Neubauer, Andreas B.; Voelkle, Manuel C.; Voss, Andreas; Mertens, Ulf K.

Estimating reliability of within-person couplings in a multilevel framework

formal und inhaltlich überarbeitete Version der Originalveröffentlichung in:
formally and content revised edition of the original source in:
Journal of personality assessment 102 (2020) 1, S. 10-21, 10.1080/00223891.2018.1521418

Bitte verwenden Sie in der Quellenangabe folgende URN oder DOI /
Please use the following URN or DOI for reference:
urn:nbn:de:0111-dipfdocs-190535
10.25657/02:19053

https://nbn-resolving.org/urn:nbn:de:0111-dipfdocs-190535
https://doi.org/10.25657/02:19053

Estimating Reliability of Within-Person Couplings in a Multilevel Framework

Andreas B. Neubauer[1,2,3], Manuel C. Voelkle[4,5], Andreas Voss[3], and Ulf K. Mertens[3]

[1]German Institute for International Educational Research (DIPF), Frankfurt am Main, Germany

[2]Center for Research on Individual Development and Adaptive Education of Children at Risk (IDeA), Frankfurt am Main, Germany

[3]Heidelberg University, Heidelberg, Germany

[4]Humboldt University Berlin, Berlin, Germany

[5]Max-Planck Institute for Human Development, Berlin, Germany

Author Note

Correspondence concerning this article should be addressed to Andreas B. Neubauer, Education and Human Development, German Institute for International Educational Research (DIPF), Frankfurt am Main. E-Mail: neubauer.andreas@dipf.de

Part of this research has been presented at the 50th meeting of the German Psychological Society in Leipzig, Germany, and the 13th Meeting of the Section for Methods and Evaluation of the German Psychological Society in Tübingen, Germany.

The reported study was not pre-registered. All data and analyses script necessary to reproduce the reported results can be retrieved from https://osf.io/bdw5r/

Neubauer, A. B., Voelkle, M. C., Voss, A., & Mertens, U. K. (in press). Estimating reliability of within-person couplings in a multilevel framework. *Journal of Personality Assessment102*, 10-21. doi: 10.1080/00223891.2018.1521418

Abstract

Within-person couplings play a prominent role in psychological research and previous studies have shown that inter-individual differences in within-person couplings predict future behavior. For example, stress reactivity – operationalized as the within-person coupling of stress and positive or negative affect – is an important predictor of various (mental) health outcomes and has often been assumed to be a more or less stable personality trait. However, issues of reliability of these couplings have been largely neglected so far. In this work, we present an estimate for the reliability of within-person couplings that can be easily obtained using the user-modifiable R code accompanying this work. Results of a simulation study show that this index performs well even in the context of unbalanced data due to missing values. We demonstrate the application of this index in a measurement burst study targeting the reliability and test-retest correlation of stress reactivity estimates operationalized as within-person couplings. Reliability and test-retest correlations of stress reactivity estimates were rather low, challenging the implicit assumption of stress reactivity as a stable person-level variable. We highlight key factors that researchers planning studies targeting inter-individual differences in within-person couplings should consider to maximize reliability.

Keywords: Monte Carlo simulation, intra-individual variability, within-person process, within-person effect, ambulatory assessment

Estimating Reliability of Within-Person Couplings in a Multilevel Framework

The increase in the application of intensive longitudinal designs (ILDs, e.g., daily diary designs, ecological momentary assessment) has been paralleled by an increasing interest in within-person effects, that is, effects that unfold within individuals across time. By investigating people's feelings, thoughts, and behaviors repeatedly across many measurement occasions in their daily lives (e.g., via ambulatory assessment; Trull & Ebner-Priemer, 2013), within-person effects can be observed as within-person (intra-individual) couplings of variables in a natural environment, providing insights into people's lives as they are lived (Bolger, Davis, & Rafaeli, 2003). In a nutshell, these couplings address the question whether within-person changes in one variable are associated ("coupled") with changes in another variable. The benefit of investigating within-person couplings has been acknowledged in a plethora of empirical studies in different fields of psychological research. For example, for working memory performance within-person couplings have been reported with negative affect (Brose, Schmiedek, Lövdén, & Lindenberger, 2012), motivation (Brose, Schmiedek, Lövdén, Molenaar, & Lindenberger, 2010), and sleep quality (Könen, Dirk, & Schmiedek, 2015). That is, individuals' working memory performance was lower on days with higher negative affect, less motivation, and poorer sleep quality. In psychotherapy research, within-person couplings have been investigated to improve understanding of the mechanisms underlying positive therapeutic outcome. Rubel, Rosenbaum, and Lutz (2017) observed within-person couplings between session-specific coping skills and symptom improvements in the next therapy session: When patients experienced more coping skills in a session than they usually do, they showed a stronger decline in symptom severity to the next session. Within-person couplings have also been investigated in studies targeting the effects of need fulfillment on well-being (Neubauer, Lerche, & Voss, 2018), the association of snack craving and snack consumption (Richard, Meule, Reichenberger, & Blechert, 2017), or the link between momentary anger and symptom severity in patients with asthma or rheumatoid arthritis (Russell, Smith, & Smyth, 2016), to name just a few examples.

In addition to examining the average within-person coupling of two variables, the multilevel model (MLM) framework, which is typically used to analyze intensive longitudinal data, further allows investigating inter-individual differences in the strength of these couplings. Such differences have also been reported in some of the studies mentioned above. For example, children differed in the degree to which reported sleep quality was associated with working memory performance (Könen et al., 2015). One interesting question arising from these findings is whether these inter-individual differences in the strength of intra-individual couplings can be used as predictors of future behavior. For example, if findings suggest that child A profits more from sleep quality than child B, an intervention targeting sleep quality should have a larger impact on child A than on child B. The benefit of using inter-individual differences in within-person couplings as predictors of future behavior hinges, however, on their reliability.

The issue of reliability has long been neglected in ILD research, but recent approaches have made this topic more easily accessible for researchers working with intensive longitudinal data. Of particular note, reliability in ILDs needs to be examined separately on the between-person level (addressing the question whether it is possible to reliably separate individuals with high vs. low scores on a measure) and on the within-person level (addressing the question whether it is possible to reliably separate situations with high vs. low scores on a measure within individuals). Recent work has made the topic of reliability in ILDs more easily accessible for empirical researchers. For example, Cranford et al. (2006) provide equations based on the framework of generalizability theory to compute reliability estimates for both the within- and between-person level (see also Shrout & Lane, 2012). Wilhelm and Schoebi (2007) use variance decomposition in a multilevel context to compute these estimates. Bulut, Davison, and Rodriguez (2017) propose a reliability index in the framework of profile analysis that has not been developed for, but could also be applied to, intensive longitudinal data. Finally, Geldhof, Preacher, and Zyphur (2014) provide Mplus code to compute reliability estimates (Cronbach's α and McDonald's ω) separately for the within- and between-person level.

Reliability is an important topic since using unreliable measures may result in substantial attenuation of statistical power (i.e., a failure to detect an effect that exists in the population, but is obscured by unreliability of the measurement). This will also lower chances for replication, and thus threaten cumulative knowledge gain in psychological science (LeBel & Paunonen, 2011). Whereas providing reliability estimates for the scales used in a study is common practice, estimates for the reliability of within-person couplings are usually not reported. Although such an estimate has been suggested (Raudenbush & Bryk, 2002) it has not been picked up by empirical researchers so far, most likely because previous elaborations on this issue have been rather technical and did not provide clear instructions or software how such an estimate can be obtained from empirical data. We aim at filling this gap in the literature by providing a practical introduction of this estimate, along with adaptable R code and an Excel sheet to estimate this reliability index from raw data or from published data in aggregated form. We will exemplify the application of this estimate in the context of one prominent within-person coupling: stress reactivity.

A Motivating Example—Reliability of Inter-Individual Differences in Stress Reactivity

Previous research has investigated whether individuals differ in the degree to which their well-being is affected by stressful events and whether these inter-individual differences can be used to predict future behavior. For example, O'Neill, Cohen, Tolpin, and Gunthert (2004) showed that stress reactivity (the individuals' estimates of the within-person coupling of stress and negative affect) predicted change in depressive symptoms in college students:

Participants with higher stress reactivity experienced an increase in depressive symptoms over the course of two months. More recently, data have been presented showing that stress reactivity parameters affect a broad range of outcomes such as sleep quality (Ong et al., 2013), depression (Charles, Piazza, Mogle, Sliwinski, & Almeida, 2013), and even mortality (Mroczek et al., 2015) longitudinally. Hence, there is a wealth of data suggesting that stress reactivity predicts a variety of highly relevant outcomes up to ten years later.

While the cited findings suggest very promising prospects for diagnostics and interventions (i.e., if we know an individual's stress reactivity estimate, we might be able to use this estimate to predict this individual's future outcomes and to build tailored interventions) it remains an open question whether these parameters can live up to this promise. Note that inherent to using stress reactivity estimates as predictor of future behavior is the idea that these parameters represent more or less stable, person-level variables that can be assessed with sufficient reliability. For example, if a researcher is interested in testing the effects of a tailored intervention targeting specifically participants high in stress reactivity, she will need to know whether it is at all possible to identify these individuals based on within-person couplings of external strain and negative emotionality. Importantly, the previously discussed approaches to reliability in the ILD context (Bulut et al., 2017; Cranford et al., 2006; Geldhof et al., 2014; Wilhelm & Schoebi, 2007) cannot be applied here. Whereas those approaches analyze whether measures can capture inter-individual differences in mean levels (between-person reliability) or intra-individual fluctuations (within-person reliability) in a construct, the present work targets inter-individual differences in the within-person association (coupling) of two variables. Hence, the required index to estimate this reliability would need to capture the between-person reliability of the individuals' within-person couplings (and thereby address the question whether we can reliably separate individuals with high vs. low couplings). However, in none of these studies (Charles et al., 2013; Gunthert, Cohen, Butler, & Beck, 2005; Mroczek et al., 2015; O'Neill et al., 2004; Ong et al., 2013) reliability estimates of these coupling parameters haven been reported. Most likely, the absence of such reliability estimates from empirical research is due to the unfamiliarity of researchers with these estimates. An index for the required reliability estimate has been proposed by Raudenbush and Bryk (2002), but it has not been utilized by applied researchers in the field of intensive longitudinal data.

Reliability of Within-person Couplings

A typical setup to investigate stress reactivity is to collect data on individuals' current stress level and current negative affect in an ILD, and analyze the intra-individual coupling of these two variables in a MLM framework to accommodate the nested data structure (repeated observations nested within individuals). Let data from *N* participants be collected in an ILD for *T* repeated measurement occasions, then the prediction of person *i*'s negative affect at time *t* (Y_{it}) from (time-varying) stress level (X_{it})[1] is represented by the following equations:

Level 1:

$$Y_{it} = \beta_{0i} + \beta_{1i}X_{it} + \varepsilon_{it} \tag{1}$$

Level 2:

$$\beta_{0i} = \gamma_{00} + \upsilon_{0i} \tag{2}$$

$$\beta_{1i} = \gamma_{10} + \upsilon_{1i} \tag{3}$$

where γ_{00} and γ_{10} are the fixed intercept and fixed slope, respectively, υ_{0i} represents person *i*'s deviation from the fixed intercept, υ_{1i} is person *i*'s deviation from the fixed slope, and ε_{it} is the person and measurement-occasion specific error term. In this simple model, we can obtain an overall slope (γ_{10}, the effect of stress on negative affect for the whole sample) as well as individual slopes (β_{1i}, person *i*'s effect of stress on negative affect). The variance of

[1] In order to obtain a valid estimator for the within-person association of stress and negative affect, the (continuous) time-varying predictor should be centered on the person mean (Wang & Maxwell, 2015). Unless otherwise noted, time-varying predictors have all been centered on the person mean in the present work.

υ_{1i} (denoted as random slope variance) captures inter-individual differences in the within-person effect "stress reactivity", and β_{1i} represents person *i*'s within-person coupling. For the present work, the variables Y_{it} and X_{it} are conceptualized as continuous variables that can be measured repeatedly using either single item measures or scales.[2]

In principle, person *i*'s within-person coupling (β_{1i}) can be estimated in three ways: First, we could run a linear regression based only on the data provided by person i. In this model, the ordinary least squares (OLS) estimator for β_{1i} (denoted as $\hat{\beta}_{1i}$) can be obtained. Second, we could disregard inter-individual differences in within-person couplings and estimate a single coupling parameter which is constant for each individual: the fixed effect across all individuals (γ_{10}). The third (and usually most efficient) option is to use a weighted combination of OLS estimates and the fixed effect estimate (for details on this weighting procedure, also referred to as shrinkage, see e.g., Raudenbush & Bryk, 2002, pp.45-49). This latter estimate is the best linear unbiased predictor (BLUP) of person *i*'s true within-person coupling and is often referred to as the empirical Bayes (EB) estimate. Following the notation introduced by Raudenbush and Bryk (2002) we further denote the EB estimate of β_{1i} as β^*_{1i}.

In their influential book on hierarchical linear modeling, Raudenbush and Bryk (2002, p. 49) proposed a formula to estimate the reliability of the OLS estimate of β_{1i}. Using the terminology of the present work[3], reliability of person *i*'s estimated within-person coupling $\hat{\beta}_{1i}$ as defined by Raudenbush and Bryk (2002) is:

[2] Note that the model can easily be expanded to include more than one predictor. However, when entering several predictors simultaneously, the meaning of the within-person coupling changes, and these couplings have to be interpreted as partial regression coefficients (see Discussion section for further details). An alternative approach to examine the within-person couplings of Y with several predictors is to run separate models including only one predictor at a time and examine the couplings from these models. This approach results in bivariate within-person couplings (that is, zero-order within-person associations not controlling for other variables). Whether a simultaneous (including all predictors at once) or sequential approach (including only one predictor at a time) is more appropriate depends on the research question. In both approaches, reliability of within-person couplings can be estimated with the procedure introduced in the current work.

[3] Note that the original formulation by Raudenbush & Bryk (2002) looks different from Equation (4); we have adjusted the equation to better fit the notation used in the present manuscript. Further details can be found in Appendix A (Supplemental Online Material).

$$\text{reliability}(\hat{\beta}_{1i}) = \frac{\tau_{11}^2}{\tau_{11}^2 + \frac{\sigma_e^2}{SS_i(X)}} \tag{4}$$

with τ_{11}^2 being the variance of υ_{1i} and σ_e^2 being the residual variance at Level-1. The term $SS_i(X)$ represents person *i*'s sum of squares of the predictor X, that is, person *i*'s sum of the squared deviances from his or her mean in X ($\bar{X}_i$).

$$SS_i(X) = \sum_{t=1}^{T_i} (X_{it} - \bar{X}_i)^2 \tag{5}$$

$$\bar{X}_i = \frac{1}{T_i} \sum_{t=1}^{T_i} X_{it} \tag{6}$$

Note that this term is also required for the computation of person *i*'s (within-person) variance of the predictor X:

$$\text{Var(X)}_i = \hat{\sigma}_{X_i}^2 = \frac{\sum_{t=1}^{T_i}(X_{it} - \bar{X}_i)^2}{T_i - 1} = \frac{SS_i(X)}{T_i - 1} \tag{7}$$

It follows that $SS_i(X)$ is the product of person *i*'s number of repeated measurement occasions minus 1 (*T_i*-1) and person *i*'s estimate of the within-person variance of the predictor X ($\hat{\sigma}_{X_i}^2$). This yields:

$$\text{reliability}(\hat{\beta}_{1i}) = \frac{\tau_{11}^2}{\tau_{11}^2 + \frac{\sigma_e^2}{(T_i - 1) \cdot \hat{\sigma}_{X_i}^2}} \tag{8}$$

Note that the reliability index, the number of repeated measurement occasions, and the within-person variance carry an index *i*, that is, they can vary across Level-2 units (participants). This allows for participants to vary in their reliability indices (each individual has his or her own reliability estimate). In order to obtain a sample reliability, Raudenbush and Bryk (2002) suggested to average the person-specific reliability estimates.[4]

[4] In the context of structural equation modeling (SEM) based latent growth curve modeling, Rast and Hofer (2014) proposed a similar measure for estimating the reliability of inter-individual differences in rate of change

For the present work, we will refer to the reliability estimate introduced in Equation (8) as within-person coupling reliability (WPCR). Following suggestions by Raudenbush and Bryk (2002) we propose to compute a person-specific WPCR for each individual and average these estimates to obtain a sample-level reliability estimate that can be reported alongside other psychometric information in the method sections of empirical research. In our simulation study and the empirical example, we will, however, hold the within-person variance of the predictor (σ_X^2) constant across individuals. The reason for this choice is that estimating inter-individual differences in these variances can introduce substantial error. As shown by Estabrook, Grimm, and Bowles (2012), reliable estimates of intra-individual variability often require 50 repeated measurement occasions or more (see also Wang and Grimm, 2012). Consequently, the person-specific reliability index for inter-individual differences in within-person couplings we propose in the present work is defined as:

$$\mathrm{WPCR}_i = \frac{\tau_{11}^2}{\tau_{11}^2 + \frac{\sigma_e^2}{(T_i - 1)\cdot\sigma_X^2}} \tag{9}$$

and the sample reliability estimate, WPCR, is defined as the average of all person-specific reliability estimates:

$$\mathrm{WPCR} = \frac{1}{N}\sum_{i=1}^{N}\mathrm{WPCR}_i \tag{10}$$

As other reliability indices such as Cronbach's alpha or McDonald's omega, WPCR can take values between 0 and 1, with higher values indicating higher reliability. Before this estimate can be applied in empirical settings, two issues need to be addressed: First, prior research using inter-individual differences in within-person couplings has primarily used EB rather than OLS estimates as estimate for β_{1i}, following the general recommendation in the literature emphasizing the relatively smaller standard errors of the former (Hox, 2010;

(growth rate reliability). As we demonstrate in Appendix A (Supplemental Online Material), this measure is a special case of the measure proposed in the present work.

Snijders & Bosker, 1999; Raudenbush & Bryk, 2002). EB and OLS estimates are closely related but (everything else being equal) their association is attenuated with smaller Level-1 sample size (i.e., less measurement occasions; under these conditions, shrinkage is more pronounced). This raises the question if / under which conditions the OLS reliability index introduced above can be used as a proxy for the reliability of the EB estimates.

Second, the three parameters τ_{11}^2, σ_e^2, and σ_X^2 required to compute WPCR are usually unknown and need to be estimated from the data (see below for details). Biases in these estimates can, in turn, bias the WPCR estimate. We will address these two issues and investigate the performance of the WPCR as an estimate for the reliability of the EB estimate of β_{1i}, as well as the amount of bias in the estimation via a simulation study before we will turn to a demonstration of the application of the WPCR in real data.

Simulation Study

This simulation study was conducted to investigate the performance of the WPCR in Equation (10) as an estimate for the reliability of inter-individual differences in within-person couplings, operationalized as EB estimates. We varied the four parameters identified in Equation (9) that should impact the empirical reliability (which we defined as the squared correlation of the person's true coupling and the person's estimated coupling). Additionally, we varied the number of Level-2 units (participants) in order to determine whether this parameter affects the bias in WPCR estimates (see Table 1 for the population values of the parameters). We sought to simulate realistic scenarios that researchers working with intensive longitudinal designs might be confronted with. That is, we focused on a realistic sample size at Level-1 (5 to 100 repeated measurement occasions) and Level-2 (30 to 100 participants). Additionally, we introduced missing values in the data set to simulate the realistic scenario of an unbalanced design due to non-perfect compliance of study participants. Data were simulated in R (version 3.4.0), MLMs were estimated with the lme4 package (version 1.1-13;

Bates, Maechler, Bolker, & Walker, 2015) in R (R Core Team, 2018). Simulation scripts and data analysis scripts can be downloaded from https://osf.io/bdw5r/.

Method

We varied the four parameters that should impact the empirical reliability:

(1) Number of repeated measurement occasions for each participant (T_i)

(2) Level-1 residual variance (σ_e^2)

(3) True random slope standard deviation (τ_{11})

(4) Level-1 variance of the predictor X (σ_X^2)

Additionally, we varied (5) the number of Level-2 units (study participants) in order to determine whether this parameter affects the bias in WPCR estimates (because the accuracy of the variance components should increase with increasing number of participants). For each of these five design characteristics, we realized between three and six values (see Table 1). There were a total of 972 cells in our design. For each cell, 300 data sets were simulated, resulting in a total of 291,600 simulated data sets. Each participant was at first assigned *T* values (5, 10, 25, 40, 60 or 100, dependent on the condition) for the independent (time-varying) variable. These values were drawn from a normal distribution with mean 0 and variance σ_X^2 (0.5, 1 or 2; see Table 1). Then, each participant was assigned one random intercept score (υ_{0i}) and one random slope score (υ_{1i}) from a multivariate normal distribution with means of zero; the variance of the random intercept was set to 1, variances of the random slope varied across conditions as specified in Table 1. The correlation of random intercept and random slope was drawn from a uniform distribution and ranged from -.5 to .5. Values on the dependent variable were then predicted with error variance σ_e^2; values of fixed intercept and fixed slope were drawn from uniform distributions (see Table 1). To simulate unbalanced data we randomly distributed missing values for each individual; the number of missing values was drawn from a positively skewed distribution[5] to simulate typical missing data patterns in empirical data. The maximum rate of missing data per participant was 60%. That is, in the simulated data, we had two effects (intercept and the predictor X). Both effects were allowed to vary across Level-2 units (random-intercept, random-slope model with correlated random effects).

After data had been simulated, model parameters were re-estimated with the lme4 package (version 1.1-13; Bates et al., 2015) in R (R Core Team, 2018) using restricted maximum likelihood (REML) estimation. REML was chosen over the alternative full information maximum likelihood estimator because the latter estimator leads to attenuated estimates of the random variances (Raudenbush & Bryk, 2002).

The simulation approach allowed us to determine the empirical reliability of the EB estimates in each sample as the squared correlation of person *i*'s true coupling, υ_{1i}, and the EB estimate of person *i*'s coupling, υ_{1i}^{*} (this squared correlation yields the proportion of variance in the EB estimate that is accounted for by the true score variance; see Estabrook et al., 2012; Wang & Grimm, 2012).[6] In order to assess whether the WPCR estimate aligns with this sample reliability, we computed the WPCR for each sample. To that end, three parameters need to be estimated from the data (τ_{11}^2, σ_e^2, and σ_X^2). The parameters τ_{11}^2 and σ_e^2 are readily provided in the output of the lme4 package (and also by all other conventional statistical software packages); the within-person variance of X can be obtained as the Level-1 residual variance of a two-level model with X as the *dependent* variable and no predictors (referred to as the “intercept-only model”; Hox, 2010).

[5] Specifically, we created a distribution with the following expected proportions of missing values: 0% missing values were expected for 18% of the participants, 10% missing values for 30% of the participants, 20% missing values for 20% of the participants, 30% missing values for 11% of the participants, 40% missing values for 9% of the participants, 50% missing values for 7% of the participants, and 60% missing values for 5% of the participants.

[6] We also analyzed empirical reliabilities of the OLS estimates, $\hat{\upsilon}_{1i}$. Across all conditions, the reliabilities of OLS and EB estimates were nearly perfectly correlated, *r* = .96, and – importantly – empirical reliability was generally higher for EB estimates compared to OLS estimates; see Appendix B in the Supplemental Online Material for detailed results regarding the OLS estimates.

Results

There were convergence warnings for a total of 191 models (.07% of all models). Almost all (97.9%) of these data sets were in the conditions with the lowest random slope standard deviation (τ_{11}=.05; *n* = 147) and/or the fewest number of measurement occasions (*T* = 5; *n* = 50) and/or the fewest number of participants (*N* = 30; *n* = 88). These models were removed from further analyses. Negative correlations between υ_{1i} and υ_{1i}^{*} were observed for 5.7% of all data sets. Most of these cases (79.9%) appeared in the condition with the lowest true score variation in the random slopes (τ_{11}= .05). We set sample reliability for these cases to zero.

Figure 1 shows that the estimated reliability (WPCR) was highly correlated with the empirical reliability; the correlation became larger with increasing Level-2 sample size, but even for the smallest sample size (*N*=30) the correlation exceeded .95. Predicting the empirical reliability from the four factors associated with WPCR (*T*, σ_e^2, τ_{11}^2, and σ_X^2) and their interactions explained 95.0% of the variance of the true reliability. Adding the number of Level-2 units (*N*) as additional factor did, as expected, not noticeably improve the prediction of the sample reliability, Δ_{R^2} < .001. Re-running these analyses after replacing empirical reliability by WPCR as criterion yielded essentially the same results (R^2 = 95.1%; Δ_{R^2} < .001).

We next evaluated potential biases in the estimation. To that end, we first computed bias of the WPCR as the difference between WPCR and empirical reliability for each simulated data set (positive values represent overestimation of the sample reliability via the WPCR; results on relative bias can be found in Appendix B in the Supplemental Online Material). Bias was dependent on sample size in the expected direction (closer to zero with increasing Level-2 sample size; *N* = 100: *M* = .003, *SD* = .070, *Mdn* = .001, *skew* = -.297, *kurtosis* = 3.18; *N* = 60: *M* = .007, *SD* = .082, *Mdn* = .001, *skew* = -.004, *kurtosis* = 3.34; *N* = 30: *M* = .013, *SD* = .107, *Mdn* = .001, *skew* = .209, *kurtosis* = 3.34).

Next, we summarized the simulated studies into bins according to their estimated reliability, with a bin width of .05. Hence, studies with WPCR between 0 and .05 were summarized into the first bin, estimated reliabilities between .05 (excluded) and .10 into the second bin and so on. Median bias for each bin is depicted in Figure 2; this figure shows that bias is attenuated with increasing WPCR and with increasing Level-2 sample size. For example, for the reliability bin .275 (all simulated data sets with estimated reliability between .25 and .30), median bias for the conditions with 60 participants was .037, while this bias was reduced to .010 in the reliability bin .525. To further investigate the sources of the observed bias, we computed biases for τ_{11}^2, σ_e^2, and σ_X^2 analogously to biases in WPCR (as the difference between the estimated parameters and the true parameters). Across all simulated samples, bias in WPCR was correlated with bias in τ_{11}^2, *r* = .30, and with bias in σ_e^2, *r* = -.14, but uncorrelated with bias in σ_X^2, *r* = .00. These results suggest that part of the bias in WPCR can be attributed to biased estimates of τ_{11}^2 and σ_e^2 (see Appendix B in the Supplemental Online Material for more detailed results).

In summary, results showed that the WPCR performed adequately well as an estimate for the reliability of within-person couplings, obtained as empirical Bayes estimates: It slightly overestimated the empirical reliability, and the amount of overestimation was attenuated with increasing Level-2 sample size. Notably, overestimation was generally small if the reliability was .50 or higher. Although the estimate is biased for low reliabilities this bias is likely negligible from an applied perspective: reliability estimates below .50 might be biased in an upward direction but they will be evaluated as low regardless of a potentially positive bias. With reliability estimates above .55, bias is only small (less than .025) and arguably negligible from an applied perspective (e.g., there will be hardly any meaningful difference between an estimated reliability of .76 and the “true” sample reliability of .75). Having established the adequacy of the WPCR as an estimate for reliability of within-person couplings we now demonstrate its application in two scenarios: estimation of the WPCR based on published results and estimation of the WPCR from raw data.

Estimating WPCR from Published Results

Although, as mentioned in the introduction, previous research has not reported estimates of the reliability of within-person couplings, this reliability can be estimated post-hoc even without access to the raw data if four parameters are reported: the random slope variance (τ_{11}^2), the Level-1 residual variance (σ_e^2), the (average) number of repeated measurement occasions per person (*T)*, and the within-person variance of the focal predictor (σ_X^2). Of the above cited literature on predictive validity of stress reactivity, no study reported all of these four parameters, with the notable exception of Mroczek et al. (2015), who provided enough information to compute a rough estimate of the WPCR of stress reactivity. In this study, 181 men (58-88 years) participated in a daily diary design for eight consecutive days. Stress reactivity was assessed for both negative affect and positive affect (differences in negative affect and positive affect, respectively, between days with at least one stressor vs. without any stressor). Person-specific reactivity measures (within-person couplings of stressor occurrence and affect) were extracted and used as predictors for mortality risk over a ten year follow-up period. The core finding reported was that stress reactivity with regard to positive affect (but not negative affect) predicted mortality: Participants who showed larger decreases in positive affect on stressor days had higher mortality risk than participants with smaller positive affect stress reactivity did. However, it remains unclear whether these couplings of stress and positive affect were estimated with sufficient reliability. This would be an important prerequisite if these parameters were to be used as diagnostic tools or targets for intervention studies. This psychometric property is an important piece of information that has not been reported in this study (or any of the other cited studies utilizing stress reactivity as predictor for future outcomes). The WPCR can be used to estimate this reliability based on the information provided by Mroczek et al. (2015).

To compute the WPCR, we first need to know the variance of the within-person coupling estimates; the reported standard deviation was 1.15, resulting in a variance of 1.32. The second relevant parameter – the Level-1 residual variance – was 18.6. To approximate the number of repeated measurement occasions, there needs to be an indication of the average compliance and the maximum number of measurement occasions. Mroczek et al. (2015; p. 40) report a compliance of 99% for their study period of eight days (We suggest using the average number of measurement occasions in this case because we cannot estimate each participant's WPCR estimate without access to the raw data). Finally, we need the within-person variance of the focal predictor, which was the dichotomous variable indicating whether or not a stressor had occurred on the current day. Although this variance has not been reported, it has to be between 0 and 0.25 (the standard deviation of a dichotomous variable is always between 0 and 0.5). For the present example, we assume the largest possible variance of the predictor (0.25) which likely results in a too optimistic WPCR estimate (note that the variance would only be 0.25 if there were exactly as many stressor days as there were non-stressor days).

Substituting these values into Equation (9) yields an estimated WPCR of[7]

$$\mathrm{WPCR} \leq \frac{1.32}{1.32 + \frac{18.6}{(.99 \cdot 8 - 1) \cdot 0.25}} = .11 \tag{11}$$

Note that even this very poor reliability represents the upper limit of the estimate since we assumed an equal distribution of stressor and non-stressor days; a violation of this assumption will further attenuate the WPCR estimate (hence the less or equal sign in Equation (11)). We consider the WPCR estimate presented in the current work as a useful tool that researchers

[7] For ease of computation, the Excel sheet in the Supplemental Online Material can be used to arrive at this estimate.

can apply to report this psychometric information in their own work, allowing them and the readers of their work to discuss issues of measurement reliability more directly. In the following real-data example we demonstrate how the attached R code can be used to estimate the sample reliability from the raw data.

Empirical Example—Reliability of Stress Reactivity

In this section, we present an empirical example, demonstrating the application of the WPCR index to determine the reliability of stress reactivity estimates in a measurement burst study (for an introduction to measurement burst studies see Sliwinski, 2008).

Participants and Procedure

Data from 135 participants (103 female; M_{age} = 22.6 years, SD_{age} = 3.2) were collected over the course of eight weeks. Participants were instructed to complete an online-questionnaire at the end of the day for 21 consecutive days (Burst 1); after that, there was a break of two weeks before the study continued for another 21 consecutive days (Burst 2). Almost all (97.8%) of the participants were students at a large German university. Four participants dropped out after the first burst, and one participant provided only one measurement in the second burst. These five individuals were excluded from the analyses resulting in a final sample of $N = 130$. In the online-questionnaire participants were asked whether they had experienced one of seven daily hassles (e.g., “Today I was criticized or insulted”) and, if they had, to what extent this was perceived as distressing on a four point scale (ranging from “not at all” to “very much”). Responses on each of the seven items were coded as 1, if the participant had not encountered this hassle on this day. If they had experienced the hassle, but stated that this was not at all perceived as a burden, the response was also scored as 1; scores were set to 2, 3, and 4, if the respective hassle was perceived as slightly distressing, rather distressing, or very distressing, respectively. Responses on the seven items were averaged and log-transformed (in its original metric, the average was positively skewed and we reduced skewness by taking the logarithm). Further, participants completed a short version of the multidimensional mood state questionnaire (Steyer, Schwenkmezger, Notz, & Eid, 1997) which consists of 12 items assessing three dimensions of current mood (good-bad, awake-tired, calm-nervous) by four items each. Only the dimension good-bad is relevant for the current work.

Data were analyzed separately for the two measurement bursts. For each burst, the effect of (logarithmized) daily stress on mood was added as fixed and random effect and the covariance between random intercept and random slope was freely estimated. The predictors were centered on the person-mean of the first and second burst, respectively, to allow for an estimation of an unconfounded within-person effect (Wang & Maxwell, 2015). All data and analyses scripts can be retrieved from https://osf.io/bdw5r/.

Results

On average, participants provided data on 18.9 (*min* = 14, *max* = 21) and 17.8 (*min* = 7, *max* = 21) days in the first and second burst, respectively. Results of the multilevel models are depicted in Table 2. There was substantial random slope variance for the effect of day-to-day stress on daily mood, $\tau_{11}^2 \geq 1.06$, indicating that participants differed in the degree to which their mood was affected by today's stress level. We computed person-specific reliability estimates as introduced in Equation (9) using each participant's number of measurement occasions and averaged these estimates to obtain a sample reliability estimate (see Equation (10)) via the R code in the Supplemental Online Material.

Following the four steps described in Appendix C (Supplemental Online Material) with the burst 1 data set revealed an average reliability of WPCR = .489. Further information on the distribution of the person-specific WPCR estimates can be obtained by summarizing these parameters into other statistics such as the standard deviation, median, minimum, maximum and so forth. For the burst 1 data, these analyses revealed a standard deviation of the person-specific WPCR scores of .028, a median of .491, a minimum of .411, and a maximum of .517. Repeating these steps with burst 2 data showed that the average reliability dropped slightly (*M* = .445, *SD* = .060, *Mdn* = .468, *min* = .227, *max* = .494) but remained in a comparable range.

Finally, we correlated each individual's stress reactivity estimates for the two bursts (these were obtained as the EB estimates). The resulting correlation was not statistically significant, *r* = .14, *p* = .11. Overall, this finding suggests that estimates of inter-individual differences in stress reactivity are not stable over a break of two weeks, which dovetails with previous research showing substantial burst-to-burst variation in stress reactivity (Sliwinski, Almeida, Smyth, & Stawski, 2009), challenging the tacit and often untested assumption of stress reactivity being a stable, trait-like variable. Part of this low test-retest reliability might be attributed to low within-burst reliability, but it remains a task for future research to investigate whether and to what extent unreliability might account for this low stability.

Discussion

In the present work, we demonstrated the estimation of within-person coupling reliability delineated from a measure that has been proposed by Raudenbush and Bryk (2002) to estimate the reliability of inter-individual differences in within-person couplings. Prior research has shown that such parameters can predict a wide array of future outcomes (e.g., Gunthert et al., 2005; Mroczek et al., 2015), but so far it is unclear, whether these parameters are reliable measures of the true within-person coupling. In order to better judge the results from these empirical findings, it is important to have indicators about the reliability of inter-individual differences in these within-person couplings. The WPCR proposed in the present work shows satisfactory properties in that it converges well to the empirical reliability. Moreover, it can easily be computed using information provided from a multilevel model, and it can also be estimated individually for each Level-2 unit (participant) in the case of unbalanced data via the accompanying R code.

Improving Reliability of Inter-individual Differences in Within-person Couplings

Based on Equation (9), we can identify four parameters that affect the reliability of within-person couplings: the number of repeated measurement occasions (*T*), the intra-individual variance of the predictor (σ_X^2), the amount of variability of the random slopes (τ_{11}^2), and the within-person variance of the dependent variable that is not accounted for by the predictors in the model (Level-1 residual variance, σ_e^2).

Increasing the number of repeated measurement occasions will lead to an increase in WPCR (holding all other factors constant). In fact, by solving Equation (9) for T_i, the WPCR formula can be used to approximate the number of repeated measurement occasions needed to obtain a requested level of reliability (given that reasonable estimates for τ_{11}^2, σ_X^2 and σ_e^2 are available, for example form previous research or pilot studies):[8]

$$T_i = \frac{WPCR_i \cdot \sigma_e^2}{(1 - WPCR_i) \cdot \sigma_X^2 \cdot \tau_{11}^2} + 1 \qquad (12)$$

However, in many situations, considerations about *T* and σ_X^2 go hand in hand. For example, assessing infrequent behavior via random time interval sampling (questionnaires are triggered at random time points during the day) will result in low within-person variance if only few repeated measurements are taken (due to the infrequent nature, random prompts have a high probability of missing instances of this behavior). Increasing the frequency of assessment not only increases *T* but can also increase the within-person variance. For this type of research, using event-contingent assessment (instructing participants to fill in a questionnaire once the behavior is present) in combination with random sampling might, however, be a better choice since this will increase σ_X^2 without overburdening research participants: If participants are, for

[8] The Excel sheet (tab “Estimating T”) in the Supplemental Online Material can be used to compute the required number of measurement occasions.

example, instructed to report their negative affect specifically in situations of high stress (hence increasing the number of highly stressful occasions in the obtained data) in addition to random samples taken throughout their daily routine, this likely increases the within-person variance of stress.

Additionally, the temporal dynamics of the variables under study need to be considered. For example, if the research question targets within-person fluctuations in a variable subjected to circadian rhythms (e.g., body temperature), assessing these variables at the same time of the day for several days obscures a large part of the within-person variation that occurs within the day. In this example, adding more days to the assessment will increase WPCR, but a possibly more efficient way to increase both σ_X^2 and *T* would be to take repeated measures within days. Hence, it needs to be considered that not only *T*, but also the temporal design of the study, will likely influence the WPCR estimate: Adding more measurement occasions without considering the temporal dynamics of the variables under study will be a very inefficient way of improving WPCR.

Furthermore, reliability is expected to increase with increasing variance of within-person couplings (τ_{11}^2). Given that reliability is defined as the proportion of true score variance to total variance (i.e., the sum of true score variance and error variance), this effect is an algebraic necessity (holding the error variance constant). Unfortunately, researchers have only limited control over the size of this variability. However, researchers planning a study that focuses on inter-individual differences in within-person couplings should keep in mind to draw a sample from a population that is expected to show meaningful inter-individual differences in this variable.

A final factor affecting WPCR is Level-1 residual variance. Level-1 residual variance is conceptualized as within-person variance in the dependent variable that cannot be accounted for by the predictors in the model. Including additional Level-1 predictors that account for within-person fluctuations of the dependent variable can decrease this residual variance and, hence, increase the reliability of the within-person couplings. However, two caveats should be noted: First, including additional predictors changes the meaning of the couplings because these are partial regression coefficients. For example, if in addition to stress we also include time-varying inter-personal conflict into the model predicting negative affect, the person-specific regression coefficients of stress predicting negative affect need to be interpreted as the amount of change in negative affect associated with change in stress and holding inter-personal conflict constant (i.e., the effect of stressors other than inter-personal conflict). The interpretation of inter-individual differences in this within-person coupling, thus, becomes less straightforward with more Level-1 predictors in the model. Second, reduction in σ_e^2 only increases WPCR if it does not come at the expense of a reduction in τ_{11}^2. Staying with the current example, if including inter-personal conflict as predictor of negative affect reduces the random slope variance of stress to the same degree as it reduces the residual variance, WPCR remains unchanged.

In conclusion, multiple factors need to be taken into account to improve the reliability of within-person coupling estimates. While increasing the number of repeated measurement occasions is the most straightforward way to do so, constraints regarding study funding and participant burden often render this possibility unfavorable. Considering the expected temporal dynamics associated with the study variables can be helpful to maximize the within-person variability of these variables and in turn, to increase WPCR. Additionally, including heterogeneous samples at Level-2 increases the random slope variance (τ_{11}^2) and thereby WPCR as well. Finally, although adding Level-1 predictors into the model might reduce Level-1 residual variance, one needs to be aware that this may alter the meaning of the within-person couplings. Hence, caution is indicated when including additional predictors.

Reliability and Stability of Stress Reactivity

In the empirical part of the present work, we investigated both reliability of inter-individual differences in stress reactivity within bursts (across 21 days), as well as rank-order stability of these parameters across bursts. Our findings showed that these inter-individual differences could only be assessed with—at best—moderate reliability (below .50). Although WPCR can be interpreted in a similar fashion as other reliability indices, we are hesitant whether the various 'cut-offs' for interpreting reliability that are discussed as 'sufficient' in the literature should be applied to WPCR as well. Certainly, the standards for adequate reliability depend on the context in which the measure is used. Nunnally (1978), for example suggested that for reasons of efficiency in early stages of research reliabilities of .70 will suffice, while in applied settings much stricter criteria should be applied (>.90). The estimates reported in the empirical part of the present manuscript are certainly selective and not necessarily representative of all empirical research utilizing within-person couplings but they need to be considered as indicating rather poor reliability of these estimates in the investigated contexts (at least when compared to the standards conventionally applied to cross-sectional data).

Regarding stability, the test-retest correlation of stress reactivity over a short break of 14 days was very low. This suggests that stress reactivity – at least in the present sample of a student population – should not offhandedly be conceived of as a stable, trait-like person level variable. It should be noted that the empirical data presented in the current work are primarily intended as an example for the application of the WPCR index. Particularities regarding the sample (primarily consisting of students) potentially led to constrained inter-individual differences in stress reactivity. More heterogeneous samples are necessary to draw more definite conclusions about the reliability and stability of stress reactivity. However, we also estimated the reliability of stress reactivity in a different sample based on the results reported by Mroczek et al. (2015). Our analyses of the reliability of the stress reactivity parameters showed, however, that also in this study reliability was rather poor (estimated as

.11). Although based only on the empirical study reported in the present work and a single published study, these findings should be understood as a reminder that estimating the reliability of person-specific estimates of stress reactivity (and other within-person couplings) is a necessary step to provide insights into the meaning and potential malleability of these parameters.

One-Step Approach: Multilevel Structural Equation Modeling

In the present work we employed a two-step approach when using inter-individual differences in within-person couplings as person-level variables (which, based on the literature cited in the present work, seems to be the dominant approach taken by empirical researchers). That is, within-person couplings were saved in a first step (separately for each burst) and then the correlation between couplings at burst 1 and burst 2 was computed in a second step. An alternative one-step approach can be implemented in Mplus in the framework of multilevel structural equation modeling (MSEM; see Stapelton, 2013). In this one-step approach, inter-individual differences in within-person couplings are estimated as latent variables on the between-person level, which can be used as outcomes and predictors of other person-level variables. In one-step approaches, uncertainties in the estimation of latent variables are explicitly considered when estimating their association with other variables, which can be advantageous in various situations (for similar reasoning in the framework of criterion profile analysis see Davison, Chang, & Davenport, 2014).

The MSEM approach is highly valuable in that it accounts for unreliability in within-person couplings. However, researchers (and reviewers) typically want to know not only what the association of within-person couplings with external criteria (e.g., depression or mortality) is on the latent level. Instead, information on the reliability of within-person couplings is of crucial importance. This is particularly true when these parameters are intended to be used as diagnostic tools. For example, a researcher who aims at developing interventions specifically tailored to participants who are particularly high in stress reactivity needs to know if it is at all possible to "diagnose" these participants in her sample. If reliability is low then such person-specific interventions might fail because participants cannot be reliably separated. The WPCR index provides this information. We hasten to add that estimating the WPCR in combination with a one-step approach in an MSEM context is, of course, possible, because the relevant information to estimate this parameter can be obtained from the output of an MSEM as well.

Limitations and Future Directions

The present findings show that, although the WPCR can be a useful tool to estimate the reliability of inter-individual differences in within-person coupling estimates, the precision of this estimation depends on the factors associated with the precision in the estimation of τ_{11}^2, σ_e^2, and σ_X^2. Small sample size at Level-2 can be expected to reduce the precision of these variance estimates (Maas & Hox, 2004), resulting in less precise reliability estimates. However, results from the simulation study showed that in the present context this bias was negligible if 60 or more participants were included. That said, data were simulated to conform to "optimal" conditions regarding distributions (variables were drawn from normal distributions without floor or ceiling effects) and missing data patterns (missing data were missing completely at random). Future studies need to explore the effect of violations of distributional assumptions and missing data mechanisms on WPCR estimates.

Furthermore, inter-individual differences in σ_X^2 could be targeted by future studies to further account for inter-individual differences in WPCR. Those individuals who show more within-person variation of the predictor (e.g., more variation in stress across the observation period) will provide within-person coupling estimates (e.g., stress reactivity) that are more reliable (holding the other variables in Equation (9) constant). Estimating inter-individual differences in these variances can, however, introduce substantial error. Reliable estimates of inter-individual differences in intra-individual variability often require 50 measurement occasions or more (Estabrook et al., 2012; Wang & Grimm, 2012). In particular with few measurement occasions, estimating this variability in a Bayesian framework (Wiley, Bei, Trinder, & Manber, 2014) could be a fruitful alternative for future research, further supplementing the analysis of inter-individual differences in reliability estimates.

Finally, our findings provide important information for future research in personality. Baumert et al. (2017) have recently called for a better integration of three research foci in personality research: structure, process, and development. Inter-individual differences in within-person couplings of stress and negative affect (or other time-varying variables) can be understood as a way to more closely investigate personality processes (in fact, these couplings have also been referred to as capturing "within-person processes"; Bolger et al., 2003). Previous research linking personality structure (e.g., self-reported neuroticism) and process (stress reactivity) has shown evidence of convergence between these two approaches, generally linking neuroticism to higher stress reactivity (e.g., Mroczek & Almeida, 2004). Supplementing this prior research by longitudinal elements in future studies (e.g., examining the lead-lag association between structure and process in measurement burst designs) could further our understanding of the dynamic interplay between structure and process that could give rise to the development of personality.

Conclusions

Processes unfolding within individuals across time lie at the heart of a great amount of psychological theories. Under realistic conditions, the analysis of intensive longitudinal data is necessary to approach such within-person effects empirically (Hamaker, 2012). In the current study, we present a way to estimate the reliability of inter-individual differences in estimates of these within-person couplings. Results from our simulation study show that the proposed reliability index converges to the empirical reliability to a reasonable degree. Our findings can help researchers in two ways: First, these reliability estimates can be computed easily using the accompanying R code and can thus be reported along other psychometric information in an empirical study. Second, researchers can also plan the required number of measurement occasions in order to obtain a desired level of reliability. The empirical data provided in the present study suggest that estimates of a very prominent within-person effect – stress reactivity – can be estimated with only moderate reliability and that the test-retest correlation over two weeks is only very low. In conjunction with secondary analyses from published results, these findings suggest that assessment of person-specific stress reactivity might be not as reliable as often implicitly assumed.

References

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48. https://doi.org/10.18637/jss.v067.i01

Baumert, A., Schmitt, M., Perugini, M., Johnson, W., Blum, G., Borkenau, P., . . . Mõttus, R. (2017). Integrating personality structure, personality process, and personality development. *European Journal of Personality*, *31*, 503–528. https://doi.org/10.1002/per.2115

Bolger, N., Davis, A., & Rafaeli, E. (2003). Diary methods: Capturing life as it is lived. *Annual Review of Psychology*, *54*, 579–616. https://doi.org/10.1146/annurev.psych.54.101601.145030

Brose, A., Schmiedek, F., Lövdén, M., & Lindenberger, U. (2012). Daily variability in working memory is coupled with negative affect: The role of attention and motivation. *Emotion)*, *12*, 605–617. https://doi.org/10.1037/a0024436

Brose, A., Schmiedek, F., Lövdén, M., Molenaar, P. C. M., & Lindenberger, U. (2010). Adult age differences in covariation of motivation and working memory performance: Contrasting between-person and within-person findings. *Research in Human Development*, *7*, 61–78. https://doi.org/10.1080/15427600903578177

Bulut, O., Davison, M. L., & Rodriguez, M. C. (2017). Estimating between-person and within-person subscore reliability with profile analysis. *Multivariate Behavioral Research*, *52*, 86–104. https://doi.org/10.1080/00273171.2016.1253452

Charles, S. T., Piazza, J. R., Mogle, J., Sliwinski, M. J., & Almeida, D. M. (2013). The wear and tear of daily stressors on mental health. *Psychological Science*, *24*, 733–741. https://doi.org/10.1177/0956797612462222

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

Davison, M. L., Chang, Y.-F., & Davenport, E. C. (2014). Modeling configural patterns in latent variable profiles: Association with an endogenous variable. *Structural Equation Modeling: a Multidisciplinary Journal*, *21*, 81–93. https://doi.org/10.1080/10705511.2014.859507

Estabrook, R., Grimm, K. J., & Bowles, R. P. (2012). A Monte Carlo simulation study of the reliability of intraindividual variability. *Psychology and Aging*, *27*, 560–576. https://doi.org/10.1037/a0026669

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

Gunthert, K. C., Cohen, L. H., Butler, A. C., & Beck, J. S. (2005). Predictive role of daily coping and affective reactivity in cognitive therapy outcome: Application of a daily process design to psychotherapy research. *Behavior Therapy*, *36*, 77–88. https://doi.org/10.1016/S0005-7894(05)80056-5

Hamaker, E. L. (2012). Why researchers should think "within-person". A paradigmatic rationale. In M. R. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 43–61). New York: Guilford Press.

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2. ed.). *Quantitative methodology series*. New York, NY: Routledge.

Könen, T., Dirk, J., & Schmiedek, F. (2015). Cognitive benefits of last night's sleep: Daily variations in children's sleep behavior are related to working memory fluctuations. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *56*, 171–182. https://doi.org/10.1111/jcpp.12296

Lebel, E. P., & Paunonen, S. V. (2011). Sexy but often unreliable: the impact of unreliability on the replicability of experimental findings with implicit measures. *Personality & Social Psychology Bulletin*, *37*, 570–583. https://doi.org/10.1177/0146167211400619

Maas, C. J. M., & Hox, J. J. (2004). Robustness issues in multilevel regression analysis. *Statistica Neerlandica*, *58*, 127–137. https://doi.org/10.1046/j.0039-0402.2003.00252.x

Mroczek, D. K., & Almeida, D. M. (2004). The effect of daily stress, personality, and age on daily negative affect. *Journal of Personality*, *72*, 355–378.

Mroczek, D. K., Stawski, R. S., Turiano, N. A., Chan, W., Almeida, D. M., Neupert, S. D., & Spiro, A. (2015). Emotional reactivity and mortality: Longitudinal findings from the VA Normative Aging Study. *The Journals of Gerontology. Series B, Psychological Sciences and Social Sciences*, *70*, 398–406. https://doi.org/10.1093/geronb/gbt107

Neubauer, A. B., Lerche, V., & Voss, A. (2018). Inter-individual differences in the intra-individual association of competence and well-being: Combining experimental and intensive longitudinal designs. *Journal of Personality*, *86*, 698–713. https://doi.org/10.1111/jopy.12351

O'Neill, S. C., Cohen, L. H., Tolpin, L. H., & Gunthert, K. C. (2004). Affective reactivity to daily interpersonal stressors as a prospective predictor of depressive symptoms. *Journal of Social and Clinical Psychology*, *23*, 172–194. https://doi.org/10.1521/jscp.23.2.172.31015

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

R Core Team. (2018). *R: A language and environment for statistical computing [Software]*. Vienna, Austria: R Foundation for Statistical Computing.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks: Sage Publications.

Richard, A., Meule, A., Reichenberger, J., & Blechert, J. (2017). Food cravings in everyday life: An EMA study on snack-related thoughts, cravings, and consumption. *Appetite*, *113*, 215–223. https://doi.org/10.1016/j.appet.2017.02.037

Rubel, J. A., Rosenbaum, D., & Lutz, W. (2017). Patients' in-session experiences and symptom change: Session-to-session effects on a within- and between-patient level. *Behaviour Research and Therapy*, *90*, 58–66. https://doi.org/10.1016/j.brat.2016.12.007

Russell, M. A., Smith, T. W., & Smyth, J. M. (2016). Anger expression, momentary anger, and symptom severity in patients with chronic disease. *Annals of Behavioral Medicine*, *50*, 259–271. https://doi.org/10.1007/s12160-015-9747-7

Shrout, P. E., & Lane, S. P. (2012). Psychometrics. In M. R. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 302–320). New York: Guilford Press.

Sliwinski, M. J. (2008). Measurement-burst designs for social health research. *Social and Personality Psychology Compass*, *2*, 245–261. https://doi.org/10.1111/j.1751-9004.2007.00043.x

Sliwinski, M. J., Almeida, D. M., Smyth, J., & Stawski, R. S. (2009). Intraindividual change and variability in daily stress processes: findings from two measurement-burst diary studies. *Psychology and Aging*, *24*, 828–840. https://doi.org/10.1037/a0017925

Stapelton, L. M. (2013). Multilevel structural equation modeling with complex sample data. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling. A second course* (pp. 521–562). Charlotte, NC: Information Age Publ.

Steyer, R., Schwenkmezger, P., Notz, P., & Eid, M. (1997). *Der mehrdimensionale Befindlichkeitsfragebogen (MDBF) [The multidimensional mood questionnaire (MDMQ)]*. Göttingen: Hogrefe, Verl. für Psychologie.

Trull, T. J., & Ebner-Priemer, U. (2013). Ambulatory assessment. *Annual Review of Clinical Psychology*, *9*, 151–176. https://doi.org/10.1146/annurev-clinpsy-050212-185510

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

Wang, L., & Maxwell, S. E. (2015). On disaggregating between-person and within-person effects with longitudinal data using multilevel models. *Psychological Methods*, *20*, 63–83. https://doi.org/10.1037/met0000030

Wiley, J. F., Bei, B., Trinder, J., & Manber, R. (2014). Variability as a predictor: A Bayesian variability model for small samples and few repeated measures. Retrieved from https://arxiv.org/abs/1411.2961

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

Table 1

Simulation conditions.

Factor	Levels
Level-1 sample size (*T*)	5, 10, 25, 40, 60, 100
True random slope standard deviation of the predictor X (τ_{11})	.05, .20, .40
Level-1 residual variance (σ_e^2)	.001, .05, .15, .50, 1, 4
Within-person variance of the predictor X (σ_X^2)	.50, 1, 2
Level-2 sample size (*N*)	30, 60, 100
Fixed effect for predictor	U (-.50, .50)
Fixed intercept	U (3, 7)
Correlation of random intercept and random slope	U (-.50, .50)

Note. U = values were drawn from a uniform distribution.

Table 2

Results of the multilevel models.

	Burst 1	Burst 2
	Fixed Effects	
Intercept	5.30 (.08)	5.22 (.09)
Stress	-2.38 (.14)	-2.00 (.15)
	Random Variances	
Intercept	.94	1.06
Stress	1.16	1.06
Level-1 Residual	1.01	.93
	Within-person Variance	
Stress	.047	.043
	Reliability for Random Slope (Stress)	
WPCR	.489	.445

Note. Table depicts fixed effect estimates (standard errors in parentheses) and variances of the random variables. WPCR = within-person coupling reliability.

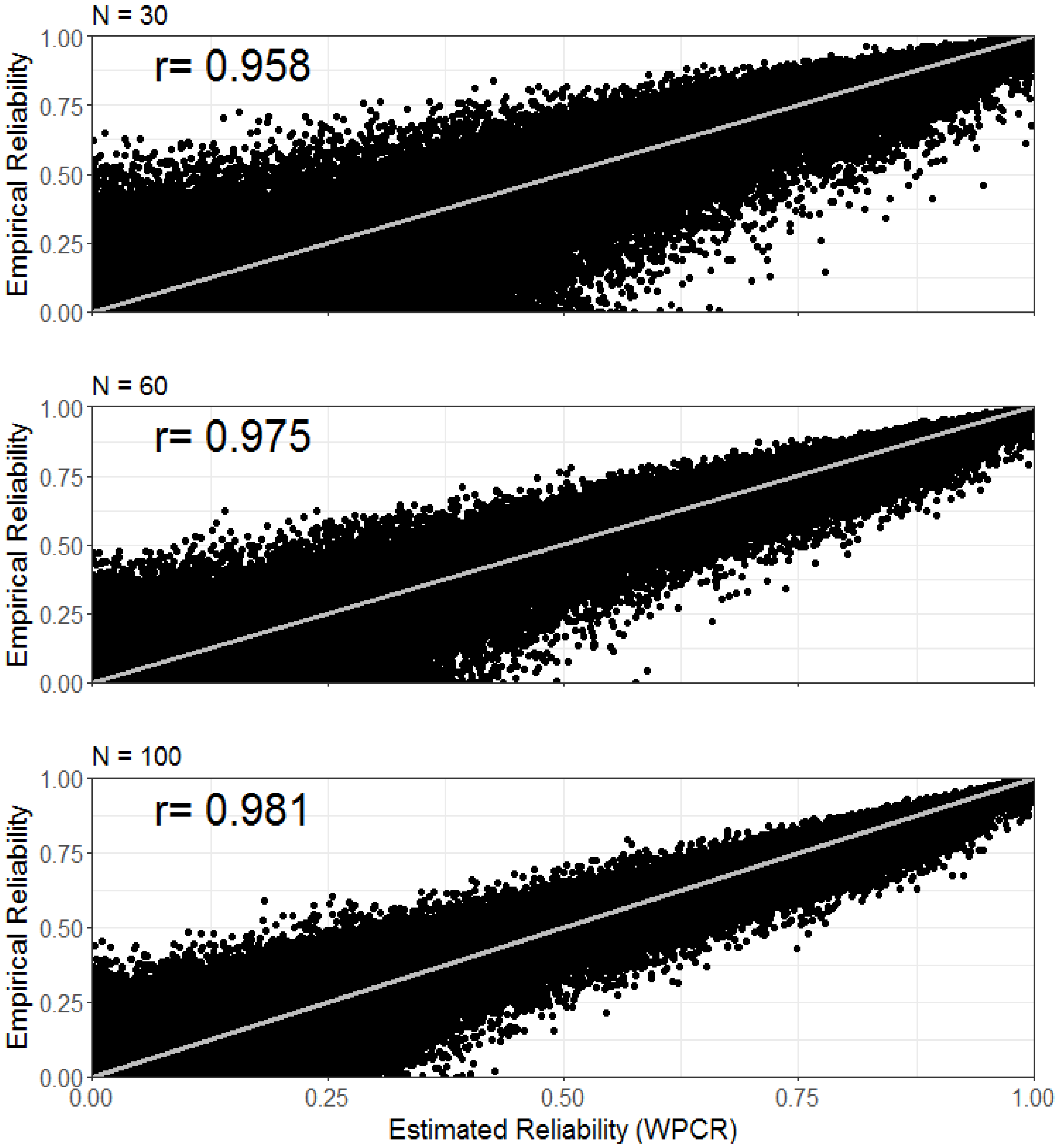

Figure 1. Empirical reliability plotted against the estimated reliability, separately for Level-2 sample size. WPCR = within-person coupling reliability. *N* = 291,409.

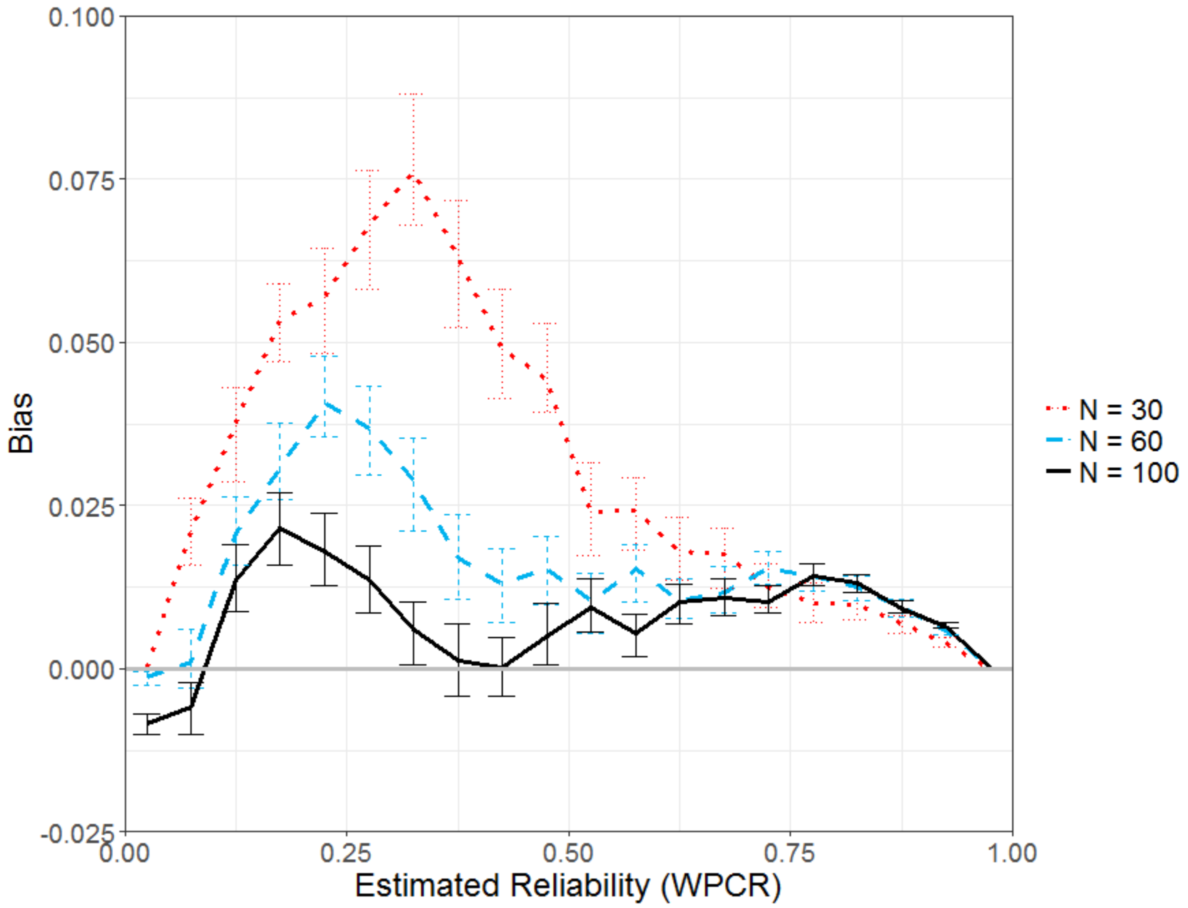

Figure 2. Median bias of the WPCR estimates plotted against the estimated reliability. Bias is depicted separately for the number of participants. Error bars indicate 95% bootstrap confidence intervals.