

Assessing the Internal Consistency Reliability of Ecological Momentary Assessment Measures: Insights From the WARN-D Study

Sebastian Castro-Alvarez¹, Di Jody Zhou¹, Laura F. Bringmann², Rayyan Tutunji³,
Ricarda K. K. Proppert³, Carlotta L. Rieble³, Eiko I. Fried³, and Siwei Liu¹

¹ Department of Human Ecology, University of California, Davis

² Department of Psychometrics and Statistics, University of Groningen

³ Department of Clinical Psychology, Leiden University

Intensive longitudinal research has become increasingly popular in the social and clinical sciences in recent years. However, this popularity has brought about many challenges for both methodological and empirical researchers, including challenges regarding measurement. In this preregistered study, we are particularly interested in the assessment of the reliability when multiple items are used to measure the same construct in intensive longitudinal data. This is important because reliability estimates are necessary (albeit not sufficient) to evaluate the quality of measures. Here, we evaluate the internal consistency reliability of scales used during Stage 2 of the WARN-D study, a 3-month period of daily and weekly measurements. The WARN-D study is a prospective 2-year study of approximately 1,750 students conducted in the Netherlands, aiming at building an early warning system for depression. Stage 2 includes 3 months of data on positive and negative affect measured four times a day and depression and anxiety measured once a week. To assess the reliability of each scale, we use six different statistical approaches including three simpler approaches that estimate the reliability at the between-person and within-person levels and three idiographic approaches that estimate person-specific reliability coefficients. This article also serves as a tutorial guide for substantive researchers, providing annotated code to facilitate estimating and reporting the reliability of ecological momentary assessment measures. We encourage all researchers to report the reliability of their data when applying the introduced statistical approaches, contributing to a collaborative effort toward developing more reliable measures in psychological and behavioral science.

Public Significance Statement

In this article, we present a tutorial for empirical researchers on estimating the internal consistency reliability of ecological momentary assessment data. We briefly introduce six approaches, which have been suggested for estimating this kind of reliability. To illustrate these approaches, we used data on depression, anxiety, and mood from the WARN-D study. The data include 1,167 participants, with up to 340 observations per person.

Keywords: internal consistency, reliability, ecological momentary assessment, intensive longitudinal data, early warning system

Supplemental materials: <https://doi.org/10.1037/pas0001410.supp>

Kasey Stanton served as action editor.

Sebastian Castro-Alvarez  <https://orcid.org/0000-0002-1326-0827>

Di Jody Zhou  <https://orcid.org/0000-0003-2118-2096>

Laura F. Bringmann  <https://orcid.org/0000-0002-8091-9935>

Rayyan Tutunji  <https://orcid.org/0000-0002-3537-9888>

Ricarda K. K. Proppert  <https://orcid.org/0000-0002-4225-2439>

Carlotta L. Rieble  <https://orcid.org/0000-0002-4764-3906>

Eiko I. Fried  <https://orcid.org/0000-0001-7469-594X>

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

Sebastian Castro-Alvarez and Di Jody Zhou contributed equally to this study and share first authorship. Eiko I. Fried and Siwei Liu contributed equally to this work and share senior authorship.

Sebastian Castro-Alvarez played an equal role in conceptualization, formal

analysis, visualization, and writing—original draft. Di Jody Zhou played an equal role in formal analysis, visualization, and writing—original draft. Laura F. Bringmann played an equal role in conceptualization, supervision, and writing—review and editing. Rayyan Tutunji played an equal role in data curation, investigation, and writing—review and editing. Ricarda K. K. Proppert played an equal role in data curation, investigation, and writing—review and editing. Carlotta L. Rieble played an equal role in data curation, investigation, and writing—review and editing. Eiko I. Fried played an equal role in conceptualization, funding acquisition, investigation, supervision, and writing—review and editing. Siwei Liu played an equal role in conceptualization, supervision, and writing—review and editing.

Correspondence concerning this article should be addressed to Sebastian Castro-Alvarez or Siwei Liu, Department of Human Ecology, University of California, Davis, 301 Shields Avenue, Davis, CA 95616, United States. Email: secastroal@gmail.com or sweliu@ucdavis.edu

The popularity of intensive longitudinal methods has seen an exponential growth during the last decade in the social and clinical sciences (Fritz et al., 2024; Hamaker & Wichers, 2017). Technological advancements have made it possible for researchers to easily collect large amounts of different types of data (e.g., self-report, biomarkers, and geolocation) in real time, which allows for studying dynamic psychological processes in great detail. Intensive longitudinal methods are also referred to in the literature as ecological momentary assessment (EMA), experience sampling methods, or ambulatory assessment; we will use EMA in the remainder of the article. In the clinical sciences, these methods are frequently used to measure mood and mental health symptoms. Typically, participants are asked to fill in short questionnaires multiple times a day for several weeks into how their experiences fluctuate throughout the day and for the duration of the study. This can be especially useful in clinical applications to monitor persons while in treatment (e.g., Riese et al., 2021) and to prevent relapse episodes of mental illness (e.g., Fried et al., 2023; Helmich et al., 2021).

While EMA methods have been received with enthusiasm by social and clinical researchers, they have also come with new methodological challenges. For example, measurement of constructs can be particularly difficult as the EMA scales need to be short to manage participant burden. In addition, there is likely time dependency among repeated measures. These data characteristics make the evaluation of measurement in this context challenging because traditional psychometric theories were developed to study (long) multi-item scales and independent observations across individuals (Crocker & Algina, 1986; Embretson & Reise, 2000). These issues have led to calls for a detailed description of the item properties of EMA data in this research area (for a tutorial, see Siepe et al., 2024), a necessary first step toward a comprehensive understanding of the validity of EMA measures.

One important psychometric property to study is the reliability of the scales. Reliability is commonly defined as the ratio of true to total variance. In other words, it indicates the proportion of variance of the observed scores that is explained by the underlying latent construct. Reliable measurements are a minimum requirement to ensure consistency and accuracy of the data and to make valid inferences. In EMA research, reliability coefficients have been defined for different levels of analysis because researchers are often interested in making valid inferences at both the between- and within-person levels (e.g., Castro-Alvarez, Bringmann, Back, & Liu, 2024; Neubauer & Schmiedek, 2020; Shrout & Lane, 2012).

Unfortunately, in the field of psychological dynamics, the reliability of the scales used in EMA is typically overlooked (Brose et al., 2020; Castro-Alvarez, Bringmann, Back, & Liu, 2024; Huang et al., 2023). Most EMA studies do not report the reliability of their measures or use approaches that are inadequate for EMA data. For example, researchers commonly estimate Cronbach's α while ignoring the nested structure of the data. It is also common to estimate the reliability based upon averages of the item responses taken over persons (Nezlek, 2017). These approaches are problematic because important assumptions of the methods are being ignored (e.g., independent observations) or because the estimated reliability coefficients do not align well with the purpose of the data analysis. Specifically, when tackling the estimation of the reliability of EMA scales, it is crucial to distinguish between-person reliability and within-person reliability (Neubauer & Schmiedek, 2020), which serve different purposes and answer conceptually different

questions. While the between-person reliability indices are measures of the reliability of the persons' means and are relevant to between-person comparisons, the within-person reliability indices indicate to what extent the measurement procedure is able to capture true intraindividual variability over time (Neubauer & Schmiedek, 2020). Furthermore, the large amount of data per individual allows for estimating person-specific reliability coefficients using an idiographic approach (e.g., Hu et al., 2016; Schuurman & Hamaker, 2019; Xiao et al., 2023), which may be important when researchers want to focus on the individuals rather than the whole sample. Such idiographic approaches may also be useful for identifying participants with unusual data, for example, for the potential detection of careless responding.

In this article, we examine and apply different methods that have been proposed to estimate the internal consistency reliability of EMA scales. We also include code in the [Supplemental Materials](#) to help applied researchers fit these models to their own EMA data. We used six different approaches, which were recently reviewed by Castro-Alvarez, Bringmann, Back, and Liu (2024), to estimate the reliability of the scales used during Stage 2 of the WARN-D study. The WARN-D study (Fried et al., 2023) is a multicohort longitudinal study conducted in the Netherlands, which collected rich data on depression, anxiety symptoms, and daily emotions. The goal of this article was to serve as an easy-to-read tutorial of the different approaches for applied researchers.

Using multiple approaches to estimate the reliability of EMA scales allows us to (a) obtain a more detailed and robust understanding of the EMA scales used in WARN-D, at the group level, idiographic level, between-subjects level, and within-subjects level. Simultaneously, it (b) enables us to compare with what extent the results across the different statistical approaches are consistent.

In what follows, we introduce the WARN-D data and the EMA scales in more detail and briefly describe each of the approaches used to estimate the internal consistency reliability. After presenting the results, we then discuss the implications of these results for the WARN-D study and highlight the importance of consistently assessing and reporting the reliability of EMA scales in research.

Method

WARN-D Data

The WARN-D study is a multicohort and multistage study with the goal to build an early warning system for depression in students. Participants were current students of Dutch institutions of higher education (i.e., vocational schools, technical universities, and universities) and were given the option to fill in the surveys in either English or Dutch. Additional inclusion criteria were being older than 18 years old and living in the Netherlands, Germany, or Belgium. Participants were excluded from the study during screening if they had moderate levels of depression, mania, thought disorders, substance use disorders, were currently under treatment for mental health problems, or found it overwhelming to have their burned calories tracked. The study consisted of four cohorts, each with approximately 500 students.

For this article, we took a subsample of 1,167 participants (approximately equally distributed across the four cohorts) and analyzed data collected during Stage 2. The rest of the sample is a preregistered holdout sample for predictive empirical work using

cross-validation techniques (Tutunji et al., 2024). From the selected subsample, 84.5% were female, and 52.4% filled in the surveys in English. In Stage 2, participants completed an EMA study for 85 (86 in the third cohort) days on their smartphone. Participants received a prompt to fill in the EMA surveys four times a day, for a maximum of 340 (344 in the third cohort) prompts. Additionally, participants completed another survey once a week on Sundays, for a maximum of 12 weeks. In this study, we are interested in the scales used to measure positive and negative affect in the EMA surveys and depression and anxiety in the weekly surveys. For more details and a full description of the study protocol and questionnaires used in the different stages, see Fried et al. (2023) and <https://osf.io/2jd9h/> (Fried et al., 2025a).

Note that WARN-D data collection is ongoing, and we want to avoid having different small parts of the data shared across many projects. We will therefore make data available (excluding potentially identifiable data) on the WARN-D project hub at <https://osf.io/frqdv/> (Fried et al., 2025b) when all data are collected, cleaned, and checked. We share the participant IDs used for this article at <https://osf.io/wupyv/> (Castro-Alvarez et al., 2024a) to make the article reproducible in the future. Data collection was approved by the Leiden University Research Ethics Committee (2021-09-06-E.I.FriedV2-3406). The project is funded by the European Research Council in the Horizon 2020 research and innovation program (Grant 949059).

Weekly Measures

Depression

An adapted version of the nine-item Patient Health Questionnaire (PHQ-9; Kroenke & Spitzer, 2002) was used to measure depression symptoms once a week. In the weekly surveys, four of the original items were disaggregated into two independent items to obtain more detailed information on the nature of symptoms. For example, the PHQ-9 item “This week I was bothered by the following problems: trouble falling or staying asleep, or sleeping too much” was split into two items: one focusing on “insomnia” symptoms and the other one on “hypersomnia” symptoms. Moreover, additional symptoms were asked for a total of 15 items. However, for the analyses in this article, the scores of the original nine items were constructed back by taking the maximum score between the two disaggregated items and by excluding the additional symptoms, as similarly done by Siepe et al. (2024). This scale was measured on a 4-point Likert scale ranging from 1 (*not at all*), 2 (*several days*), 3 (*more than half the days*), to 4 (*nearly every day*).

Anxiety

The seven-item measure to assess generalized anxiety disorder (GAD-7; Spitzer et al., 2006) was used to measure anxiety disorder symptoms. Due to the very strict limit of items feasible when asking participants to fill out EMA surveys for 3 months, two of the original GAD items—“Being so restless that it is hard to sit still” and “Becoming easily annoyed or irritable”—were excluded from the WARN-D study because they overlapped with questions already asked in the PHQ-9. Similar to the measure of depression symptoms, this scale was also measured on a 4-point Likert scale ranging from

1 (*not at all*), 2 (*several days*), 3 (*more than half the days*), to 4 (*nearly every day*).

EMA Measures

Positive and Negative Affect

To measure positive affect (PA) and negative affect (NA), the EMA surveys included the following 10 emotions and related states: cheerful, motivated, relaxed, sad, stressed, overwhelmed, nervous, ruminate, irritable, and tired. The first three emotions were aimed to measure PA, and the last seven were aimed to measure NA. Most items of the EMA survey were presented as “I feel sad right now.” However, for the item ruminate, the wording was “I am experiencing negative thoughts right now.” All emotions were measured on a 7-point Likert scale ranging from 1 (*not at all*) to 7 (*very much*).

Statistical Analyses

To estimate the reliability of the different scales, we used up to six different approaches suitable for intensive longitudinal data, which have been described in a recent overview by Castro-Alvarez, Bringmann, Back, and Liu (2024). These approaches are (a) the generalizability theory (GT; Cranford et al., 2006), (b) the multilevel modeling framework (MLM; Nezlek, 2017), (c) the multilevel confirmatory factor analysis model (ML-CFA; Geldhof et al., 2014; Lai, 2021), (d) the p-technique factor analysis (PT-FA; Hu et al., 2016), (e) the two-level random dynamic model-based approach (2RDM; Xiao et al., 2023), and (f) the mixed-effects state-trait-occasion model (ME-TSO; Castro-Alvarez et al., 2022). For the weekly measures PHQ-9 and GAD-7, we only used the first three approaches as a maximum of 12 data points is not suitable to fit the remaining models. For the scales of PA and NA, all approaches were used. When estimating the reliability based on GT, MLM, and ML-CFA, all available data were used, regardless of the compliance rate of the participants. By contrast, when estimating the reliability based on the PT-FA, 2RDM, and ME-TSO models, we only included participants who completed at least 100 prompts of the EMA study and whose variances were different from 0 in all items. Furthermore, for the analyses, we divided the sample into two subsamples by language and analyzed each subsample independently. We did this because the scales in different languages are technically different and it might be the case that there is measurement noninvariance across language groups. Next, we briefly describe each of the six approaches. Analyses for this study were preregistered (see <https://doi.org/10.17605/OSF.IO/QEY9C>; Castro-Alvarez et al., 2024b).

GT

Cranford et al. (2006) suggested a procedure to assess the reliability of daily diary measures based on the GT to distinguish among different sources of variability and error. These sources of variability are the persons, the items, the measurement occasions, and their two-way interactions. To identify these sources and how much they contribute to the total variance, a linear random-effects model is typically used. Then, the estimated variances for each component are used to compute meaningful reliability coefficients. These coefficients are defined as variance ratios wherein the denominator (weighted sum of variances) represents the “total variance” and the

numerator is a fraction of this “total variance.” Five reliability coefficients have been suggested within this approach (Cranford et al., 2006; Shrout & Lane, 2012). In these analyses, we report the reliability of the average of measures taken over the planned random occasions (R_{kr}) and the reliability of change (R_c). R_{kr} indicates the reliability of the persons’ means and can be seen as a between-person reliability coefficient. R_c indicates the reliability of the change over time within the persons and can be seen as a within-person reliability coefficient.

MLM

Multilevel modeling techniques have also been suggested as an approach to estimate the reliability of intensive longitudinal data (Nezlek, 2017; Nezlek & Gable, 2001). In this approach, the procedure consists of fitting an unconditional three-level multilevel model to the data in which the items are the Level 1 units, the occasions are the Level 2 units, and the persons are the Level 3 units. This model estimates three variance components, one per level, which are used to compute two reliability coefficients: between-person and within-person reliability. In this approach, the reliability coefficients are also defined as variance ratios. On the one hand, the between-person reliability indicates how reliable the mean responses of the persons are. On the other hand, the within-person reliability indicates how consistent the responses of the persons are on a given occasion. According to Nezlek (2017), the within-person reliability is equivalent to the traditional Cronbach’s α while accounting for the nested structure.

ML-CFA

A popular approach to estimate the reliability in intensive longitudinal settings is the multilevel confirmatory factor model as presented by Geldhof et al. (2014). This procedure was later revisited and improved by Lai (2021). In this approach, a two-level confirmatory factor model is fitted to the data. The model has the following characteristics: (a) Factor loadings are assumed to be equal across levels, (b) the variance of the within-level latent factor is fixed at 1 for identification purposes, and (c) the variance of the between-level latent factor is freely estimated. Then, the estimated parameters are used to compute the reliability coefficients at each level. In this model, the reliability coefficients are defined based on McDonald’s ω coefficient adjusted for each level. Briefly, McDonald’s ω is defined as the ratio of the variability explained by the items to the total variance of the scale (McNeish, 2018), where the total variance is computed based on the estimated factor variance, error variances, and factor loadings. Thus, in this approach, each of the suggested coefficients has a similar formula to McDonald’s ω but using the parameter estimates corresponding to each level. First, the between-person reliability coefficient is a measure of the reliability of the between-person composite scores, meaning the persons’ means. By contrast, the within-person reliability coefficient is a measure of the reliability of the within-person composite scores, which are defined as the deviations from the persons’ means.

PT-FA

Taking an idiographic approach, PT-FA (Molenaar, 1985) also allows for estimating the reliability of psychological time series data

(Hu et al., 2016; Shrout & Lane, 2012). Here, the procedure consists of fitting a confirmatory factor model to the time series data of each person and computing McDonald’s ω based on the estimated parameters of each fitted model. As a result, there are as many reliability coefficients as there are participants in the study. Each estimated reliability coefficient is a measure of the internal consistency of the responses of a given participant.

2RDM

The two-level random dynamic model-based approach (Xiao et al., 2023) is a Bayesian multilevel dynamic factor model encompassed within the dynamic structural equation modeling framework (Asparouhov et al., 2018). This model is characterized by (a) being a two-level confirmatory factor model, (b) incorporating lagged effects likely to be found in intensive longitudinal data, and (c) allowing parameters at the within-level model (i.e., factor loadings, lagged effects, measurement error variances, and dynamic error variances) to vary randomly across persons. Based on this model, one can estimate one reliability coefficient at the between-level and person-specific reliability coefficients for each person. The between-person reliability is computed using McDonald’s ω on the between-level estimated parameters and has the same interpretation as the between-person reliability computed with Geldhof et al.’s (2014) approach. The person-specific reliability coefficients are also computed based on McDonald’s ω but adjusted to account for the lagged effects and the dynamic error variances by incorporating these parameters into their computation. These person-specific coefficients describe to what extent the items in the scale fluctuate in the same direction at a given measurement occasion.

ME-TSO

The last approach is the so-called mixed-effects trait-state-occasion model (Castro-Alvarez et al., 2022). This model is encompassed within the latent state-trait theory (Eid et al., 2017; Steyer et al., 2015) and was especially developed to analyze intensive longitudinal data. The ME-TSO can also be described as a Bayesian multilevel confirmatory dynamic factor model, which was implemented within the dynamic structural equation modeling framework (Asparouhov et al., 2018). In the ME-TSO, single-indicator latent variables are used at the between level, and a unidimensional latent structure is used at the within level. Moreover, the within-level structure also incorporates lagged effects that vary randomly across persons. By contrast, the factor loadings, measurement error variances, and dynamic error variances are fixed parameters. As a latent state-trait theory model, in the ME-TSO, several variance coefficients are defined for each item, one of which is the reliability coefficient. Due to the random lagged effects, the reliability coefficients are also computed for each person. As a result, when estimating the reliability based on the ME-TSO, there are $I \times N$ reliability estimates, with I the number of items and N the number of persons. In this approach, the reliability coefficient is a variance ratio of the variability explained by the reliable source of variance (true score) to the total variance of each indicator or item. The reliable sources of variance include both the time-invariant (trait) and the time-varying (state) components.

For the interpretation of the point estimates of the different reliability coefficients, we follow the guidelines proposed by Shrout (1998). In these guidelines, reliability point estimates from 0.00 to

0.10 indicate virtually no reliability, from 0.11 to 0.40 indicate slight, from 0.41 to 0.60 indicate fair, from 0.61 to 0.80 indicate moderate, and from 0.81 to 1.00 indicate substantial/high reliability. Note, however, that these kinds of guidelines should not be taken as ground truths as their usefulness also depends on the context.

Results

In the following, we present the reliability of the weekly and EMA questionnaires based on the different approaches. Detailed results including descriptive statistics, item distributions, compliance rates, and code to perform the analyses are available on the [Supplemental Materials](#) and the Open Science Framework at <https://osf.io/wupyv> (Castro-Alvarez et al., 2024a).

Weekly Questionnaires

Out of the initial sample of 1,167 participants in this study, 1,127 (591 in the English subsample and 536 in the Dutch subsample) completed at least one of the weekly measurements. In the English subsample, the median number of complete measurements was 10, with Quartile 1 = 6 and Quartile 3 = 12. In the Dutch subsample, the median number of complete measurements was 10, with Quartile 1 = 7 and Quartile 3 = 12.

For the adapted PHQ-9 and adapted GAD-7 scales, we estimated the reliability based on the GT, MLM, and ML-CFA approaches. We only used these approaches because there were not enough repeated measures to consider applying the more complex models. The estimated between- and within-person reliability coefficients for both scales and both subsamples given each approach are presented in [Table 1](#). First, for the estimated between-person reliabilities of the PHQ-9, the results show that the three approaches had a very high between-person reliability in both subsamples (varying between 0.83 and 0.97). While the between-person reliability based on the ML-CFA approach was lower by more than 0.1, the estimated coefficient was still very high. This indicates that the persons' mean scores are a good estimate of their average depression severity

Table 1
Estimated Reliability Coefficients of the Weekly Questionnaires

Sample language and level of analysis	GT	MLM	ML-CFA
PHQ-9			
English			
Between	0.96	0.96	0.84
Within	0.70	0.45	0.73
Dutch			
Between	0.97	0.97	0.83
Within	0.64	0.22	0.66
GAD-7			
English			
Between	0.95	0.95	0.86
Within	0.77	0.65	0.78
Dutch			
Between	0.95	0.95	0.84
Within	0.69	0.46	0.72

Note. GT = generalizability theory; MLM = multilevel modeling framework; ML-CFA = multilevel confirmatory factor analysis model; PHQ-9 = nine-item Patient Health Questionnaire; GAD-7 = seven-item measure to assess generalized anxiety disorder.

during the Stage 2 of the WARN-D study. By contrast, there are some differences in the estimated within-person reliability coefficients. For example, in both subsamples, the within-person reliabilities based on the GT and ML-CFA were relatively similar (between 0.64 and 0.73), suggesting that the scale's within-person reliability is moderate. However, the within-person reliability based on MLM was lower than the other methods for both subsamples, especially in the Dutch subsample, where the estimate was 0.22. Generally, the within-person reliability of the PHQ-9 in the Dutch subsample was lower compared with the English subsample across methods.

The estimated between- and within-person reliabilities of the GAD-7 followed similar trends to those observed for the PHQ-9. The between-person reliability coefficients were high and similar across methods (between 0.84 and 0.95), and the within-person reliabilities were generally moderate (mostly larger than 0.65). Differences of results across methods and subsamples were relatively minor, especially for the within-person reliabilities whose differences in the GAD-7 across methods were less pronounced than those observed in the PHQ-9. Overall, the within-person reliability estimates of the GAD-7 were generally higher than the within-person reliability estimates of the PHQ-9. This is somewhat surprising considering that the GAD-7 is a shorter scale than the PHQ-9.

EMA Questionnaires

For the EMA questionnaires, there were 611 participants in the English and 556 in the Dutch subsample. All participants completed at least one EMA measurement, and the maximum number of completed assessments was 338. The distribution of the compliance for each subsample appears to be flat. In the English subsample, the median number of measurements completed was 190, with Quartile 1 = 92.5 and Quartile 3 = 257. Similarly, in the Dutch subsample, the median number of measurements completed was 185, with Quartile 1 = 98.75 and Quartile 3 = 254.25. As explained in detail in the Method section, some participants were excluded when estimating the reliability based on the PT-FA, 2RDM, and ME-TSO approaches. Therefore, for these approaches, the number of participants was reduced to 441 and 401 for the English and the Dutch subsamples, respectively.

The results from estimating the reliability of the items of PA and NA based on each of the six approaches are presented in [Table 2](#). The estimated between- and within-person reliability coefficients for the GT, MLM, and ML-CFA methods are presented, similar to how they are shown in [Table 1](#). For the PT-FA and 2RDM approaches, the within-person reliabilities shown are the medians of the estimated person-specific reliabilities for each subsample. The median person-specific reliability estimated by PT-FA is based on smaller sample sizes due to convergence issues or improper solutions (English PA: $N = 436$, English NA: $N = 430$, Dutch PA: $N = 398$, Dutch NA: $N = 383$). No results are shown for the ML-CFA when fitted to the items of PA in both subsamples because the solution was improper. No results are shown for the 2RDM when fitted to the items of NA in the English subsample as this model did not converge. Last, for the ME-TSO, the within-person reliability shown is the median across items after taking the median of the person-specific reliabilities across persons per item.

For both sets of items and in both subsamples, the between-person reliabilities estimated by the GT and MLM were 1, indicating that the persons' mean scores were perfectly consistent. This result can be counterintuitive under the premise that measurement error is

Table 2
Estimated Reliability Coefficients of the EMA Questionnaires

Sample language and level of analysis	GT	MLM	ML-CFA	PT-FA	2RDM	ME-TSO
PA						
English						
Between	1.00	1.00	—	N/A	0.88	N/A
Within	0.63	0.50	—	0.62 ^a	0.65 ^a	0.49 ^b
Dutch						
Between	1.00	1.00	—	N/A	0.85	N/A
Within	0.61	0.43	—	0.61 ^a	0.63 ^a	0.50 ^b
NA						
English						
Between	1.00	1.00	0.93	N/A	—	N/A
Within	0.79	0.68	0.80	0.76 ^a	—	0.64 ^b
Dutch						
Between	1.00	1.00	0.93	N/A	0.95	N/A
Within	0.74	0.53	0.75	0.70 ^a	0.67 ^a	0.64 ^b

Note. — = not available due to model not converging to a proper solution. EMA = ecological momentary assessment; GT = generalizability theory; MLM = multilevel modeling framework; ML-CFA = multilevel confirmatory factor analysis model; PT-FA = p-technique factor analysis; 2RDM = two-level random dynamic model-based approach; ME-TSO = mixed-effects state-trait-occasion model; PA = positive affect; NA = negative affect; N/A = not applicable.

^a Median of the estimated person-specific reliability coefficients. ^b Median of the medians over persons of the person- and item-specific reliability coefficients.

always present. However, in the context of EMA data, having “perfectly consistent” scores at the between-level can be justified due to the large number of measurement occasions. Necessarily, the persons’ means tend to be measured more precisely as one increases the number of measurement occasions. By contrast, when available, the between-person reliabilities estimated by the ML-CFA and 2RDM were lower, although their estimated coefficients still tended to be high, with values ranging from 0.85 and 0.95.

Regarding the within-person reliabilities estimated for the PA scale, there were no major differences when comparing the two subsamples across methods. Generally, the within-person reliabilities were higher according to the GT and ML-CFA approaches compared with the MLM approach. For the items of PA, the estimates according to MLM were 0.50 and 0.43 for the English and the Dutch subsamples, respectively, indicating that the reliability was fair. By contrast, the estimates according to GT were 0.63 and 0.61 for the English and Dutch subsamples, indicating that the reliability was moderate. The results based on the PT-FA, 2RDM, and ME-TSO approaches are not directly comparable as they estimate person-specific reliability coefficients, but they still provide a general idea about the performance of the scale for each subsample. When looking at these summary statistics, there are no major differences between subsamples; the largest difference between languages is 0.02. The values of these medians are generally similar to the ones estimated by the simpler methods, varying between 0.49 and 0.65.

The within-person reliability estimates of the NA scale followed similar patterns to those observed for PA. Overall, NA within-person

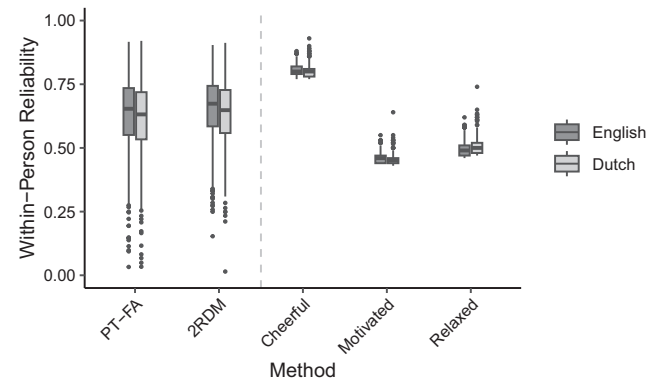
reliabilities were higher than the PA within-person reliabilities, indicating moderate levels of reliability based on the GT, MLM, and ML-CFA approaches. These coefficients varied between 0.53 and 0.8. This is not surprising given that the NA scale had more items than the PA scale. Similarly, the median person-specific reliabilities according to the PT-FA, 2RDM, and the ME-TSO had moderate values between 0.64 and 0.76. Furthermore, a slight difference between the two subsamples emerges, in which the within-person reliability of the scores of the English subsample seems to be higher than those of the Dutch subsample for almost all approaches.

Person-Specific Reliabilities

The advantage of the PT-FA, 2RDM, and ME-TSO approaches is that they allow zooming in at the individual level, providing a more fine-grained picture of the scales’ performances in the subsamples. Figure 1 presents the boxplots of the person-specific reliabilities for PA according to these three methods. The ME-TSO method yields person-specific reliabilities per item, each represented by a boxplot. The within-person reliabilities according to PT-FA varied between 0.03 and 0.92 with similar negatively skewed distributions for the two subsamples. The within-person reliabilities of the 2RDM varied between 0.01 and 0.91, with negatively skewed distributions. The correlation of the within-person reliabilities between the PT-FA and 2RDM approaches was 0.88 and 0.89 for the English and the Dutch subsamples, respectively. By contrast, the distributions of the item-specific reliabilities estimated by ME-TSO tended to be narrow and positively skewed. There are clear differences among items, where *cheerful* is most reliable with a median of 0.80 in both subsamples. The correlations of the person-specific reliabilities ranged from 0.26 to 0.37 between ME-TSO and PT-FA and from 0.34 to 0.47 between ME-TSO and 2RDM.

The boxplots of the within-person reliabilities of NA according to the three approaches are presented in Figure 2. For the PT-FA, the distributions were negatively skewed and slightly different across subsamples, ranging from 0.09 to 0.95 in the English subsample and

Figure 1
Positive Affect Person-Specific Reliabilities Based on PT-FA, 2RDM, and Mixed-Effects State-Trait-Occasion Model

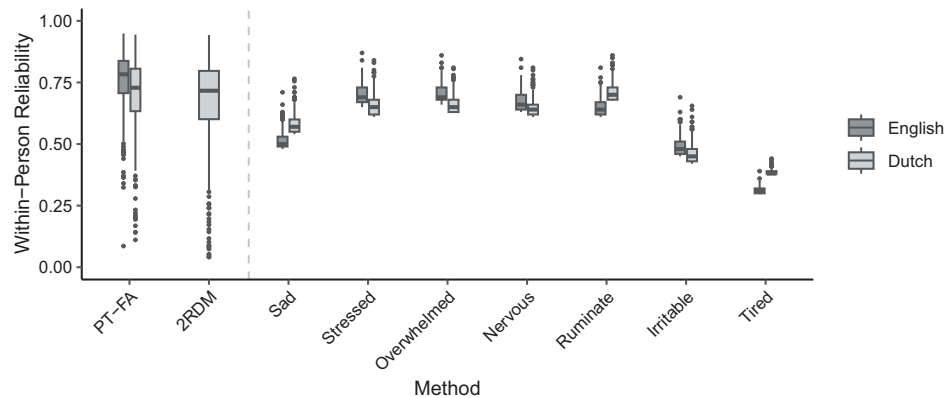


Note. PT-FA = p-technique factor analysis; 2RDM = two-level random dynamic model-based approach.

This document is copyrighted by the American Psychological Association or one of its allied publishers. This article is intended solely for the personal use of the individual user and is not to be disseminated broadly. All rights, including for text and data mining, AI training, and similar technologies, are reserved.

Figure 2

Negative Affect Person-Specific Reliabilities Based on the PT-FA, 2RDM, and Mixed-Effects State-Trait-Occasion Model



Note. NA = negative affect; PT-FA = p-technique factor analysis; 2RDM = two-level random dynamic model-based approach.

from 0.11 to 0.94 in the Dutch subsample. For the 2RDM, only the estimated within-person reliabilities for the Dutch subsample are available given that the model did not converge for the English subsample, which varied between 0.04 and 0.94. Its correlation with the PT-FA estimates was 0.95. The results based on the ME-TSO show some variability in the reliability estimates between items. Within items, there is little variability, and the distributions are positively skewed. Furthermore, there are noticeable differences between subsamples; the item with the lowest median reliability was *tired* for both subsamples (0.31 for the English and 0.38 for the Dutch). The items with the largest median reliability for the English subsample were *stressed* and *overwhelmed*, both with a median of 0.69, and the item with the largest median reliability for the Dutch subsample was *ruminative* with a median of 0.70. In particular, with the items *sad*, *ruminative*, and *tired*, it seems there are differences in the reliability estimates by language. The correlations between the estimates of the ME-TSO and the PT-FA varied between 0.50 and 0.62, and the correlations between the estimates of the ME-TSO and the 2RDM varied between 0.49 and 0.63.

Discussion

In this article, we studied the reliability of four scales assessed during Stage 2 of the WARN-D study (Fried et al., 2023) using different approaches suitable for longitudinal (nested) data. In total, we estimated the reliabilities of four scales: PHQ-9, GAD-7, PA, and NA. The first two were filled in weekly up to 12 times; the last two were filled in multiple times a day, up to 340 times. To estimate the reliability, we used six different methods with varying complexity, including three simpler approaches (i.e., GT, MLM, and ML-CFA) that provided reliability estimates at the between- and the within-person levels and three idiographic approaches (i.e., PT-FA, 2RDM, and ME-TSO) that provided person-specific reliability estimates.

In the field of psychological dynamics, the persons' means have been typically conceptualized as the trait scores (see Hamaker & Grasman, 2015; Nezlek, 2007), meaning that they represent the stable component of the psychological construct that characterizes

the individual. From a practical point of view, high between-person reliability coefficients are important when comparing trait scores across individuals or examining long-term durable changes in the constructs. For example, burst designs (Reis et al., 2024; Schricker et al., 2023; Stawski et al., 2015), where participants are measured intensively pre- and postintervention, are highly useful to assess the effectiveness of a treatment. Verifying that the measures have adequate levels of between-person reliability is essential for establishing the validity of the findings in these studies.

In Stage 2 of the WARN-D study, the between-person reliability estimates were generally high for all weekly and EMA scales. In short, these high coefficients indicate that the persons' means were accurately measured. As a consequence, the persons' means of the different scales can be safely used to relate them with other variables and to make statistical inferences in future research with the WARN-D data. Moreover, if the scores of the scales are shown to be equally reliable in Stage 4, which is a repetition of Stage 2 two years later, differences between persons' means between these stages can be attributed to "real" changes.

The within-person reliability coefficients indicate the consistency of individuals' responses at a given occasion (Nezlek, 2017), as well as the precision of capturing the true intraindividual variability (Cranford et al., 2006; Neubauer & Schmiedek, 2020). Having adequate levels of within-person reliability is particularly important to ascertain that the fluctuations at the intraindividual level are more than random error. In clinical settings, where the interest is, for example, to identify early warning signals and to develop time-sensitive preventive mental health care (Fried et al., 2023; Helmich et al., 2021; Olthof et al., 2020), having reliable measurements at the within-person level is essential.

In the empirical data used for this study, we observed more diverse results across the different scales regarding the within-person reliability estimates. For the PHQ-9 and GAD-7, the within-person reliabilities were more variable across methods and subsamples and indicated fair or moderate levels of reliability of the persons' scores. For PA, the point estimates were more homogeneous, indicating fair or moderate levels of within-person reliability. Last, for NA, the

within-person reliabilities tended to be moderate. This means that at any given time point, there is nonignorable measurement error in the scores of both the weekly and the daily scales. These results should be considered when, for example, studying the psychological dynamics of these constructs in the WARN-D data because some bias might be present in the parameters of interest (e.g., auto- and cross-regressive effects) if measurement errors are not taken into account in the analysis (Oh et al., 2025). Moreover, there seemed to be differences by language in the within-person reliabilities of NA, with the reliabilities of the Dutch subsample being slightly lower than those of the English subsample. Even though these differences by language did not seem to be too large, such differences may imply that there was measurement non-invariance between subsamples. This suggests that dedicated analyses of measurement invariance¹ might be needed, and caution should be taken when attempting to pool results across both subsamples.

Person-specific reliabilities can be particularly useful in clinical EMA research concerning early warning signals (Fried et al., 2023; Helmich et al., 2021; Olthof et al., 2020) to identify which participants have more consistent responses throughout the study. It is likely that the identification of early warning signals is more accurate for participants with higher reliability. However, more research is needed to support this claim. Moreover, approaches such as the ME-TSO, which additionally provide detailed information per item, can be useful for the development and validation of EMA scales (Cloos et al., 2023). Overall, validated EMA scales for studying psychological dynamics are lacking, and applied clinical researchers interested in EMA methods would benefit from such developments.

When using the idiographic approaches to estimate the person-specific reliabilities of the EMA scores of the WARN-D data, the results were more diverse. While the reliabilities based on the PT-FA and 2RDM showed that there was considerable heterogeneity across participants, the results based on the ME-TSO were more homogeneous. This is not surprising considering how different the approaches are conceptually and statistically. To conclude, we can think of using the person-specific reliability estimates as quality checks for the data of each participant, for example, participants could be excluded from future analyses based on their estimated reliability, or the recommendations and feedback to the participants can be more nuanced depending on their reliability estimate. Nonetheless, this is uncharted territory that needs more methodological and empirical research.

Advantages and Limitations

Strengths of this study are the combination of a high-powered EMA data set that captures scales in different languages at different time frames and applying six different methods to estimate internal consistency reliability. There are also several limitations worth highlighting. One important limitation is that all the approaches considered were developed for continuous responses, which is a mismatch with the Likert scales analyzed in this study. This may be especially relevant in the analyses of the NA scale, which is rather skewed and suffers from floor effects (Siepe et al., 2024). Therefore, measurement models capable of studying the psychometric properties of Likert scale items in EMA research need to be developed; some seminal work on this topic has been suggested by

Castro-Alvarez, Bringmann, Meijer, and Tendeiro (2024), Hecht et al. (2019), and Vogelsmeier et al. (2024).

Another limitation is that using different approaches, it is unclear what to do when contradictory results are found. For example, with the results of the PHQ-9, the within-person reliability based on MLM was considerably lower than the within-person reliabilities based on GT and ML-CFA. Therefore, shall we conclude that the within-level reliability of the PHQ-9 was bad or fair? To answer such a question, further research is needed to compare these different approaches in estimating the reliability of EMA data. In particular, it is important to investigate how different data characteristics (such as sample size, time series length, and sample heterogeneity) influence the reliability measures of these approaches. There might be unknown relations between data characteristics and these approaches that might explain the observed differences in the reliability measures. However, methodological research on this topic is lacking. To get a comprehensive understanding regarding under which circumstances each approach might be preferred, simulation studies are needed to systematically compare these approaches.

We also observed some convergence issues for some methods when analyzing the EMA scales. To solve these issues, researchers can try to simplify or constrain the models (e.g., fix all the factor loadings at 1, fix residual variances in the between level at 0, and remove some random effects). We tried some of these solutions with our data, but the final models, which converged to a proper solution, had an extremely poor fit. Therefore, we did not find it worth reporting. These issues of nonconvergence and improper solutions probably indicate some problems with the scales. For example, with PA, the ML-CFA did not reach a satisfactory solution; by considering the assumptions of this model, we can hypothesize that the factor structure might be misspecified or that there is measurement noninvariance across participants (see Jak et al., 2013, 2014). Moreover, the generally low within-person reliabilities for PA might indicate that not enough items were included in the scale. For NA, the 2RDM model did not converge for the English subsample, and the median person-specific reliabilities were quite diverse across items based on the ME-TSO. These issues might be caused by model misspecification. For example, we suspect that the item “tired” might be the reason for these problems as its within-person reliability has the lowest median values for both subsamples (see Figure 2), thus fitting the target construct NA somewhat less than some of the other items. This calls for careful examination of scale items in relation to the purported target constructs they measure.

Last, in terms of *constraints on generality*, the six approaches showcased here for estimating the internal consistency require stationary data wherein the mean, variance, and covariance structure of a process remain relatively stable over time (Chatfield, 2003). For data wherein long-term systematic change trajectories such as gradual improvements during treatment for mental health problems or developmental changes can be expected, the methods may not be well suited. One way forward is to segment the data into relatively

¹ Measurement invariance is an assumption required to make meaningful comparisons of the scores of a test from different groups, time points, or persons (Meredith, 1993; Meredith & Teresi, 2006). If measurement invariance does not hold, the scores from different groups are not directly comparable. Procedures to test for measurement invariance in nested data have been suggested by Jak et al. (2013, 2014) and to account for changes of the measurement models over time have been suggested by Vogelsmeier et al. (2019).

stable periods and estimate the reliability within each time window. Perhaps a better way is to develop dedicated new models, or to extend existing ones, for the estimation of the reliability of non-stationary time series data.

Final Remarks

To conclude, in this study, we exemplified how to estimate the reliability of the scales used in EMA research. As shown by analyzing the weekly and the EMA scales of the WARN-D study, the methods that can be used also depend on the amount of data available per participant. While simpler methods offer an easy way to estimate the between- and within-person reliability of any kind of longitudinal nested data, the more complex models allow zooming in on each participant's dynamic processes, thus fully endorsing the idiographic perspective. Using a variety of approaches allows for a more comprehensive assessment of the performance of the scales, which could motivate important improvements to the measurement of psychological dynamics.

The analyses presented in this study, alongside the documented code provided in the [Supplemental Material](#), represent a key contribution to the field of measurement in clinical EMA research as they can serve as a template for the study of the internal consistency reliability of EMA scales. We encourage empirical researchers in the field to analyze and report the reliability of their scales. This can be extremely beneficial in the near future to start developing cumulative science about the quality of the measurements in EMA research. Furthermore, considering that affect, depression, and anxiety symptoms are popular constructs typically assessed in clinical EMA research, the collaborative effort of reporting the reliability of these scales can lead to meta-analytical research on the measurement properties of such scales.

References

- Asparouhov, T., Hamaker, E. L., & Muthén, B. (2018). Dynamic structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, 25(3), 359–388. <https://doi.org/10.1080/10705511.2017.1406803>
- Brose, A., Schmiedek, F., Gerstorff, D., & Voelkle, M. C. (2020). The measurement of within-person affect variation. *Emotion*, 20(4), 677–699. <https://doi.org/10.1037/emo0000583>
- Castro-Alvarez, S., Bringmann, L. F., Meijer, R. R., & Tendeiro, J. N. (2024). A time-varying dynamic partial credit model to analyze polytomous and multivariate time series data. *Multivariate Behavioral Research*, 59(1), 78–97. <https://doi.org/10.1080/00273171.2023.2214787>
- Castro-Alvarez, S., Bringmann, L., Back, J., & Liu, S. (2024). *The many reliabilities of psychological dynamics: An overview of statistical approaches to estimate the internal consistency reliability of intensive longitudinal data*. PsyArXiv. <https://doi.org/10.31234/osf.io/qyk2r>
- Castro-Alvarez, S., Tendeiro, J., de Jonge, P., Meijer, R. R., & Bringmann, L. (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. <https://doi.org/10.31234/osf.io/4ext3>
- Castro-Alvarez, S., Zhou, D. J., Bringmann, L. F., Tutunji, R., Proppert, R. K. K., Rieble, C. L., Fried, E. I., & Liu, S. (2024a, August). *07_Assessing the internal consistency reliability of ecological momentary assessment measures: Insights from the WARN-D study*. Retrieved June 10, 2025, from <https://osf.io/wupyv/>
- Castro-Alvarez, S., Zhou, D. J., Bringmann, L. F., Tutunji, R., Proppert, R. K. K., Rieble, C. L., Fried, E. I., & Liu, S. (2024b, August). *Assessing the internal consistency reliability of ecological momentary assessment measures: Insights from the WARN-D study*. <https://doi.org/10.17605/OSF.IO/QEY9C>
- Chatfield, C. (2003). *The analysis of time series: An introduction* (6th ed.). Chapman & Hall/CRC Press. <https://doi.org/10.4324/9780203491683>
- Cloos, L., Ceulemans, E., & Kuppens, P. (2023). Development, validation, and comparison of self-report measures for positive and negative affect in intensive longitudinal research. *Psychological Assessment*, 35(3), 189–204. <https://doi.org/10.1037/pas0001200>
- 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. <https://doi.org/10.1177/0146167206287721>
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart and Winston.
- Eid, M., Holtmann, J., Santangelo, P., & Ebner-Priemer, U. (2017). On the definition of latent-state-trait models with autoregressive effects: Insights from LST-R theory. *European Journal of Psychological Assessment*, 33(4), 285–295. <https://doi.org/10.1027/1015-5759/a000435>
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum.
- Fried, E. I., Proppert, R. K. K., & Rieble, C. L. (2023). Building an early warning system for depression: Rationale, objectives, and methods of the WARN-D study. *Clinical Psychology in Europe*, 5(3), 1–25. <https://doi.org/10.32872/cpe.10075>
- Fried, E. I., Proppert, R. K. K., Rieble, C. L., Platania, N., & Tutunji, R. (2025a, January). *01_warn-D protocol paper*. <https://doi.org/10.17605/OSF.IO/2JD9H>
- Fried, E. I., Proppert, R. K. K., Rieble, C. L., Platania, N., & Tutunji, R. (2025b, March). *WARN-D project hub*. <https://doi.org/10.17605/OSF.IO/FRQDV>
- Fritz, J., Piccirillo, M. L., Cohen, Z. D., Frumkin, M., Kirtley, O., Moeller, J., Neubauer, A. B., Norris, L. A., Schuurman, N. K., Snippe, E., & Bringmann, L. F. (2024). So you want to do ESM? 10 essential topics for implementing the experience-sampling method. *Advances in Methods and Practices in Psychological Science*, 7(3), Article 7912. <https://doi.org/10.1177/25152459241267912>
- 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. <https://doi.org/10.1037/a0032138>
- Hamaker, E. L., & Grasman, R. P. P. P. (2015). To center or not to center? Investigating inertia with a multilevel autoregressive model. *Frontiers in Psychology*, 5, Article 1492. <https://doi.org/10.3389/fpsyg.2014.01492>
- 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. <https://doi.org/10.1177/0963721416666518>
- Hecht, M., Hardt, K., Driver, C. C., & Voelkle, M. C. (2019). Bayesian continuous-time Rasch models. *Psychological Methods*, 24(4), 516–537. <https://doi.org/10.1037/met0000205>
- Helmich, M. A., Smit, A. C., Bringmann, L. F., Schreuder, M. J., Oldehinkel, A. J., Wichers, M., & Snippe, E. (2021). *Detecting impending symptom transitions using early warning signals in individuals receiving treatment for depression*. <https://doi.org/10.31234/osf.io/vf86s>
- Hu, Y., Nesselroade, J. R., Erbacher, M. K., Boker, S. M., Burt, S. A., Keel, P. K., Neale, M. C., Sisk, C. L., & Klump, K. (2016). Test reliability at the individual level. *Structural Equation Modeling: A Multidisciplinary Journal*, 23(4), 532–543. <https://doi.org/10.1080/10705511.2016.1148605>
- Huang, D., Susser, E., Rudolph, K. E., & Keyes, K. M. (2023). Depression networks: A systematic review of the network paradigm causal assumptions. *Psychological Medicine*, 53(5), 1665–1680. <https://doi.org/10.1017/S0033291723000132>
- Jak, S., Oort, F. J., & Dolan, C. V. (2013). A test for cluster bias: Detecting violations of measurement invariance across clusters in multilevel data.

- Structural Equation Modeling: A Multidisciplinary Journal*, 20(2), 265–282. <https://doi.org/10.1080/10705511.2013.769392>
- Jak, S., Oort, F. J., & Dolan, C. V. (2014). Measurement bias in multilevel data. *Structural Equation Modeling: A Multidisciplinary Journal*, 21(1), 31–39. <https://doi.org/10.1080/10705511.2014.856694>
- Kroenke, K., & Spitzer, R. L. (2002). The PHQ-9: A new depression diagnostic and severity measure. *Psychiatric Annals*, 32(9), 509–515. <https://doi.org/10.3928/0048-5713-20020901-06>
- Lai, M. H. C. (2021). Composite reliability of multilevel data: It's about observed scores and construct meanings. *Psychological Methods*, 26(1), 90–102. <https://doi.org/10.1037/met0000287>
- McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*, 23(3), 412–433. <https://doi.org/10.1037/me0000144>
- Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58(4), 525–543. <https://doi.org/10.1007/BF02294825>
- Meredith, W., & Teresi, J. A. (2006). An essay on measurement and factorial invariance. *Medical Care*, 44(11), S69–S77. <https://doi.org/10.1097/01.mlr.0000245438.73837.89>
- Molenaar, P. C. M. (1985). A dynamic factor model for the analysis of multivariate time series. *Psychometrika*, 50(2), 181–202. <https://doi.org/10.1007/BF02294246>
- Neubauer, A. B., & Schmiedek, F. (2020). Studying within-person variation and within-person couplings in intensive longitudinal data: Lessons learned and to be learned. *Gerontology*, 66(4), 332–339. <https://doi.org/10.1159/000507993>
- Nezlek, J. B. (2007). A multilevel framework for understanding relationships among traits, states, situations and behaviours. *European Journal of Personality*, 21(6), 789–810. <https://doi.org/10.1002/per.640>
- Nezlek, J. B. (2017). A practical guide to understanding reliability in studies of within-person variability. *Journal of Research in Personality*, 69, 149–155. <https://doi.org/10.1016/j.jrp.2016.06.020>
- Nezlek, J. B., & Gable, S. L. (2001). Depression as a moderator of relationships between positive daily events and day-to-day psychological adjustment. *Personality and Social Psychology Bulletin*, 27(12), 1692–1704. <https://doi.org/10.1177/01461672012712012>
- Oh, H., Hunter, M. D., & Chow, S.-M. (2025). Measurement model misspecification in dynamic structural equation models: Power, reliability, and other considerations. *Structural Equation Modeling: A Multidisciplinary Journal*, 32(3), 511–528. <https://doi.org/10.1080/10705511.2025.2452884>
- Olthof, M., Hasselman, F., Strunk, G., Van Rooij, M., Aas, B., Helmich, M. A., Schiepek, G., & Lichtwarck-Aschoff, A. (2020). Critical fluctuations as an early-warning signal for sudden gains and losses in patients receiving psychotherapy for mood disorders. *Clinical Psychological Science*, 8(1), 25–35. <https://doi.org/10.1177/2167702619865969>
- Reis, D., Hart, A., Krautter, K., Prestele, E., Lehr, D., & Friese, M. (2024). Mindfulness and cognitive-behavioral strategies for psychological detachment: Comparing effectiveness and mechanisms of change. *Journal of Occupational Health Psychology*, 29(4), 258–279. <https://doi.org/10.1037/ocp0000381>
- Riese, H., von Klipstein, L., Schoevers, R. A., van der Veen, D. C., & Servaas, M. N. (2021). Personalized ESM monitoring and feedback to support psychological treatment for depression: A pragmatic randomized controlled trial (Therap-i). *BMC Psychiatry*, 21(1), Article 143. <https://doi.org/10.1186/s12888-021-03123-3>
- Schricker, I. F., Nayman, S., Reinhard, I., & Kuehner, C. (2023). Reactivity toward daily events: Intraindividual variability and change in recurrent depression—A measurement burst study. *Behaviour Research and Therapy*, 168, Article 104383. <https://doi.org/10.1016/j.brat.2023.104383>
- Schuurman, N. K., & Hamaker, E. L. (2019). Measurement error and person-specific reliability in multilevel autoregressive modeling. *Psychological Methods*, 24(1), 70–91. <https://doi.org/10.1037/met0000188>
- Shrout, P. E. (1998). Measurement reliability and agreement in psychiatry. *Statistical Methods in Medical Research*, 7(3), 301–317. <https://doi.org/10.1177/096228029800700306>
- 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). Guilford Press.
- Siepe, B. S., Rieble, C. L., Tutunji, R., Rimpler, A., März, J., Proppert, R. K. K., & Fried, E. I. (2024, January). *Understanding EMA data: A tutorial on exploring item performance in ecological momentary assessment data*. <https://doi.org/10.31234/osf.io/dvj8g>
- Spitzer, R. L., Kroenke, K., Williams, J. B. W., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: The GAD-7. *Archives of Internal Medicine*, 166(10), 1092–1097. <https://doi.org/10.1001/archinte.166.10.1092>
- Stawski, R. S., MacDonald, S. W. S., & Sliwinski, M. J. (2015). Measurement burst design. In S. K. Whitbourne (Ed.), *The encyclopedia of adulthood and aging* (pp. 1–5). Wiley. <https://doi.org/10.1002/9781118521373.wbeaa313>
- Steyer, R., Mayer, A., Geiser, C., & Cole, D. A. (2015). A theory of states and traits—Revised. *Annual Review of Clinical Psychology*, 11(1), 71–98. <https://doi.org/10.1146/annurev-clinpsy-032813-153719>
- Tutunji, R., Proppert, R. K. K., Rieble, C. L., & Fried, E. I. (2024, April). *Defining a generic holdout sample for combined exploratory and predictive analyses in the WARN-D dataset*. Retrieved October 22, 2024, from <https://osf.io/3afkj>
- Vogelsmeier, L. V. D. E., Jongerling, J., & Ulitzsch, E. (2024, September). *Accounting for measurement invariance violations in careless responding detection in intensive longitudinal data: Exploratory vs. partially constrained latent Markov factor analysis*. <https://doi.org/10.31234/osf.io/rab7u>
- Vogelsmeier, L. V. D. E., Vermunt, J. K., van Roekel, E., & De Roover, K. (2019). Latent Markov factor analysis for exploring measurement model changes in time-intensive longitudinal studies. *Structural Equation Modeling*, 26(4), 557–575. <https://doi.org/10.1080/10705511.2018.1554445>
- 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>

Received November 4, 2024

Revision received June 10, 2025

Accepted June 11, 2025 ■