centered score (WPCS). Conceptually, AGGS and WPCS represent the interindividual differences and the intraindividual differences, respectively. The reliability coefficients for AGGS and WPCS were derived using a multilevel factor model with a serial correlation structure framework. Point estimates and confidence intervals of these coefficients were obtained using Mx (Neale, Boker, Xie, & Maes, 2004). A simulation study showed that the proposed procedure performed well empirically. Diary data from Huang (2009), which recorded daily joy level of 110 undergraduate students for 8 days, was used to illustrate the applicability of the proposed method.

Over the past 2 decades, ecological momentary assessment (EMA; Stone & Shiffman, 1994) has been adopted in several areas of psychology, including clinical psychology (Trull & Ebner-Priemer, 2009; Wenze & Miller, 2010), health psychology (Shiffman & Stone, 1998), and personality psychology (Tennen, Affleck, & Armeli, 2005). According to Stone and Shiffman (1994), EMA is not

Correspondence concerning this article should be addressed to Li-Jen Weng, Department of Psychology, National Taiwan University, No. 1, Sec. 4, Roosevelt Road, Taipei, 10617 Taiwan. E-mail: ljweng@ntu.edu.tw

a single method but rather a collection of methods that share the following four characteristics: (a) data are collected in real-world environments, (b) assessments focus on individuals' current states or behaviors, (c) assessment moments are strategically selected (e.g., event-based, time-based, or randomly prompted), and (d) participants complete multiple assessments over time. These characteristics indicate that EMA may increase ecological validity, avoid biased recollections of memories, and allow researchers to explore within-person changes in experiences or behaviors over time as well as across contexts (for a detailed discussion, see Bolger, Davis, & Rafaeli, 2003; Hektner, Schmidt, & Csikszentmihalyi, 2007; Shiffman, Stone, & Hufford, 2008; Stone, Shiffman, Atienza, & Nebeling, 2007).

Although the use of EMA is promising, it also presents several methodological challenges (e.g., Hufford, 2007; Schwarz, 2007; Shiffman, 2007; Shiyko & Ram, 2011). In this article, the issue of reliability estimation with EMA-related test scores is considered. Based on classical test theory (Lord & Novick, 1968), several popular procedures have been proposed to estimate test score reliability (see Feldt & Brennan, 1989; Haertel, 2006, for a review). However, these procedures may not be appropriate in EMA studies. The first reason for this is that EMA measurements are repeatedly measured, which usually implies the violation of the independence assumption among observations (Raudenbush & Bryk, 2002). When the independence assumption does not hold, Type I error rate may be incorrect (Kenny & Judd, 1986; Muthén & Satorra, 1995). Second, in EMA studies, the most commonly used test score, the composite score—defined as the unweighted sum of item scores for a given person at specific timepoint—carries two types of information: the information of *interindividual differences* (i.e., the typical response) and the information of *intraindividual differences* (i.e., the deviations from the typical response). These two types of information reflect different aspects of human behaviors and should not be treated in the same way (see Molenaar, 2004, for a discussion). Unfortunately, the traditional procedures for estimating reliability do not make this distinction and may result in ambiguous interpretations. To our knowledge, only a few studies have considered the reliability issue of EMA scores (Csikszentmihalyi & Larson, 1987; Hektner et al., 2007). Yet, these studies simply applied comprehensive evaluations based on traditional reliability estimation procedures, such as the coefficient α (Cronbach, 1951; Guttman, 1945) and the test-retest reliability coefficient, to EMA studies; the characteristics of EMA measurement have yet to be taken into consideration. As a consequence, both the nonindependence issue remains untouched and the meaning of the obtained coefficients remains ambiguous.

Some studies have come close to investigating the reliability of measurements with data structure similar to those in EMA but ultimately fall short. Several studies have focused on the reliability of repeated measures (Biemer, Christ, &

Wiesen, 2009; Laenen, Alonso, & Molenberghs, 2007; Laenen, Alonso, Molenberghs, & Vangeneugden, 2009; Laenen, Vangeneugden, Geys, & Molenberghs, 2006). In these works, the nonindependence problem was overcome by modeling the dependence structure directly, but none of these studies consider the reliability of test scores that represent intraindividual differences. Because EMA data can be approached from a multilevel perspective, reliability estimation procedures proposed under multilevel measurement models are legitimate alternatives (Raudenbush, Rowan, & Kang, 1991; Raykov & du Toit, 2005; Snijders & Bosker, 1999). The main strength of these procedures is that they allow for the estimation of test score reliability at different levels (Raudenbush et al., 1991). However, the dependency structures considered in these works are relatively simple and do not incorporate the essential property of serial correlation that is embedded in EMA measures. Serial correlation is defined as the correlation between two measurements on a single person in a longitudinal study (Everitt, 2002). Modeling such a correlation structure in EMA studies has been recommended (Schwartz & Stone, 1998, 2007) and ignoring them may result in biased variance estimates (Laenen, 2008).

In this study, a procedure for estimating the reliability of EMA-related scores is proposed. The proposed procedure is based on the framework of a *multilevel factor model with serial correlation structure* (MFM-SCS). MFM-SCS is an extension of the traditional multilevel factor model (MFM; Lee, 1990; McDonald & Goldstein, 1989; Muthén, 1989) and accounts for the serial correlation among repeated measures. Two types of EMA-related test scores are considered in this study: the *aggregated score* (AGGS) and the *within-person centered score* (WPCS). AGGS is the arithmetic average of the composite score from repeated measurements for a given person and WPCS is the deviation score from AGGS for a given person measured at a certain time point. AGGS and WPCS are widely used in multilevel modeling to represent the information of interindividual differences and the information on intraindividual differences, respectively (see Curran & Bauer, 2011, for a discussion). Because an important advantage of using EMA data is that it allows researchers to model interindividual and intraindividual processes simultaneously, the choice of AGGS and WPCS as the target test scores for further reliability estimation is straightforward. Reliability coefficients for AGGS and WPCS are to be derived using the model-based approach, which has been used pervasively (e.g., Bentler, 2009; Jöreskog, 1971; Raykov & Shrout, 2002).

The proposed method can be implemented in Mx (Neale, Boker, Xie, & Maes, 2004). Supplemental materials, including Mx scripts and artificial data sets, are available on the Web (http://homepage.ntu.edu.tw/~f97227110/Rel_Est_EMA.zip). Diary data from Huang (2009), which recorded daily joy level of 110 undergraduate students for 8 days, is used to illustrate the applicability of the proposed method.

This article is organized as follows: it begins with the introduction of the MFM-SCS model followed by a proposal for a procedure for estimating the reliabilities of AGGS and WPCS. A simulation study to evaluate the empirical performance of the proposed procedure is presented. This is followed by the real data example. Finally, merits, cautions, and further directions concerning the procedure are discussed.

MULTILEVEL FACTOR MODEL WITH SERIAL CORRELATION STRUCTURE

This section specifies the MFM-SCS model and explains its implications in the context of EMA studies. Without loss of generality, all the items are assumed to be congeneric (i.e., all the items are measuring the same latent construct). Let y_{ijk} denote the response for person $i (i = 1, 2, \ldots, N)$ on item $k (k = 1, 2, \ldots, K)$ at time $j (j = 1, 2, \ldots, J)$, MFM-SCS decomposes y_{ijk} into the sum of the between-person component v_{ik} and the within-person component v_{ijk}

$$y_{ijk} = v_{ik} + v_{ijk}, \tag{1}$$

In Equation 1, v_{ik} represents the typical response of person i to item k and contains the information of interindividual differences. Similarly, v_{ijk} represents the deviation from v_{ik} and contains the information of intraindividual differences. In the literature of multilevel modeling, v_{ik} and v_{ijk} are also called the Level 2 component and Level 1 component, respectively. MFM-SCS describes the linear relationship between the components and the corresponding latent factors as

$$\begin{aligned} v_{ik} &= \alpha_k + \lambda_{bk}\eta_i + \varepsilon_{ik}, \\ v_{ijk} &= \lambda_{wk}\eta_{ij} + \varepsilon_{ijk}, \end{aligned} \tag{2}$$

where α_k is the intercept for item k, η_i is the Level 2 latent factor, λ_{bk} is the Level 2 factor loading, ε_{ik} is the Level 2 residual, η_{ij} is the Level 1 latent factor, λ_{wk} is the Level 1 factor loading, and ε_{ijk} is the Level 1 residual. For EMA measurements, η_i and η_{ij} can be regarded as the same psychological attribute at different levels to make the distinction between trait and occasion effects (e.g., Cole, Martin, & Steiger, 2005). In this sense, η_i represents the latent trait score for person i, and η_{ij} represents the occasional score for person i at time j. For example, suppose the items measure people's current level of joy, then η_i represents the trait joy level for person i and η_{ij} represents the occasional joy level for person i at time j. The factor loadings λ_{bk} and λ_{wk} describe the degrees of association between the corresponding components and the factors.

The residuals ε_{ik} and ε_{ijk} represent measurement errors that cannot be explained by the latent factors.

Equations 1 and 2 can be written in a compact matrix form as

$$\mathbf{y}_{ij} = \mathbf{v}_i + \mathbf{v}_{ij} = (\boldsymbol{\alpha} + \boldsymbol{\lambda}_b \eta_i + \boldsymbol{\varepsilon}_i) + (\boldsymbol{\lambda}_w \eta_{ij} + \boldsymbol{\varepsilon}_{ij}), \tag{3}$$

where $\mathbf{y}_{ij}$, $\mathbf{v}_i$, $\mathbf{v}_{ij}$, $\boldsymbol{\alpha}$, $\boldsymbol{\lambda}_b$, $\boldsymbol{\varepsilon}_i$, $\boldsymbol{\lambda}_w$, and $\boldsymbol{\varepsilon}_{ij}$ are all $K \times 1$ vectors that stack the elements in Equation 2, and η_i and η_{ij} are scalars in the one-factor model. The following assumptions are made for the MFM-SCS model:

(A1) For each i, the Level 2 component $\mathbf{v}_i$ is uncorrelated with Level 1 components $\{\mathbf{v}_{i1}, \mathbf{v}_{i2}, \ldots, \mathbf{v}_{iJ}\}$.
(A2) Level 2 components $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_N\}$ are independently and identically distributed with $\mathbf{v}_i \sim N(\boldsymbol{\alpha}, \boldsymbol{\Sigma}_b)$, where $\boldsymbol{\alpha}$ is a $K \times 1$ mean vector and $\boldsymbol{\Sigma}_b$ is a $K \times K$ covariance matrix.
(A3) Let $\mathbf{V}_i = [\mathbf{v}_{i1} \quad \mathbf{v}_{i2} \quad \mathbf{v}_{i3} \cdots \mathbf{v}_{iJ}]$; $\mathbf{V}_i$ follows a matrix normal distribution independently, that is, $\mathbf{V}_i \sim MN_{K \times J}(\mathbf{0}, \boldsymbol{\Sigma}_w, \boldsymbol{\Phi}_i)$ (see Kollo & von Rosen, 2005), where $\boldsymbol{\Sigma}_w$ is a $K \times K$ covariance matrix and $\boldsymbol{\Phi}_i$ is a $J \times J$ serial correlation matrix illustrating the dependence structure among observations across time.
(A4) For each i, the Level 2 latent factor η_i is uncorrelated with the Level 2 residual $\boldsymbol{\varepsilon}_i$. In addition, $\eta_i \sim N(0, \psi_b)$ and $\boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \boldsymbol{\Theta}_b)$, where $\boldsymbol{\Theta}_b$ is a $K \times K$ Level 2 residual covariance matrix.
(A5) For each i and j, the Level 1 latent factor η_{ij} is uncorrelated with the Level 1 residual $\boldsymbol{\varepsilon}_{ij}$. In addition, $\eta_{ij} \sim N(0, \psi_w)$ and $\boldsymbol{\varepsilon}_{ij} \sim N(\mathbf{0}, \boldsymbol{\Theta}_w)$, where $\boldsymbol{\Theta}_w$ is a $K \times K$ Level 1 residual covariance matrix.

(A1) is the standard assumption in multilevel modeling that allows for variance decomposition. (A2)–(A3) state the independence/dependence and distribution assumptions for Level 2 and Level 1 components. The independence assumptions for both components are essential and can be achieved when the participants are sampled independently. The dependence structure made in (A3) is the main difference between MFM-SCS and traditional MFM. Traditional MFM assumes that the Level 1 components $\{\mathbf{v}_{i1}, \mathbf{v}_{i2}, \ldots, \mathbf{v}_{iJ}\}$ are independently and identically distributed with $\mathbf{v}_{ij} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_w)$. Note that MFM is a special case of MFM-SCS with the serial correlation matrix $\boldsymbol{\Phi}_i$ being a $J \times J$ identity matrix. The normality assumptions in (A2)–(A3) are not necessary for deriving the model-implied mean and covariance structure but will be crucial when maximum likelihood estimation is used. (A4)–(A5) make further assumptions on latent factors and residuals. Because latent factors are often treated as true score and residuals are treated as measurement errors, the independence between latent factors and residuals is an analogy of independence between true scores and residuals in classical test theory. Again, the distribution assumptions in (A4)–(A5) are only important for maximum likelihood estimation.

When (A1)–(A5) are satisfied, the model-implied mean and covariance structure for $\mathbf{y}_{ij}$ is

$$\begin{aligned} \boldsymbol{\mu}(\boldsymbol{\theta}) &= \boldsymbol{\alpha}, \\ \boldsymbol{\Sigma}(\boldsymbol{\theta}) &= \boldsymbol{\Sigma}_b + \boldsymbol{\Sigma}_w = (\boldsymbol{\lambda}_b \psi_b \boldsymbol{\lambda}_b^T + \boldsymbol{\Theta}_b) + (\boldsymbol{\lambda}_w \psi_w \boldsymbol{\lambda}_w^T + \boldsymbol{\Theta}_w), \end{aligned} \tag{4}$$

where $\boldsymbol{\theta}$ is a Q-dimensional vector of model parameters. Equation 4 shows that the total covariance is decomposed into a between-person covariance $\boldsymbol{\Sigma}_b$ and a within-person covariance $\boldsymbol{\Sigma}_w$. Each of the two covariance matrices can be further decomposed into two parts: the covariance associated with true score ($\boldsymbol{\lambda}_b \psi_b \boldsymbol{\lambda}_b^T$ and $\boldsymbol{\lambda}_w \psi_w \boldsymbol{\lambda}_w^T$) and the covariance associated with measurement errors ($\boldsymbol{\Theta}_b$ and $\boldsymbol{\Theta}_w$). This decomposition is crucial for deriving the reliability coefficients.

Notice that the mean and covariance structures in Equation 4 are identical for MFM-SCS and traditional MFM. The difference will appear if we consider the covariance structure for the complete response vector, $\mathbf{y}_i^* = [\mathbf{y}_{i1}^T, \mathbf{y}_{i1}^T, \ldots, \mathbf{y}_{iJ}^T]^T$ (see Appendix A). In such case, the serial correlation matrix $\boldsymbol{\Phi}_i$ is involved. $\boldsymbol{\Phi}_i$ is always structured by a P-dimensional serial correlation parameter $\boldsymbol{\beta}$ and the fixed time separation $d_{ijj'}$. Here, $d_{ijj'} = |t_{ij} - t_{ij'}|$, where t_{ij} denotes the time coding given to person i at time j. An element of $\boldsymbol{\Phi}_i$, $\phi_{ijj'}$, describes the magnitude of serial correlation between times j and j' for person i and is assumed to be a function of $\boldsymbol{\beta}$ and $d_{ijj'}$. For $\boldsymbol{\beta} > 0$, the associated function value must be decreasing with respect to $d_{ijj'}$ and tends toward zero as $d_{ijj'}$ goes to infinity with appropriate order. This restriction reflects the fact that the serial correlation between two measurements will disappear as the time separation becomes longer. There are several commonly used serial correlation structures in the literature (see Verbeke & Molenberghs, 2000, p. 99). For example, if the intervals between measurements are discrete with equal spacing, the first-order autoregressive structure can be adopted

$$\boldsymbol{\Phi}_i^{AR1}(\beta) = \begin{bmatrix} 1 & & & & \\ \beta & 1 & & Sym. & \\ \beta^2 & \beta & 1 & & \\ \vdots & \vdots & \vdots & \ddots & \\ \beta^{J-1} & \beta^{J-2} & \beta^{J-3} & \cdots & 1 \end{bmatrix}, \tag{5}$$

where β is the so-called autoregressive parameter that represents the correlation between times j and $j+1$ for the considered variable. Taking another example, when the intervals of observations are continuous with unequal spacing, we may consider the exponential structure

$$\mathbf{\Phi}_i^{EXP}(\beta) = \begin{bmatrix} 1 & & & & \\ \exp(-|d_{i21}|/\beta) & 1 & & Sym. & \\ \exp(-|d_{i31}|/\beta) & \exp(-|d_{i32}|/\beta) & 1 & & \\ \vdots & \vdots & \vdots & \ddots & \\ \exp(-|d_{iJ1}|/\beta) & \exp(-|d_{iJ2}|/\beta) & \exp(-|d_{iJ3}|/\beta) & \cdots & 1 \end{bmatrix}. \tag{6}$$

With this structure, the serial correlation structure for persons i and i' may be different when the time separations for them are not equal.

The parameters in MFM-SCS can be estimated by maximum likelihood estimation (see Appendix A). The statistical software Mx (Neale et al., 2004), characterized by its flexibility in defining complex mean and covariance structures, is used to obtain the maximum likelihood estimates. The "individual" serial correlation matrix can also be modeled by using the definition variable approach in Mx (see Mehta & West, 2000).

RELIABILITY ESTIMATION FOR AGGS AND WPCS

This section illustrates the procedure for estimating the reliability of AGGS and WPCS. First, three relevant indices ρ^{ICC}, ρ^{LV1}, and ρ^{LV2} are defined. Let $\mathbf{1}$ denote the K-dimensional vector of 1's, then $y_{ij}^{+} = \mathbf{1}^T\mathbf{y}_{ij}$ is the composite score of the K items for person i at time j. Conceptually, the composite score y_{ij}^{+} is used to measure the state level of some psychological attribute that is a linear combination of the trait level η_i and the occasional level η_{ij} (see Cole et al., 2005). Similarly, $v_i^{+} = \mathbf{1}^T\mathbf{v}_i$ and $v_{ij}^{+} = \mathbf{1}^T\mathbf{v}_{ij}$ denote the composite scores of Level 2 and Level 1 components. The associated true scores are $\tau_i^{+} = \mathrm{E}(v_i^{+}|\eta_i) = \mathbf{1}^T\boldsymbol{\alpha} + \mathbf{1}^T\boldsymbol{\lambda}_b\eta_i$ and $\tau_{ij}^{+} = \mathrm{E}(v_{ij}^{+}|\eta_{ij}) = \mathbf{1}^T\boldsymbol{\lambda}_w\eta_{ij}$, respectively, where $\mathrm{E}(\cdot|\cdot)$ denotes the operator of conditional expectation. The three indices are defined as

$$\begin{aligned}
\rho^{ICC} &= \frac{Var(v_i^{+})}{Var(y_{ij}^{+})} = \frac{\mathbf{1}^T\boldsymbol{\Sigma}_b\mathbf{1}}{\mathbf{1}^T(\boldsymbol{\Sigma}_b + \boldsymbol{\Sigma}_w)\mathbf{1}},\\
\rho^{LV1} &= \frac{Var(\tau_{ij}^{+})}{Var(v_{ij}^{+})} = \frac{\mathbf{1}^T\lambda_w\psi_w\boldsymbol{\lambda}_w^T\mathbf{1}}{\mathbf{1}^T\boldsymbol{\Sigma}_w\mathbf{1}},\\
\rho^{LV2} &= \frac{Var(\tau_i^{+})}{Var(v_i^{+})} = \frac{\mathbf{1}^T\lambda_b\psi_b\lambda_b^T\mathbf{1}}{\mathbf{1}^T\boldsymbol{\Sigma}_b\mathbf{1}}.
\end{aligned} \tag{7}$$

ρ^{ICC} is a multivariate extension of the intraclass correlation coefficient (ICC) in multilevel models. ICC is defined as the proportion of the total variance based upon differences between clusters (Raudenbush, & Bryk, 2002). In EMA studies, ρ^{ICC} is used to describe the relative importance between the variances of the Level 1 and Level 2 components. A higher ρ^{ICC} indicates y_{ij}^{+} being more stable across time. Coefficients ρ^{LV1} and ρ^{LV2} can be taken as the reliability coefficients for v_i^{+} and v_{ij}^{+}, respectively. The values of ρ^{LV1} and ρ^{LV2} depend on the magnitudes of the factor loadings and the total number of items. When the magnitudes of the factor loadings or the total number of items increase, the values of ρ^{LV1} and ρ^{LV2} also increase. Notice that there is no ordinal relationship between the values of these indices. It is possible for a measurement to have a reasonable ρ^{LV1} but tiny values for ρ^{LV2} and ρ^{ICC}.

For deriving the reliability coefficient of AGGS, AGGS and its true score should be represented formally. Let $y_i^{AGG} = J^{-1}\Sigma_{j=1}^{J} y_{ij}^{+}$ denote the AGGS for person i. y_i^{AGG} can be interpreted as the typical response for person i on a given measurement. As a test score, y_i^{AGG} is used to measure the latent trait η_i. Therefore, the true score for y_i^{AGG} will be $\tau_i^{AGG} = \mathrm{E}(y_i^{AGG}|\eta_i) = \tau_i^{+}$. From the derivations given in Appendix B, the reliability coefficient for y_i^{AGG} is

$$\rho_i^{AGG} = \frac{Var(\tau_i^{ACC})}{Var(y_i^{AGG})} = \frac{\rho^{ICC}}{\rho^{ICC} + (1-\rho^{ICC})J^{-2}\Sigma_{j=1}^{J}\Sigma_{j'=1}^{J}\phi_{ijj'}}\rho^{LV2}. \tag{8}$$

As a reliability coefficient, ρ_i^{AGG} can be understood as the proportion of variance that τ_i^{+} can account for in y_i^{AGG}. The subscript i in ρ_i^{AGG} shows that the value of ρ_i^{AGG} depends on which person is being considered. Several factors influence the value of ρ_i^{AGG}, including ρ^{LV2}, J, and $\phi_{ijj'}$. The higher the ρ^{LV2}, the larger the value of ρ_i^{AGG} will be. J and $\phi_{ijj'}$ influence ρ_i^{AGG} through the second term of the denominator in Equation 8. This term tends toward zero when J is large, and the magnitude of $\phi_{ijj'}$ influences the rate of such convergence. When there is no serial correlation, Equation 8 reduces to

$$\rho_i^{AGG} = \frac{\rho^{ICC}}{\rho^{ICC} + (1-\rho^{ICC})J^{-1}}\rho^{LV2}. \tag{9}$$

This simplification shows that the Level 2 reliability coefficients proposed by Raudenbush et al. (1991) and Snijders and Bosker (1999) are special cases of ρ_i^{AGG}.

For deriving the reliability coefficient of WPCS, let $y_{ij}^{WPC} = y_{ij}^{+} - y_i^{AGG}$ denote the WPCS for person i at time j. In contrast to y_i^{AGG}, y_{ij}^{WPC} concerns a different aspect of human behavior: the deviation from the typical response for person i at time j. This type of score is often used in multilevel modeling to represent the pure Level 1 information. As a measure for η_{ij}, the true score for

y_{ij}^{WPC} will be $\tau_{ij}^{WPC} = \mathrm{E}(y_i^{WPC}|\eta_{ij}) = \frac{J-1}{J}\tau_{ij}^{+}$. The reliability coefficient for y_{ij}^{WPC} as derived in Appendix B will be

$$\rho_{ij}^{WPC} = \frac{Var(\tau_{ij}^{WPC})}{Var(y_{ij}^{WPC})} = \frac{(J-1)^2}{(J^2 + \Sigma_{l=1}^{J}\Sigma_{l'=1}^{J}\phi_{ill'} - 2J\Sigma_{l=1}^{J}\phi_{ijl})}\rho^{LV1}. \quad (10)$$

Similarly, ρ_{ij}^{WPC} quantifies the explained variance of y_{ij}^{WPC} by τ_{ij}^{WPC}. Clearly, the value of ρ_{ij}^{WPC} is mainly determined by ρ^{LV1}. In addition, the value of ρ_{ij}^{WPC} depends on not only which person is considered but also which time is selected. The selected time j influences the value of ρ_{ij}^{WPC} through the third term of the denominator in Equation 10. Because this term becomes larger when j is close to $\frac{J}{2}$, so does ρ_{ij}^{WPC}. If serial correlation is zero, Equation 10 reduces to

$$\rho_{ij}^{WPC} = \frac{J-1}{J}\rho^{LV1}. \quad (11)$$

The simplified expression has a form similar to that of the Level 1 reliability coefficient proposed by Raudenbush et al. (1991).

The invariance property of maximum likelihood estimation is used to obtain the maximum likelihood estimate (MLE) of ρ_i^{AGG} and ρ_{ij}^{WPC}. If $\hat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$, then for any function $g(\boldsymbol{\theta})$, the MLE of $g(\boldsymbol{\theta})$ is $g(\hat{\boldsymbol{\theta}})$ (Cassella & Berger, 2002). Therefore, given specific i and j, we can substitute the obtained MLEs of parameters into Equations 8 and 10 to obtain $\hat{\rho}_i^{AGG}$ and $\hat{\rho}_{ij}^{WPC}$. In addition to the point estimates, the confidence intervals (CIs) of ρ_i^{AGG} and ρ_{ij}^{WPC} can also be constructed by the likelihood-based approach.[1] The likelihood-based $1-\alpha$ CI for ρ is the set $\mathcal{P}_\alpha$ that satisfies

$$\mathcal{P}_\alpha = \left\{ \rho | 2\log\frac{L(\hat{\rho})}{L(\rho)} < \chi^2_{1,1-\alpha} \right\}, \quad (12)$$

where $L(\hat{\rho})$ is the value of the likelihood function of Equation A3 in Appendix A evaluated at the MLE $\hat{\rho}$ and $\chi^2_{1,1-\alpha}$ is the critical value corresponding to the chosen α for chi-square variable with one degree of freedom. The likelihood-based CIs for the ρ's can also be obtained in the output of Mx.

Taking into account the fact that the values of ρ_i^{AGG} and ρ_{ij}^{WPC} depend on which person or time is considered, a comprehensive way to evaluate the reliabilities of y_i^{AGG} and y_{ij}^{WPC} is to estimate ρ_i^{AGG} and ρ_{ij}^{WPC} for each combination of i and j. However, this is a cumbersome task in practice. Therefore, to avoid such

[1]To construct the CIs for the reliability coefficients, several studies used the delta method (e.g., Laenen et al., 2009; Raykov & Shrout, 2002). The likelihood-based approach is another method to construct CIs in maximum likelihood estimation. The relative merits and drawbacks of the two approaches were discussed by Cheung (2009) and Neale and Miller (1997).

a cumbersome task, we propose "pooled" versions of ρ_i^{AGG} and ρ_{ij}^{WPC}, denoted by ρ_p^{AGG} and ρ_p^{AGG}, as an alternative method. Let $\bar{t}_j$ denote the expected time coding at time j, then the pooled serial correlation matrix $\mathbf{\Phi}_p = \{\bar{\phi}_{jj'}\}$ can be constructed by using the expected separation $\bar{d}_{jj'} = |\bar{t}_j - \bar{t}_{j'}|$. We define ρ_p^{AGG} and ρ_p^{WPC} as

$$\rho_p^{AGG} = \frac{\rho^{ICC}}{p^{ICC} + (1-\rho^{ICC})J^{-2}\Sigma_{j=1}^{J}\Sigma_{j'=1}^{J}\bar{\phi}_{jj'}}\rho^{LV2},$$
$$\rho_p^{WPC} = \frac{(J-1)^2}{(J^2 - \Sigma_{j=1}^{J}\Sigma_{j'=1}^{J}\bar{\phi}_{jj'})}\rho^{LV1}. \quad (13)$$

When $\mathbf{\Phi}_i$'s are heterogeneous or J is large, researchers can use ρ_p^{AGG} and ρ_p^{WPC} to obtain pooled information about the ρ_i^{AGG}'s and ρ_{ij}^{WPC}'s.

A SIMULATION STUDY

A small simulation study was conducted to evaluate the empirical performance of the proposed procedure. The parameter values of $\mathbf{\Sigma}(\boldsymbol{\theta})$ were set as $\boldsymbol{\alpha} = [0,0,0,0]^T$, $\boldsymbol{\lambda}_b = [0.8, 0.6, 0.8, 0.6]^T$, $\mathbf{\Theta}_b = \text{diag}[0.108, 0.192, 0.108, 0.192]$, $\psi_b = 0.3$, $\boldsymbol{\lambda}_w = [0.8, 0.6, 0.8, 0.6]^T$, and $\mathbf{\Theta}_w = \text{diag}[0.252, 0.448, 0.252, 0.448]$, and $\psi_w = 0.7$. According to Equation 7, we have $\rho^{ICC} = .3$ and $\rho^{LV1} = \rho^{LV2} = .7967$. Two types of serial correlation structures were considered: the first-order autoregressive structure (AR1) and the exponential structure (EXP). The serial correlation matrix in AR1 was set as $\mathbf{\Phi}_i(\beta) = \{\beta^{|j-j'|}\}$, where $\beta = .3$ and $j = 1, 2, \ldots, 8$. The corresponding true values of the pooled reliability coefficients are $\rho_p^{AGG} = .5322$ and $\rho_p^{WPC} = .7751$. For EXP, $\mathbf{\Phi}_i(\beta) = \{\exp(-|t_{ij} - t_{ij'}|/\beta)\}$, where $\beta = .8$, $j = 1, 2, \ldots, 8$, and $d_{i(j+1)j} = |t_{i(j+1)} - t_{ij}|$ followed a uniform distribution in the interval of [0.5, 1.5] independently. The EXP shown here took into account the heterogeneity of sampling intervals. In this case, $\rho_p^{AGG} = .5365$ and $\rho_p^{WPC} = .7700$. The simulated data sets were generated in R (R Development Core Team, 2010). The sample size N was set to be 100 with 500 replications in both conditions. After data generation, the data sets were analyzed by Mx. All the parameters were set to be free with the exception of ψ_b and ψ_w, which were fixed at 0.3 and 0.7 for identification. In addition to evaluating the empirical performance of the proposed procedure, we also evaluated the data under traditional MFM to assess the consequences of ignoring serial correlations.

Several criteria were used to evaluate the empirical performance of the proposed procedure. First, the convergence rate was used to assess the quality of optimization. Second, the empirical biases of estimates were used to evaluate the performances of parameter estimates. Here, the empirical bias of the estimates is defined as the difference between the mean of estimates ($\bar{\hat{\theta}}$) and the true value θ_0. Third, the ratios of the means of estimated standard errors ($\overline{\widehat{SE}(\theta)}$) to the empirical standard deviations of estimates ($SD(\hat{\theta})$) were examined with the expectation that they were to be around 1. Finally, the coverage rate of the 95% likelihood-based CIs for ρ_p^{AGG} and ρ_p^{WPC} were examined. Under the assumption that the CIs behave well, the coverage rate should be close to 95%.

In general, the proposed procedure performed well under MFM-SCS with serial correlations considered. The convergence rate was acceptable with four data sets in AR1 (0.8%) and three data sets in EXP (0.6%) yielding a nonconvergence problem. New data sets were generated to replace the original data sets for further analysis. As shown in Tables 1 and 2, both the model parameter estimates and their associated standard errors performed well. Table 3 shows

TABLE 1
Results of Parameter Estimations for the First-order Autoregressive (AR1) Structure

Parameter	*MFM-SCS*		*MFM*	
θ_0	$\bar{\hat{\theta}} - \theta_0$	$\overline{\widehat{SE}(\theta)}/SD(\hat{\theta})$	$\bar{\hat{\theta}} - \theta_0$	$\overline{\widehat{SE}(\theta)}/SD(\hat{\theta})$
$\lambda_{b1} = .8$	−.0040	.9758	.0870	.9739
$\lambda_{b2} = .6$	−.0121	.9624	.0596	.9575
$\lambda_{b3} = .8$	−.0039	.9882	.0883	.9843
$\lambda_{b4} = .6$	.0012	.9796	.0694	.9624
$\theta_{b1} = .108$	−.0093	.9658	.0178	.9265
$\theta_{b2} = .192$	−.0029	1.0203	.0429	.9961
$\theta_{b3} = .108$	−.0096	1.0236	.0169	1.0405
$\theta_{b4} = .192$	−.0081	1.0021	.0397	1.0025
$\alpha_1 = 0$	.0052	.9384	.0046	.9311
$\alpha_2 = 0$	.0074	1.0631	.0071	1.0581
$\alpha_3 = 0$	.0103	1.0280	.0097	1.0246
$\alpha_4 = 0$	.0071	.9921	.0067	.9871
$\lambda_{w1} = .8$	.0018	.9590	−.0403	1.0322
$\lambda_{w2} = .6$	.0029	1.0415	−.0287	1.1131
$\lambda_{w3} = .8$	.0014	1.0385	−.0408	1.1193
$\lambda_{w4} = .6$	−.0007	.9526	−.0321	.9991
$\theta_{w1} = .252$	.0004	1.0334	−.0253	1.1266
$\theta_{w2} = .448$	−.0014	1.0255	−.0469	1.0717
$\theta_{w3} = .252$	.0006	1.0111	−.0251	1.0619
$\theta_{w4} = .448$	.0006	.9176	−.0456	.9729
$\beta = .3$	.0024	.9487	—	—

Note. MFM-SCS = multilevel factor model with serial correlation structure; MFM = multilevel factor model; '—' indicates that MFM did not estimate β.

TABLE 2
Results of Parameter Estimations for the Exponential (EXP) Structure

Parameter	*MFM-SCS*		*MFM*	
θ_0	$\bar{\hat{\theta}} - \theta_0$	$\overline{\widehat{SE}(\theta)}/SD(\hat{\theta})$	$\bar{\hat{\theta}} - \theta_0$	$\overline{\widehat{SE}(\theta)}/SD(\hat{\theta})$
$\lambda_{b1} = .8$	−.0149	1.0100	.0780	.9971
$\lambda_{b2} = .6$	−.0237	1.0216	.0472	1.0181
$\lambda_{b3} = .8$	−.0055	1.0648	.0873	1.0709
$\lambda_{b4} = .6$	−.0072	.9916	.0609	.9908
$\theta_{b1} = .108$	−.0057	.9978	.0215	.9854
$\theta_{b2} = .192$	−.0024	.9937	.0455	1.0039
$\theta_{b3} = .108$	−.0092	1.0151	.0180	.9958
$\theta_{b4} = .192$	−.0065	.8722	.0403	.8785
$\alpha_1 = 0$	−.0012	1.0531	−.0020	1.0594
$\alpha_2 = 0$	−.0053	1.0609	−.0057	1.0582
$\alpha_3 = 0$	−.0034	1.0279	−.0040	1.0338
$\alpha_4 = 0$	−.0032	1.0085	−.0037	1.0109
$\lambda_{w1} = .8$	−.0010	.9964	−.0441	1.0666
$\lambda_{w2} = .6$	.0042	1.0212	−.0266	1.0681
$\lambda_{w3} = .8$	.0021	1.0078	−.0397	1.0872
$\lambda_{w4} = .6$	−.0001	.9582	−.0322	1.0486
$\theta_{w1} = .252$	−.0010	1.0055	−.0264	1.0619
$\theta_{w2} = .448$	−.0012	.9977	−.0470	1.0160
$\theta_{w3} = .252$	−.0006	.9732	−.0271	1.0675
$\theta_{w4} = .448$	−.0038	1.0250	−.0494	1.0929
$\beta = .8$	−.0028	1.0411	—	—

Note. MFM-SCS = multilevel factor model with serial correlation structure; MFM = multilevel factor model; '—' indicates that MFM did not estimate β.

that the empirical biases of ρ's were also close to zero and the coverage rates of the CIs for ρ_p^{AGG} and ρ_p^{WPC} were acceptable in both conditions.

However, when the data were analyzed under traditional MFM, nearly all the parameters were biased empirically, as presented in Tables 1 and 2. The parameters in $\boldsymbol{\Sigma}_b$ were overestimated and those in $\boldsymbol{\Sigma}_w$ were underestimated. As functions of parameter estimates, $\hat{\rho}^{ICC}$, $\hat{\rho}_p^{AGG}$, and $\hat{\rho}_p^{WPC}$ were also biased. Furthermore, Table 3 shows that the coverage rates of the CIs under MFM were far from the nominal level.

A REAL DATA EXAMPLE

Data from Huang (2009) were adopted for illustration. In Huang's study, 110 undergraduate students recorded their daily positive/negative affects and thoughts

TABLE 3
Empirical Biases and Coverage Rates of 95% CIs for ρs

a. Empirical Bias

	AR1 Structure		*EXP Structure*	
Parameter	*MFM-SCS*	*MFM*	*MFM-SCS*	*MFM*
ρ^{ICC}	−.0045	.0675	−.0075	.0650
ρ^{LV2}	.0005	.0008	.0058	−.0047
ρ^{LV1}	.0003	.0002	.0010	.0009
ρ_p^{AGG}	.0053	.1229	−.0057	.1182
ρ_p^{WPC}	.0010	−.0777	.0009	−.0725

b. Coverage rate

	AR1 Structure		*EXP Structure*	
Parameter	*MFM-SCS*	*MFM*	*MFM-SCS*	*MFM*
ρ_p^{AGG}	.9640	.4180	.9600	.4380
ρ_p^{WPC}	.9540	.0000	.9380	.0000

Note. AR1 = first-order autoregressive structure; EXP = exponential; MFM-SCS = multilevel factor model with serial correlation structure; MFM = multilevel factor model.

for 8 days. This data collection method, called the diary method, was a special case of EMA. Daily positive affect, the piece of data chosen for present illustration, was measured by three items from the Positive and Negative Affect Schedule (PANAS; Watson, Clark, & Tellegen, 1988): excited, proud, and enthusiastic. These items were chosen based on the study by Egloff, Schmukle, Kohlman, Burns, & Hock (2003), which argued that the three items reflected the feeling of joy. Every day, participants rated the items according to their daily average joy level by using a 5-point Likert-type scale, with 1 representing *very slightly* and 5 representing *extremely.* All the factor loadings, intercepts, and measurement error variances were set free with measurement errors assumed to be independent. For identification, the variance of the latent factor "joy" was set to one at each level. AR1 was chosen to model serial correlation. The estimates and CIs of ρ's are reported in Table 4. The significance of $\hat{\rho}^{ICC}$ indicates salient interindividual differences on the composite score among participants. Nearly 19% of the daily composite score variance can be attributed to interindividual differences. This significance also justifies the necessity of using multilevel modeling. $\hat{\rho}^{LV2} = 0.90$ and $\hat{\rho}^{LV1} = 0.78$ gave the "ideal" values of $\hat{\rho}_i^{AGG}$ and $\hat{\rho}_{ij}^{WPC}$ as J goes to infinity. When J is small, the estimated values may

TABLE 4
Parameter Estimates and Confidence Intervals (CI) of ρs for the Illustrated Example

Coefficient	$\hat{\rho}$	*95% CI*
ρ^{ICC}	0.19	[0.12, 0.26]
ρ^{LV2}	0.90	[0.81, 0.95]
ρ^{LV1}	0.78	[0.75, 0.81]
ρ_i^{AGG}	0.54	[0.39, 0.66]
$\rho_{i1}^{WPC} = \rho_{i8}^{WPC}$	0.69	[0.66, 0.71]
$\rho_{i2}^{WPC} = \rho_{i7}^{WPC}$	0.71	[0.68, 0.73]
$\rho_{i3}^{WPC} = \rho_{i6}^{WPC}$	0.71	[0.68, 0.74]
$\rho_{i4}^{WPC} = \rho_{i5}^{WPC}$	0.71	[0.68, 0.74]
ρ_p^{AGG}	0.54	[0.39, 0.66]
ρ_p^{WPC}	0.70	[0.68, 0.73]

be quite different from the ideal ones. Because the serial correlation matrices are homogeneous in this example, all participants share the same ρ_i^{AGG} and $\rho_i^{AGG} = \rho_p^{AGG}$. The value of $\hat{\rho}_i^{AGG}$ indicates that about 54% of the variance of AGGS can be explained by its true score, the trait level of joy. If a researcher is unsatisfied with such a relatively low value, it is possible to obtain a more reliable AGGS by adding more observation timepoints. By simple calculation, it can be shown that $\hat{\rho}_i^{AGG}$ becomes 0.73 when $J = 24$ under the current parameter estimates. Intuitively, if J becomes larger, we will have more behavior samples across time, thereby allowing AGGS to average out the occasional effect and become a more reliable score for the latent trait. Although the serial correlation matrices are homogeneous, the values of $\hat{\rho}_{ij}^{WPC}$'s still depend on the timepoint j due to the significance of the serial correlation parameter ($\hat{\beta} = 0.1$, 95% CI = [0.50, 1.50]). The values of $\hat{\rho}_{ij}^{WPC}$'s range from 0.69 to 0.71, which represents only a slight difference. The values of $\hat{\rho}_{ij}^{WPC}$'s indicated that the occasional joy level can account for about 70% of the variance of WPCS. The pooled version $\hat{\rho}_p^{WPC} = 0.70$ provided almost the same information about the reliability of WPCS in this example. In conclusion, WPCS seems to be relatively reliable, but the reliability of AGGS can be further improved by adding observation timepoints.

One may wonder about the consequences of ignoring serial correlation in this example. The same data were analyzed by traditional MFM and the reliability coefficients were calculated according to Equation 9 and Equation 11. The resulting estimates were $\hat{\rho}_i^{AGG} = 0.59$ and $\hat{\rho}_{ij}^{WPC} = 0.68$. Compared with the result of MFM-SCS, the difference was only slight. This is due to the small amount of serial correlation in this example.

DISCUSSION

In this study, we extend previous works in two directions. First, MFM-SCS was proposed as an extension of traditional MFM (Lee, 1990; McDonald & Goldstein, 1989; Muthén, 1989). Recently, some researchers started using traditional MFM to model trait and state variation (e.g., Merz & Roesch, 2011; Roesch et al., 2010). In their studies, the serial correlation was assumed to be zero implicitly. However, our simulation study shows that the ignorance of serial correlation may result in biased parameter estimates. We believe that MFM-SCS will be a more suitable alternative. Second, two reliability coefficients, ρ_i^{AGG} and ρ_{ij}^{WPC}, and their pooled versions, ρ_p^{AGG} and ρ_p^{WPC}, are proposed to quantify how reliable AGGS and WPCS are in a given EMA study. The interpretation of these coefficients is basically the same as that of the traditional reliability coefficients. From a theoretical perspective, our work extends reliability estimation methods to test scores that reflect intraindividual differences, which had been thought to be measurement errors (Biemer et al., 2009; Laenen et al., 2007; Laenen et al., 2009; Laenen et al., 2006). The proposed reliability coefficient can also be seen as a more general form of some previous ones to accommodate the presence of serial correlation (Raudenbush et al., 1991; Snijders & Bosker, 1999). We recommend that researchers evaluate the reliabilities of AGGS and WPCS before these scores are used for further statistical analysis. A detailed analytic strategy has been demonstrated in the real data example provided. When several measurements are used in an EMA study, the proposed method can be implemented separately for each measure. Whether both reliability coefficients should be reported depends on the purpose of the EMA study. Shiffman et al. (2008) have mentioned that EMA data are collected for four purposes: (a) characterizing individual differences, (b) describing natural history, (c) assessing contextual associations, and (d) documenting temporal sequences. Clearly, the first purpose is related to interindividual differences whereas the others are related to intraindividual differences. Therefore, if only one type of individual difference is involved, it is sufficient to report only the corresponding reliability coefficient. Of course, if both types are involved, then both coefficients should be reported.

When the proposed procedure is used, some cautions should be taken. First, it should be recognized that the proposed procedure only concerns the reliabilities of AGGS and WPCS. Therefore, the procedure is only suitable for EMA studies that use AGGS and WPCS and is not applicable to EMA studies that use other types of test scores, such as scores that represent the instability of human behavior (Jahng, Wood, & Trull, 2008). Second, the total number of items K plays an important role in our procedure. $K \geq 3$ is necessary for model identification.[2] When the items are all congeneric, the value of ρ_i^{AGG} and ρ_{ij}^{WPC} are both increasing functions of K. However, a large K may be impractical in EMA settings. Researchers should choose appropriate values of K in their studies. Third, before estimating ρ_i^{AGG} and ρ_{ij}^{WPC}, the overall model fit should be evaluated carefully. If the overall model fit is poor, the parameter estimates may be meaningless and the resulting reliability coefficient estimates will be biased (Yang & Green, 2010). A strategy for model evaluation in multilevel settings has been discussed by Ryu and West (2009).

Furthermore, sample size is critical for the quality of estimates. Our simulation shows that the proposed procedure performs well when sample size is 100. In real EMA studies, data sets may be of smaller sample sizes. Previous simulations have shown that the estimated standard errors for the Level 2 parameters may be biased when the sample size is smaller than 100 in multilevel modeling (Maas & Hox, 2005). Therefore, researchers should interpret their results carefully when the sample size is small. Finally, the MLE is established on the normality assumption. In practice, data are seldom normally distributed. When the normal assumption is violated but the model is correctly specified, the consistency of parameter estimators can still be achieved (Shapiro, 1984). However, the estimated standard errors and the limiting chi-square distribution of the test statistic will become invalid. In such a case, the nonparametric bootstrap CIs for the proposed coefficients can be considered as an alternative. Bootstrap is a data-based simulation method for statistical inference (Efron & Tibshirani, 1993) and has been used to construct CIs for reliability coefficients (Raykov & Shrout, 2002).

Although our procedure represents an improvement over those of previous studies, future research in several directions can further improve this procedure. First, only the one-factor model was considered in this study. MFM-SCS can be extended to multiple-factor models to accommodate more complicated factor structures. Second, although the proposed procedure performed well in our simulation study, whether the same conclusion can be reached under other simulation conditions awaits further evaluation. Third, the simulation study shows that ignoring serial correlation may result in biased estimates. However, the degree of bias seems negligible when the magnitude of serial correlation is of a small value as it is in the real data example. Further studies can be conducted to explore the possible influences of the size of the serial correlation and to understand what magnitudes of serial correlation should be considered nonignorable.

[2]If $K = 2$, further constraints on parameters are needed. For example, the factor loadings at each level could be constrained to be equal.

REFERENCES

Bentler, P. M. (2009). Alpha, dimension-free, and model-based internal consistency reliability. *Psychometrika, 74,* 137–143.

Biemer, P. P., Christ, S. L., & Wiesen, C. A. (2009). A general approach for estimating scale score reliability for panel survey data. *Psychological Methods, 14,* 400–412.

Bolger, N., & Davis, A., & Rafaeli, E. (2003). Diary methods: Capturing life as it is lived. *Annual Review of Psychology, 54,* 579–616.

Casella, G., & Berger, R. L. (2002). *Statistical Inference* (2nd ed.). New York, NY: Duxbury.

Cheung, M. W. L. (2009). Constructing approximate confidence intervals for parameters with structural equation models. *Structural Equation Modeling, 16,* 267–294.

Cole, D. A., Martin, N. C., & Steiger, J. H. (2005). Empirical and conceptual problems with longitudinal trait-state models: Introducing a trait-state-occasion model. *Psychological Methods, 10,* 3–20.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297–334.

Csikszentmihalyi, M., & Larson, R. (1987). Validity and reliability of the experience sampling method. *The Journal of Nervous and Mental Disease, 175,* 526–536.

Curran, P. J., & Bauer, D. J. (2011). The disaggregation of within-person and between-person effects in longitudinal models of change. *Annual Review of Psychology, 62,* 583–619.

Efron, B., & Tibshirani, R. (1993). *An Introduction to the bootstrap.* New York, NY: Chapman & Hall.

Egloff, B., Schmukle, S. C., Kohlman, C. W., Burns, L. R., & Hock, M. (2003). Facets of dynamic positive affect: Differentiation joy, interest and anticipation in the positive and negative affect schedule (PANAS). *Journal of Personality and Social Psychology, 85,* 528–540.

Everitt, B. S. (2002). *Cambridge dictionary of statistics* (2nd ed.). Cambridge, UK: Cambridge University Press.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105–146). New York, NY: American Council on Education and Macmillan.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 10,* 255–282.

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65–110). Westport, CT: American Council on Education and Praeger.

Hektner, J. M., Schmidt, J. A., & Csikszentmihalyi, M. (2007). *Experience sampling method: Measuring the quality of everyday life.* Thousand Oaks, CA: Sage.

Huang, P. H. (2009). *Depression and mood congruent thoughts in everyday life: A diary study.* (Master's thesis, National Chengchi University, Taipei, Taiwan).

Hufford, M. R. (2007). Special methodological challenges and opportunities in ecological momentary assessment. In A. A. Stone, S. Shiffman, A. Atienza, & L. Nebeling (Eds.), *The science of real-time data capture: Self-reports in health research* (pp. 54–75). New York, NY: Oxford University Press.

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,* 354–375.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika, 36,* 109–133.

Kenny, D. A., & Judd, C. M. (1986). Consequences of violating the independence assumption in analysis of variance. *Psychological Bulletin, 99,* 422–431.

Kollo, T., & von Rosen, D. (2005). *Advanced multivariate statistics with matrices.* Dordrecht, The Netherlands: Springer.

Laenen, A. (2008). *Psychometric validation of continuous rating scales from complex data.* Doctoral thesis, Universiteit Hasselt, Diepenbeek, Belgium).

Laenen, A., Alonso, A., & Molenberghs, G. (2007). A measure for the reliability of a rating scale based on longitudinal clinical trial data. *Psychometrika, 72,* 443–448.

Laenen, A., Alonso, A., Molenberghs, G., & Vangeneugden, T. (2009). Reliability of a longitudinal sequence of scale ratings. *Psychometrika, 74,* 49–64.

Laenen, A., Vangeneugden, T., Geys, H., & Molenberghs, G. (2006). Generalized reliability estimation using repeated measurements. *British Journal of Mathematical and Statistical Psychology, 59,* 113–131.

Lee, S. Y. (1990). Multilevel analysis of structural equation models. *Biometrika, 77,* 763–772.

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

Maas, C., & Hox, J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 1,* 86–92.

McDonald, R. P., & Goldstein, H. (1989). Balanced versus unbalanced designs for linear structural relations in two-level data. *British Journal of Mathematical and Statistical Psychology, 42,* 215–232.

Mehta, P. D., & West, S. G. (2000). Putting the individual back into individual growth curves. *Psychological Methods, 5,* 23–43.

Merz, E. L., & Roesch, S. C. (2011). Modeling trait and state variation using multilevel factor analysis with PANAS daily diary data. *Journal of Research in Personality, 45,* 2–9.

Molenaar, P. C. M. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement, 2,* 201–218.

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika, 54,* 557–585.

Muthén, B., & Satorra, A. (1995). Complex sample data in structural equation modeling. In P. V. Marsden (Ed.), *Sociological methodology* (pp. 267–316). Washington, DC: American Sociological Association.

Neale, M. C., Boker, S. M., Xie, G., & Maes, H. H. (2004). *Mx: Statistical modeling* (5th ed.). Richmond: Virginia Commonwealth University, Department of Psychiatry.

Neale, M. C., & Miller, M. B. (1997). The use of likelihood-based confidence intervals in genetic models. *Behavior Genetics, 27,* 113–120.

R Development Core Team. (2010). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models* (2nd ed.). Newbury Park, CA: Sage.

Raudenbush, S. W., Rowan, B., & Kang, S. J. (1991). A multilevel, multivariate model for studying school climate in secondary schools with estimation via the EM algorithm. *Journal of Educational Statistics, 16,* 295–330.

Raykov, T., & du Toit, S. H. C. (2005). Evaluation of reliability for multiple component measuring instruments in hierarchical designs. *Structural Equation Modeling, 12,* 536–550.

Raykov, T., & Shrout, P. E. (2002). Reliability of scales with general structure: Point and interval estimation using a structural equation modeling approach. *Structural Equation Modeling, 9,* 195–212.

Roesch, S. C., Aldridge, A. A., Stocking, S. N., Villodas, F., Leung, Q., Bartley, C. E., & Black, L. J. (2010). Multilevel factor analysis and structural equation modeling of daily diary coping data: Modeling trait and state variation. *Multivariate Behavioral Research, 45,* 767–789.

Ryu, E., & West, S. G. (2009). Level-specific evaluation of model fit in multilevel structural equation modeling. *Structural Equation Modeling, 16,* 583–601.

Schwartz, J. E., & Stone, A. A. (1998). Data analysis for EMA studies. *Health Psychology, 17,* 6–16.

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). New York, NY: Oxford University Press.

Schwarz, N. (2007). Retrospective and concurrent self-reports: The rationale for real-time data capture. In A. A. Stone, S. S. Shiffman, A. Atienza, & L. Nebeling (Eds.), *The science of real-time data capture: Self-reports in health research* (pp. 11–26). New York, NY: Oxford University Press.
Shapiro, A. (1984). A note on the consistency of estimators in the analysis of moment structures. *British Journal of Mathematical and Statistical Psychology, 37,* 84–88.
Shiffman, S. (2007). Designing protocols for ecological momentary assessment. In A. A. Stone, S. Shiffman, A. Atienza, & L. Nebeling (Eds.), *The science of real-time data capture: Self-reports in health research* (pp. 27–53). New York, NY: Oxford University Press.
Shiffman, S. M., & Stone, A. A. (1998). Introduction to the special section: Ecological momentary assessment in health psychology. *Health Psychology, 17,* 3–5.
Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological momentary assessment. *Annual Review of Clinical Psychology, 4,* 1–32.
Shiyko, M. P., & Ram, N. (2011). Conceptualizing and estimating process speed in studies employing ecological momentary assessment designs: A multilevel variance decomposition approach. *Multivariate Behavioral Research, 46,* 875–899.
Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling.* London, UK: Sage.
Stone, A. A., & Shiffman, S. (1994). Ecological momentary assessment (EMA) in behavioral medicine. *Annals of Behavioral Medicine, 16,* 199–202.
Stone, A. A., Shiffman, S., Atienza, A., & Nebeling, L. (Eds.) (2007). *The science of real-time data capture: Self-reports in health research.* New York, NY: Oxford University Press.
Tennen, H., Affleck, G., & Armeli, S. (2005). Personality and daily experience revisited. *Journal of Personality, 73,* 1–19.
Trull, J. T., & Ebner-Priemer, U. W. (2009). Using experience sampling methods/ecological momentary assessment (ESM/EMA) in clinical assessment and clinical research: Introduction to the special section. *Psychological Assessment, 21,* 457–462.
Verbeke, G., & Molenberghs, G. (2000). *Linear mixed models for longitudinal data.* New York, NY: Springer.
Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology, 54,* 1063–1070.
Wenze, S. J., & Miller, I. V. (2010). Use of ecological momentary assessment in mood disorders research. *Clinical Psychology Review, 30,* 794–804.
Yang, Y., & Green, S. B. (2010). A note on structural equation modeling estimates of reliability. *Structural Equation Modeling, 17,* 66–81.

APPENDIX A

The Derivation of the Likelihood Function of Multilevel Factor Model with Serial Correlation Structure (MFM-SCS)

Let $\mathbf{y}_i^* = [\mathbf{y}_{i1}^T, \mathbf{y}_{i1}^T, \ldots, \mathbf{y}_{iJ}^T]^T$ denote the complete response vector and $vec(\mathbf{V}_i) = [\mathbf{v}_{i1}^T, \mathbf{v}_{i2}^T, \ldots, \mathbf{v}_{iJ}^T]^T$ denote the "vector form" of $\mathbf{V}_i$. Then $\mathbf{y}_i^*$ can be written as

$$\mathbf{y}_i^* = \mathbf{1} \otimes \mathbf{v}_i + vec(\mathbf{V}_i), \tag{A1}$$

where $\mathbf{1}$ is the J-dimensional vector of 1's and $\otimes$ is the Kronecker product operator. Based on the conditions that (a) $\mathbf{V}_i \sim MN_{K\times J}(\mathbf{0}, \mathbf{\Sigma}_w, \mathbf{\Phi}_i)$ implies $vec(\mathbf{V}_i) \sim N(\mathbf{0}, \mathbf{\Phi}_i \otimes \mathbf{\Sigma}_w)$ (see Kollo & von Rosen, 2005) and (b) $\mathbf{v}_i \sim N(\boldsymbol{\alpha}, \mathbf{\Sigma}_b)$, the model-implied mean and covariance structure for $\mathbf{y}_i^*$ will be

$$\begin{aligned} \boldsymbol{\mu}_i^*(\boldsymbol{\theta}) &= \mathbf{1} \otimes \boldsymbol{\alpha}, \\ \mathbf{\Sigma}_i^*(\boldsymbol{\theta}, \boldsymbol{\beta}) &= \mathbf{1}\mathbf{1}^T \otimes \mathbf{\Sigma}_b + \mathbf{\Phi}_i \otimes \mathbf{\Sigma}_w. \end{aligned} \tag{A2}$$

We can substitute the mean and covariance structures into the multivariate normal density function to construct the likelihood function

$$\log L(\boldsymbol{\theta}, \boldsymbol{\beta}) = -\frac{1}{2}\sum_{i=1}^{N}\{\log|\mathbf{\Sigma}_i^*(\boldsymbol{\theta}, \boldsymbol{\beta})| + [\mathbf{y}_i^* - \boldsymbol{\mu}_i^*(\boldsymbol{\theta})]^T \mathbf{\Sigma}_i^*(\boldsymbol{\theta}, \boldsymbol{\beta})^{-1}[\mathbf{y}_i^* - \boldsymbol{\mu}_i^*(\boldsymbol{\theta})]\}. \tag{A3}$$

The maximum likelihood estimates of $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ can be obtained by maximizing this function with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$.

APPENDIX B

The Derivations of ρ_i^{AGG} and ρ_{ij}^{WPC}

Because $y_i^{AGG} = J^{-1}\Sigma_{j=1}^{J}\mathbf{1}^T\mathbf{y}_{ij}$ and $\tau_i^{AGG} = \mathbf{1}^T\boldsymbol{\alpha} + \mathbf{1}^T\boldsymbol{\lambda}_b\eta_i$, the reliability coefficient for y_i^{AGG} is

$$\begin{aligned} \rho_i^{AGG} &= \frac{Var(\tau_i^{AGG})}{Var(y_i^{AGG})} \\ &= \frac{Var(\mathbf{1}^T\boldsymbol{\alpha} + \mathbf{1}^T\boldsymbol{\lambda}_b\eta_i)}{Var(\mathbf{1}^T\mathbf{v}_i) + Var(J^{-1}\Sigma_{j=1}^{J}\mathbf{1}^T\mathbf{v}_{ij}) + 2Cov(\mathbf{1}^T\mathbf{v}_i, J^{-1}\Sigma_{j=1}^{J}\mathbf{1}^T\mathbf{v}_{ij})} \\ &= \frac{\mathbf{1}^T\boldsymbol{\lambda}_b\psi_b\boldsymbol{\lambda}_b^T\mathbf{1}}{\mathbf{1}^T\Sigma_b\mathbf{1} + J^{-2}\Sigma_j\Sigma_{j'}\phi_{ijj'}\mathbf{1}^T\mathbf{\Sigma}_w\mathbf{1}}. \end{aligned} \tag{B1}$$

After dividing both the numerator and the denominator of Equation B1 by $\mathbf{1}^T(\mathbf{\Sigma}_b + \mathbf{\Sigma}_w)\mathbf{1}$, ρ_i^{AGG} becomes

$$\rho_i^{AGG} = \frac{\rho^{ICC}}{\rho^{ICC} + (1 - \rho^{ICC})J^{-2}\Sigma_{j=1}^{J}\Sigma_{j'=1}^{J}\phi_{ijj'}}\rho^{LV2}. \tag{B2}$$

Similarly, because $y_{ij}^{WPC} = \mathbf{1}^T \mathbf{y}_{ij} - J^{-1} \Sigma_{j'=1}^{J} \mathbf{1}^T \mathbf{y}_{ij'}$ and $\tau_{ij}^{WPC} = \frac{J-1}{J} \mathbf{1}^T \boldsymbol{\lambda}_w \eta_{ij}$, the reliability coefficient for y_{ij}^{WPC} will be

$$
\begin{aligned}
\rho_{ij}^{WPC} &= \frac{Var(\tau_{ij}^{WPC})}{Var(y_{ij}^{WPC})} \\
&= \frac{Var\left(\frac{J-1}{J}\mathbf{1}^T \lambda_w \eta_{ij}\right)}{Var(\mathbf{1}^T \mathbf{y}_{ij}) + Var(J^{-1}\Sigma_{l=1}^{J}\mathbf{1}^T \mathbf{y}_{il}) - 2Cov(\mathbf{1}^T \mathbf{y}_i, J^{-1}\Sigma_{l=1}^{J}\mathbf{1}^T \mathbf{y}_{il})} \\
&= \frac{\left(\frac{J-1}{J}\right)^2 \mathbf{1}^T \boldsymbol{\lambda}_\mathbf{w} \psi_\mathbf{w} \boldsymbol{\lambda}_\mathbf{w}^T \mathbf{1}}{\mathbf{1}^T \Sigma_\mathbf{w} \mathbf{1} + J^{-2} \Sigma_l \Sigma_{l'} \phi_{ill'} \mathbf{1}^T \boldsymbol{\Sigma}_\mathbf{w} \mathbf{1} - 2J^{-1} \Sigma_l \phi_{ijl} \mathbf{1}^T \boldsymbol{\Sigma}_\mathbf{w} \mathbf{1}} \\
&= \frac{\left(\frac{J-1}{J}\right)^2 \mathbf{1}^T \boldsymbol{\lambda}_\mathbf{w} \psi_\mathbf{w} \boldsymbol{\lambda}_\mathbf{w}^T \mathbf{1}}{(1 + J^{-2} \Sigma_l \Sigma_{l'} \phi_{ill'} - 2J^{-1} \Sigma_l \phi_{ijl}) \mathbf{1}^T \boldsymbol{\Sigma}_\mathbf{w} \mathbf{1}} \\
&= \frac{(J-1)^2}{(J^2 + \Sigma_l \Sigma_{l'} \phi_{ill'} - 2J \Sigma_l \phi_{ijl})} \rho^{LV1}
\end{aligned}
\tag{B3}
$$