Estimation of Test Reliability in Intensive Longitudinal Studies: A Multilevel Structure and Dynamic Characteristics Perspective

Authors: Luo Xiaohui, Liu Hongyun, Liu Hongyun

Date: 2023-11-28T00:00:00+00:00

Abstract

With the extensive application of intensive longitudinal research in psychology and other social science fields, the estimation of test reliability in intensive longitudinal contexts has also attracted increasing attention from researchers. Early methods that followed the reliability estimation ideas from cross-sectional research or were based on generalizability theory have numerous limitations and are not suitable for intensive longitudinal contexts. Addressing the two key characteristics of intensive longitudinal data—multilevel structure and dynamic nature—test reliability in intensive longitudinal research can be estimated based on multilevel confirmatory factor analysis, dynamic factor analysis, and dynamic structural equation modeling. Through demonstration and comparison with empirical data, the characteristics and applicable contexts of the three estimation methods are discussed. Future research could explore the estimation of test reliability based on other intensive longitudinal models and should also emphasize the examination and reporting of test reliability.

Full Text

Estimating Test Reliability in Intensive Longitudinal Studies: Perspectives on Multilevel Structure and Dynamic Nature

LUO Xiaohui, LIU Hongyun
Beijing Key Laboratory of Applied Experimental Psychology, National Demonstration Center for Experimental Psychology Education (Beijing Normal University), Faculty of Psychology, Beijing Normal University, Beijing 100875, China

Abstract

With the widespread application of intensive longitudinal studies in psychology and other social sciences, the estimation of test reliability in intensive longitudinal contexts has attracted increasing attention from researchers. Early approaches that borrowed reliability estimation concepts from cross-sectional studies or relied on generalizability theory suffer from numerous limitations and are ill-suited for intensive longitudinal designs. Considering the two key characteristics of intensive longitudinal data—multilevel structure and dynamic nature—test reliability in intensive longitudinal studies can be estimated using multilevel confirmatory factor analysis, dynamic factor analysis, and dynamic structural equation modeling. Through demonstration and comparison with empirical data, this paper discusses the features and applicable contexts of these three estimation methods. Future research should explore reliability estimation based on other intensive longitudinal models and place greater emphasis on the testing and reporting of test reliability.

Keywords: intensive longitudinal study, reliability, multilevel structure, dynamic nature, dynamic structural equation modeling

1. Introduction

Intensive longitudinal studies have become increasingly prevalent in social science fields such as psychology, education, and management. These studies typically employ diary methods, experience sampling methods, and ecological momentary assessment to collect data from individuals at multiple time points in natural settings (e.g., more than 20 time points; Collins, 2006). Compared to traditional retrospective surveys and laboratory research, intensive longitudinal designs offer advantages including reduced recall bias and enhanced ecological validity (Bolger et al., 2003; Shiffman et al., 2008; Trull & Ebner-Priemer, 2013). More importantly, the high-frequency, repeated measurements enable researchers to capture fine-grained temporal changes in individuals' behaviors and states, facilitating deeper exploration of dynamic processes and interaction mechanisms among variables (Zheng et al., 2021; Hamaker & Wichers, 2017; Zhou et al., 2021).

Despite these benefits, intensive longitudinal studies present significant methodological challenges (Hamaker & Wichers, 2017), particularly regarding variable measurement and test evaluation (Mielniczuk, 2023). Many intensive longitudinal studies use self-report measures to assess daily behaviors and states, typically adapting items from trait measures by adding temporal cues such as "today" or "since the last response." However, most such studies fail to adequately evaluate the psychometric properties, including reliability, of their measures (Stone et al., 2023). For instance, Brose et al. (2020) reviewed 50 emotion-related intensive longitudinal studies published in *Emotion* between 2005 and September 2017, finding that 29 reported test reliability but only 10 explicitly mentioned

that reliability estimates were based on within-person variation. Similarly, Trull and Ebner-Priemer (2020) reviewed 63 intensive longitudinal studies published in major psychopathology journals from 2012 to 2018, revealing that only 30% reported psychometric information (e.g., reliability and validity). Furthermore, Horstmann and Ziegler (2020) examined 24 intensive longitudinal studies on personality states and found that most simply adapted items or adjectives from trait measures without pre-testing their psychometric properties. The most common reliability estimation method involved averaging each individual's responses across all time points for each item and then computing inter-item consistency for the entire sample. This approach, however, cannot capture the reliability of state scores and is inappropriate for intensive longitudinal contexts (Horstmann & Ziegler, 2020). Given that reliability assessment is a critical step in data analysis and reporting and serves as an important basis for evaluating result credibility (Ye et al., 2012; Scherer & Teo, 2020), it is essential to develop and adopt appropriate reliability estimation methods tailored to the characteristics of intensive longitudinal data.

Early explorations of reliability estimation in intensive longitudinal contexts primarily fell into two categories. The first category borrowed reliability estimation concepts from cross-sectional studies by aggregating or splitting intensive longitudinal data to resemble cross-sectional patterns, then applying conventional reliability indices (e.g., Cronbach's alpha). Specifically, three approaches have been used (Nezlek, 2017): (1) aggregating each individual's scores across all time points for each item (e.g., computing means) and then calculating reliability using these aggregated scores; (2) splitting data by time point and computing reliability separately for each time point, then averaging across time points; and (3) splitting data by individual and computing reliability separately for each individual, then averaging across individuals. These methods suffer from notable limitations: the first assesses between-person reliability rather than within-person dynamic change reliability; the second ignores that different individuals are assessed at different time points, making it inappropriate to combine reliability estimates across time; and the third violates the assumption of independence of observations required for reliability calculation by ignoring the dependency among repeated measures from the same individual. Consequently, none of these methods are suitable for intensive longitudinal reliability estimation.

The second category of approaches is based on generalizability theory (Cronbach et al., 1963). These methods first identify measurement facets (e.g., person, time, item) to examine sources of measurement error, then use analysis of variance to estimate variance components attributable to each facet and their interactions, from which different reliability coefficients are computed. For example, Cranford et al. (2006) proposed that observed score variation in intensive longitudinal studies could be attributed to person, time, and item facets, and developed multiple reliability formulas based on different assumptions about fixed versus random effects for each facet. Subsequent researchers extended this approach to contexts with additional measurement facets (Schönbrodt et al.,

2021). However, generalizability theory-based methods also have limitations (Scherer & Teo, 2020). They require strong assumptions such as equal factor loadings across individuals and time-invariant error variances—conditions rarely met in practice, potentially leading to inaccurate reliability estimates (Lane & Shrout, 2010). Therefore, these methods are also unsuitable for intensive longitudinal contexts, and previous research has advised against their application in longitudinal studies (Ye et al., 2012).

As understanding of intensive longitudinal data has deepened, researchers have developed more targeted reliability estimation methods that address the unique characteristics of these data, particularly their multilevel structure and dynamic nature (Hamaker & Wichers, 2017; Lafit et al., 2021). The multilevel structure refers to repeated measurements (Level 1) nested within individuals (Level 2), while the dynamic nature reflects that observations from adjacent time points are not independent but correlated. Focusing on these two characteristics, new approaches to reliability estimation have emerged.

2. Reliability Estimation Methods Focusing on Multilevel Structure

Multilevel confirmatory factor analysis (MCFA; Geldhof et al., 2014) provides a reliability estimation approach that focuses on the multilevel structure of intensive longitudinal data by estimating reliability separately at within-person and between-person levels. This method has been widely applied in intensive longitudinal research across developmental (Eltanamly et al., 2023; Xu & Zheng, 2022), educational (Hausen et al., 2023; Neubauer et al., 2022), social (Di Sarno et al., 2020; Koval et al., 2019), clinical health (Gerstberger et al., 2023; Van Der Tuin et al., 2023; Wright et al., 2017), and organizational management (Reis et al., 2016; Schmitt et al., 2017) domains.

The MCFA-based reliability estimation method applies to both unidimensional and multidimensional measurement structures. This paper uses the unidimensional case as an example (see Di Sarno et al., 2020; Neubauer et al., 2022; Wright et al., 2017 for multidimensional cases). When both within-person and between-person levels have unidimensional structures [Figure 1: see original paper], MCFA decomposes the observed score Y_{jti} for item j from person i at time t ($j = 1, 2, \ldots, q$; $t = 1, 2, \ldots, T$; $i = 1, 2, \ldots, n$) into between-person (Y_{ji}) and within-person ($Y_{jti}^{(w)}$) components:

$$Y_{jti} = Y_{ji} + Y_{jti}^{(w)}$$

The within-person component is further decomposed into a true score ($S_{ti}^{(w)}$) and error ($\epsilon_{jti}^{(w)}$):

$$Y_{jti}^{(w)} = \lambda_j^{(w)} S_{ti}^{(w)} + \epsilon_{jti}^{(w)}$$

where $S_{ti}^{(w)}$ is the latent state factor for person i at time t, $\lambda_j^{(w)}$ is the item j factor loading at the within-person level (assumed equal across persons and time-invariant), and $\epsilon_{jti}^{(w)}$ is random measurement error, assumed normally distributed ($\epsilon_{jti}^{(w)} \sim N(0, \sigma^2_{\epsilon_j^{(w)}})$) with zero covariance between items ($\text{cov}(\epsilon_{jti}^{(w)}, \epsilon_{j'ti}^{(w)}) = 0$ for $j \neq j'$).

The between-person component is decomposed as:

$$Y_{ji} = \nu_j + \lambda_j^{(B)} T_i + \epsilon_{ji}^{(B)}$$

where T_i is the latent trait factor for person i, ν_j is the item j intercept, $\lambda_j^{(B)}$ is the item j factor loading at the between-person level, and $\epsilon_{ji}^{(B)}$ is measurement error, assumed normally distributed ($\epsilon_{ji}^{(B)} \sim N(0, \sigma^2_{\epsilon_j^{(B)}})$) with zero covariance between items ($\text{cov}(\epsilon_{ji}^{(B)}, \epsilon_{j'i}^{(B)}) = 0$ for $j \neq j'$).

Based on this model, reliability can be computed for each item and each dimension at both levels. At the within-person level, item-specific within-person reliability is defined as the ratio of variance in the item's state component explained by the latent state factor to the total variance of the item's state component. Dimension-level within-person reliability is the ratio of total variance explained by the latent state factor across all items in the dimension to the total variance of the state components for those items. With the latent state factor variance fixed to 1, item reliability ($\text{Rel}_j^{(w)}$) and dimension reliability ($\text{Rel}^{(w)}$) are:

$$\text{Rel}_j^{(w)} = \frac{(\lambda_j^{(w)})^2}{(\lambda_j^{(w)})^2 + \text{var}(\epsilon_{jti}^{(w)})}$$

$$\text{Rel}^{(w)} = \frac{\sum_{j=1}^{q} (\lambda_j^{(w)})^2}{\sum_{j=1}^{q} (\lambda_j^{(w)})^2 + \sum_{j=1}^{q} \text{var}(\epsilon_{jti}^{(w)})}$$

Similarly, at the between-person level, item-specific between-person reliability is the ratio of variance in the item's trait component explained by the latent trait factor to the total variance of the item's trait component. Dimension-level between-person reliability is the ratio of total variance explained by the latent trait factor across all items in the dimension to the total variance of the trait components for those items. With the latent trait factor variance fixed to 1, item reliability ($\text{Rel}_j^{(B)}$) and dimension reliability ($\text{Rel}^{(B)}$) are:

$$\text{Rel}_j^{(B)} = \frac{(\lambda_j^{(B)})^2}{(\lambda_j^{(B)})^2 + \text{var}(\epsilon_{ji}^{(B)})}$$

$$\text{Rel}^{(B)} = \frac{\sum_{j=1}^{q}(\lambda_j^{(B)})^2}{\sum_{j=1}^{q}(\lambda_j^{(B)})^2 + \sum_{j=1}^{q}\text{var}(\epsilon_{ji}^{(B)})}$$

Although widely used, the MCFA-based approach has limitations. It assumes equal factor loadings and residual variances across all individuals, yielding only an overall assessment of within-person reliability. This assumption may not hold in practice, as test reliability likely varies across individuals in intensive longitudinal studies (Hu et al., 2016). Additionally, MCFA ignores the temporal dependencies among consecutive observations, neglecting the dynamic nature of intensive longitudinal data, which may compromise the accuracy of reliability estimates.

3. Reliability Estimation Methods Focusing on Dynamic Nature

Dynamic factor analysis (DFA) offers another important reliability estimation approach for intensive longitudinal studies. Originally proposed by Molenaar (1985), DFA extends P-technique factor analysis (Cattell et al., 1947) by incorporating time series analysis, enabling the modeling of person-specific dynamic processes. Researchers have applied this method to reliability estimation in intensive longitudinal studies (Fuller-Tyszkiewicz et al., 2017; Lane & Shrout, 2010). DFA accounts for the autoregressive processes of variables, thereby capturing the dynamic nature of intensive longitudinal data. It also allows for person-specific reliability estimation based on individual data, helping researchers understand between-person differences in test reliability.

The DFA-based reliability estimation method builds a separate dynamic factor model for each individual to compute person-specific reliability. Like MCFA, DFA applies to both unidimensional and multidimensional structures; this paper uses the unidimensional case as an example (see Fuller-Tyszkiewicz et al., 2017 for multidimensional cases). Person i's dynamic factor model consists of measurement and structural components [Figure 2: see original paper]. The measurement model is:

$$Y_{jti} = \nu_{ji} + \lambda_{ji}F_{ti} + \epsilon_{jti}$$

where Y_{jti} is the observed score for item j from person i at time t ($j = 1, 2, \ldots, q$; $t = 1, 2, \ldots, T$; $i = 1, 2, \ldots, n$), ν_{ji} is the item j intercept for person i, λ_{ji} is the item j factor loading for person i, F_{ti} is the latent factor for person i at time t, and ϵ_{jti} is measurement error, assumed normally distributed ($\epsilon_{jti} \sim N(0, \sigma^2_{\epsilon_{ji}})$) with zero covariance between items ($\text{cov}(\epsilon_{jti}, \epsilon_{j'ti}) = 0$ for $j \neq j'$).

The structural component assumes the latent factor follows a first-order autoregressive process:

$$F_{ti} = \phi_i F_{t-1,i} + \zeta_{ti}$$

where ϕ_i is the person-specific autoregressive effect (also called inertia or carry-over effect), describing how the previous time point's latent factor level influences the current level, and ζ_{ti} is dynamic error at time t, assumed normally distributed ($\zeta_{ti} \sim N(0, \sigma^2_{\zeta_i})$).

Based on this model, person-specific reliability can be computed for each item and dimension. Person-specific item reliability is defined as the ratio of variance explained by the latent factor to total variance for that item. Person-specific dimension reliability is the ratio of total variance explained by the latent factor across items in the dimension to total variance across those items. For person i, item reliability (Rel_{ji}) and dimension reliability (Rel_i) are:

$$\text{Rel}_{ji} = \frac{\lambda^2_{ji}\text{var}(F_{ti})}{\lambda^2_{ji}\text{var}(F_{ti}) + \text{var}(\epsilon_{jti})}$$

$$\text{Rel}_i = \frac{\sum_{j=1}^{q} \lambda^2_{ji}\text{var}(F_{ti})}{\sum_{j=1}^{q} \lambda^2_{ji}\text{var}(F_{ti}) + \sum_{j=1}^{q} \text{var}(\epsilon_{jti})}$$

where $\text{var}(F_{ti})$ is variance explained by the latent factor, equal to the product of the latent factor variance ($\text{var}(F_{ti})$) and squared factor loading (λ^2_{ji}), and $\text{var}(\epsilon_{jti})$ is unexplained variance (measurement error variance, $\sigma^2_{\epsilon_{ji}}$).

From equation (9), the latent factor variance satisfies:

$$\text{var}(F_{ti}) = \phi_i^2 \text{var}(F_{t-1,i}) + \text{var}(\zeta_{ti})$$

Under the weak stationarity assumption for the first-order autoregressive process, the latent factor variance is time-invariant ($\text{var}(F_{ti}) = \text{var}(F_{t-1,i})$), allowing equation (12) to be rewritten as:

$$\text{var}(F_{ti}) = \frac{\sigma^2_{\zeta_i}}{1 - \phi_i^2}$$

Despite its advantages, the DFA-based approach has limitations. First, DFA confounds trait components (general levels of a construct across repeated observations) with state components (deviations from general levels at specific occasions), potentially biasing person-specific reliability estimates. Second, it ignores between-person measurement structure, precluding estimation of between-person reliability. Third, relying solely on single-person repeated measures without incorporating information from other individuals or the entire sample may

lead to convergence difficulties for some individual models, preventing reliability estimation for certain persons (see Fuller-Tyszkiewicz et al., 2017, or the empirical example in this paper).

4. Integrating Multilevel Structure and Dynamic Nature

While MCFA and DFA each address only partial features of intensive longitudinal data, Asparouhov et al. (2018) proposed dynamic structural equation modeling (DSEM) as an integrative framework. DSEM combines multilevel modeling, time series analysis, and structural equation modeling (McNeish & Hamaker, 2020). It enables factor modeling at both within-person and between-person levels to account for measurement structures at different levels, thereby capturing the multilevel structure of intensive longitudinal data. Simultaneously, it models autoregressive processes at the within-person level to account for temporal dependencies among consecutive observations, thereby capturing the dynamic nature. Moreover, DSEM employs Bayesian estimation, which more flexibly estimates random effects of parameters (e.g., between-person differences) compared to traditional multilevel models using maximum likelihood estimation (McNeish & Hamaker, 2020; Muthén & Asparouhov, 2012). Like DFA, DSEM can estimate person-specific reliability, making it an extension of DFA to multilevel contexts (Asparouhov et al., 2018). In summary, DSEM simultaneously captures both multilevel structure and dynamic nature while examining individual differences in reliability, offering a comprehensive approach to reliability estimation in intensive longitudinal studies (Luo et al., under review; Xiao et al., 2023).

Like the previous methods, DSEM-based reliability estimation applies to both unidimensional and multidimensional structures; this paper uses the unidimensional case as an example (see Xiao et al., 2023 for multidimensional cases). For a unidimensional construct, a common two-level DSEM [Figure 3: see original paper] first decomposes observed scores into between-person (trait) and within-person (state) components:

$$Y_{jti} = Y_{ji} + Y_{jti}^{(w)}$$

where Y_{jti} is the observed score for item j from person i at time t ($j = 1, 2, \ldots, q$; $t = 1, 2, \ldots, T$; $i = 1, 2, \ldots, n$), Y_{ji} is person i's latent mean for item j across all time points (the between-person component representing trait level), and $Y_{jti}^{(w)}$ is the deviation of person i's observed score from their latent mean at time t (the within-person component representing state level).

Next, a within-person model is specified for the within-person component [FIGURE:3, lower left]. The measurement model decomposes the within-person component as:

$$Y_{jti}^{(w)} = \lambda_{ji}^{(w)} S_{ti}^{(w)} + \epsilon_{jti}^{(w)}$$

where $S_{ti}^{(w)}$ is the latent state factor for person i at time t, $\lambda_{ji}^{(w)}$ is the within-person factor loading for item j (randomly estimated across persons, assumed time-invariant), and $\epsilon_{jti}^{(w)}$ is random measurement error, assumed normally distributed ($\epsilon_{jti}^{(w)} \sim N(0, \sigma^2_{\epsilon_{ji}^{(w)}})$) with zero covariance between items ($\text{cov}(\epsilon_{jti}^{(w)}, \epsilon_{j'ti}^{(w)}) = 0$ for $j \neq j'$) and random error variances across persons.

The structural component assumes the latent state factor follows a first-order autoregressive process:

$$S_{ti}^{(w)} = \phi_i S_{t-1,i}^{(w)} + \zeta_{ti}^{(w)}$$

where ϕ_i is the person-specific autoregressive effect and $\zeta_{ti}^{(w)}$ is dynamic error at time t, assumed normally distributed ($\zeta_{ti}^{(w)} \sim N(0, \sigma^2_{\zeta_i^{(w)}})$).

Subsequently, a between-person model is specified for the between-person component [FIGURE:3, lower right]. The measurement model decomposes the between-person component as:

$$Y_{ji} = \nu_j + \lambda_j^{(B)} T_i + \epsilon_{ji}^{(B)}$$

where ν_j is the item j intercept, T_i is person i's latent trait factor, $\lambda_j^{(B)}$ is the between-person factor loading for item j, and $\epsilon_{ji}^{(B)}$ is measurement error, assumed normally distributed ($\epsilon_{ji}^{(B)} \sim N(0, \sigma^2_{\epsilon_j^{(B)}})$) with zero covariance between items ($\text{cov}(\epsilon_{ji}^{(B)}, \epsilon_{j'i}^{(B)}) = 0$ for $j \neq j'$).

In the random effects component, within-person factor loadings ($\lambda_{ji}^{(w)}$), natural logarithms of random measurement error variances ($\ln(\sigma^2_{\epsilon_{ji}^{(w)}})$), autoregressive effects (ϕ_i), and natural logarithms of dynamic error variances ($\ln(\sigma^2_{\zeta_i^{(w)}})$) are decomposed into fixed and random parts:

$$\lambda_{ji}^{(w)} = \lambda_j^{(w)} + u_{ji}$$
$$\phi_i = \phi + e_i$$
$$\ln(\sigma^2_{\epsilon_{ji}^{(w)}}) = \upsilon_j + \upsilon_{ji}$$
$$\ln(\sigma^2_{\zeta_i^{(w)}}) = \varphi + \varphi_i$$

The fixed parts represent population means, while random parts represent individual deviations. Random effects are assumed normally distributed: $u_{ji} \sim$

$N(0, \sigma^2_{u_j})$, $e_i \sim N(0, \sigma^2_e)$, $\upsilon_{ji} \sim N(0, \sigma^2_{\upsilon_j})$, and $\varphi_i \sim N(0, \sigma^2_\varphi)$. Natural logarithms ensure positive variance estimates and facilitate examination of correlations between log-transformed variances and other random effects (Hamaker et al., 2018).

Based on this model, reliability can be computed for each item and dimension at both levels. At the within-person level, person-specific item reliability is the ratio of variance in the item's state component explained by the latent state factor to total variance of the state component. Person-specific dimension reliability is the ratio of total variance explained by the latent state factor across items to total variance of state components. For person i, item reliability ($\text{Rel}^{(w)}_{ji}$) and dimension reliability ($\text{Rel}^{(w)}_{i}$) are:

$$\text{Rel}^{(w)}_{ji} = \frac{(\lambda^{(w)}_{ji})^2 \text{var}(S^{(w)}_{ti})}{(\lambda^{(w)}_{ji})^2 \text{var}(S^{(w)}_{ti}) + \text{var}(\epsilon^{(w)}_{jti})}$$

$$\text{Rel}^{(w)}_{i} = \frac{\sum_{j=1}^{q} (\lambda^{(w)}_{ji})^2 \text{var}(S^{(w)}_{ti})}{\sum_{j=1}^{q} (\lambda^{(w)}_{ji})^2 \text{var}(S^{(w)}_{ti}) + \sum_{j=1}^{q} \text{var}(\epsilon^{(w)}_{jti})}$$

where $\text{var}(S^{(w)}_{ti})$ is variance explained by the latent state factor, equal to the product of latent state factor variance ($\text{var}(S^{(w)}_{ti})$) and squared within-person factor loading ($(\lambda^{(w)}_{ji})^2$), and $\text{var}(\epsilon^{(w)}_{jti})$ is unexplained variance (random measurement error variance, $\sigma^2_{\epsilon^{(w)}_{ji}}$). Notably, the latent state factor variance formula mirrors that in DFA:

$$\text{var}(S^{(w)}_{ti}) = \frac{\sigma^2_{\zeta^{(w)}_i}}{1 - \phi^2_i}$$

By aggregating person-specific reliability estimates across all individuals, overall within-person reliability for each item and dimension can be obtained as a descriptive index of within-person reliability (see the empirical example for computational details).

At the between-person level, item reliability is the ratio of variance in the item's trait component explained by the latent trait factor to total variance of the trait component. Dimension reliability is the ratio of total variance explained by the latent trait factor across items to total variance of trait components. Item reliability ($\text{Rel}^{(B)}_{j}$) and dimension reliability ($\text{Rel}^{(B)}$) are:

$$\text{Rel}^{(B)}_{j} = \frac{(\lambda^{(B)}_{j})^2 \text{var}(T_i)}{(\lambda^{(B)}_{j})^2 \text{var}(T_i) + \text{var}(\epsilon^{(B)}_{ji})}$$

$$\text{Rel}^{(B)} = \frac{\sum_{j=1}^{q}(\lambda_j^{(B)})^2\text{var}(T_i)}{\sum_{j=1}^{q}(\lambda_j^{(B)})^2\text{var}(T_i) + \sum_{j=1}^{q}\text{var}(\epsilon_{ji}^{(B)})}$$

where $\text{var}(T_i)$ is variance explained by the latent trait factor, equal to the product of latent trait factor variance ($\text{var}(T_i)$) and squared between-person factor loading ($(\lambda_j^{(B)})^2$), and $\text{var}(\epsilon_{ji}^{(B)})$ is unexplained variance (measurement error variance, $\sigma^2_{\epsilon_j^{(B)}}$).

5. Empirical Demonstration

5.1 Data and Analytic Methods

This section demonstrates how to estimate item- and scale-level reliability (using the unidimensional case where scale reliability equals dimension reliability) in intensive longitudinal studies using MCFA, DFA, and DSEM (Mplus syntax and R code are available at https://osf.io/n2gw7/?view_{only}=44938b711ff3425a8e65a87cf523a49c). The empirical data consist of daily procrastination reports from 252 female college students over 34 consecutive days. Following previous research on daily procrastination measurement (Kühnel et al., 2016; Kühnel et al., 2022; Maier et al., 2021; Van Eerde & Venus, 2018), we adapted Tuckman's (1991) procrastination scale by adding the temporal cue "today" (e.g., "Today, I unnecessarily delayed completing work, even when it was important") to assess daily procrastination. The scale comprised six items rated from 1 ("strongly disagree") to 7 ("strongly agree") each evening. The average response rate was 94.89%.

MCFA-based reliability estimation was conducted in Mplus using robust maximum likelihood estimation (the default for two-level models). Based on equations (4)-(7), the MODEL CONSTRAINT command directly provided reliability estimates and standard errors for items and the overall scale at both levels.

DFA-based reliability estimation was performed in R using the MplusAutomation package (Hallquist & Wiley, 2018) to call Mplus and fit separate dynamic factor models for each individual's daily procrastination data. Bayesian estimation was used (10,000 iterations; convergence assessed via PSR and trace plots as recommended by Hamaker et al., 2018). Parameter posterior distributions (200 plausible values) were saved using SAVEDATA. Person-specific reliability posterior distributions (200 plausible values per person) were then computed in R using equations (10) and (11). The median of each posterior distribution served as the point estimate for person-specific reliability, and the distribution of these estimates across persons described the variability in person-specific reliability.

DSEM-based reliability estimation required both Mplus and R. In Mplus, Bayesian estimation (10,000 iterations) provided parameter estimates. Between-person item and scale reliability estimates with 95% Bayesian credible intervals

were obtained directly using MODEL CONSTRAINT and equations (25) and (26). To estimate person-specific reliability, parameter posterior distributions (200 plausible values) were saved and used to compute person-specific reliability posterior distributions in R via equations (22) and (23). As with DFA, medians provided point estimates and distributions described between-person variability.

For both DFA and DSEM, within-person reliability was also estimated as an overall descriptive index. After saving parameter posterior distributions, person-specific reliability was computed for each iteration (yielding 200 person-specific reliability values per person), then averaged across persons to obtain a posterior distribution of within-person reliability (200 values). The median served as the point estimate, with 2.5% and 97.5% percentiles forming the 95% credible interval.

When computing reliability via DFA or DSEM, some iterations may produce negative estimates of latent (state) factor variance. Following Xiao et al. (2023), these problematic iterations were excluded by replacing corresponding person-specific reliability values with missing values.

5.2 Results and Discussion

Table 1 presents between-person and within-person reliability estimates for items and the overall scale across the three methods. For the overall scale, MCFA and DSEM produced similar between-person reliability estimates, while their within-person reliability estimates differed substantially and were both lower than DFA's within-person estimate. Item-level reliability estimates also varied across methods. MCFA and DSEM yielded relatively similar between-person and within-person item reliability, whereas DFA produced higher within-person item reliability than DSEM. Notably, during DFA estimation, 145 individuals' models failed to converge (due to non-positive definite variance-covariance matrices, among other issues), so reliability estimates were based on only 107 individuals (42.46%). This suggests comparisons between DFA and other methods may be problematic due to different sample bases, and researchers should interpret such results cautiously. More importantly, this highlights that DFA may encounter difficulties fitting some individual models, preventing estimation of person-specific reliability for certain individuals and potentially biasing within-person reliability estimates.

Examination of item-level reliability revealed that Item 2 ("Today, I postponed making difficult decisions") showed the lowest between-person and within-person reliability across all three methods. Further inspection of person-specific reliability distributions from DFA and DSEM showed that Item 2's median and mean person-specific reliability were markedly lower than other items, indicating poorer internal consistency with other items when measuring the state component of procrastination. Content analysis provides a plausible explanation. In Tuckman's (1991) original scale, Item 2 assessed general

tendencies to postpone difficult decisions. By adding "today," we measured daily occurrences of this behavior. However, individuals do not necessarily face difficult decisions daily, potentially causing confusion or difficulty responding, which reduces consistency with other items.

6. Discussion

6.1 Comparative Analysis of Three Methods

To help researchers select appropriate reliability estimation methods, Table 3 summarizes the features and limitations of the three intensive longitudinal reliability estimation methods. From the perspectives of data fit, estimable reliability coefficients, and estimation approach, DSEM integrates the strengths of MCFA and DFA. It captures both multilevel structure and dynamic nature, estimates reliability at person-specific, within-person, and between-person levels, and uses Bayesian estimation to flexibly model random effects and examine individual differences. However, from the perspectives of software requirements and computational time, DSEM requires both Mplus and additional software (e.g., R), and its complexity demands longer runtime. In contrast, MCFA can be completed entirely within Mplus with concise syntax, direct output, and high efficiency, offering simplicity advantages.

Table 3 also outlines each method's primary limitations. MCFA makes strong assumptions, cannot examine individual differences in reliability, and ignores dynamic nature. DFA confounds trait and state components, neglects multilevel structure (preventing between-person reliability estimation), and may fail to fit some individual models. DSEM is relatively complex, time-consuming, and less straightforward.

Given these characteristics, we propose a decision flowchart for selecting reliability estimation methods [Figure 4: see original paper]. If researchers are not interested in person-specific reliability or individual differences, but rather in overall within-person and between-person reliability, or if they have already verified measurement invariance across individuals using appropriate methods (e.g., cross-classified models; McNeish et al., 2021), MCFA provides a relatively simple way to examine and report reliability. If researchers believe trait factors do not influence item responses, focus on person-specific measurement models and reliability, or have small samples (even single-person time series) insufficient for examining between-person performance, DFA can provide person-specific and within-person reliability, though adequate time points and model convergence must be considered. In most other cases, we recommend DSEM to obtain person-specific, within-person, and between-person reliability. Many intensive longitudinal studies adapt trait measure items to assess temporal dynamics (Horstmann & Ziegler, 2020; Trull & Ebner-Priemer, 2020) without quantitative support for item selection or adaptation effectiveness. DSEM can thoroughly evaluate whether adapted measures reliably capture between-person differences and within-person dynamic processes. Moreover, as researchers call for devel-

oping measures specifically suited for intensive longitudinal studies (Dietrich et al., 2022; Horstmann & Ziegler, 2020; Mielniczuk, 2023) and such development efforts increase (Blanke & Brose, 2017; Engyel et al., 2022; Ringwald et al., 2022), reliability estimation during scale development should employ methods that fit intensive longitudinal data and can estimate multiple reliability types (i.e., DSEM) to better evaluate new measures' reliability.

6.2 Key Considerations in Intensive Longitudinal Reliability Estimation

6.2.1 Item-Level Reliability Item-level reliability warrants attention in intensive longitudinal studies. While many researchers use multiple items (e.g., three or more) to measure temporal dynamics, most report only overall scale reliability without examining individual items (Eltanamly et al., 2023; Koval et al., 2019; Van Der Tuin et al., 2023; Wright et al., 2017). Researchers note that items selected from trait measures may not be directly suitable for intensive state measurement (Horstmann & Ziegler, 2020; Mielniczuk, 2023). Our empirical application further demonstrates that some adapted trait items exhibit lower reliability at all levels compared to other items, with content analysis suggesting they may be unsuitable for intensive longitudinal contexts. Thus, beyond overall scale reliability, researchers should examine item-level reliability. Item reliability estimates and comparisons can help identify items unsuitable for intensive longitudinal studies, which is particularly important for research adapting trait measures to assess states. Additionally, given recommendations to use brief scales (e.g., 3-6 items) in intensive longitudinal studies to balance measurement quality and response burden (Mielniczuk, 2023), evaluating item reliability can inform appropriate scale reduction and improve measurement efficiency.

6.2.2 Individual Differences in Reliability Individual differences in reliability represent another important consideration. Early reliability research emphasized that reliability is a test characteristic specific to the administered sample (Mellenbergh, 1996; Wilkinson, 1999), with estimates from one sample not necessarily generalizing to others. Similarly, in intensive longitudinal contexts, researchers examine within-person dynamic processes and score reliability, but individuals' behaviors and states may change over time in idiosyncratic ways (Schuurman & Hamaker, 2019). Consequently, test score reliability likely varies across individuals (Fisher et al., 2018; Stone et al., 2023), necessitating consideration of person-specific reliability and its between-person variation. This not only deepens understanding of a measure's reliability and applicability to the sample but also provides supportive or cautionary information for interpreting results. High person-specific reliability for most individuals supports the credibility of within-person findings, whereas low reliability for many individuals warrants caution in interpretation and generalization.

6.2.3 Reporting Reliability Results Based on the above, we offer two recommendations for reporting reliability in intensive longitudinal studies. First,

given the importance of item-level reliability, researchers using MCFA or DSEM should report within-person and between-person reliability for each item and the overall scale (or each dimension). Researchers using DFA should report within-person reliability for each item and the overall scale. Each reliability estimate should include point estimates and 95% (Bayesian) credible intervals (see Table 1). These results reflect overall performance of items and scales at both levels, help identify items unsuitable for intensive longitudinal contexts, and provide primary evidence for evaluating reliability.

Second, for individual differences in reliability, studies using DFA or DSEM that examine person-specific reliability should additionally report distributions of person-specific reliability. Specifically, researchers can present distribution plots of person-specific reliability for each item and the overall scale (see Xiao et al., 2023, Figure 2) or report descriptive statistics (e.g., median, mean, standard deviation; see Table 2) to examine item and scale applicability to each individual, providing supplementary evidence for reliability evaluation.

6.3 Other Methods and Future Directions

Beyond the methods discussed, other approaches warrant exploration in intensive longitudinal reliability estimation. Inspired by traditional test-retest reliability, Dejonckheere et al. (2022) randomly repeated an emotion item in an intensive longitudinal measure and estimated reliability using squared differences between the two scores. Hu et al. (2016) proposed creating parallel forms in intensive longitudinal studies and estimating person-specific reliability by correlating scores on parallel forms within each individual.

Additionally, latent state-trait theory (LST; Steyer et al., 1999, 2015) offers alternative reliability estimation frameworks (Castro-Alvarez, Tendeiro, Meijer, & Bringmann, 2022; Castro-Alvarez, Tendeiro, & de Jonge et al., 2022). LST theory includes three key coefficients (Steyer et al., 2015): consistency (proportion of variance from stable trait components), occasion specificity (proportion from occasion-specific state components), and reliability (sum of consistency and occasion specificity, representing proportion of variance from trait and state components combined, i.e., variance not due to random measurement error). Various models can estimate reliability under this framework, including multistate-singletrait (MSST; Steyer et al., 2015), common and unique trait–state (CUTS; Hamaker et al., 2017), and trait-state-occasion (TSO; Eid et al., 2017) models. These models relate to those discussed herein. For example, the multilevel CUTS model is statistically equivalent to MCFA (Roesch et al., 2010), and the mixed-effects TSO model is statistically equivalent to the two-level DSEM presented here (Castro-Alvarez, Tendeiro, & de Jonge et al., 2022). However, because LST-based models differ in variance decomposition and reliability definition, their reliability estimates and interpretations may differ. Interested readers may consult Castro-Alvarez, Tendeiro, Meijer, and Bringmann (2022) and Castro-Alvarez, Tendeiro, and de Jonge et al. (2022).

As intensive longitudinal research continues to develop, reliability issues in this context merit greater attention from methodological and applied researchers. Methodological studies often estimate reliability based on specific models, limiting the applicability of resulting indices (Laenen et al., 2009). Future research should explore reliability definitions and estimation methods based on other models, such as continuous time structural equation modeling (CTSEM; Driver et al., 2017). In applied research, reliability testing has not received sufficient attention (Brose et al., 2020; Horstmann & Ziegler, 2020; Stone et al., 2023; Trull & Ebner-Priemer, 2020). Future studies should incorporate reliability testing as a necessary step in data analysis and select appropriate estimation methods based on specific research contexts to ensure more robust conclusions.

References

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

Blanke, E. S., & Brose, A. (2017). Mindfulness in daily life: A multidimensional approach. *Mindfulness, 8*, 737–750.

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

Bolger, N., Davis, A., & Rafaeli, E. (2003). Diary methods: Capturing life as it is lived. *Annual Review of Psychology, 54*(1), 579–616.

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

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

Castro-Alvarez, S., Tendeiro, J. N., Meijer, R. R., & Bringmann, L. F. (2022). Using structural equation modeling to study traits and states in intensive longitudinal data. *Psychological Methods, 27*(1), 17–43.

Cattell, R. B., Cattell, A. K. S., & Rhymer, R. M. (1947). P-technique demonstrated in determining psychophysiological source traits in a normal individual. *Psychometrika, 12*, 267–288.

Collins, L. M. (2006). Analysis of longitudinal data: The integration of theoretical model, temporal design, and statistical model. *Annual Review of Psychology, 57*, 505–528.

Cranford, J. A., Shrout, P. E., Iida, M., Rafaeli, E., Yip, T., & Bolger, N. (2006). A procedure for evaluating sensitivity to within-person change: Can

mood measures in diary studies detect change reliably? *Personality and Social Psychology Bulletin, 32*(7), 917–929.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology, 16*(2), 137–163.

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

Di Sarno, M., Zimmermann, J., Madeddu, F., Casini, E., & Di Pierro, R. (2020). Shame behind the corner? A daily diary investigation of pathological narcissism. *Journal of Research in Personality, 85*, 103924.

Dietrich, J., Schmiedek, F., & Moeller, J. (2022). Academic motivation and emotions are experienced in learning situations, so let's study them [Special issue]. *Learning and Instruction, 81*, 101623.

Driver, C. C., Oud, J. H., & Voelkle, M. C. (2017). Continuous time structural equation modeling with R package ctsem. *Journal of Statistical Software, 77*, 1–35.

Eid, M., Holtmann, J., Santangelo, P., & Ebner-Priemer, U. (2017). On the definition of latent-state-trait models with autoregressive effects. *European Journal of Psychological Assessment, 33*(4), 285–295.

Eltanamly, H., Leijten, P., Van Roekel, E., Mouton, B., Pluess, M., & Overbeek, G. (2023). Strengthening parental self-efficacy resilience: A within-subject experimental study with refugee parents adolescents. *Child Development, 94*(1), 187–201.

Engyel, M., de Ruiter, N. M., & Urbán, R. (2022). Momentarily narcissistic? Development of a short, state version of the Pathological Narcissism Inventory applicable in momentary assessment. *Frontiers in Psychology, 13*, 1009839.

Fisher, A. J., Medaglia, J. D., & Jeronimus, B. F. (2018). Lack of group-to-individual generalizability is a threat to human subjects research. *Proceedings of the National Academy of Sciences, 115*(27), E6106-E6115.

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

Geldhof, G. J., Preacher, K. J., & Zyphur, M. J. (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological Methods, 19*(1), 72–91.

Gerstberger, L., Blanke, E. S., Keller, J., & Brose, A. (2023). Stress buffering after physical activity engagement: An experience sampling study. *British*

Journal of Health Psychology, 28(3), 876–892.

Hallquist, M. N., & Wiley, J. F. (2018). MplusAutomation: an R package for facilitating large-scale latent variable analyses in Mplus. *Structural Equation Modeling: A Multidisciplinary Journal, 25*(4), 621–638.

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

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

Hamaker, E. L., Schuurman, N. K., & Zijlmans, E. A. O. (2017). Using a few snapshots to distinguish mountains from waves: Weak factorial invariance in the context of trait-state research. *Multivariate Behavioral Research, 52*(1), 47–60.

Hausen, J. E., Möller, J., Greiff, S., & Niepel, C. (2023). Morningness and state academic self-concept in students: Do early birds experience themselves as more competent in daily school life? *Contemporary Educational Psychology, 74*, 102199.

Horstmann, K. T., & Ziegler, M. (2020). Assessing personality states: What to consider when constructing personality state measures. *European Journal of Personality, 34*(6), 1037–1059.

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

Koval, P., Holland, E., Zyphur, M. J., Stratemeyer, M., Knight, J. M., Bailen, N. H., … Haslam, N. (2019). How does it feel to be treated like an object? Direct and indirect effects of exposure to sexual objectification on women's emotions in daily life. *Journal of Personality and Social Psychology, 116*(6), 885–898.

Kühnel, J., Bledow, R., & Feuerhahn, N. (2016). When do you procrastinate? Sleep quality and social sleep lag jointly predict self-regulatory failure at work. *Journal of Organizational Behavior, 37*(7), 983–1002.

Kühnel, J., Bledow, R., & Kuonath, A. (2022). Overcoming procrastination: Time pressure and positive affect as compensatory routes to action. *Journal of Business and Psychology, 38*(4), 803–819.

Laenen, A., Alonso, A., Molenberghs, G., & Vangeneugden, T. (2009). A family of measures to evaluate scale reliability in a longitudinal setting. *Journal of the Royal Statistical Society Series A: Statistics in Society, 172*(1), 237–253.

Lafit, G., Adolf, J. K., Dejonckheere, E., Myin-Germeys, I., Viechtbauer, W., & Ceulemans, E. (2021). Selection of the number of participants in intensive longitudinal studies: A user-friendly shiny app and tutorial for performing power

analysis in multilevel regression models that account for temporal dependencies. *Advances in Methods and Practices in Psychological Science, 4*(1), 1–24.

Lane, S. P., & Shrout, P. E. (2010). Assessing the reliability of within-person change over time: A dynamic factor analysis approach. *Multivariate Behavioral Research, 45*(6), 1027–1059.

Luo, X., Hu, Y., & Liu, H. (under review). Assessing between- and within-person reliabilities of items and scale for daily procrastination: A multilevel and dynamic approach. *Assessment*.

Maier, T., Kühnel, J., & Zimmermann, B. (2021). How did you sleep tonight? The relevance of sleep quality and sleep-wake rhythm for procrastination at work. *Frontiers in Psychology, 12*, 785154.

McNeish, D., & Hamaker, E. L. (2020). A primer on two-level dynamic structural equation models for intensive longitudinal data in Mplus. *Psychological Methods, 25*(5), 610–635.

McNeish, D., Mackinnon, D. P., Marsch, L. A., & Poldrack, R. A. (2021). Measurement in intensive longitudinal data. *Structural Equation Modeling: A Multidisciplinary Journal, 28*(5), 807–822.

Mellenbergh, G. J. (1996). Measurement precision in test score and item response models. *Psychological Methods, 1*(3), 293–299.

Mielniczuk, E. (2023). Call for new measures suitable for intensive longitudinal studies: Ideas and suggestions. *New Ideas in Psychology, 68*, 100983.

Molenaar, P. C. (1985). A dynamic factor model for the analysis of multivariate time series. *Psychometrika, 50*(2), 181–202.

Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: a more flexible representation of substantive theory. *Psychological Methods, 17*(3), 313–335.

Neubauer, A. B., Schmidt, A., Schmiedek, F., & Dirk, J. (2022). Dynamic reciprocal relations of achievement goals with daily experiences of academic success and failure: An ambulatory assessment study. *Learning and Instruction, 81*, 101617.

Nezlek, J. B. (2017). A practical guide to understanding reliability in studies of within-person variability. *Journal of Research in Personality, 69*, 149–155.

Reis, D., Arndt, C., Lischetzke, T., & Hoppe, A. (2016). State work engagement and state affect: Similar yet distinct concepts. *Journal of Vocational Behavior, 93*, 1–10.

Ringwald, W. R., Manuck, S. B., Marsland, A. L., & Wright, A. G. (2022). Psychometric evaluation of a Big Five personality state scale for intensive longitudinal studies. *Assessment, 29*(6), 1301–1319.

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*(5), 767–789.

Scherer, R., & Teo, T. (2020). A tutorial on the meta-analytic structural equation modeling of reliability coefficients. *Psychological Methods, 25*(6), 747–775.

Schmitt, A., Belschak, F. D., & Den Hartog, D. N. (2017). Feeling vital after a good night's sleep: The interplay of energetic resources and self-efficacy for daily proactivity. *Journal of Occupational Health Psychology, 22*(4), 493–504.

Schönbrodt, F. D., Zygar-Hoffmann, C., Nestler, S., Pusch, S., & Hagemeyer, B. (2021). Measuring motivational relationship processes in experience sampling: A reliability model for moments, days, and persons nested in couples. *Behavior Research Methods, 54*(4), 1869–1888.

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

Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological momentary assessment. *Annual Review of Clinical Psychology, 4*, 1–32.

Steyer, R., Mayer, A., Geiser, C., & Cole, D. A. (2015). A theory of states and traits-Revised. *Annual Review of Clinical Psychology, 11*, 71–98.

Steyer, R., Schmitt, M., & Eid, M. (1999). Latent state-trait theory and research in personality and individual differences. *European Journal of Personality, 13*(5), 389–408.

Stone, A. A., Schneider, S., & Smyth, J. M. (2023). Evaluation of pressing issues in ecological momentary assessment. *Annual Review of Clinical Psychology, 19*, 107–131.

Trull, T. J., & Ebner-Priemer, U. (2013). Ambulatory assessment. *Annual Review of Clinical Psychology, 9*, 151–176.

Trull, T. J., & Ebner-Priemer, U. W. (2020). Ambulatory assessment in psychopathology research: A review of recommended reporting guidelines and current practices. *Journal of Abnormal Psychology, 129*(1), 56–63.

Tuckman, B. W. (1991). The development and concurrent validity of the procrastination scale. *Educational and Psychological Measurement, 51*(2), 473–480.

Van Der Tuin, S., Booij, S. H., Oldehinkel, A. J., Van Den Berg, D., Wigman, J. T. W., Lång, U., & Kelleher, I. (2023). The dynamic relationship between sleep and psychotic experiences across the early stages of the psychosis continuum. *Psychological Medicine*. Advance online publication. https://doi.org/10.1017/S0033291723001459

Van Eerde, W., & Venus, M. (2018). A daily diary study on sleep quality and procrastination at work: The moderating role of trait self-control. *Frontiers in Psychology, 9*, 2029.

Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*(8), 594–604.

Wright, A. G., Stepp, S. D., Scott, L. N., Hallquist, M. N., Beeney, J. E., Lazarus, S. A., & Pilkonis, P. A. (2017). The effect of pathological narcissism on interpersonal and affective processes in social interactions. *Journal of Abnormal Psychology, 126*(7), 898–910.

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

Xu, J., & Zheng, Y. (2022). Links between shared and unique perspectives of parental psychological control and adolescent emotional problems: A dyadic daily diary study. *Child Development, 93*(6), 1649–1662.

Ye, B., Wen, Z., & Chen, Q. (2012). Estimating test reliability in longitudinal studies. *Advances in Psychological Science, 20*(3), 467–474.

Zheng, S., Zhang, L., Qiao, X., & Pan, J. (2021). Intensive longitudinal data analysis: Models and applications. *Advances in Psychological Science, 29*(11), 1949–1960.

Zhou, L., Wang, M., & Zhang, Z. (2021). Intensive longitudinal data analyses with dynamic structural equation modeling. *Organizational Research Methods, 24*(2), 219–250.

Note: Figure translations are in progress. See original paper for figures.