Journal of Business and Psychology (2022) 37:1141–1156
https://doi.org/10.1007/s10869-022-09803-5

ORIGINAL PAPER

Optimizing Measurement Reliability in Within-Person Research: Guidelines for Research Design and R Shiny Web Application Tools

Liu-Qin Yang[1] · Wei Wang[2] · Po-Hsien Huang[3] · Anthony Nguyen[1]

Accepted: 27 February 2022 / Published online: 14 March 2022

Abstract

Within-person research has become increasingly popular over recent years in the field of organizational studies for its unique theoretical and methodological advantages for studying dynamic intrapersonal processes (e.g., Dalal et al., Journal of Management 40:1396–1436, 2014; McCormick et al., Journal of Management 46:321–350, 2020). Despite the advancements, there remain serious challenges for many organizational researchers to fully appreciate and appropriately implement within-person research—more specifically, to correctly conceptualize and compute the within-person measurement reliability, as well as navigate key within-person research design factors (e.g., number of measurement occasions, *T*; number of participants, *N*; and scale length, *I*) to optimize within-person reliability. By conducting a comprehensive Monte Carlo simulation with 3240 data conditions, we offer a practical guideline table showing the expected within-person reliability as a function of key design factors. In addition, we provide three easy-to-use, free R Shiny web applications for within-person researchers to conveniently (a) compute expected within-person reliability based on their customized research design, (b) compute observed validity based on the expected reliability and hypothesized within-person validity, and (c) compute observed within-person (as well as between-person) reliability from collected within-person research datasets. We hope these much-needed evidence-based guidelines and practical tools will help enhance within-person research in organizational studies.

Keywords Within-person research · Level-specific alpha · Reliability · R shiny web application · Validity

As the focus of organizational research shifts toward dynamic, intrapersonal processes of organizational phenomena that unfold over time, within-person research has increased substantially in recent years (C. Fisher & To, 2012; Gabriel et al., 2019). Within-person research refers to studies that use repeated measures of focal phenomena and examine the intrapersonal variability of the phenomena and/or the dynamic relations among these phenomena over time. For example, within-person research is able to study how an employee's increased positive mood enhances their job performance over time, and to examine how a new employee's teamwork skills change in their first year of tenure. Such research typically utilizes repeated measures or trials during an experimental study in a lab or a survey study in a field setting, using methods like experience sampling methods (e.g., multiple surveys within 1 workweek). As such, within-person research differs from between-person research that is most commonly done in organizational research and focuses on the variation among individuals, such as studying how between-person differences relate to each other at a single time point or across few time points.

Notably, there are both theoretical and methodological advantages of within-person research over between-person research (Dalal et al., 2014; Gabriel et al., 2019; McCormick et al., 2020). Theoretically, a significant advantage

Liu-Qin Yang and Wei Wang contributed equally to this article

✉ Liu-Qin Yang
lyang@pdx.edu

✉ Wei Wang
wwang@gc.cuny.edu

Po-Hsien Huang
psyphh@nccu.edu.tw

Anthony Nguyen
adn2@pdx.edu

[1] Department of Psychology, Portland State University, P.O. Box 751, Portland, OR 97207, USA

[2] The Psychology Program, CUNY Graduate Center, 365 5th Avenue, New York, NY 10016, USA

[3] National Chengchi University, Taipei, Taiwan

of within-person research is the capacity to investigate the dynamic *processes* through which a focal phenomenon unfolds and relates to another phenomenon over time. Insights from such investigations afford possibilities to challenge existing theories that were developed based on between-person research and further build new theories that are focused on dynamic intrapersonal processes and temporal contributions. For example, examining why and how new employees' task performance fluctuates more during their first year of job tenure relative to subsequent years makes novel theoretical contributions to the job performance literature. Methodologically, repeated measurements of focal phenomena better capture the manifestation of the phenomena over a specified time period, minimizing recall biases that are more prevalent in between-person research using a single measurement (Beal & Weiss, 2003; Robinson & Clore, 2002).

Despite such appealing advantages, there remain significant challenges for many organizational researchers to fully appreciate and appropriately implement within-person research. More specifically, one of the most important challenges involves (a) how to correctly conceptualize and compute the within-person measurement reliability and (b) how to optimize such reliability through appropriate within-person research design. We believe this challenge is largely due to two reasons. First, within-person research methods are often missing in typical research methods textbooks or graduate training curricula in the fields of organizational studies, leading to an incomplete or non-systematic understanding of the conceptualization and computation of within-person reliability. Relatedly, there has also been a lack of clear guidelines for within-person research design in the literature that can practically help organizational researchers navigate key research design factors (e.g., scale length, number of measurement occasions, number of participants) in order to optimize within-person reliability and ultimately maximize their research validity. Second, there lack practical, user-friendly, and analytical tools that facilitate the computation of within-person reliability, as such tools are not readily available in popular statistical packages, such as SPSS, Stata, SAS, or R.

With these obstacles in mind, we conduct this study intending to achieve four goals. First, we elaborate on the concept of within-person reliability, demystifying misconceptions and discussing the potential impact of various research design factors on this reliability. Second, using Monte Carlo simulations, we systematically examine how representative within-person research design factors—the number of measurement occasions, sample size, and scale length—along with an important psychometric property (e.g., item factor loading) collectively impact within-person reliability. Third, based on the simulation results, we provide practical guideline tables to help researchers from many research areas make informed decisions on within-person research designs to optimize within-person reliability. Fourth, we develop R shiny web applications[1] to facilitate easy computation of the expected within-person reliability based on a researcher's customized planned research design and the observed within-person reliability after collecting within-person research data.

Within-Person Research

Within-person research in psychological and organizational research utilizes a variety of methods, including experience sampling methods (ESM; i.e., diary methods or ecological momentary assessments), longitudinal methods with three or more time points and longer intervals than in ESM, and experimental methods with three or more repeated measurements. ESM includes time-based designs, such as interval-based designs (e.g., a survey is administered at the end of each workday for 2 consecutive workweeks) and event-based designs (e.g., employees fill out a survey whenever a target event like a conflict with a customer occurs). ESM has advantages to capturing experiences and examining processes as they occur in situ (Beal, 2015; C. Fisher & To, 2012; Gabriel et al., 2019) and it has become increasingly popular over the past two decades in organizational research. Longitudinal methods with three or more time points and a longer time interval than that in ESM (e.g., 1 month or 6 months) have been strongly recommended to examine research questions focused on the change in organizational phenomena in the field settings (Ployhart & Vandenberg, 2010; Zapf et al., 1996). Experimental methods with three or more repeated measurements in the laboratory or field settings have been deemed to afford strong causal inference in relationships between organizational phenomena (Shadish et al., 2002).

Compared to the between-person approach, within-person organizational research possesses numerous scientific advantages. Theoretically, within-person organizational research is advantageous to addressing temporally focused research questions, such as how newcomers' task performance fluctuates during their first year of job tenure, and how their involvement in an ongoing mentoring program accounts for their fluctuation in performance levels. Addressing such temporally intensive questions advances understanding of the dynamic processes that shape employees' task performance during their early job tenure and further informs the development of within-person theory that is distinct from theories developed based on between-person research (Dalal

[1] The R shiny web applications are available through URL https://psychmethods.shinyapps.io/WithinPersonResearch or https://tinyurl.com/LevelAlpha.

et al., 2014; Ployhart & Vandenberg, 2010). For example, it is likely that new employees tend to have less fluctuation in performance during their first year of tenure when the quality of their relationship with a mentor is higher, mainly because the mentoring relationship offers a source of emotional and instrumental support that play pivotal roles in these new employees' adjustment to new job roles, through processes like managing their anxiety and boosting their self-efficacy under situations of learning new and challenging job tasks (e.g., Ellis et al., 2015). Therefore, such evidence from within-person research may point to the need to integrate the tenets of negative feedback loops with those of broaden-and-build theory (see Dalal et al., 2014, for a review of performance theories) to understand how and why a workplace mentoring program benefits new employees' task performance over time.

Methodologically, within-person organizational research is advantageous in multiple ways (Beal, 2015; Gabriel et al., 2019). First, such research designs enhance confidence in causal inference about intrapersonal change, especially among studies using experimental designs with multiple measurements and those using longitudinal methods with three or more time points that allow explicit modeling of change (e.g., using latent change score modeling or latent growth modeling; Ferrer & McArdle, 2010; Ployhart & Vandenberg, 2010). Second, within-person research reduces common method variance or biases, through approaches like separating the measurements of predictors and outcomes and removing same-source biases (e.g., self-reports of all variables) via person-mean centering during within-person analyses. Additionally, within-person research reduces recall biases in survey responses, through the usage of shorter timeframes in survey instructions (e.g., today, this week) that allow participants to recall more recent episodic experiences as opposed to semantic knowledge about their general experiences (Beal, 2015; Robinson & Clore, 2002).

Within-Person Reliability

Despite the many appealing advantages of within-person research in organizational studies, there remain serious challenges for organizational researchers to fully appreciate and appropriately implement this method, one of which is within-person reliability. Specifically, misconceptions regarding within-person reliability still prevail in the community, knowledge about how different aspects of within-person research design impact within-person reliability is still lacking, and easy-to-use analytical tools to compute within-person reliability are missing. In the following sections, we focus on addressing these compelling issues.

In general, measurement reliability concerns the extent to which a measurement scale or test consistently measures the underlying construct of a focal phenomenon despite the random errors rooted in test-related, occasion-related, and personal factors. From the classic test theory (CTT; Allen & Yen, 1979) perspective, reliability is conceptualized as the ratio of the variance of true scores (standings in the underlying construct) to the variance of observed/test scores. Extending the CTT reliability conceptualization to the multilevel setting of within-person research, we define level-specific reliability—namely between- and within-person reliability—as the ratio of true score variance to observed score variance at the between-person and within-person levels, respectively (Geldhof et al., 2014). As such, within-person and between-person reliabilities represent reliability in measuring between-person differences and within-person variability of a focal phenomenon. We believe this level-specific reliability is the most applicable to within-person research focused on episodic variance or variance of shorter-term states across time. Such research is most concerned with questions about antecedents and consequences of dynamic phenomena and/or the relations between these dynamic phenomena, such as the relation between daily time pressure at work and employee burnout over time or whether participation in a peer-based workplace wellness program accounts for less daily fluctuation in burnout.

With respect to level-specific reliability, unfortunately, there remain serious challenges due to incomplete understanding and misconceptions of what it is and how it is relevant to within-person research. Such challenges account for the improper practices of calculating and reporting level-specific reliability, in particular, within-person reliability (Gabriel et al., 2019; Ohly et al., 2010), which can often lead to inaccurate and misleading conclusions in within-person research (Curran & Bauer, 2011; Geldhof et al., 2014). In this section, we will first discuss and demystify common misconceptions related to within-person reliability. Then, we will elaborate on research design factors that are critical for within-person reliability and address how to optimize within-person reliability through appropriate within-person research designs.

Common Misconceptions About Reliability in Within-Person Research

To better illustrate various misconceptions related to the reliability in within-person research, we will use two hypothetical research studies as examples. The first study is assumed to examine how both environmental and personal factors influence employees' daily fluctuations in work engagement levels, and we assume the researcher recruited 60 employees to fill out a daily diary survey for 10 days during 2 workweeks, thus obtaining 600 data points or observations. In the second hypothetical study, researchers recruited 100 newly promoted leaders across multiple organizations to fill

out three surveys during their first 3 months of promotion, to examine how their leadership skills fluctuate and what organizational initiatives (e.g., diversity training) influence their skill fluctuation.

Misconception #1: There Is Only One Level of Reliability for a Measure

A common misconception in the within-person research literature is that there is only one level of reliability for a measure. Accordingly, it is the gold standard to report the overall reliability of all observations (i.e., single-level reliability). This misconception partially stems from the common practice in the between-person research literature where only a single reliability index calculated from all observations is reported. In the context of the first hypothetical example illustrated above, all 600 observations are used to compute a single estimate of α for the work engagement measure. However, this approach to conceptualize and calculate reliability is problematic (see Nezlek, 2017, for a detailed review), as it mixes the variabilities of item scores that originate from both between-person differences and within-person changes/fluctuations, thus ignoring the multilevel nature of the data. In addition, previous research indicates that single-level reliability is generally more biased than level-specific reliabilities (Geldhof et al., 2014). Indeed, as illustrated in Eq. 1 below for reliability α, the single-level reliability is a weighted average of the within-person level reliability (i.e., α_w) and between-person level reliability (i.e., α_b).

$$\alpha = \rho_{ICC}\alpha_b + \left(1 - \rho_{ICC}\right)\alpha_w \quad (1)$$

where ρ_{ICC} is the intra-class correlation, ICC(1), defined as the ratio of between-person variance to total test score variance, or the nesting/clustering effect (Raudenbush et al., 1991). If the test score reliability is different across the two levels—which is typical for within-person research (e.g., Bakker et al., 2015; Brose et al., 2015), the single-level reliability reflects neither within- or between-level reliability.

Misconception #2: Reporting an Averaged Reliability Across Time Points Is Just Fine

Another common misconception in the within-person research literature is that reporting an averaged reliability across all time points is just fine. Owing to this misconception, researchers calculate a reliability for each time point; thus, multiple reliabilities are calculated and reported. In the context of the first hypothetical example, researchers may calculate an α for each day and then compute and report the mean of 10 α's, hoping to represent the measurement reliability of dynamic engagement levels. Conceptually, averaged reliability is equivalent to single-level reliability, when data meet the assumption that test reliability is equal across time points. However, as argued by Nezlek and others (C. Fisher & To, 2012; Nezlek, 2017), this averaged reliability approach—similar to single-level reliability—also mixes the variabilities of item scores that originated from both between-person differences and within-person fluctuations and thus also ignores the multilevel nature of the data.

Ignoring the multilevel nature of the data, using a single-level or averaged reliability, is consequential in within-person research, as it leads to biased reliability estimates and even misleading study conclusions under many circumstances (Curran & Bauer, 2011; Geldhof et al., 2014; Nezlek, 2017). For example, in cases where between-level reliability is much higher than within-person reliability (e.g., 0.90 versus 0.60) for within-person research like the work engagement study, the single-level or averaged reliability (i.e., approximately 0.76, assuming ICC(1) = 0.54 as reported in the review by McCormick et al., 2020) innevitably overestimates the within-person reliability, leading researchers to conclude that their study had reliable measurements to address within-person research questions. Yet, the relatively low within-person reliability may be a factor attenuating focal effect sizes and, in turn, limiting the statistical power of their study (Spearman, 1904; Wänström, 2009), which partially accounts for the non-significant findings on focal within-person or cross-level relations—an erroneous conclusion.

Misconception #3: Increasing the Number of Time Points Is Optimal to Reliability

Still, many within-person researchers may mistakenly believe that increasing the number of time points is optimal to measurement reliability in within-person research. Although there are often conceptual advantages for within-person research to have more measurement time points (e.g., to generalize findings on dynamic relationships based on more frequent sampling of time/experiences), having more time points is not necessarily optimal for within-person reliabilities. The key idea behind the reliability of the within-person measures, such as leadership skills among newly promoted leaders, lies in the ratio of true variability (variability of true scores or the underlying skills) to noise levels (measurement errors) across multiple time points, such that a higher ratio corresponds to a higher within-person measurement reliability—ratio of true variability to total test score variability (Allen & Yen, 1979; Geldhof et al., 2014). Having a larger number of time points broadens the sampling of time but does not guarantee a higher ratio of true variability to noise levels. Using the aforementioned leadership study as an example, it is likely that having the new leaders complete six surveys (vs. three surveys) during the first 3 months of their leadership tenure might introduce

more noise in their responses yet not necessarily increase the true variability in the leadership skills, because participants are too overwhelmed by the new role to fill out this many surveys carefully.

Misconception #4: Shorter Scales Can Always Measure Within-Level Phenomena Reliably

Yet, another misconception in within-person research is that shorter scales work well in reliably measuring phenomena that fluctuate. This misconception partially stems from the common practice of using shorter scales in within-person research, in line with the argument and effort to reduce survey fatigue in repeated surveys of employees for organizational within-person research, particularly ESM research (C. Fisher & To, 2012; Gabriel et al., 2019; Ohly et al., 2010). Owing to this misconception, it has been common for researchers to use shorter scales or shorten existing longer scales in within-person organizational research (Gabriel et al., 2019; Ohly et al., 2010), such as using a short leadership skill scale in the example leadership study in order to reduce these new leaders' survey fatigue and increase their response rate. Conceptually, we may infer that shorter scales used to measure within-person fluctuations of leadership skills may have lower reliabilities than their longer-version counterparts, due to the more significantly reduced variability of true scores reflecting the underlying skills (e.g., as indicated by the more significantly reduced shared variability or covariances between fewer items) relative to the reduction in noise levels (item unique variances; Cronbach, 1951). Little empirical research has been completed, however, to examine the effect that scale length has on within-person reliability of dynamic phenomena relative to other design factors in informing research design.

Demystifying Within-Person Reliability

Within the broader reliability literature, research has developed the two most efficacious (least biased) methods for operationalizing and computing level-specific reliability in multilevel settings (Geldhof et al., 2014; Huang & Weng, 2012): level-specific α and level-specific ω. Conceptually, these two level-specific reliability methods can be integrated into the framework of multilevel confirmatory factor analysis (MCFA), where α is a special case of ω. Specifically, the α method assumes essential tau equivalence across items (equal true score variance or equal factor loadings), whereas the ω method assumes congeneric items (i.e., the items are measurements of the same latent construct). In the MCFA framework, between- and within-person α and ω coefficients are both derived from the item covariance matrix at the between-person and within-person level, respectively (Geldhof et al., 2014; McDonald, 1999). In terms of computational approach, however, α and ω coefficients are somewhat distinct, in that ω captures the ratio of true score variance to observed score variance more directly using a model-based approach yet α is more sensitive to item homogeneity by using a formula focused on the averaged item covariance.

In the present research, we will focus on level-specific α, for two reasons. First, evidence to date (e.g., Geldhof et al., 2014) suggests that both level-specific α and level-specific ω methods are equally effective under multilevel settings, in terms of estimating level-specific reliability with little bias for unidimensional scales under a variety of research conditions.[2] Second, since α is the reliability method that psychologists and organizational scholars are most familiar with (Cho, 2016; Cortina et al., 2020; Geldhof et al., 2014), level-specific α (as an extension of single-level α applied in conventional between-person organizational research) can be comprehended and accepted more easily than level-specific ω in the organizational research literature. Accordingly, we will describe below how between-person and within-person α are defined and computed.

Level-Specific α

Extending the classic Cronbach's α, an internal consistency reliability (Cronbach, 1951), Geldhof et al. (2014) proposed level-specific α that is appropriate for multilevel research, including for within-person research. The level-specific α method adapts Cronbach's α formula by using level-specific variance components to compute between-person α and within-person α (shown in Eq. 2),

$$\alpha_w = \frac{I}{I-1}\left(1 - \frac{\sum_{i=1}^{I}\left(\sigma_i^{(w)}\right)^2}{\left(\sigma_{sum}^{(w)}\right)^2}\right), \tag{2}$$

where $\left(\sigma_i^{(w)}\right)^2$ and $\left(\sigma_{sum}^{(w)}\right)^2$ represent the variance of item i and sum score at the within-person level, respectively, and I is the number of items in a scale or scale length. In the context of within-person organizational research, we argue

[2] Our research team also conducted a separate set of simulations using realistic research design conditions based on the empirical review of 103 unique organizational ESM studies published in 10 representative organizational research journals (e.g., scale length of 2–5 items, 10 vs. 25 measurement occasions, 60 vs. 90 vs. 120 participants), to compare the averaged biases and root mean squared errors (RMSE) of the alpha and omega reliability estimates relative to the true reliabilities. We found that the alpha and omega methods were equally efficacious in terms of having little biases and RMSE in estimating within- and between-person reliabilities for unidimensional scales.

that within-person reliability is of utmost concern for the following reason. The vast majority of temporal-focused research questions addressed in within-person organizational research are concerned with the intrapersonal processes underlying focal phenomena over time (Dalal et al., 2014; McCormick et al., 2020). Thus, assessing these processes over time with reliable measures is a key prerequisite for the validity of measuring these focal phenomena and for the validity of establishing within-person theories focused on the intrapersonal variability of these phenomena over time. Indeed, Dalal and colleagues argued (2014) as follows: "The theories of within-person performance variability we reviewed previously would by definition be falsified… if the within-person variability in observed performance scores is attributable primarily to measurement error related to the items in the instrument and/or the occasion of measurement" (p. 1424).

Impact of Research Design Factors on Within-Person Reliability

There are many factors to consider in designing within-person research, including representative research design factors—the number of observations/time points, the number of participants, and the number of items/scale length—and psychometric property factors, such as the magnitude of item factor loading. Although item factor loading is a key psychometric property of a within-person scale, we also include it as a design factor in the current study, especially under situations where the researchers can obtain such information (i.e., factor loadings) from the past literature or from their own pilot data (e.g., validating a new or newly adapted scale for within-person research), and take it into account in research design.

First, the number of observations per person (i.e., *T*) may influence the intrapersonal variability of focal phenomena (e.g., engagement levels). In the context of organizational ESM research, cumulative evidence has shown that studies with a higher number of measurement occasions reported a higher percentage of within-person variance over the total variance across variables like attitude and behavior (54% for 11 or more occasions), relative to those with a lower number of occasions (e.g., 43% for five or fewer occasions; for a review, see McCormick et al., 2020). Relatedly, three or more time points are recommended to capture any linear or nonlinear trend or change of focal phenomena over time (Ployhart & Vandenberg, 2010; Shadish et al., 2002). The choice of the observation number is determined by both the frequency of the measurement and the intended study length. Both theoretical and practical considerations can go into this decision. Theoretically, it may take longer for some phenomena to manifest change than others; and practically, it is more feasible to administer more surveys within a shorter time window for some study settings than for others. Given a reasonable study length (e.g., 1 or a few months in total), we contend that having a higher (vs. lower) number of measurement occasions will increase (vs. decrease) the capacity to detect meaningful and statistically significant within-person variability and relations (Dalal et al., 2014; McCormick et al., 2020). As discussed earlier in "Misconception #3: Increasing the Number of Time Points Is Optimal to Reliability," however, empirical research is sorely needed to systematically examine the impact of *T* on within-person reliability.

Second, the number of participants (i.e., sample size, or *N*) is another important factor to consider in designing organizational within-person research. For within-person research questions (e.g., within-person relations between phenomena over time), a higher number of participants (*N*) contribute to a higher total number of observations (i.e., $N \times T$), which in turn reduces standard errors of the estimates of within-person relations and increases statistical power to detect the between-person boundary conditions of within-person relations or cross-level interactions (Hox, 2002; Wänström, 2009). Yet, it is not clear to what extent *N* may impact within-person reliability.

Third, the scale length (i.e., number of items in a scale, or *I*) is another critical research design factor to consider. Due to the need for participants to complete multiple surveys, having shorter scales in organizational within-person research (especially ESM research) is a common practice, with a goal to reduce survey fatigue and increase the quality and rate of participant responses (Beal, 2015; Gabriel et al., 2019). Indeed, a recent review of organizational ESM research indicated that the majority of the literature (56% of the cases) utilized four or fewer items to measure a focal construct (McCormick et al., 2020). As discussed earlier in "Misconception #4: Shorter Scales Can Always Measure Within-Level Phenomena Reliably," conceptually, shorter scales may have a lower within-person reliability. Yet, empirical research is critically needed to examine the extent to which scale length impacts within-person reliability, in order to better inform within-person research design and to guide common practices like shortening scales.

Fourth and lastly, the magnitude of item factor loadings (i.e., λ) may play an important role in determining the magnitude of within-person reliability. Conceptually, the size of within- and between-person item factor loadings represents the extent to which each item captures the content of the underlying phenomenon (e.g., leadership skills) at the respective within- and between-person level. Thus, scales with higher within- and between-person item factor loadings tend to produce a higher within- and between-person reliability, respectively (Geldhof et al., 2014). Unfortunately, few published organizational within-person studies to date report the item factor loadings of focal scales at the within- and

between-person levels based on multilevel confirmatory factor analyses (MCFA; Gabriel et al., 2019). As a result, simulation research is sorely needed to empirically examine how within-person factor loadings influence within-person reliability, in order to inform within-person research design and guide practices like shortening scales (e.g., efforts to validate a shortened scale).

Navigating Research Design Factors to Optimize Within-Person Reliability

To optimize within-person reliability, we need to consider the trade-offs and feasibility in choosing the aforementioned research design factors (i.e., *T*, *N*, *I*, and λ). Given the challenges described above (Gabriel et al., 2019), within- and between-person item factor loadings were often neither calculated nor reported in the existing literature; thus, it may not be feasible to adjust the item factor loading while designing organizational within-person research. Accordingly, we will primarily focus on the trade-offs among the other three design factors, while still including the item factor loading in the simulation to examine the impact of each of four factors.

As discussed earlier, *T*, *N*, and *I* are three key research design factors that collectively influence the quality of estimates of within-person and cross-level relations concerning focal phenomena, and/or the statistical power to detect a meaningful intrapersonal change and within-person relations. These factors can also independently and jointly influence the magnitude of within-person reliability and can be strategically chosen to optimize reliability. Unfortunately, to the best of our knowledge, there has been no empirical research to examine the effects of *T*, *N* and, *I* and on within-person reliability in within-person research. Thus, our current study is one of the first ones to systematically examine how within-person research design factors, *T*, *N*, and *I*—both independently and interactively—impact within-person reliability in within-person research, which is expected to offer practical insights on within-person research design.

> Research question: How shall we navigate within-person research design factors (i.e., T, N, and I) to optimize within-person reliability?

Method

We conducted a comprehensive Monte Carlo simulation study to examine how the magnitude of the estimated within-person reliability is impacted by three research design factors: (a) the number of measurement occasions, *T*; (b) the number of participants, *N*; and (c) scale length, *I*. We examined each of these research design factors with a wide range of possible values. Specifically, we set the number of measurement occasions, *T*, ranging from 3 to 30 occasions with 12 specifications: 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25, and 30; set the sample size, *N*, ranging from 5 to 200 with 10 specifications: 5, 10, 15, 20, 25, 50, 75, 100, 150, and 200; and set scale length, *I*, ranging from 2 to 10 items with an increment of 1. In addition, we also manipulated a psychometric property—the magnitude of the item factor loadings, λ, to examine if and how the factor loading might play a role beyond the three research design factors examined in the study. Although item factor loadings at the within-person level are typically unknown to researchers beforehand, according to the past literature (e.g., Bakker et al., 2015; Brose et al., 2015), 0.70 was a common value for item factor loadings at the within-person levels. Thus, in the simulation, we examined three levels of item factor loading: 0.50, 0.70, and 0.90, corresponding to the low, medium, and high levels. We set the loading the same for all items at within- and between-person levels. We also set ICC(1) at 0.50, a typical ICC(1) based on the organizational ESM literature (McCormick et al., 2020). Finally, the data simulated in this study are continuous data with a normal distribution. Past research suggests that response data based on a Likert scale of 5 points or higher, most commonly used in organizational research, could be treated as continuous with respect to estimation quality (e.g., DiStefano, 2002; Rhemtulla et al., 2012); thus, findings from this study are readily applicable to organizational research.

We fully crossed all the specifications/levels of the four factors, resulting in 3240 conditions (i.e., 12 [*T*] × 10 [*N*] × 9 [*I*] × 3 [λ]) of data. In addition, we simulated 1000 Monte Carlo iterations for each condition. For each iteration, we simulated response data by using the specified data parameters and then computed the within-person reliability. We calculated the average reliability across the 1000 iterations as the estimated reliability for each condition.

We believe that the 3240 studied conditions cover the vast majority of within-person research design scenarios in organizational research. However, for those beyond the 3240 conditions (e.g., $N > 200$, measurement occasions $T > 30$), we have also developed an R shiny web application, where researchers may freely customize research design by entering their specific parameters to compute the corresponding level-specific reliability (between- and within-person reliabilities). The use of the R shiny web application is straightforward, and we have also provided a user guide in Appendix B of the online supplemental materials (Fig. S1).

We conducted the data simulation and the computation of the level-specific reliability in R Version 4.0.2 (R Core Team, 2021) by using the R package *OpenMx* (Neale et al., 2016). After computing the reliability at both the within- and between-person levels, we ran ANOVAs for the estimated reliability computed from the 3240 studied conditions and also calculated eta-squared to assess the effect size of each

factor (along with all two-way interactions) for impacting the magnitude of the within-person reliability.

Results

We present the estimated within-person reliability as a function of the number of participants, scale length, and the number of measurement occasions across the 1080 conditions for medium factor loadings (0.70) in Table 1. The corresponding results for high (0.90) and low (0.50) factor loadings are presented in Tables S1 and S2, in Appendix A of the online supplemental materials. These tables are organized by panels based on scale length, ranging from 2 items (the first panel) to 10 items (the last panel). In each panel, the expected reliability is presented in a matrix, in which the rows represent the number of participants ranging from 5 to 200, and the columns represent the number of measurement occasions ranging from 3 to 30. The simulation results revealed a wide range of reliability at the within-person level. For medium factor loadings, the estimated within-person reliability ranged from the lowest 0.57 for a simple design with 2 items, 5 participants, and 3 measurement occasions, up to the highest 0.91 for a more complex design with 10 items, 200 participants, and 30 measurement occasions (see Table 1). This table provides a practical guideline for within-person researchers to plan their research design from the within-person reliability perspective.

To better understand the result patterns comparatively, we first visualized within-person reliability for 240 selected representative data conditions ($\lambda = .70$) and then ran ANOVA analyses to quantitatively determine the effect of each factor along with all the two-way interactions between factors using all 3240 conditions. The results of the within-person reliability ($\lambda = .70$) shown in Table 1 and Fig. 1 indicate that there is a strong positive effect of scale length on the within-person reliability: The more items included in a scale, the higher the within-person reliability can be achieved. In addition, the within-person reliability estimates showed a clear ceiling effect, primarily determined by the scale length (I). That is, the scale length determined the upper limit of within-person reliability, or the highest possible within-person reliability given a certain scale length planned in a research design with varying number of participants and/or the number of measurement occasions. For items with medium item factor loadings, this upper limit was 0.66 with a 2-item scale, and it increased to 0.79 for a 4-item scale, and it further increased to 0.85, 0.89, and 0.91 for a 6-, 8-, and 10-item scale, respectively. However, the incremental effect on the within-person reliability decreased when adding one item to a longer scale (e.g., a 9-item scale) relative to adding one item to a shorter scale (e.g., a 2-item scale).

Similar to the effect of the scale length, the number of participants and the number of measurement occasions showed a positive—yet much weaker—effect on the within-person reliability. That is, the within-person reliability increases as the number of participants or measurement occasions increases to a smaller extent than the case when scale length increases.

Most interestingly, as comparing within-person reliability with medium factor loadings (Table 1) to the ones with high and low factor loadings (Tables S1 and S2 in Appendix A of the online supplemental materials), we noticed the factor loading also had a strong effect on the within-person reliability: the within-person reliability became much higher when factor loadings increased from 0.70 to 0.90. In addition, the results together revealed a compensatory effect between factor loading and scale length. That is, using a shorter scale with higher factor loadings could achieve the same level of within-person reliability as a longer scale but with medium factor loadings. For example, with medium factor loading items (i.e., $\lambda = 0.70$), for a sample size of 15 and 4 measurement occasions, nine items were required in the scale to achieve within-person reliability of 0.89; yet, with high factor loading items (i.e., $\lambda = 0.90$), only two items were required to achieve the same within-person reliability.

All the patterns we discovered from the tables and figures were supported by the effect size analyses shown in Table 2, in which η^2 refers to the effect size of the corresponding factor in influencing the estimated magnitude of level-specific reliability, whereas partial η^2 refers to the effect size after the effects of all other factors were partialled out. The effect size results clearly showed that factor loading was the strongest factor that impacted the reliability at within-person level ($\eta^2 = 0.706$). Similarly, the number of items also showed a strong effect on the within-person reliability estimates ($\eta^2 = 0.201$). In contrast, the number of time points and the sample size showed little impact on within-person reliability ($\eta^2 = 0.000$ for both factors). In addition, there showed a small but significant interaction effect of scale length and factor loading for within-person reliability ($\eta^2 = 0.056$), suggesting there was a compensatory effect between the scale length and factor loading.

Discussion

Using Monte Carlo simulations, the present research systematically examines the effects of a few key within-person research design factors (the number of measurement occasions, T; the number of participants, N; scale length, I; and item factor loading, λ) on the within-person reliability. Based on the findings, we offer practical guidelines on how to balance the specifications of these factors in designing within-person organizational research, while ensuring adequate

Table 1 Expected within-person reliability as a function of number of measurement occasions (T), sample size (N), number of items (I), with medium item factor loadings (λ = .70)

Number of participants (*N*)	Number of measurement occasions (*T*)											
	3	4	5	6	7	8	9	10	15	20	25	30
Scale length (*I*) = 2												
5	.565	.603	.612	.632	.639	.634	.640	.637	.649	.650	.653	.649
10	.630	.630	.634	.643	.638	.646	.652	.651	.652	.651	.656	.656
15	.630	.638	.648	.646	.648	.653	.654	.652	.654	.655	.655	.657
20	.642	.645	.651	.651	.649	.653	.652	.652	.658	.658	.656	.655
25	.643	.652	.649	.651	.652	.654	.656	.656	.654	.655	.656	.657
50	.652	.653	.655	.654	.658	.655	.657	.656	.657	.658	.657	.658
75	.654	.654	.654	.656	.657	.657	.658	.657	.656	.657	.657	.658
100	.653	.654	.655	.656	.657	.657	.657	.657	.656	.658	.657	.658
150	.656	.658	.656	.658	.657	.657	.657	.659	.658	.657	.657	.658
200	.658	.658	.657	.657	.657	.657	.657	.657	.658	.658	.657	.658
Scale length (*I*) = 3												
5	.681	.699	.711	.718	.725	.731	.734	.728	.735	.738	.737	.739
10	.711	.721	.728	.733	.734	.737	.734	.738	.738	.740	.740	.741
15	.723	.731	.730	.733	.735	.736	.738	.739	.740	.739	.742	.741
20	.728	.730	.737	.739	.739	.738	.739	.741	.741	.740	.742	.742
25	.731	.733	.737	.738	.738	.740	.739	.740	.741	.740	.741	.744
50	.738	.737	.738	.741	.740	.741	.741	.741	.741	.742	.742	.743
75	.739	.740	.740	.740	.743	.742	.743	.742	.742	.742	.742	.742
100	.742	.741	.741	.742	.742	.741	.741	.742	.742	.742	.742	.742
150	.739	.741	.742	.742	.742	.742	.742	.742	.742	.742	.743	.743
200	.741	.742	.741	.743	.742	.742	.741	.742	.743	.742	.742	.743
Scale length (*I*) = 4												
5	.740	.762	.768	.774	.779	.783	.783	.783	.788	.788	.789	.792
10	.769	.782	.786	.787	.785	.789	.790	.788	.791	.793	.792	.792
15	.784	.787	.787	.787	.790	.789	.790	.790	.791	.793	.793	.793
20	.784	.787	.789	.788	.789	.792	.791	.791	.792	.793	.793	.793
25	.783	.790	.788	.790	.791	.791	.792	.792	.792	.792	.793	.793
50	.789	.792	.792	.792	.792	.792	.792	.792	.793	.793	.793	.793
75	.791	.791	.794	.792	.792	.792	.792	.793	.793	.793	.794	.794
100	.792	.792	.793	.793	.792	.793	.793	.793	.793	.793	.794	.793
150	.792	.793	.793	.793	.793	.793	.792	.793	.794	.793	.793	.793
200	.792	.793	.794	.793	.793	.794	.793	.793	.794	.793	.793	.794
Scale length (*I*) = 5												
5	.776	.805	.804	.813	.819	.815	.819	.821	.821	.823	.823	.825
10	.804	.818	.817	.821	.823	.823	.824	.824	.826	.826	.828	.828
15	.813	.818	.822	.822	.823	.824	.825	.823	.826	.826	.827	.827
20	.818	.822	.820	.826	.824	.826	.826	.826	.826	.826	.828	.827
25	.821	.822	.824	.824	.824	.825	.826	.827	.827	.828	.827	.827
50	.823	.825	.826	.825	.827	.827	.826	.827	.827	.828	.828	.828
75	.825	.827	.827	.827	.827	.827	.827	.827	.828	.827	.828	.828
100	.825	.826	.827	.827	.827	.828	.827	.827	.827	.828	.828	.827
150	.827	.828	.827	.827	.827	.827	.828	.827	.828	.828	.828	.828
200	.827	.827	.827	.828	.828	.828	.828	.828	.828	.828	.828	.828
Scale length (*I*) = 6												
5	.812	.830	.834	.838	.845	.844	.842	.845	.847	.847	.850	.849
10	.840	.842	.843	.846	.848	.847	.850	.849	.850	.851	.851	.851
15	.843	.844	.845	.847	.850	.850	.849	.850	.851	.850	.850	.852
20	.846	.847	.849	.848	.850	.851	.851	.851	.851	.852	.851	.851
25	.846	.849	.848	.850	.849	.851	.852	.851	.852	.851	.852	.852
50	.849	.850	.851	.851	.851	.851	.851	.852	.852	.852	.852	.852

Table 1 (continued)

Number of participants (*N*)	Number of measurement occasions (*T*)											
	3	4	5	6	7	8	9	10	15	20	25	30
75	.850	.851	.851	.852	.852	.852	.851	.852	.852	.852	.852	.852
100	.850	.851	.852	.852	.852	.852	.852	.852	.852	.852	.852	.852
150	.851	.852	.852	.851	.852	.852	.852	.852	.852	.852	.852	.852
200	.851	.852	.852	.852	.852	.852	.852	.852	.852	.852	.852	.852
Scale length (*I*) = 7												
5	.844	.850	.856	.859	.863	.863	.865	.864	.867	.868	.869	.869
10	.858	.861	.864	.865	.867	.867	.867	.868	.870	.869	.869	.870
15	.864	.864	.866	.867	.867	.868	.868	.869	.870	.870	.870	.870
20	.865	.865	.866	.867	.868	.868	.868	.870	.870	.870	.870	.870
25	.868	.868	.868	.868	.869	.869	.869	.869	.869	.870	.870	.870
50	.868	.869	.869	.869	.870	.870	.870	.870	.870	.870	.870	.870
75	.869	.870	.869	.869	.870	.870	.870	.870	.870	.871	.870	.871
100	.870	.870	.870	.870	.871	.871	.870	.871	.870	.871	.870	.870
150	.870	.870	.870	.871	.870	.870	.870	.871	.870	.870	.871	.871
200	.870	.871	.870	.870	.870	.870	.870	.870	.871	.870	.871	.870
Scale length (*I*) = 8												
5	.854	.866	.875	.875	.876	.879	.878	.880	.880	.882	.882	.884
10	.872	.877	.879	.880	.881	.882	.883	.883	.883	.884	.884	.884
15	.877	.879	.879	.881	.883	.882	.883	.883	.884	.884	.884	.884
20	.877	.881	.881	.882	.883	.883	.884	.883	.884	.884	.884	.885
25	.879	.881	.884	.883	.883	.884	.884	.884	.884	.884	.884	.885
50	.883	.884	.883	.884	.884	.884	.884	.884	.885	.884	.885	.885
75	.883	.883	.884	.884	.885	.884	.884	.885	.885	.885	.885	.885
100	.884	.884	.885	.884	.884	.885	.885	.884	.885	.885	.885	.885
150	.885	.884	.885	.884	.885	.885	.885	.885	.885	.885	.885	.885
200	.884	.884	.885	.885	.885	.885	.885	.885	.885	.885	.885	.885
Scale length (*I*) = 9												
5	.872	.881	.884	.888	.890	.892	.891	.890	.893	.894	.895	.895
10	.884	.888	.891	.893	.894	.893	.893	.894	.895	.896	.895	.896
15	.889	.893	.892	.893	.894	.895	.895	.895	.895	.895	.896	.896
20	.890	.893	.893	.895	.895	.895	.894	.896	.896	.895	.896	.896
25	.893	.894	.894	.895	.896	.895	.895	.895	.896	.896	.896	.896
50	.894	.895	.896	.895	.896	.896	.896	.896	.896	.896	.896	.896
75	.895	.896	.896	.895	.896	.896	.896	.896	.896	.896	.896	.896
100	.895	.896	.896	.896	.896	.896	.896	.896	.896	.896	.896	.896
150	.896	.896	.896	.896	.896	.896	.896	.896	.896	.896	.896	.896
200	.896	.896	.896	.896	.896	.896	.896	.896	.896	.896	.896	.896
Scale length (*I*) = 10												
5	.883	.890	.895	.898	.900	.899	.901	.901	.903	.903	.904	.904
10	.894	.899	.902	.902	.901	.903	.903	.903	.904	.904	.905	.905
15	.899	.903	.902	.904	.903	.903	.904	.905	.905	.906	.905	.905
20	.901	.902	.903	.904	.903	.904	.904	.905	.905	.905	.905	.905
25	.901	.904	.904	.904	.905	.905	.904	.905	.905	.905	.906	.905
50	.904	.905	.906	.905	.905	.905	.906	.905	.906	.906	.905	.905
75	.904	.905	.904	.905	.905	.906	.905	.906	.906	.906	.906	.906
100	.904	.905	.905	.906	.905	.906	.905	.905	.906	.906	.906	.906
150	.905	.905	.905	.905	.905	.906	.906	.905	.905	.905	.906	.906
200	.905	.905	.905	.905	.906	.906	.905	.906	.906	.906	.906	.906

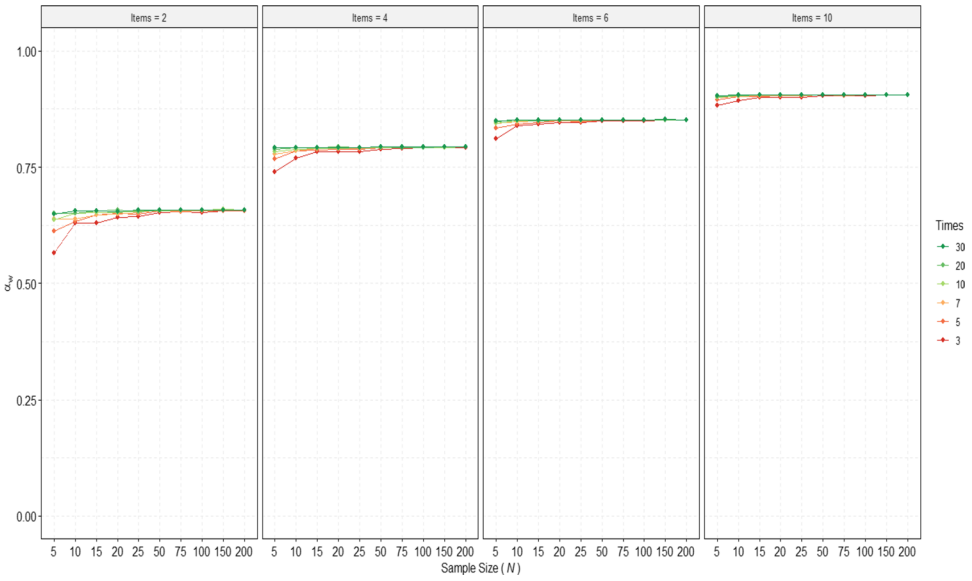

Fig. 1 Expected within-person reliability for selected conditions ($\lambda = .70$)

Table 2 Effect size (η^2) of research design factors impacting the magnitude of the estimated reliability at the within-person level

Factor	Within-person reliability	
	η^2	partial η^2
Timepoint	.000	.008
Items	***.201***	***.844***
N	.000	.005
Factor loading	***.706***	***.950***
Timepoint × items	.000	.001
Timepoint × *N*	.000	.003
Timepoint × factor loading	.000	.003
Items × *N*	.000	.001
Items × factor loading	***.056***	***.599***
N × factor loading	.000	.002
Adjusted R^2	.9627	

Timepoint, number of time points or measurement occasions, Items, number of items or scale length, *N*, sample size, and Factor loading, within- and between-person item factor loading. Effect sizes greater than .05 are italicized and boldfaced. Partial η^2 is the eta squared effect size after other independent variables and interactions are partialled out

within-person reliabilities; we also provide three interactive R shiny web applications to allow within-person research design a much more straightforward practice. In this section, we will discuss the implications of our research findings for the methodologies and theorization of within-person organizational research and describe the extended applications of our research findings for consideration of validity in within-person research.

Within-Person Research Design Factors and Within-Person Reliability

As discussed earlier, systematic and practical guidelines are sorely needed for designing within-person research in organizational studies. In this section, we will discuss the implications of our simulation study findings for within-person organizational research design.

Dominant Effects of Scale Length and Item Factor Loading

Our results (Table 1, Fig. 1) indicate that scale length determines the upper limit of the within-person reliability across a variety of study design conditions we examined. The strong effect of scale length was substantiated by the results from the effect size analysis (Table 2). Furthermore, item factor loading and its interaction with scale length were also strong predictors of the magnitude of within-person reliability (Table 2), such that higher within-person factor loadings contribute to higher within-person reliability and they can compensate lower scale length to ensure adequate reliability (Table 1, Tables S1 and S2 in Appendix A of the online supplemental materials). Therefore, in the context of adapting longer scales established in between-person organizational research for use in organizational within-person research (a common practice; Gabriel et al., 2019; Ohly et al., 2010),

researchers need to be cognizant of the implications of choosing fewer items for within-person reliability and can benefit from choosing items with higher within-person item factor loadings whenever such information is available. For example, 2-item scales tend to have within-person reliabilities at 0.66 or lower across conditions, given medium-level, within-person item factor loadings (0.70); the reliabilities of such scales can go up to 0.87 or higher when the item factor loadings go up to 0.90. Notably, the compensatory effect of item factor loading for scale length is stronger than the other way around, in terms of achieving a similar level of reliability, holding the other factors constant.

Compensatory Effects of the Numbers of Participants and Measurement Occasions

This finding has the most implication for the costs and feasibility of within-person research. When designing and implementing within-person research, it is often a trade-off between resources expended on recruiting more participants and those expended on retaining the participating individuals during repeated surveys/study sessions. For example, it may be equally costly and effortful to recruit 100 participants to participate in a week-long daily diary study (seven daily surveys in total), relative to recruiting 50 participants to complete a 2-week-long daily diary study (fourteen daily surveys in total), for a maximal total of 700 observations. If the researcher believes that representative sampling of both time and participants is important for the generalizability of their findings from the within-person research on a prevalent phenomenon (e.g., quality of social relationships at work), our results suggest that they can balance both aspects of their within-person research design by choosing a modest number of participants and a higher number of measurement occasions. Using the aforementioned scenario (up to 700 observations for a study on social relationships) as an example, these researchers would be better off choosing to recruit 50 participants with relatively diverse backgrounds for 2-week-long daily surveys, since the literature on ESM suggests that 2 weeks is a good representation of an individual's social life (Reis & Gable, 2000).

In contrast, if a researcher is concerned more about generalizing the within-person research findings to a broader population while retaining the methodological rigor in examining within-person processes, then it might be a better trade-off for them to recruit 100 participants to participate in a week-long daily diary study. In summary, in the context of ensuring adequate within-person reliability, researchers have some flexibility in balancing the choice of the number of participants versus the number of measurement occasions during research design, while considering the nature of their research questions and resources available to carry out the study.

Within-Person Reliability and Validity of Within-Person Research

The issue of reliability is closely tied to the issue of validity, including the validity of focal measures and the validity of study conclusions. The primary goal of our present research is to offer new insights on how common research design factors can be combined to optimize within-person reliability and to provide systematic guidelines for researchers to design their within-person research more efficiently. Additionally, we have two other goals essential to within-person organizational research: to demonstrate the implications of within-person reliability for the within-person validity of focal phenomena and to promote more usage of level-specific reliability in organizational within-person research.

Compute Expected/Observed Within-Person Validity

We aim to demonstrate the implications of within-person reliability for the within-person validity of focal phenomena (and thereafter the validity of within-person research findings). We address this goal by offering another separate free R shiny web application/app (see Fig. S2 in the supplementary materials) for researchers to compute the expected or observed within-person validity, given the expected within-person reliability under the choice of research design. Through this effort, we will further raise the awareness among within-person researchers that within-person reliability is a critical aspect of within-person research quality.

Compute One's Own Within-Person Reliability

We also aim to facilitate the usage of level-specific reliability in within-person organizational research. We address this goal by offering another free R shiny web app (see Fig. S3 in Appendix B of the online supplemental materials) which allows researchers to easily compute the within- and between-person reliabilities of focal scales by uploading a dataset they collected themselves. Through this R shiny web app, we hope organizational researchers will be able to compute and report within-person (and between-person) reliabilities of focal measures more frequently in future within-person research, as opposed to single-level reliabilities or averaged reliabilities across occasions.

Recommendations

Recommendation #1: a Lower CutOff Value Should Be Considered for Within-Person Reliability Under Certain Study Design Conditions

Although a single cutoff value of 0.70 is commonly used for between-person reliability reported in conventional

organizational research across all study conditions, for evaluating within-person reliability, we recommend two different cutoff values, depending on the study design conditions. Specifically, when three or more items are used for a scale in within-person research, we recommend 0.70 as the cutoff for within-person reliability, since our simulation results indicate that the estimated within-person reliability is generally above 0.70 when item factor loadings are at 0.70 (representing a typical good quality scale), across the vast majority of study design conditions (see Table 1). Furthermore, when within-person researchers use a two-item scale based on theoretical justifications and considerations of adequate construct validity (e.g., content validity) afforded by such a short scale, we recommend 0.60 as a more relaxed cutoff for within-person reliability. We caution researchers to be cognizant that having scales with within-person reliability of 0.60 or lower may potentially lead to an inadequate statistical power (e.g., less than 0.70) in detecting the focal effect size/validity. Accordingly, we also recommend researchers to compute statistical power during their research design phase, based on expected study parameters (e.g., within-person reliability of 0.60 for a 2-item scale, theoretical validity/effect size, sample size) by using our "Observed Validity" Shiny app.

Consistent with our recommendation, past research has suggested that reliability values within a range of 0.50–0.70 may be acceptable, in some contexts, such as for scales with fewer items or for studies focused on within- and between-person variations (e.g., G. Fisher et al., 2016; Fleeson, 2001). Similarly, Shrout (1998) proposed 0.61–0.80 as a range to indicate moderate reliability.

Recommendation #2: Scales with 2–4 Items Work for Organizational Within-Person Research, If Item Factor Loadings Are High

In the past, shorter scales (e.g., with four or fewer items) have been used in the majority of past organizational ESM research literature (McCormick et al., 2020). The past literature has observed a common practice in shortening longer scales established in between-person research literature based on published between-person item factor loadings; scholars have questioned its appropriateness yet offered no empirically based recommendation (Gabriel et al., 2019; Nguyen et al., 2019). Our simulation offers empirical evidence to support a clear recommendation that scales with four or fewer items can be reliable in measuring within-person variability, if the items (or those chosen from a longer scale) have moderate or high within-person factor loadings (0.70-0.90). Yet, if the items are expected to have lower within-person factor loadings (e.g., 0.50 or lower), we recommend using five or more items for each scale. Inevitably, one may ask how to obtain evidence on within-person item factor loadings in designing organizational within-person research. In line with the recommendation from the recent literature (e.g., Gabriel et al., 2019), we urge researchers to conduct their own pilot study to validate a newly developed or adapted/shortened measure to ensure its within-person reliability and construct validity before employing it in their within-person research, whenever resources allow. Additionally, we make "Recommendation #3: Organizational Within-Person Researchers Should Report Both Between- and Within-Person Item Factor Loadings" below. Notably, we neither include the 1-item condition in our simulations nor recommend it from psychometrical perspectives, because it is impossible to distinguish the true score and error score for the within-person component of the test score without further model assumptions or external information like correlating the single-item test with a multiple-item test of the same construct (e.g., Wanous & Reichers, 1996).

Recommendation #3: Organizational Within-Person Researchers Should Report Both Between- and Within-Person Item Factor Loadings

As suggested by the recent literature (e.g., Gabriel et al., 2019; Geldhof et al., 2014), multilevel confirmatory factor analysis (MCFA) is an effective approach to evaluate the level-specific reliability and validity of focal measures in studies of a multilevel nature, including organizational within-person research. In the context of shifting the current practice of estimating and reporting reliability in organizational within-person research toward level-specific reliability, we urge future organizational researchers to conduct MCFAs of their focal scales and report/publish the between- and within-person item factor loadings from such analyses, so that others can use these factor loadings to better inform within-person research design.

Recommendation #4: Organizational Within-Person Researchers Should Report Both Between- and Within-Person Reliabilities

As reviewed earlier, the organizational within-person research literature can significantly benefit from clearer understanding on level-specific reliability and a newer practice of estimating and reporting level-specific reliability. To facilitate the establishment of such a scientific practice focused level-specific measurement, we urge organizational researchers to estimate and report both between- and within-person reliabilities by using our third R Shiny app in their future work. Doing so would effectively add to the cumulating evidence on the within- and between-person psychometrical properties of many measurement tools that were initially developed and validated in between-person research. In the event that measures of certain phenomena

were reported to have consistently poor within-person reliability (relative to adequate between-person reliability), it would offer opportunities for theoretical development of such phenomena; that is, conceptualization of these phenomena might not be equivalent across the between- and within-person levels, representing a non-isomorphic process supported by the past multilevel literature (Bliese et al., 2007; Tay et al., 2014).

Recommendation #5: Use Free Online Analytical Tools to Conveniently Aid Within-Person Research Design and Level-Specific Reliability Computation

As discussed earlier, our R Shiny web apps are designed for researchers to estimate expected within- (and between-person) reliability based on the key design factors they have chosen and expected (first app), as well as the expected within-person validity based on chosen study design factors and the expected within-person reliability of a focal scale (second app). We urge organizational researchers to utilize these free R Shiny web applications to effectively plan their within-person research, by simultaneously considering the level-specific reliability and validity of their focal phenomena. In particular, if many researchers followed "Recommendation #3: Organizational Within-Person Researchers Should Report Both Between- and Within-Person Item Factor Loadings" above, it would be easier for future researchers to apply these apps during research design through retrieving level-specific item factor loadings of pertinent scales from the past literature. Furthermore, we urge organizational researchers to regularly report both within- and between-person reliabilities from their within-person research through using our third R Shiny web application, so that it would be easier for future researchers to retrieve such reliability evidence to use in their research design (e.g., applying our second web application to estimate validity). Lastly, given the many study design parameters to consider in within-person research, we encourage organizational researchers to pre-register their study design in line with the Open Science Framework. Specifically, we suggest future organizational within-person researchers follow Logg and Dorison's (2021) general guidelines in creating pre-registration files documenting their research design parameters selected based on the nature of their research questions and the evidence-based recommendations we have offered in the current study.

Limitations and Future Research Directions

Although this study conducted systematic simulations across a wide range of research designs and offers useful guidelines with respect to choosing research design to optimize measurement reliability (and validity) in within-person research, it did come with limitations. The first limitation is the estimation method for level-specific reliability. In this study, we chose the level-specific α method as α is the most commonly used estimation method in industrial and organizational psychology, and it has been widely accepted in the literature of management and business administration. However, we also acknowledge other methods in the literature that could be used to estimate level-specific reliability, and level-specific α is most suitable when the focal scale is unidimensional at both between- and within-person levels. For example, one may also adopt the level-specific ω method (Geldhof et al., 2014).

Nevertheless, the ω and α methods are mathematically equivalent or similar when scale items have equal or similar factor loadings. Future research may explore alternative reliability estimation methods for within-person organizational research. Furthermore, given that the present research is focused on issues of reliability and research design for within-person studies using unidimensional scales, we urge future research to examine how to choose appropriate within-person research design in optimizing within-person reliability for studies focusing on multidimensional constructs (for a review see Cho, 2016).

Furthermore, our study is limited, because our operationalization of within-person reliability may not represent all aspects of reliability for within-person research, particularly for research focused on growth issues. In studies focused on growth issues, we recommend researchers to estimate reliability indices that focus on change scores or the intercept/slope aspect of growth curves, such as reliability of change scores and growth curve reliability (e.g., Cranford et al., 2006; Meredith & Tisak, 1990; Rogosa et al., 1982; Willett, 1989).

Another limitation is related to the number of data conditions we examined in the study. Although our simulated 3240 data conditions have covered most of the designs in within-person research, they were not exhaustive. When a specific design is not covered in Table 1, precise information regarding the expected within-person reliability would not be available in the table. In such a case, researchers may use our R shiny web app to compute the expected reliability by entering the customized research design; they may also approximate reliability based on a similar design available in the table. Nevertheless, future research may provide simulation results based on more design conditions. For example, future simulation research may vary the levels of item factor loadings across within- and between-person levels, such as 0.90 for between-person level and 0.70 for within-person level, in an effort to examine the implications of that research design condition for optimizing within-person reliability.

Finally, a promising direction for future research is to examine the impact that the nature or type of phenomena (e.g., more or less variable) may have on within-person

reliability estimates. Using ESM research as an example, meta-analytical evidence from the organizational literature indicates that the percentage of within-person variation to total test score variation tends to be higher among phenomena like negative mood and sleep behavior (with the average percent of variation at approximately 55% and 64%, respectively), than among job situations like social support and job autonomy (at about 40% and 43%, respectively), based on their confidence intervals (McCormick et al., 2020). Thus, future research may be fruitful by examining the impact of varied ICC(1)—equal to 1 minus percent of within-person variation—on the magnitude and biases of within-person reliability estimates.

Conclusion

Addressing the challenges of lacking systematic guidelines on research design and incomplete understanding on conceptualization and computation of within-person reliability in the organizational within-person research literature, we conducted a simulation study, offered guidelines for research design that optimizes within-person reliability, provided three R shiny web apps, and made practical recommendations for future within-person research. Findings and tools from this research enable organizational researchers to optimize within-person reliability (and validity) through designing within-person research with easy-to-use R shiny web apps, and to better understand and report level-specific reliabilities in within-person research. Moving forward, we hope this research will facilitate the shift of the common practice of primarily reporting single-level reliability or averaged reliability across occasions in the organizational within-person research literature, toward the practice of reporting level-specific reliabilities that correspond to the level of analysis researchers' focal research questions lie in.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10869-022-09803-5.

Acknowledgements We thank Dr. Jason Newsom for the feedback on the earlier versions of this manuscript. We also thank Stefanie Fox, M.S., for her proofreading of the most recent version of this manuscript.

Funding This research was supported by the grant T03OH008435 awarded to Portland State University, funded by the Centers for Disease Control and Prevention, National Institute for Occupational Safety and Health. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of NIOSH, CDC, or HHS. This research is also supported by the National Science Foundation under Grant awarded to Wei Wang (No. 16406229). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do notnecessarily reflect the views of the National Science Foundation.

References

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Waveland Press.

Bakker, A. B., Sanz-Vergel, A. I., Rodríguez-Muñoz, A., & Oerlemans, W. G. M. (2015). The state version of the recovery experience questionnaire: A multilevel confirmatory factor analysis. *European Journal of Work and Organizational Psychology, 24*(3), 350–359. https://doi.org/10.1080/1359432X.2014.903242

Beal, D. J. (2015). ESM 2.0: State of the art and future potential of experience sampling methods in organizational research. *Annual Review of Organizational Psychology and Organizational Behavior, 2*(1), 383–407. https://doi.org/10.1146/annurev-orgpsych-032414-111335

Beal, D. J., & Weiss, H. M. (2003). Methods of ecological momentary assessment in organizational research. *Organizational Research Methods, 6*(4), 440–464. https://doi.org/10.1177/1094428103257361

Bliese, P. D., Chan, D., & Ployhart, R. E. (2007). Multilevel methods: Future directions in measurement, longitudinal analyses, and nonnormal outcomes. *Organizational Research Methods, 10*(4), 551–563. https://doi.org/10.1177/1094428107301102

Brose, A., Voelkle, M. C., Lövdén, M., Lindenberger, U., & Schmiedek, F. (2015). Differences in the between–person and within–person structures of affect are a matter of degree. *European Journal of Personality, 29*(1), 55–71. https://doi.org/10.1002/per.1961

Cho, E. (2016). Making reliability reliable: A systematic approach to reliability coefficients. *Organizational Research Methods, 19*(4), 651–682. https://doi.org/10.1177/1094428116656239

Cortina, J. M., Sheng, Z., Keener, S. K., Keeler, K. R., Grubb, L. K., Schmitt, N., Tonidandel, S., Summerville, K. M., Heggestad, E. D., & Banks, G. C. (2020). From alpha to omega and beyond! A look at the past, present, and (possible) future of psychometric soundness in the Journal of Applied Psychology. *Journal of Applied Psychology, 105*(12), 1351–1381.

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

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. https://doi-org.proxy.lib.pdx.edu/https://doi.org/10.1007/BF02310555

Curran, P. J., & Bauer, D. J. (2011). The disaggregation of within-person and between-person effects in longitudinal models of change. *Annual Review of Psychology, 62*(1), 583–619. https://doi.org/10.1146/annurev.psych.093008.100356

Dalal, R. S., Bhave, D. P., & Fiset, J. (2014). Within-person variability in job performance. *Journal of Management, 40*(5), 1396–1436. https://doi.org/10.1177/0149206314532691

DiStefano, C. (2002). The impact of categorization with confirmatory factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal, 9*(3), 327–346. https://doi.org/10.1207/S15328007SEM0903_2

Ellis, A. M., Bauer, T. N., Mansfield, L. R., Erdogan, B., Truxillo, D. M., & Simon, L. S. (2015). Navigating uncharted waters: Newcomer socialization through the lens of stress theory. *Journal of Management, 41*(1), 203–235. https://doi.org/10.1177/0149206314557525

Ferrer, E., & McArdle, J. J. (2010). Longitudinal modeling of developmental changes in psychological research. *Current Directions in Psychological Science, 19*(3), 149–154. https://doi.org/10.1177/0963721410370300

Fisher, C. D., & To, M. L. (2012). Using experience sampling methodology in organizational behavior. *Journal of Organizational Behavior, 33*(7), 865–877. https://doi.org/10.1002/job.1803

Fisher, G. G., Matthews, R. A., & Gibbons, A. M. (2016). Developing and investigating the use of single-item measures in organizational research. *Journal of Occupational Health Psychology, 21*(1), 3–23. https://doi.org/10.1037/a0039139

Fleeson, W. (2001). Toward a structure- and process integrated view of personality: Traits as density distributions of states. *Journal of Personality and Social Psychology, 80*(6), 1011–1027. https://doi.org/10.1037/0022-3514.80.6.1011

Gabriel, A. S., Podsakoff, N. P., Beal, D. J., Scott, B. A., Sonnentag, S., Trougakos, J. P., & Butts, M. M. (2019). Experience sampling methods: A discussion of critical trends and considerations for scholarly advancement. *Organizational Research Methods, 22*(4), 969–1006. https://doi.org/10.1177/1094428118802626

Geldhof, G. J., Preacher, K. J., & Zyphur, M. J. (2014). Reliability estimation in a multi-level confirmatory factor analysis framework. *Psychological Methods, 19*(1), 72–91. https://doi.org/10.1037/a0032138

Huang, P.-H., & Weng, L.-J. (2012). Estimating the reliability of aggregated and within-person centered scores in ecological momentary assessment. *Multivariate Behavioral Research, 47*(3), 421–441. https://doi.org/10.1080/00273171.2012.673924

Hox, J. (2002). *Multi-level analysis: Techniques and applications*. Lawrence Erlbaum Associates.

Logg, J. M., & Dorison, C. A. (2021). Pre-registration: Weighing costs and benefits for researchers. *Organizational Behavior and Human Decision Processes, 167*, 18–27. https://doi.org/10.1016/j.obhdp.2021.05.006

McCormick, B. W., Reeves, C. J., Downes, P. E., Li, N., & Ilies, R. (2020). Scientific contributions of within-person research in management: Making the juice worth the squeeze. *Journal of Management, 46*(2), 321–350. https://doi.org/10.1177/0149206318788435

McDonald, R. P. (1999). *Test theory: A unified treatment*. Lawrence Erlbaum Associates Publishers.

Meredith, W., & Tisak, J. (1990). Latent curve analysis. *Psychometrika, 55 (1)*, 107–122. https://doi-org.proxy.lib.pdx.edu/https://doi.org/10.1007/BF02294746

Neale, M. C., Hunter, M. D., Pritikin, J. N., Zahery, M., Brick, T. R., Kirkpatrick, R. M., et al. (2016). OpenMx 2.0: Extended structural equation and statistical modeling. *Psychometrika, 81*, 535–549. https://doi.org/10.1007/s11336-014-9435-8

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

Nguyen, A. N., Yang, L.-Q., Wang, W., & Huang, P.-H. (November 2019). *Is your diary "Reliable"?: A review of current measurement practices in ESM research.* Paper presented at the Biennnial conference of Conference of Work, Stress, and Health, Philadelphia, PA.

Ohly, S., Sonnentag, S., Niessen, C., & Zapf, D. (2010). Diary studies in organizational research. *Journal of Personnel Psychology, 9*(2), 79–93. https://doi.org/10.1027/1866-5888/a000009

Ployhart, R. E., & Vandenberg, R. J. (2010). Longitudinal research: The theory, design, and analysis of change. *Journal of Management, 36*(1), 94–120. https://doi.org/10.1177/0149206309352110

Raudenbush, S. W., Rowan, B., & Kang, S. J. (1991). A multilevel, multivariate model for studying school climate with estimation via the EM algorithm and application to US high-school data. *Journal of Educational Statistics, 16*(4), 295–330. https://doi.org/10.2307/1165105

R Core Team (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/. Accessed 3/8/2020

Reis, H. T., & Gable, S. L. (2000). Event-sampling and other methods for studying everyday experience. In H. T. [Ed Reis & C. M. [Ed Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 190–222, Chapter xii, 558 Pages). Cambridge University Press (New York, NY, US).

Rhemtulla, M., Brosseau-Liard, P. E., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods, 17*(3), 354–373. https://doi.org/10.1037/a0029315

Robinson, M. D., & Clore, G. L. (2002). Belief and feeling: Evidence for an accessibility model of emotional self-report. *Psychological Bulletin, 128*(6), 934.

Rogosa, D., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin, 92*(3), 726–748. http://dx.doi.org.proxy.lib.pdx.edu/https://doi.org/10.1037/0033-2909.92.3.726

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Miufflin Company.

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

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology, 15*, 72–101. https://doi.org/10.2307/1422689

Tay, L., Woo, S. E., & Vermunt, J. K. (2014). A conceptual and methodological framework for psychometric isomorphism: Validation of multilevel construct measures. *Organizational Research Methods, 17*(1), 77–106. https://doi.org/10.1177/1094428113517008

Wanous, J. P., & Reichers, A. E. (1996). Estimating the reliability of a single-item measure. *Psychological Reports, 78*(2), 631–634. https://doi.org/10.2466/pr0.1996.78.2.631

Wänström, L. (2009). Sample sizes for two-group second-order latent growth curve models. *Multivariate Behavioral Research, 44*(5), 588–619. https://doi.org/10.1080/00273170903202589

Willett, J. B. (1989). Some results on reliability for the longitudinal measurement of change: Implications for the design of studies of individual growth. *Educational and Psychological Measurement, 49 (3)*, 587- 602. https://doi-org.proxy.lib.pdx.edu/https://doi.org/10.1177/001316448904900309

Zapf, D., Dormann, C., & Frese, M. (1996). Longitudinal studies in organizational stress research: A review of the literature with reference to methodological issues. *Journal of Occupational Health Psychology, 1*(2), 145–169. https://doi.org/10.1037/1076-8998.1.2.145