Title

Optimizing Detection of True Within-Person Effects for Intensive Measurement Designs: A Comparison of Multilevel SEM and Unit-Weighted Scale Scores

Journal

Behavior Research Methods, 52(5)

ISSN

1554-351X

Authors

Rush, Jonathan
Rast, Philippe
Hofer, Scott M

Publication Date

2020-10-01

DOI

10.3758/s13428-020-01369-5

Peer reviewed

HHS Public Access

Author manuscript

Behav Res Methods. Author manuscript; available in PMC 2021 October 01.

Published in final edited form as:

Behav Res Methods. 2020 October ; 52(5): 1883–1892. doi:10.3758/s13428-020-01369-5.

Optimizing Detection of True Within-Person Effects for Intensive Measurement Designs: A Comparison of Multilevel SEM and Unit-Weighted Scale Scores

Jonathan Rush[1], **Philippe Rast**[2], **Scott M. Hofer**[1]

[1]University of Victoria, Davis

[2]University of California, Davis

Abstract

Intensive repeated measurement designs are frequently used to investigate within-person variation over relatively brief intervals of time. The majority of research utilizing these designs relies on unit-weighted scale scores, which assume that the constructs are measured without error. An alternative approach makes use of multilevel structural equation models (MSEM), which permit the specification of latent variables at both within-person and between-person levels. These models disaggregate measurement error from systematic variance, which should result in less biased within-person estimates and larger effect sizes. Differences in power, precision, and bias between multilevel unit-weighted and MSEM models were compared through a series of Monte Carlo simulations. Results based on simulated data revealed that precision was consistently poorer in the MSEM models than the unit-weighted models, particularly when reliability was low. However, the degree of bias was considerably greater in the unit-weighted model than the latent variable model. Although the unit-weighted model consistently underestimated the effect of a covariate, it generally had similar power relative to the MSEM model due to the greater precision. Considerations for scale development and the impact of within-person reliability are highlighted.

Keywords

Multilevel modeling; within-person effects; power; multilevel structural equation modeling; composite scores

Intensive repeated measurement designs (e.g., daily diary, ecological momentary assessment) are frequently used in psychological research to investigate within-person

Correspondence concerning this article should be addressed to Jonathan Rush, Department of Psychology, University of Victoria, P.O. Box 1700, STN CSC, Victoria, BC, Canada, V8W 2Y2. jrush@uvic.ca.
Jonathan Rush and Scott M. Hofer, Department of Psychology, University of Victoria, Canada. Philippe Rast, Department of Psychology, University of California, Davis.

Open Practices Statement

The data and materials for the experiments reported here are available upon request. None of the experiments were preregistered.

variation over relatively brief intervals of time (e.g., hours, days, or weeks). These designs allow variance to be partitioned into within-person and between-person sources of variability, enabling differential effects to be estimated at the within-person and between-person level of analysis (e.g., Curran & Bauer, 2011; Hoffman & Stawski, 2009; Sliwinski, 2008). Much research has examined within-person covariation of time-varying constructs to identify how variables travel dynamically together across time. These covariations have been examined in a variety of domains to identify reliable short-term within-person associations. For example, Hoppman and Klumb (2006) examined daily variations in personal goals as within-person predictors of daily mood and cortisol levels. Webster and Hadwin (2015) found that within-person fluctuations in positive emotions covary with goal attainment during study sessions. Rush and Grouzet (2012) examined how fluctuations in daily temporal focus accounted for daily levels of psychological well-being.

Research examining within-person associations often investigate constructs (e.g., affect, stress, rumination, etc.) that are measured through self-report scales assessed repeatedly over many occasions. These measurement scales consist of multiple items that are assumed to reflect a single construct that varies within an individual across measurement occasions. When developing measures for use in within-person intensive measurement research, a primary motivation is to limit the participant burden that results from repeatedly responding to the same questions day after day. It has become common practice to use short-form scales that have been adapted from existing cross-sectional measures. Often these shortened scales consist of just three or four items, and sometimes fewer (e.g., Kashdan et al., 2013; Morelli et al., 2015; Reynolds et al., 2016). A concern is that these measures have been designed and evaluated for identification of between-person differences rather than within-person fluctuations. As a result, many of the within-person measures used lack a proper investigation into the within-person psychometric properties and factor structure, and the reporting of within-person reliabilities are often omitted.

Furthermore, the common analytic approach of research utilizing intensive measurement designs to examine within-person associations rely on unit-weighted (UW) scale scores (i.e., composite scores), where all of the items reflecting a construct are summed (or averaged) to create a total (or mean) score. This approach weights each item of the scale equally and assumes that the constructs are measured without error. The composite score is then included as a time-varying predictor (X_{ij}) in a multilevel modeling (MLM) analysis to account for occasion-specific covariation with an outcome (Y_{ij}). The following equations display a common MLM approach for estimating a within-person effect (γ_{10}):

$$\text{Level 1:} \quad Y_{ij} = \beta_{0i} + \beta_{1i}\left(X_{cij} - X_{c\cdot i}\right) + e_{ij} \tag{1a}$$

$$\text{Level 2:} \quad \beta_{0i} = \gamma_{00} + \gamma_{01}\left(X_{c\cdot i}\right) + u_{0i} \tag{1b}$$

$$\beta_{1i} = \gamma_{10} + u_{1i}, \tag{1c}$$

where Y_{ij} is the outcome variable for person i on occasion j. β_{0i}, and β_{1i}, refer to the intercept and within-person association for person i, respectively; X_{cij} is the composite score

for person i on occasion j; $X_{c\cdot i}$ is the person-mean of X_{cij} for person i; and e_{ij} represents the within-person residual variance. At Level 2, γ_{00} represents the average intercept; γ_{10} represents the average within-person effect of X on Y; γ_{01} represents the between-person association between X and Y; and u_{0i} and u_{1i} represent individual deviations from average intercepts and slopes (i.e., random effects).

The UW composite score consists of both systematic within-person variability (i.e., true score variance) and unsystematic within-person variability (i.e., measurement error). When a composite score is computed, it cannot be determined how much variability is due to measurement error and how much is due to systematic occasion-to-occasion within-person fluctuations. The true systematic fluctuations are of substantive interest to understand the contextual circumstances when individuals are deviating from their typical levels. However, the unsystematic within-person variations that are due to measurement unreliability adds noise to statistical models attempting to capture true within-person associations.

Variance over time in the same measure has long been an indicator of measurement unreliability. Indices of test-retest reliability treats all within-person variations as scale measurement error under assumptions of stable true scores and no learning effects. Extensive research now clearly demonstrates that many constructs can systematically vary within an individual over either short or longer periods of time (e.g., Bolger, Davis, & Rafaeli, 2003; Rush, Rast, Almeida, & Hofer, 2019; Sliwinski, 2008). However, from a measurement perspective, it is often unclear how much of the within-person variability over time is due to true systematic variance in the construct and how much is due to measurement error. Short-term within-person variability may be misinterpreted as true systematic variance, when indeed it is merely the result of scale unreliability. Therefore, it is important when examining within-person effects to consider how much scale unreliability is influencing our ability to detect true within-person associations based on systematic covariation. Failing to account for such error has the potential to downwardly bias our estimates and may decrease the sensitivity to identify true within-person effects.

An alternative analytic approach, multilevel structural equation modeling (MSEM), permits the specification of latent variables at both within-person and between-person levels in order to disentangle measurement error from systematic variance. Multilevel SEM combines a measurement model and structural model across levels of analysis. This allows for the within-person variance to be disaggregated from the between-person variance, while still adjusting for measurement error at both levels. The multilevel measurement model can be expressed by the following equation (Muthén, 1991; Preacher, Zyphur, & Zhang, 2010):

$$X_{ij} = \upsilon + \lambda_w \eta_{ij} + \varepsilon_{ij} + \lambda_b \eta_i + \varepsilon_i, \quad (2a)$$

where X_{ij} is a p-dimensional vector of observed variables (i.e., scale items) for individual i on occasion j, where p is the number of observed indicators; υ is a p-dimensional vector of intercepts; λ_w is a $p \times q$ within-person factor loadings matrix, where q is the number of latent variables; λ_b is a $p \times q$ between-person factor loadings matrix; η_{ij} and η_i are q-dimensional vectors of within-person and between-person latent variables,[1] respectively; and

ε_{ij} and ε_i are p-dimensional vectors of within-person and between-person uniqueness factors (i.e., residuals), respectively.

At the between-person level, the indicators are person-means of each within-person indicator that are aggregated in order to adjust for unreliability due to sampling error (see Lüdtke et al., 2008; Marsh et al., 2009 for further details), such that the between-person indicators are represented as latent means. Both the between-person and within-person parts of the model are estimated simultaneously with the within-person factor structure representing common covariance in the indicators at each specific occasion across time and the between-person factor structure representing common covariance in the person-mean indicators across people.

The structural model permits latent variables from the measurement model to be specified as exogenous or endogenous variables within and across levels of analysis. A reduced form of the within-person (Level 1) and between-person (Level 2) structural models can be expressed by the following equations (Muthén & Asparouhov, 2009; Preacher et al., 2010):

$$\text{Level 1:} \quad \eta_{ij} = \alpha_i + \beta_i \eta_{ij} + \zeta_{ij} \tag{2b}$$

$$\text{Level 2:} \quad \eta_i = \mu + \gamma \eta_i + \zeta_i, \tag{2c}$$

where α_i, is q-dimensional vector of intercepts, β_i is q × q matrix of regression coefficients for individual i; ζ_{ij} represents level 1 residuals; μ is a q-dimensional vector of level 2 coefficient means; γ represents a q × q matrix of level 2 regression slopes; and ζ_i is a vector of level 2 residuals. Within this framework, multiple latent variables can be specified as exogenous or endogenous to one another. Specifically, latent variables at the within-person level can be included as a time-varying predictor of endogenous outcomes.

An important distinction between the multilevel UW and MSEM approach is how they deal with unsystematic within-person variations. With the specification of latent variables at the within-person level, the MSEM approach removes the occasion-to-occasion variability in the scale items that are not common across items. Common within-person covariance reflects occasion-specific variations that are common across the scale indicators. That is, on occasions when an individual deviates from their average on one scale item, it reflects the extent that they also deviate from their average in the other scale items. The remaining occasion-specific variability that is not common across the scale items is estimated as item uniqueness factors (i.e., unsystematic measurement error). Therefore, the within-person latent variable contains only occasion-to-occasion variability in the construct and not random error variance. As a result, using the true score variance to account for within-person deviations in an outcome variable should result in larger effect sizes due to a reduction in the noise and an increase in the signal compared to the UW approach that combines true score variance with measurement error. The difference between these two multilevel modeling approaches should be more pronounced when the within-person scale reliability is poor and

[1]The number of latent variables was specified to be the same across levels of analysis; however, it should be noted that the factor structure can be specified to differ across levels (i.e., at the within- and between-person levels).

possesses a greater proportion of measurement error. Of note, it has been shown in cross-sectional research that latent models often result in poorer precision (i.e., increased standard errors) of estimated mediation effects than observed score models (Ledgerwood & Shrout, 2011). The degradation of precision in latent models can result in less power than observed score models, despite the larger effect sizes. However, to date there has yet to be an investigation that compares the ability of these two modeling approaches to capture within-person effects within a multilevel modeling framework. Therefore, it is unclear the extent that relying on a multilevel UW approach impacts the estimation of true within-person associations within intensive repeated measurement research. Given that using UW composite scores is the predominant approach to examining within-person effects, it is critical to gain an understanding of how unsystematic within-person variance (i.e., measurement error) is affecting within-person estimates throughout the literature, particularly under conditions when the within-person measures may not possess adequate measurement properties (i.e., within-person reliability).

Present Study

The goal of the present study was to compare the multilevel UW composite score approach typically used in MLM with the latent variable modeling approach of MSEM. We compared these two different modeling approaches in their ability to effectively capture true within-person effects. Differences in power, precision, and bias between multilevel UW models and MSEMs were examined through a series of Monte Carlo simulations carried out in Mplus version 8.

We expected that the within-person latent variable should better capture true systematic within-person variation and would result in less biased within-person estimates of a time-varying covariate (i.e., within-person association) and larger estimated effects. It was further hypothesized that the larger estimated effects from the MSEM would result in greater power to detect a within-person effect than the UW multilevel model. However, as has been found previously, the use of latent variables often results in poorer measurement precision than manifest variables (Ledgerwood & Shrout, 2011), which reduces power. Given the two counteracting influences of larger effect sizes and poorer precision, it was plausible that the MSEM would not result in greater power than the UW models.

Method

Simulation Data

Data were generated to examine the two modeling approaches under varying conditions. Monte Carlo simulations with 5000 replications were carried out using Mplus v8 software (Muthén & Muthén, 2012). Three factors were manipulated, including the statistical model type, number of measurement occasions, and scale reliability of the predictor variable, resulting in a 2 (*Model*: Unit-Weighted vs. MSEM) × 2 (*Reliability*: High [0.9] vs. Low [0.6]) × 3 (*Measurement Occasions*: *5* vs. *7* vs. *10*) simulation design. Each condition was examined at varying sample sizes. The within-person effect of a time-varying predictor was the focus of these simulations; thus the between-person population parameters were held constant across conditions to isolate the within-person effects. The within-person effect size

of the predictor was also held constant at 0.2 across conditions. The predictor variable consisted of a four-item scale. Reliability of this scale was computed as the ratio of true-score variance to total variance (ω; see McDonald, 1999; Geldhof, Preacher, & Zyphur, 2014) and was derived by varying the amount of error in population parameters of each of the scale indicators. Population parameters for the high and low reliability conditions are presented in Table 1.[2] Number of clusters (i.e., sample size) varied from $N = 25$ to 200. Population parameter values were grounded on values found in actual datasets that examined within-person effects in self-reported scales across varying measurement occasions (i.e., Rush & Grouzet, 2012; Rush & Hofer, 2014; Rush et al., 2019). These values were selected to be within a plausible range of values commonly found in daily diary well-being research examining within-person effects. The simulation design included conditions that are less favorable (e.g., low scale reliability; five measurement occasions) to identify whether the two modeling approaches diverged under less desirable research designs.

Data Analytic Strategy

Unit-weighted multilevel models and MSEMs were both fit to the simulated data. Figure 1 displays the specification of the two models. Power, precision, and bias in detecting the effect of a within-person covariate (γ_{10}) was examined across conditions. Power was assessed as the proportion of replications (out of 5000) that yielded a statistically significant within-person effect, based on $\alpha = .05$. Precision was assessed as the variability in the within-person estimate over simulated samples (i.e., population standard error), where smaller standard errors indicate better precision. The accuracy in estimating the within-person effect was assessed with proportion of bias, which was computed as the difference of the mean estimated within-person effect across 5000 replications from the true population value, all divided by the population value (Muthén, 2002). Bias in standard error estimates was also examined by computing the difference between the population standard error and the average estimated standard error. Finally, type I error rates were assessed by setting the population within-person effect to zero and examining the proportion of statistically significant outcomes across simulation conditions.

The unit-weighted MLM was specified within an MSEM framework with constraints rather than the traditional composite score predictor included in an MLM as a time-varying covariate. In order to produce the equivalent model to the typical MLM (Equation 1), the factor loading of each item was fixed to 1 and the item uniquenesses (i.e., specific error factors) were fixed to 0 (see Fig. 1a). This specification asserts that the items are equally weighted in their contributions to the common factor ($\eta_{w\,ij}$) and that the construct is measured without error. This is the equivalent model to the traditional MLM investigating the within-person effect of a time-varying covariate that was measured as a composite variable (Equation 1). By specifying the UW MLM in this manner, it permitted a direct comparison with the freely estimated latent variable MSEM approach.

The MSEM approach specified the time-varying predictor as a latent variable at both the within-person and between-person levels, which adjusted for measurement error at both

[2]Additional population parameters with homogeneous factor loadings were also considered; however, these results did not differ much from the heterogeneous factor loading condition, so are not presented here.

levels (see Fig. 1b). Both the within-person and between-person parts of the model were estimated simultaneously with the within-person latent variable (i.e., η_{wij}) representing common covariance in the indicators at each specific occasion across time (i.e., $X1ij$ to X_{4ij}) and the between-person latent variable (i.e., η_{bij}) representing common covariance in the person-mean indicators across people (i.e., X_{1j} to X_{4j}). Item-specific measurement error variance (i.e., variance not shared with the common factor) was freely estimated at the within-person (i.e., ε_{w1ij} to ε_{w4ij}) and between-person levels (i.e., ε_{b1i} to ε_{b4i}). Therefore, the within-person latent variable reflected occasion-specific deviations from person-mean levels (i.e., within-person variation) that were common across the four scale items.

Results and Discussion

A number of findings emerged from the simulation studies. Independent of modeling approach, there were consistent main effects of sample size and number of occasions on power and precision of the estimated within-person effect. Figures 2 and 3 clearly demonstrate that larger sample sizes and more measurement occasions resulted in greater power to detect the within-person effect and more precise estimates (i.e., smaller SEs) for both modeling approaches. Conversely, accuracy of the estimated effect did not vary based on sample size nor measurement occasions (see Fig. 4). The high-reliability condition ($\omega = .90$) consistently resulted in higher power than the low-reliability condition ($\omega = .60$) across conditions. The influence of sample size and number of occasions did not vary based on scale reliability, as patterns were consistent for low or high reliability.

A comparison of the MSEM and multilevel UW modeling approaches revealed that both performed comparably in power to detect within-person effects across conditions (see Fig. 2). However, the standard errors were consistently higher in the MSEMs than the UW models, particularly when reliability was low (see Fig. 3). Additionally, the UW models were much less accurate in detecting the within-person effect compared to the MSEM approach (see Fig. 4). The UW models consistently underestimated the true effect across conditions, whereas the MSEM approach estimated the true within-person effect with minimal bias. Differences in the degree of bias between the two modeling approaches were exacerbated when the scale reliability was low. Even with low scale reliability, the MSEM approach produced minimal bias (< 1%). However, the degree of bias in the UW model increased from around 10% when scale reliability was high to over 40% when scale reliability was low. Finally, both models produced acceptable degrees of bias in standard errors and rates of type I errors across conditions. Though the UW model had slightly higher type I error rates than the MSEM model when reliability was low and sample size was small, the differences were minimal, and overall type I error rates deviated only slightly from the expected 0.05 (see Fig. 5). Similarly, bias in standard error estimates was slightly higher in the UW models where reliability was low and sample size small; however, the standard error bias was consistently less than 5% in both models across conditions (see Fig. 6).

To further inspect the impact of scale reliability on power, precision, and accuracy, additional simulations were conducted that held sample size and number of occasions constant at 75 and 7, respectively,[3] but varied the scale reliability semi-continuously from 0 to 1. As scale reliability emerged as one of the most influential elements to consider when

comparing MSEM and UW models, these series of simulations permitted a thorough examination of the impact of poor reliability of within-person measures on the two modeling approaches. Figure 7 displays the results of varying the reliability. The UW approach consistently underestimated the within-person effect across all levels of scale reliability. Only when the scale reliability of the within-person predictor variable was very high ($\omega > .90$) was the bias of the within-person effect reduced to an acceptable level (bias < 10%; see Fig. 7b). Furthermore, the 95% coverage of the true population parameter was consistently low with the UW modeling approach (see Fig. 7d). Whereas, the MSEM approach acceptably recovered the true population parameter with minimal bias (< 1%), once a minimum reliability ($\omega > .50$) permitted the models to converge properly. Conversely, the precision of the within-person estimate was dramatically poorer for the MSEM model compared to the UW models in situations when reliability was less than .70 (see Fig. 7c). The competing elements of underestimated effects combined with more precise estimates in the UW models resulted in similar levels of power between the MSEM and UW approaches (see Fig. 7a). Ledgerwood and Shrout (2011) demonstrated similar trade-offs between using a latent variable model versus an observed score model (i.e., unit-weighted) in between-person mediation analyses. The latent model improved accuracy (i.e., produced less biased estimates), but also yielded poorer precision (i.e., higher standard errors) relative to the unit-weighted model. Although the latent variable model produced larger estimates, the reduction in precision typically resulted in lower power to detect the effect.

Due to the considerable proportion of bias at moderate to low reliability, within-person effects based on UW composite scores are likely underestimated throughout the literature. Given that the reliability of within-person measures are often not considered and rarely reported, it is difficult to gauge the extent of the issue. Many measures created to capture within-person dynamics emphasize face validity and participant burden as the primary considerations and neglect within-person reliability and measurement properties (e.g., factor structure). Computing a composite score from these items obfuscates the potential multidimensionality or poor reliability of the measure. As a result, the within-person effects reported throughout the literature, which tend to be small, may be a poor representation of the true magnitude of effects (e.g., 40% bias would reduce $r = .40$ to $r = .24$). The results of this research clearly highlight the importance of considering within-person reliability of time-varying predictors when examining within-person effects, particularly when using a UW composite score approach.

The MSEM approach outperformed the UW approach in capturing true within-person effects across conditions and provides a viable modeling framework to better represent the true magnitude of within-person effects. Despite the many advantages, an MSEM approach may not always be the most appropriate choice. Multilevel SEM estimates many more parameters, and fluctuations across samples appear to lead to less consistency from one sample to the next in the estimated within-person effect (as indicated by the larger SEs). Even though on average across the 5000 replications, the MSEM approach resulted in a

[3]These values were chosen for the sample size and measurement occasions because (a) they represent design characteristics that are commonly used in intensive measurement designs, and (b) this appeared to be a point where the two modeling approaches began to diverge.

considerably less biased within-person estimate, in any given sample the estimate may not be as accurate. Precision was particularly concerning when reliability was low and sample sizes were small ($N < 75$). Larger sample sizes are typically required to produce stable estimates within an SEM framework using latent variables, compared with models that rely solely on observed variables (Kline, 2011).

Therefore, it may be advantageous at times to continue to use a UW composite score approach, but only under certain conditions and with the specific limitations in mind. First, a UW composite score should only be used after the within-person factor structure and reliability of the scale have been established. In the same way that we devote much effort to establishing acceptable measurement properties for cross-sectional between-person scales, so too should such effort be devoted to establishing measures designed to capture systematic within-person fluctuations within intensive measurement studies. It is not sufficient to assume that measures designed and validated for between-person assessment will be suitable and maintain their psychometric properties when used for within-person assessment. Research devoted to establishing and replicating the multilevel factor structure and measurement properties of within-person measurement scales should be more normative (e.g., Rush & Hofer, 2014). After establishing that the measure represents a single construct at the within-person level and possesses adequate reliability (i.e., > .90) to capture within-person fluctuations, then it may be reasonable to treat these measures as a composite score. Under these circumstances, the UW approach could be employed to cases with smaller cluster-level sample size where model complexity and convergence may create issues for the MSEM approach. In these cases, however, it should be noted that the within-person estimates based on composite score predictors will likely be smaller than their true value.

Limitations and Future Directions

Despite the straightforward goals and strengths of the current study, there are a number of limitations that should be addressed with future research. First, the study was solely focused on differences in detecting true within-person effects. In order to isolate the within-person effect, all between-person population parameters were held constant across conditions. Future research could benefit from examining the impact of these conditions on between-person effects in hierarchically structured data by varying between-person population parameters (e.g., between-person scale reliability). Furthermore, cross-level effects could be examined to determine how unsystematic within-person variance impacts between-person estimates. Second, the size of the within-person effect was held constant. Though varying effect size may seem warranted, additional analyses examining large versus small effect sizes revealed that the pattern of results did not interact with effect size. Finally, the current study held the number of scale items constant at four. It could be useful to examine how fewer or more scale items impacts results. Similar to sample size and number of occasions, it is expected that there will be main effects of number of items, where adding more items will improve power and precision. However, it is less clear whether the influence of number of items would differ across modeling frameworks.

Conclusions

The magnitude of short-term within-person effects reported throughout psychological research tends to be quite small in general. This is likely exacerbated by the frequent use of a UW composite scores modeling approach in conjunction with scales that possess moderate to low reliability. It is important to examine scale reliability and to design measures that reflect true within-person variability in the construct of interest. In doing so, it may be reasonable to utilize a UW modeling approach, as bias may be minimal and precision improved, particularly in situations when sample sizes are small and estimating a latent measurement model in addition to a structural pathway leads to convergence issues. However, the MSEM approach is a more accurate modeling approach when estimating the effect of a within-person covariate and could provide a clearer picture of the true magnitude of within-person effect sizes. Furthermore, the reliability of the predictor variable is less of a concern in capturing the magnitude of the true effect compared to the UW modeling approach. Nevertheless, the reduced precision of these estimates does warrant some consideration when the goal is to reliably detect a true within-person effect.

Acknowledgments

Research reported in this manuscript was supported by the National Institute on Aging of the National Institutes of Health, Grant R01AG050720 and P01AG043362. Jonathan Rush was supported by a Joseph Armand Bombardier Doctoral Scholarship from the Social Sciences and Humanities Research Council of Canada.

References

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

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

Geldhof GJ, Preacher KJ, & Zyphur MJ (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. Psychological Methods, 19, 72–91. [PubMed: 23646988]

Hoffman L, & Stawski RS (2009). Persons as contexts: Evaluating between-person and within-person effects in longitudinal analysis. Research in Human Development, 6, 97–120.

Hoppman CA, & Klumb PL (2006). Daily goal pursuits predict cortisol secretion and mood states in employed parents with preschool children. Psychosomatic Medicine, 68, 887–894. [PubMed: 17132838]

Kashdan TB, Farmer AS, Adams LM, Ferssizidis P, McKnight PE, & Nezlek JB (2013). Distinguishing healthy adults from people with social anxiety disorder: Evidence for the value of experiential avoidance and positive emotions in everyday social interactions. Journal of Abnormal Psychology, 122, 645–655. [PubMed: 23815396]

Kline RB (2011). Principles and practice of structural equation modeling (3rd ed.). New York: Guilford.

Ledgerwood A, & Shrout PE (2011). The trade-off between accuracy and precision in latent variable models of mediation processes. Journal of Personality and Social Psychology, 101, 1174–1188 [PubMed: 21806305]

Lüdtke O, Marsh HW, Robitzsch A, Trautwein U, Asparouhov T, Muthén B (2008). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. Psychological Methods, 13, 203–229. [PubMed: 18778152]

Marsh HW, Lüdtke O, Robitzsch A, Trautwein U, Asparouhov T, Muthén B, & Nagengast B (2009). Doubly-latent models of school contextual effects: Integrating multilevel and structural equation

approaches to control measurement and sampling error. Multivariate Behavioral Research, 44, 764–802. [PubMed: 26801796]

McDonald RP (1999). Test theory: A unified treatment. Mahwah, NJ: Erlbaum.

Morelli SA, Lee IA, Arnn ME, & Zaki J (2015). Emotional and instrumental support provision interact to predict well-being. Emotion, 15, 484–493. [PubMed: 26098734]

Muthén BO (1991). Multilevel factor analysis of class and student achievement components. Journal of Educational Measurement, 28, 338–354.

Muthén BO, & Asparouhov T (2009). Growth mixture modeling: Analysis with non-Gaussian random effects In Fitzmaurice G, Davidian M, Verbeke G, Molenberghs G (Eds.), Longitudinal data analysis (pp. 143–165). Boca Raton, FL: Chapman & Hall/CRC.

Muthén LK, & Muthén BO (1998-2012). Mplus user's guide (7th ed.). Los Angeles, CA: Muthén & Muthén.

Preacher KJ, Zyphur MJ, & Zhang Z (2010). A general multilevel SEM framework for assessing multilevel mediation. Psychological Methods, 15, 209–233. [PubMed: 20822249]

Reynolds BM, Robles TF, & Repetti RL (2016). Measurement reactivity and fatigue effects in daily diary research with families. Developmental Psychology, 52, 442–456. [PubMed: 26689757]

Rush J, & Grouzet FME (2012). It is about time: Daily relationships between temporal perspective and well-being. Journal of Positive Psychology, 7, 427–442.

Rush J, & Hofer SM (2014). Differences in within- and between-person factor structure of positive and negative affect: Analysis of two intensive measurement studies using multilevel structural equation modeling. Psychological Assessment, 20, 462–473.

Rush J, Rast P, Almeida DM, & Hofer SM (2019). Modeling long-term changes in daily within-person associations: An application of multilevel SEM. Psychology and Aging, 34, 163–176. [PubMed: 30730161]

Sliwinski MJ (2008). Measurement-burst designs for social health research. Social and Personality Psychology Compass, 2, 245–261.

Webster EA, & Hadwin AF (2015). Emotions and emotion regulation in undergraduate studying: Examining students' reports from a self-regulated learning perspective. Educational Psychology, 35, 794–818.

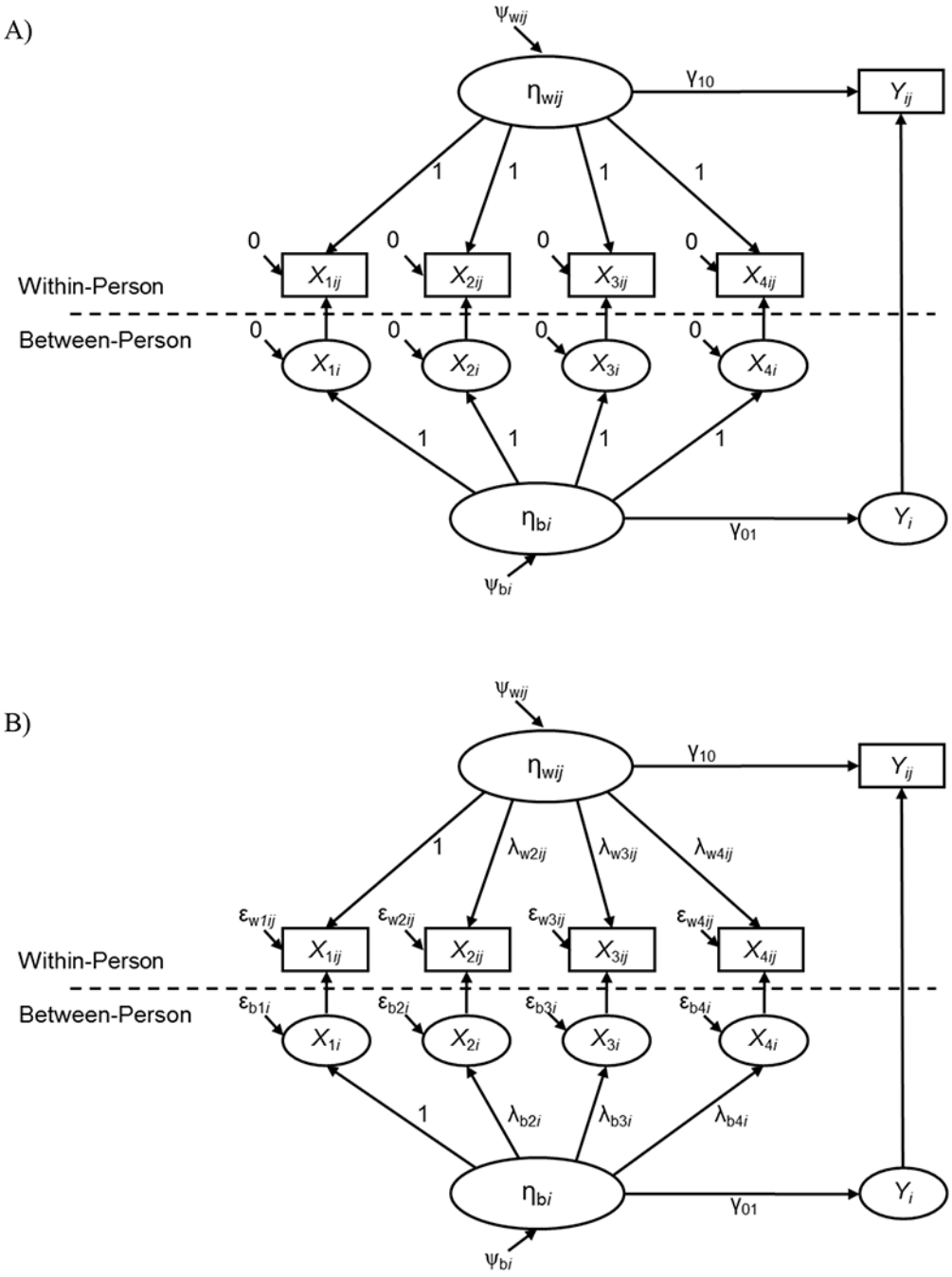

Fig. 1.

a Unit-weighted multilevel model with within-person and between-person predictor variable. **b** Multilevel SEM with within- and between-person predictor variable. *Note.* η_{wij} and η_{bij} = within- and between-person latent variables; X_{1ij} to X_{4ij} = time-varying indicators; X_{1i} to X_{4i} = person-means of indicators; ε_{w1ij} to ε_{w4ij} = within-person item residuals; ε_{b1i} to ε_{b4i} = between-person item residuals; λ_{w2ij} to λ_{w4ij} = within-person factor loadings; λ_{b2ij} to λ_{b4ij} = between-person factor loadings; ψ_{wij} and ψ_{bij} = within- and

between-person factor variance; γ_{10} = effect of within-person covariate on Y_{ij}; and γ_{01} = effect of between-person covariate on Y_j.

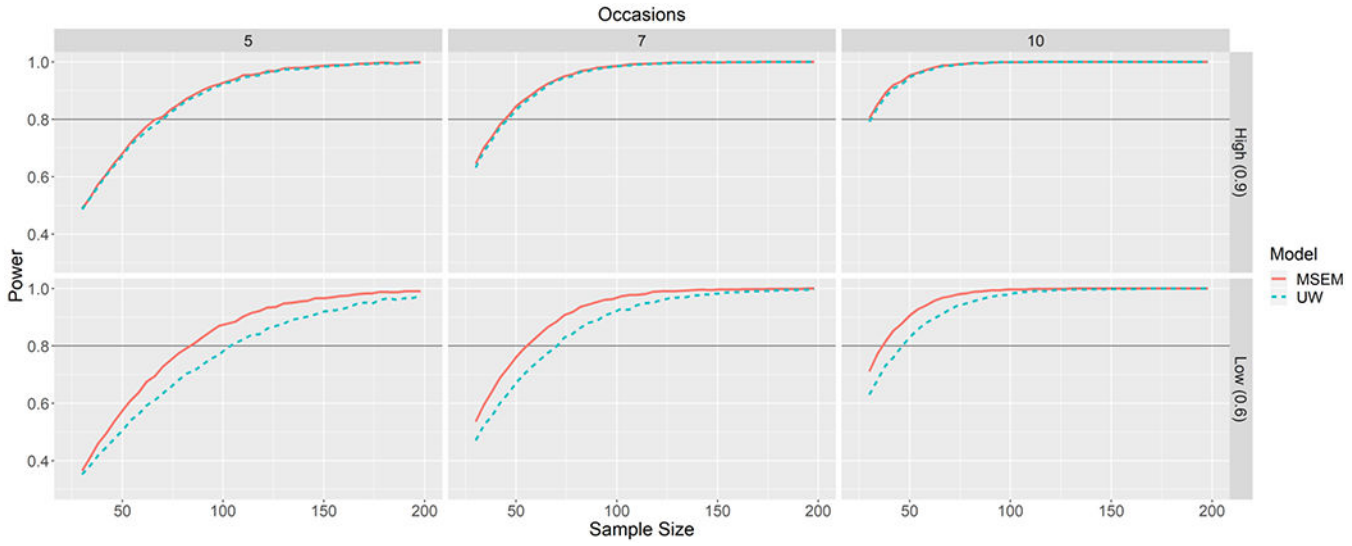

Fig. 2.

Power to detect a within-person effect across varying conditions. *Note*: Based on Monte Carlo simulation of 5000 replications. MSEM = multilevel structural equation model; UW = unit-weighted model

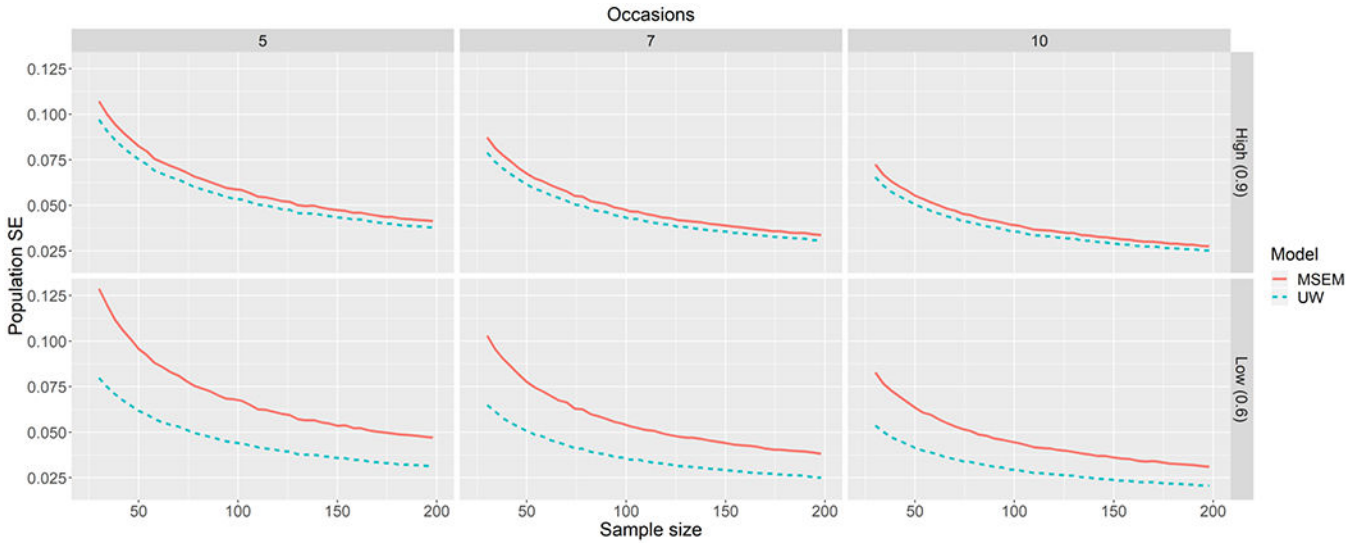

Fig. 3.

Precision of within-person estimate across varying conditions. *Note*: Based on Monte Carlo simulation of 5000 replications. MSEM = multilevel structural equation model; UW = unit-weighted model

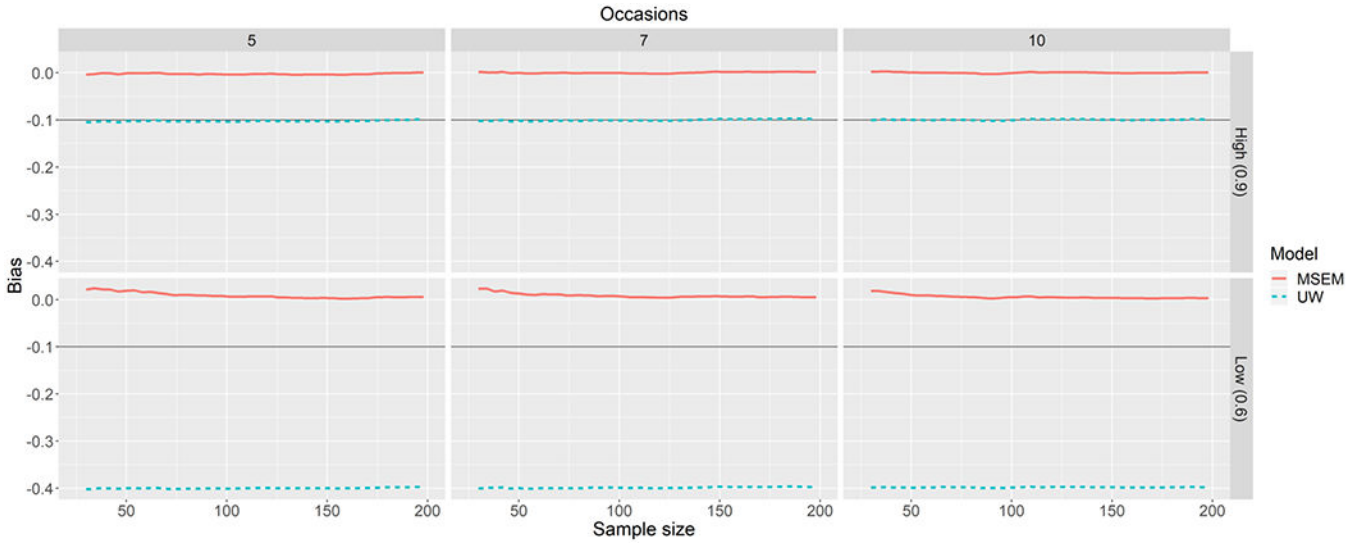

Fig. 4.

Bias in within-person estimates across varying conditions. *Note*: Based on Monte Carlo simulation of 5000 replications. Bias (mean estimated value − population value)/population value. MSEM = multilevel structural equation model; UW = unit-weighted model

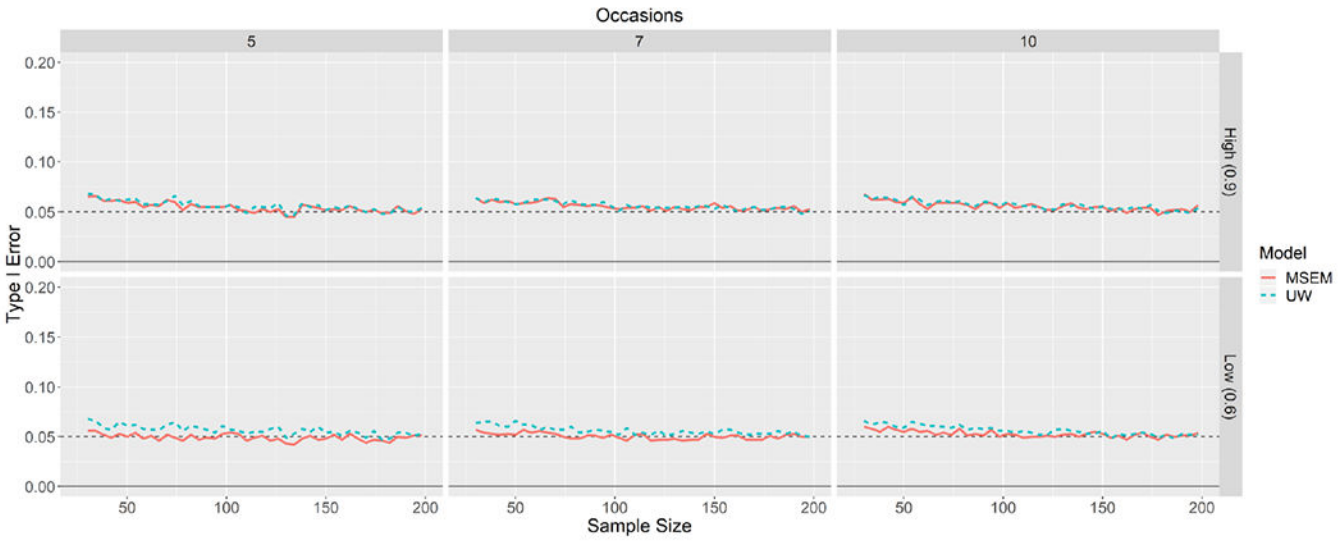

Fig. 5.

Type I error rate in within-person estimates across varying conditions. *Note*: Based on Monte Carlo simulation of 5000 replications. MSEM = multilevel structural equation model; UW = unit-weighted model

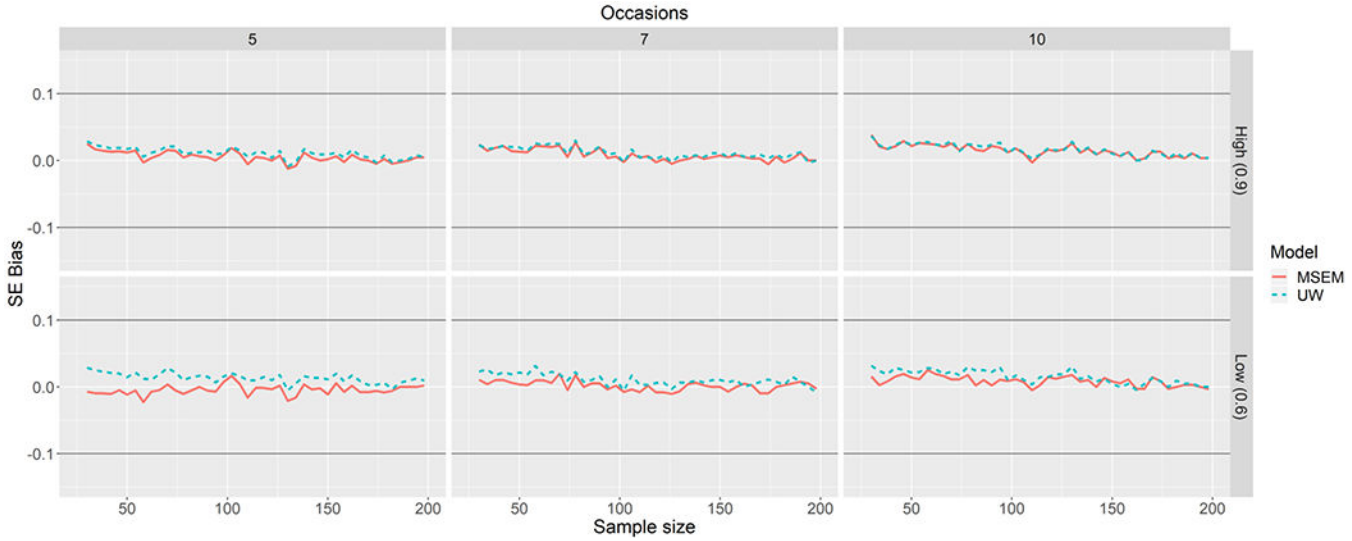

Fig. 6.

Bias in standard error (SE) of within-person estimates across varying conditions. *Note*: Based on Monte Carlo simulation of 5000 replications. Bias = (population standard error − average standard error) / population standard error. MSEM = multilevel structural equation model; UW = unit-weighted model

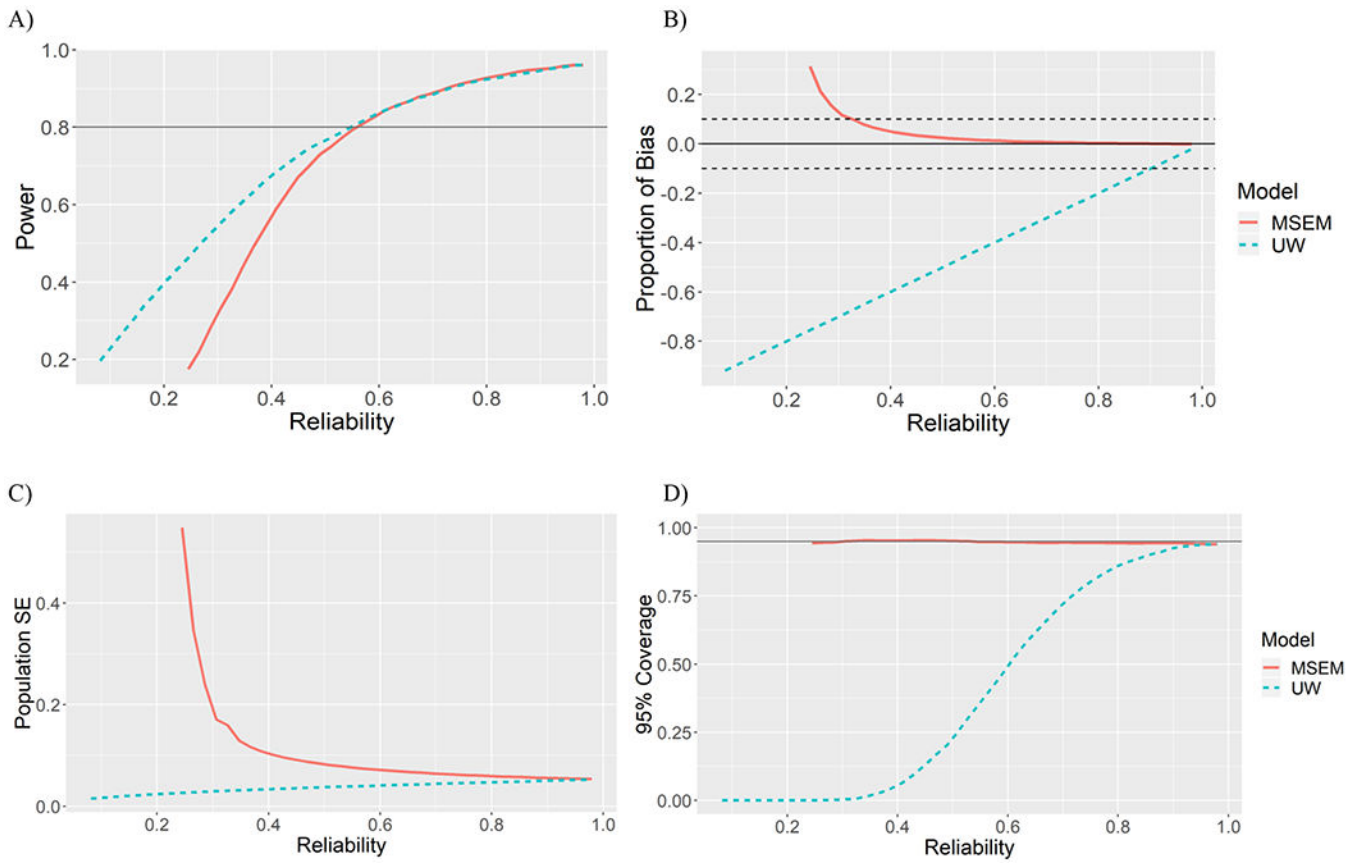

Fig. 7.

a Power, **b** bias, **c** precision, and **d** coverage across varying reliability of within-person predictor (n = 75; occasions = 7). *Note*: Based on Monte Carlo simulation of 5000 replications. Bias = (mean estimated value − population value) / population value. MSEM = multilevel structural equation model; UW = unit-weighted model

Table 1

Within-person population parameters

	Effect size	Factor variance	Factor loadings[a]	Item residuals
Scale reliability[b]				
High (0.9)	0.2	0.4	1.00, 1.36, 0.91, 0.73	0.15, 0.09, 0.21, 0.26
Low (0.6)	0.2	0.4	1.00, 1.64, 0.78. 0.58	1.07, 0.34, 1.30, 1.56

[a]Unstandardized factor loadings.

[b]Computed as true score variance / total variance.