Published in final edited form as:
Monogr Soc Res Child Dev. 2017 June ; 82(2): 67–83. doi:10.1111/mono.12299.

Design-Based Approaches for Improving Measurement in Developmental Science

Jonathan Rush and **Scott M. Hofer**
Department of Psychology, University of Victoria

Abstract

The study of change and variation within individuals, and the relative comparison of changes across individuals, relies on the assumption that observed measurements reflect true change in the construct being measured. Measurement properties that change over time, contexts, or people pose a fundamental threat to validity and lead to ambiguous conclusions about change and variation. We highlight such measurement issues from a within-person perspective and discuss the merits of measurement-intensive research designs for improving precision of both within-person and between-person analysis. In general, intensive measurement designs, potentially embedded within long-term longitudinal studies, provide developmental researchers an opportunity to more optimally capture within-person change and variation as well as provide a basis to understand changes in dynamic processes and determinants of these changes over time.

The study of change and variation within individuals, and the relative comparison of changes across individuals, relies on the assumption that observed measurements reflect true change in the construct being measured. Measurement properties that change over time within a person, across contexts, or individuals pose a fundamental threat to the validity of a study and ambiguates conclusions about change and variation. That observed measurements are a reflection of true scores is critical to the study of individual differences and in understanding change and variation within individuals over time.

A variety of research designs, including decisions about the number, frequency, and types of measurements, are used to understand developmental and aging-related processes. Various statistical models can be applied to answer specific questions regarding population average patterns of change, individual differences in level and rate of change, and multivariate dynamics of within-person variation. Each observed score carries many sources of variation that influence our models. Design features play an important role in the ability to disentangle the different sources of variation. Figure 1 shows how an individual's observed scores over time can be broken down into a representation of population average level and slope, individual deviation from level and slope, and systematic within-person deviations from the individual slope as distinct from random error. Perhaps the most important feature of longitudinal studies is the opportunity to distinguish between-person age differences from within-person age variation and change and how individuals differ in terms of within-person

Correspondence concerning this article should be addressed to Jonathan Rush, Department of Psychology, University of Victoria, Victoria, BC V8W 2Y2. jrush@uvic.ca.

age change. Notably, measurement error is often confounded with reliable within-person variation in cross-sectional and typical longitudinal designs, with such variation within and across days often the focus of intensive measurement (e.g., daily diary) designs.

Important longitudinal design features (e.g., Lerner, Schwartz, & Phelps, 2009) include whether the sample at baseline is age-heterogeneous or homogeneous (which may complicate interpretation given confounds with birth cohort and mortality selection), and the number and spacing of measurement occasions as this will affect the types of within-person models of change that can be reliably estimated (see Rast & Hofer, 2014). Design features can be combined in a number of ways to answer research questions that vary in scope from population change across birth cohorts to daily dynamics of within-person processes. Fundamental to developmental research is measurement with the question of how best to measure processes that vary and change within person?

Change in physiological, cognitive, and social functioning early in the lifespan can be relatively rapid. Measurement in different developmental periods often require the use of different measures, with different emphases given changing contexts related to home, school, work, and retirement. Kagan (1980) described the need for identifying a construct using different measures, referred to as phenotypic discontinuity, with the construct retaining its meaning across developmental periods (i.e., heterotypic continuity). There have been a number of recent advances and applications for bridging measurements across developmental periods to maintain continuity in the construct, permitting the analysis of individual change (e.g., McArdle, Grimm, Hamagami, Bowles, & Meredith, 2009).

In this chapter, we expand on these previous developments with an emphasis on measurement issues from a within-person perspective, and highlight benefits for measurement and understanding developmental dynamics from measurement-intensive research designs (e.g., Hoffman, 2007; Nesselroade, 1991; Rast, MacDonald, & Hofer, 2012; Salthouse & Nesselroade, 2010; Walls, Barta, Stawski, Collyer, & Hofer, 2012). Such designs better enable developmental researchers the opportunity to capture within-person change and variation as well as provide potentially better foundations to understand stability and change dynamics in developmental processes. While we highlight the importance of considering design features for improving measurements in this paper, we want to also point to the value of evaluating factorial invariance of measurements and subsequent measurement development for repeated measurement designs (e.g., Bontempo, Grouzet, & Hofer, 2012; Bontempo & Hofer, 2007; Ferrer, Balluerka, & Widaman, 2008; Meredith & Horn, 2001).

Sampling time: Issues for the measurement of change and variation

The majority of measurement development has been in terms of between-person differences, where the level of a construct is captured relative to other individuals. These between-person differences are predominantly captured at a single occasion in time. Conclusions about individual differences and long-term change are dependent upon accurately measuring an individual’s characteristic level at a specific time period. Both widely spaced longitudinal designs and single occasion, cross-sectional, designs are susceptible to biases (e.g., recall error) that threaten the accurate measurement of true level during a given period of time.

This subsequently obscures the accuracy of the measurement of long-term change and between-person differences.

Issues with single occasion measurements

Cross-sectional and widely spaced longitudinal measures fail to account for the potential variability around trait levels. When measures vary both within-person across time as well as between people, measuring only once forces all systematic within-person variations to be grouped together and treated as random measurement error. As a result, the cross-sectional measure carries both between person information (i.e., characteristic individual level) and within-person information (i.e., deviations from individual level) with no possibility of disentangled the two sources of variation with only a single measurement (e.g., Curran & Bauer, 2011; Hoffman & Stawski, 2009). For example, an individual could be higher than others on a measure of well-being because they are a generally a happier person, or their well-being level could be affected by them having a particularly good day, which elevates their score higher than their typical level (Schwarz & Strack, 1999). Assuming that a construct is stable can be problematic when the construct does indeed systematically vary over time and can lead to conclusions about individual differences that are confounded with within-person variance (e.g., Rush & Hofer, 2014).

Many constructs in developmental research are captured via recall of behaviors, attitudes, or experiences within a delimited period of time (e.g., well-being, victimization, substance use). These measures typically rely on self-report recall, or the recall of other informants (e.g., friends, family, teachers; Allen, Chango, Szwedo, Schad, & Marston, 2012; Jordan & Graham, 2012; Ladd & Kochenderfer-Ladd, 2002). The retrospective time-range of cross-sectional measures can vary widely from the previous months or years, to asking about global levels. When measures are derived solely from a single occasion there are a number of biases that distort the true level of the construct. Global measures are susceptible to retrospection bias, particularly when the assessment period is farther removed from the period of recall (Schwarz, Kahneman, & Xu, 2009). A potentially more problematic issue with global measures are social desirability biases, which include 1) impression management, where individuals purposefully attempt to present themselves more favorably; and 2) deceptive self-enhancement, where individuals unintentionally respond according to their self-image, rather than actual behaviors/experiences (Barta, Tennen, & Litt, 2012). An inability to accurately recall the events of the distant past (e.g., months/year) often results in the responses being based on a top-down approach of relying on a global self-perception of themselves and how someone who fits that self-perception would act (Schwarz, 2012). For example, parents who rated the enjoyment they experience while spending time with their children via a global self-report consistently rank it as among the most enjoyable things they do (Juster, 1985). However, when rating their enjoyment with their children on a particular day, through an end-of-day reconstruction, they rated it as among the least enjoyable events of the day (Kahneman, Krueger, Schkade, Schwarz, & Stone, 2004). Reporting globally that one does not enjoy time with their children would likely be in stark contrast to their self-perception as a loving parent, however reporting that on this one day they did not enjoy time with their children does not preclude them as quality parents. Aggregating across multiple daily reports would therefore reflect the parents' actual enjoyment during this time period and individual differences among parents would be based on actual differences in enjoyment rather than differences in global self-perception. Other undesirable behaviors have found similar patterns. In a study of unsafe sexual practices, it was found that participants underreported the number of unsafe sexual behaviors in general cross-sectional measures compared to daily reports (McAuliffe, DiFranceisco, & Reed, 2007).

Contrary to undesirable behaviors, global measures of life satisfaction are often negatively skewed (Diener, 2000), with most people considering themselves to be generally quite satisfied with their life. However, these responses are more likely based on their perception of themselves as a happy person, rather than on actual accounts of how satisfied they are day in and day out. Thus, aggregating over many closely spaced assessments may provide an account of an individual’s true level of a construct that is less dependent on retrospection and social desirability biases.

Sampling many points in time also addresses a number of the issues that plague cross-sectional measures. Intensive measurement designs, with frequent closely-spaced assessments (e.g., daily diary, ecological momentary assessments), enable within-person variation to be disaggregated from between-person differences. Furthermore, the lag-time between the experiencing and the reporting can be reduced to the point where retrospection bias is nearly eliminated and reports are based more on a bottom-up report of actual events rather than a top-down representation of perceived self-image. However, sampling time more frequently produces additional challenges. As the time-scale varies, the process which is being measured may also change in terms of quantitative or qualitative shifts (Birren & Schroots, 1996; Martin & Hofer, 2004). It cannot be assumed that constructs are equivalent across occasions or levels of analysis. Measures designed to capture stable between-person differences may not possess suitable sensitivity to accurately capture small increments in within-person variation.

Reactivity in studies of within-person change

Individuals often perform better on measures of performance and functional assessments with repeated testing, with the greatest gains following the first assessment. This phenomenon is known as retest, practice, exposure, learning, or reactivity, and has been reported in a number of longitudinal studies of aging (Ferrer, Salthouse, Stewart, & Schwartz, 2004; Hultsch, Hertzog, Dixon, & Small, 1998; Rabbitt, Diggle, Smith, Holland, & Mc Innes, 2001; Schaie, 1996). A further complication is that individuals are likely to differ from one another in terms of amount of gain due to retesting or learning in systematic ways related to age or developmental stage, level of ability or age and test-specific learning, such as learning content and strategies. Additionally, such gains may be due to underperformance at the first occasion, known as warm-up effects and related to initial anxiety. Sliwinski, Hoffman, and Hofer (2010) addressed retest effects in a longitudinal design based on measurement bursts, a set of closely spaced retest intervals to model practice effects and longer, for example, six month intervals to model age-related changes (Nesselroade, 1991). The pairing of multi-burst designs and informative measurement models allowed the separation of short-term (e.g., retest gains) from longtime developmental change which operate across two different time scales.

Optimizing measurement for between-person differences and within-person change and variation

Measurements that are developed for between-person differences may not be optimal for within-person research. Intensive measurement designs are now commonly utilized to account for within-person variation (e.g., Bolger & Laurenceau, 2013; Rush & Grouzet, 2012; Sliwinski, Smyth, Hofer, & Stawski, 2006), where either new measures are created to be used in these designs, or more commonly, are adapted from measures designed for between-person cross-sectional research. Though these new measures may be a better solution to account for within-person variation, little effort has been devoted to evaluating the properties of these measures to adequately account for within-person variation and between-person difference. The structure, reliability, and validity of measures used in intensive measurement designs need to be evaluated with the same rigor that is expected of cross-sectional measures. Multilevel reliability estimates and factor analysis allow for these measurement properties to be readily examined. Here we illustrate a number of measurement challenges and innovations in measurement development.

Alternative measurement models based on intensive measurement designs

Sample data from a daily diary study of 147 participants (M_{age} = 20) assessed over 14 consecutive days on measures of subjective well-being (i.e., life satisfaction, positive and negative affect) will be used to demonstrate how intensive measurement designs, which emphasize within-person measurement, can be utilized to generate a between-person measure that may be preferred over cross-sectional measures. Participants initially completed global cross-sectional measures of the 5-item Satisfaction with Life Scale (SWLS; Pavot & Diener, 1993) and the 20-item Positive and Negative Affect Schedule (PANAS; Watson, Clark, & Tellegen, 1988). The same scales were adapted slightly to be used for daily assessments and were administered over the next 14 days to capture daily levels of life satisfaction, positive affect (PA) and negative affect (NA).

The daily measures of well-being possessed considerable amounts of within-person variability. Intraclass correlation coefficients (ICC), which indicate the proportion of variability that is between-person, were around 0.5 for all three measures (see Table 1). Thus, half of the total variability was due to within-person variation, such that individuals deviated as much from their own mean level of well-being as their own mean deviated from the grand mean (see Figure 2). Relying solely on a cross-sectional measure would ignore all within-person variability and assume that these constructs were stable across time. Furthermore, all within-person variability would be confounded with between-person variability, impacting the conclusions drawn at a between-person level. Disaggregating within- and between-person variability allows the effects at both levels to be more appropriately modeled and accounted for.

As outlined above, the cross-sectional measures are more susceptible to retrospection and social desirability biases than the daily measures. This would be expected especially for the global measure of life satisfaction, where individuals tend to view themselves in an overly positive light on global measures. Comparing the cross-sectional measure of life satisfaction to the aggregated daily measure reveals that the two measures are only moderately correlated (r = 0.58, Table 1). Additionally, individuals rated their general level of life satisfaction higher than their average daily life satisfaction (t(146) = 11.71, p < .001). Figure 2 shows the comparison of the cross-sectional and daily measures of ten participants. Each participant had higher levels on the cross-sectional life satisfaction measure than the aggregated daily measure. More than 85% of the sample overestimated their global life satisfaction relative to their daily mean, providing support for the upward bias of cross-sectional measures. When reporting on typical level of life satisfaction, participants were likely using a top-down approach where they perceived themselves as more satisfied to a greater extent than was actually the case if asked to assess day-by-day. It is important to note that global perceptions of life satisfaction may be of substantive interest, however, it differs from actual experiences of life satisfaction as they occur on a daily basis.

In addition to disaggregating effects and reducing bias, repeated measurements improve the reliability of between-person estimates (Sliwinski, 2008). Calculation of reliability is often neglected in within-person measurements, or a single-level alpha is reported that does not account for the hierarchical data structure. Recent work on multilevel reliability provides practical alternatives to single-level reliability estimates (e.g., Cranford et al., 2006; Geldhof, Preacher, & Zyphur, 2014; Shrout & Lane, 2012). We suggest utilizing a multilevel omega (ω) reliability estimate derived from multilevel confirmatory factor analysis (CFA) that was outlined by Geldhof and colleagues (2014; see also McDonald, 1999; Shrout & Lane, 2012). The ω reliability utilizes the factor loadings to derive the ratio of true score variance to total variance:

$$\omega = \frac{(\sum\lambda_i)^2}{(\sum\lambda_i)^2 + \sum\Psi_i^2}, \quad (1)$$

where λ_i and Ψ_i^2 are the factor loading and residual variance for item i, respectively. This equation can be applied to both levels of the multilevel factor model to derive both within- and between-person reliability. Table 1 displays the multilevel reliability estimates for the daily measures and the single-level (between-person) reliability estimate for the cross-sectional measure. It is clear that between-person reliability is improved considerably by employing an intensive measurement design over a cross-sectional (single-occasion) design.

Multilevel factor analysis: Evaluation of structure at between- and within-person levels

Intensive repeated measure designs allow for both a within-person and between-person factor structure to be examined simultaneously through the use of multilevel factor analysis. In multilevel factor analysis, the within-person factor structure reflects common covariance in the indicators at each specific occasion, pooled across occasions and individuals. The between-person factor structure reflects common covariance in individual levels of indicators aggregated across time (i.e., person-mean level). Similar to conventional factor analysis, the quality of the model can be assessed with a variety of fit indices, which include both global fit indices (e.g., comparative fit index; CFI, root mean square error of approximation; RMSEA) and level specific fit indices (e.g., standardized root mean square residual; SRMR within/between).

Multilevel confirmatory factor analyses were employed in Mplus v.7 (Muthén & Muthén, 2012) to evaluate the optimal within-person and between-person factor structure of the SWLS (see Appendix for sample Mplus code). Multilevel CFA allows for the within-person variance to be disaggregated from the between-person variance, while still attenuating for measurement error at both levels. The multilevel measurement model can be expressed by the following equation (Muthén, 1991; Preacher, Zyphur, & Zhang, 2010):

$$Y_{ij} = v + \lambda_{\mathrm{w}}\eta_{ij} + \varepsilon_{ij} + \lambda_{\mathrm{b}}\eta_i + \varepsilon_i, \quad (2)$$

where Y_{ij} is a p-dimensional vector of observed variables for individual i on occasion j, where p is the number of observed indicators; v is a p-dimensional vector of intercepts; λ_{w} is a p X q within-person factor loadings matrix, where q is the number of latent variables; λ_{b} is a p X q between-person factor loadings matrix; η_{ij} and η_i are q-dimensional vectors of within-person and between-person latent variables, respectively; and ε_{ij} and ε_i are p-dimensional vectors of within-person and between-person specific 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 in sampling error (see Lüdtke et al., 2008 for further details), such that the between-person indicators are represented as latent means.

In the case of the SWLS, a single factor at both the within- and between-person level fit the data extremely well, with all five items loading onto this single factor (see Table 2; Figure 3). These five items reliably covary within a person across occasions (i.e., on occasions when one item deviates from typical levels, the other four items also deviate in the same direction) and between people (i.e., individuals who are higher on one item relative to others are also higher on the other items). It will not always be the case that the within-person structure is identical to the between-person structure. For example, Rush and Hofer (2014) found that the PANAS was best represented by two inversely related factors (PA and NA) at the within-person level, but independent PA and NA factors at the between-person level.

Discrepancies in model fit across levels demand additional decisions. Situations where the measure fits well at the within-person level, but not at the between-person level require decisions to be made about the utility of the measure and the appropriateness for assessing between-person differences. For example, a 7-item daily measure of competence that was adapted from the cross-sectional measure of psychological well-being (Ryff, 1989) was also included in the above sample data. Results from a multilevel CFA revealed good model fit at the within-person level (SRMR = .04; with all items loading onto a single factor), but not at the between-person (SRMR = .21; see Table 2). Specifically, items 2 and 5 did not load well onto the between-person factor of competence (loadings = .16 and −.05, respectively), but did load onto the within-person factor (loading = .34 and .33, respectively). In this situation, the same scale may not capture both within-person variation and between-person differences with the same precision and accuracy and decisions will depend on the intended use of the scale (i.e., to capture characteristic individual levels or intraindividual variations). The goal may be to develop scales that adequately measure both within-person and between-person elements. However, depending on the constructs of interest, this may not always be feasible as they may manifest themselves differently across levels of analysis. Multilevel factor analysis provides a technique to evaluate the multilevel structure of measures to ensure they adequately reflect the constructs they are intended. This technique further enables the reliability and validity to be examined across levels. A greater emphasis on the measurement properties from multilevel data will go a long way in enhancing the quality of measures, and as a result the conclusions drawn from them.

Trait as maximal performance (e.g., cognition)

Aggregation of frequent repeated measurements may not always be the optimal approach to capture an individual's characteristic level. For measures of cognitive or physical ability, repeated assessments will often lead to improvements in performance, related to learning content, strategies, or due to repeated practice or training gains. This modification of the system itself that result in improved or altered performance on subsequent assessments is an important consideration in longitudinal developmental research. For example, in the case of executive functioning, as a consequence of repeated assessment and learning the problem-solving strategy, over time the test may measure a different construct than it did originally, particularly when the test was designed to measure novel reasoning ability. In many cases, however, such retest effects can be managed through the use of a measurement burst design, permitting the estimation of maximal performance and change in maximal performance over time (Sliwinski et al., 2010; see also Hoffman, Hofer, & Sliwinski, 2011; and Thorvaldsson, Hofer, Berg, & Johansson, 2006 for critique of other approaches for correcting for retest effects).

Optimizing assessments: Planned missingness designs, adaptive tests, and web-based assessment

To be able to appropriately model within-person and between-person constructs a sufficient number of items and measurement occasions are often required. However, large item scales assessed frequently over many occasions can become burdensome on participants. Among the most frequently criticized elements of daily diary studies that participants complain about is the repetitiveness of the questionnaires. An approach to minimize the repetitiveness, while still ensuring enough items to appropriately capture and evaluate the constructs through multilevel CFAs, could utilize a planned missingness design (e.g., Graham, Hofer, & MacKinnon, 1996; McArdle, 1994; Silvia, Kwapil, Walsh, & Myin-Germeys, 2013). Planned missingness designs present a few anchor items at each occasion, while presenting a remaining set of items intermittently over the course of the assessment period. In this way, participants are presented with some different items each occasion, which enhances their interest and motivation to continue. These designs also reduce the total number of items that are required to be asked at each occasion, allowing for a greater number of overall constructs to be examined, without the risk of burning out the participants. Furthermore, the rotating items can still be included within a multilevel SEM framework to evaluate the fit, reliability, and validity of the constructs.

Adaptive testing can serve a similar purpose for cognitive and performance testing that the planned missingness design serves for survey measures. Adaptive tests that adjust to the participants ability level, provides a way for the participant to reach their maximal performance level in fewer trials or items than standard tests (Gorin & Embretson, 2012). The reduced time of each test reduces the burden and fatigue on the participants and allows for more assessments and/or greater depth or range of constructs (e.g., cognition, psychopathology, affect, behavioral dispositions) to be evaluated at each assessment (Kim-O & Embretson, 2010).

Developing novel approaches to reduce the time and burden on the participants without sacrificing the necessity for quality of measures will be paramount in the development of measures that adequately capture both within-person change and variation and between-person differences in these dynamic processes. The growing availability of web-based and mobile assessment tools will enable data to be collected more readily in remote locations (i.e., the participant's home; Intille, 2007). As intensive measurement designs become less burdensome and costly for participants and researchers, the benefits will clearly outweigh the costs. However, in order to ensure that these designs are utilized to their potential, a greater focus must be placed on developing measures that adequately and appropriately capture both within-person variation and between-person differences.

Summary and Closing Statement

An increasing number of research studies demonstrate the remarkable intraindividual variability that is present in cognitive, behavioral, and physical functioning across different time scales. The results from these studies, and that of several measurement studies using short-term repeated assessments, provides evidence that single assessments do not usually provide optimal estimates of an average or typical value of a person's functioning, and adversely affect results from between-person and within-person analysis. We have highlighted the strengths of measuring individuals more often in order to better sample the contextual and intrinsic variation of individual functioning. As we make use of existing measures and adapt them for repeated-measures designs, we are finding that some of these measures may not be optimal for such purposes. There is a need for developing measures that are sufficiently sensitive for detecting within-person variation and change. Such developments have the potential to improve measures and models of between-person differences and to understand whether such measurements can be homogeneously applied to all individuals within a population. We demonstrated how measuring individuals more often can improve the discrimination of between-person differences by disentangling true between-person differences in typical level from contextual and/or intrinsic intraindividual variation. Such designs encourage us to make our assessments more efficient and less burdensome, such as through planned missingness designs, adaptive tests, and web-based assessment.

Acknowledgments

Research reported in this publication was supported by National Institute on Aging of the National Institutes of Health under grant number P01AG043362 and R01AG050720. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Allen JP, Chango J, Szwedo D, Schad M, Marston E. Predictors of susceptibility to peer influence regarding substance use in adolescence. Child Development. 2012; 83:337–350. [PubMed: 22188526]

Barta, WD., Tennen, H., Litt, MD. Measurement reactivity in diary research. In: Mehl, MR., Conner, TS., editors. Handbook of research methods for studying daily life. New York: Guilford Press; 2012. p. 108-123.

Birren, JE., Schroots, JJF. History, concepts, and theory in the psychology of aging. In: Birren, JE.Schaie, KW.Abeles, RP.Gatz, M., Salthouse, TA., editors. Handbook of the psychology of aging. 4. San Diego, CA: Academic Press; 1996. p. 3-23.

Bolger, N., Laurenceau, JP. Intensive longitudinal methods: An introduction to diary and experience sampling research. New York: Guilford Press; 2013.

Bontempo, DE., Grouzet, FME., Hofer, SM. Measurement issues in the analysis of within-person change. In: Newsom, JT.Jones, RN., Hofer, SM., editors. Longitudinal data analysis: A practical guide for researchers in aging, health, and social sciences. New York: Routledge; 2012. p. 97-142.

Bontempo, DE., Hofer, SM. Assessing factorial invariance in cross-sectional and longitudinal studies. In: Ong, AD., van Dulmen, MHM., editors. Oxford handbook of methods in positive psychology. New York: Oxford University Press; 2007. p. 153-175.

Cranford JA, Shrout PE, Iida M, Rafaeli E, Yip T, Bolger N. A procedure for evaluating sensitivity to within-person change: Can mood measures in diary studies detect change reliably? Personality and Social Psychology Bulletin. 2006; 32:917–929. [PubMed: 16738025]

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

Diener E. Subjective well-being: The science of happiness and a proposal for a national index. American Psychologist. 2000; 55:34–43. [PubMed: 11392863]

Ferrer E, Balluerka N, Widaman KF. Factorial invariance and the specification of second-order latent growth models. Methodology. 2008; 4:22–36. [PubMed: 20046801]

Ferrer E, Salthouse TA, Stewart WF, Schwartz BS. Modeling age and retest processes in longitudinal studies of cognitive abilities. Psychology and Aging. 2004; 19:243–259. [PubMed: 15222818]

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

Gorin, JS., Embretson, SE. Using cognitive psychology to generate items and predict item characteristics. In: Gierl, MJ., Haladyna, TM., editors. Automatic item generation: Theory and practice. New York: Taylor & Francis; 2012. p. 136-156.

Graham JW, Hofer SM, MacKinnon DP. Maximizing the usefulness of data obtained with planned missing value patterns: An application of maximum likelihood procedures. Multivariate Behavioral Research. 1996; 31:197–218. [PubMed: 26801456]

Hoffman L. Multilevel models for examining individual differences in within-person variation and covariation over time. Multivariate Behavioral Research. 2007; 42:609–629.

Hoffman L, Hofer SM, Sliwinski MJ. On the confounds among retest gains and age-cohort differences in the estimation of within-person change in longitudinal studies: A simulation study. Psychology and Aging. 2011; 26:778–791. [PubMed: 21639642]

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

Hultsch, DF., Hertzog, C., Dixon, RA., Small, BJ. Memory change in the aged. Cambridge, MA: Cambridge University Press; 1998.

Intille, SS. Technological innovations enabling automatic, context-sensitive ecological momentary assessment. In: Stone, AA.Shiffman, S.Atienza, AA., Nebeling, L., editors. The science of real-time data capture: Self-report in health research. New York: Oxford University Press; 2007. p. 308-337.

Jordan LP, Graham E. Resilience and well-being among children of migrant parents in South-East Asia. Child Development. 2012; 83:1672–1688. [PubMed: 22966930]

Juster, FT. Preferences for work and leisure. In: Juster, FT., Stafford, FP., editors. Time, goods, and well-being. Ann Arbor, MI: Institute for Social Research; 1985. p. 335-351.

Kagan J. Four questions in psychological development. International Journal of Behavioral Development. 1980; 3:231–241.

Kahneman D, Krueger AB, Schkade DA, Schwarz N, Stone AA. A survey method for characterizing daily life experience: The Day Reconstruction Method. Science. 2004; 306:1776–1780. [PubMed: 15576620]

Kim-O, M-A., Embretson, SE. Item response theory and its application to measurement in behavioral medicine. In: Steptoe, A., editor. Handbook of behavioral medicine. New York: Springer; 2010. p. 113-123.

Ladd GW, Kochenderfer-Ladd B. Identifying victims of peer aggression from early to middle childhood: Analysis of cross-informant data for concordance, estimation of relational adjustment, prevalence of victimization, and characteristics of identified victims. Psychological Assessment. 2002; 14:74–96. [PubMed: 11911051]

Lerner RM, Schwartz SJ, Phelps E. Problematics of time and timing in the longitudinal study of human development: Theoretical and methodological issues. Human Development. 2009; 52:44–68. [PubMed: 19554215]

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

Martin M, Hofer SM. Intraindividual variability, change, and aging: Conceptual and analytical issues. Gerontology. 2004; 50:7–11. [PubMed: 14654720]

McArdle JJ. Structural factor analysis experiments with incomplete data. Multivariate Behavioral Research. 1994; 29:409–454. [PubMed: 26745236]

McArdle JJ, Grimm KJ, Hamagami F, Bowles RP, Meredith W. Modeling life-span growth curves of cognition using longitudinal data with multiple samples and changing scales of measurement. Psychological Methods. 2009; 14:126–149. [PubMed: 19485625]

McAuliffe TL, DiFranceisco W, Reed BR. Effects of question format and collection mode on the accuracy of retrospective surveys of health risk behavior: A comparison with daily sexual activity diaries. Health Psychology. 2007; 26:60–67. [PubMed: 17209698]

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

Meredith, W., Horn, J. The role of factorial invariance in modeling growth and change. In: Collins, LM., Sayer, AG., editors. New methods for the analysis of change. Washington, DC: American Psychological Association; 2001. p. 203-240.

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

Muthén, LK., Muthén, BO. Mplus user’s guide. 7. Los Angeles, CA: Muthén & Muthén; 1998–2012.

Nesselroade, JR. The warp and the woof of the developmental fabric. In: Downs, RM.Liben, LS., Palermo, DS., editors. Visions of aesthetics, the environment & development: The legacy of Joachim F. Wohlwill. Hillsdale, NJ: Erlbaum; 1991. p. 213-240.

Pavot W, Diener E. Review of the satisfaction with life scale. Psychological Assessment. 1993; 5:164–172.

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

Rabbitt P, Diggle P, Smith D, Holland F, Mc Innes L. Identifying and separating the effects of practice and of cognitive ageing during a large longitudinal study of elderly community residents. Neuropsychologia. 2001; 39:532–543. [PubMed: 11254936]

Rast P, Hofer SM. Substantial power to detect variance and covariance among rates of change: Results based on actual longitudinal studies and related simulations. Psychological Methods. 2014; 19:133–154. [PubMed: 24219544]

Rast P, MacDonald SW, Hofer SM. Intensive measurement designs for research on aging. GeroPsych: The Journal of Gerontopsychology and Geriatric Psychiatry. 2012; 25:45–55. [PubMed: 24672475]

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

Rush J, Hofer SM. 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. 2014; 20:462–473.

Ryff CD. Happiness is everything, or is it? Explorations on the meaning of psychological well-being. Journal of Personality and Social Psychology. 1989; 57:1069–1081.

Salthouse TA, Nesselroade JR. Dealing with short-term fluctuation in longitudinal research. The Journal of Gerontology: Psychological Sciences. 2010; 65B:698–705.

Schaie, KW. Intellectual development in adulthood: The Seattle Longitudinal Study. New York: Cambridge University Press; 1996.

Schwarz, N. Why researchers should think “real-time”: A cognitive rationale. In: Mehl, MR., Conner, TS., editors. Handbook of research methods for studying daily life. New York: Guilford Press; 2012. p. 22-42.

Schwarz, N., Kahneman, D., Xu, J. Global and episodic reports of hedonic experience. In: Belli, R.Stafford, F., Alwin, D., editors. Using calendar and diary methods in life events research. Los Angeles: Sage Publications; 2009. p. 157-174.

Schwarz, N., Strack, F. Reports of subjective well-being: Judgmental processes and their methodological implications. In: Kahneman, D.Diener, E., Schwarz, N., editors. Well-being: The foundations of hedonic psychology. New York: Russell Sage Foundation; 1999. p. 61-84.

Shrout, PE., Lane, SP. Psychometrics. In: Mehl, MR., Conner, TS., editors. Handbook of research methods for studying daily life. New York: Guilford Press; 2012. p. 302-320.

Silvia PJ, Kwapil TR, Walsh MA, Myin-Germeys I. Planned missing-data designs in experience-sampling research: Monte Carlo simulations of efficient designs for assessing within-person constructs. Behavior Research Methods. 2013:1–14. [PubMed: 23055156]

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

Sliwinski, MJ., Hoffman, L., Hofer, S. Modeling retest and aging effects in a measurement burst design. In: Molenaar, PCM., Newell, KM., editors. Individual pathways of change: Statistical models for analyzing learning and development. Washington, DC: American Psychological Association; 2010. p. 37-50.

Sliwinski MJ, Smyth JM, Hofer SM, Stawski RS. Intraindividual coupling of daily stress and cognition. Psychology and Aging. 2006; 21:545–557. [PubMed: 16953716]

Thorvaldsson V, Hofer SM, Berg S, Johansson B. Effects of repeated testing in a longitudinal age-homogeneous study of cognitive aging. The Journal of Gerontology: Psychological Sciences. 2006; 61B:P348–P354.

Walls, TA., Barta, WD., Stawski, RS., Collyer, CE., Hofer, SM. Time-scale-dependent longitudinal designs. In: Laursen, B.Little, TD., Card, NA., editors. Handbook of developmental research methods. New York: Guilford Press; 2012. p. 46-64.

Watson D, Clark LA, Tellegen A. Development and validation of brief measures of positive and negative affect: The PANAS scales. Journal of Personality and Social Psychology. 1988; 54:1063–1070. [PubMed: 3397865]

Appendix

```
TITLE: Mplus code for Daily Life Satisfaction multilevel CFA;
DATA:  FILE = WB_mfa_2012.dat;
         FORMAT = free;
         TYPE = INDIVIDUAL;
VARIABLE:
    NAMES ARE ID Session d_ls1 d_ls2 d_ls3 d_ls4 d_ls5;
    USEVARIABLES ARE ID d_ls1 d_ls2 d_ls3 d_ls4 d_ls5;
    MISSING ARE ALL (999);
    CLUSTER = ID;
ANALYSIS:  TYPE IS TWOLEVEL;
           ESTIMATOR = MLR;
MODEL:
 !!!!! Two-Level CFA: 1 within factor, 1 between factor;
   !! Level-1, day-level model;
       %WITHIN%
         fw BY d_ls1 d_ls2 d_ls3 d_ls4 d_ls5;
   !! Level-2, person-level model;
       %BETWEEN%
         fb BY d_ls1 d_ls2 d_ls3 d_ls4 d_ls5;
OUTPUT: Sampstat; STDYX;
```

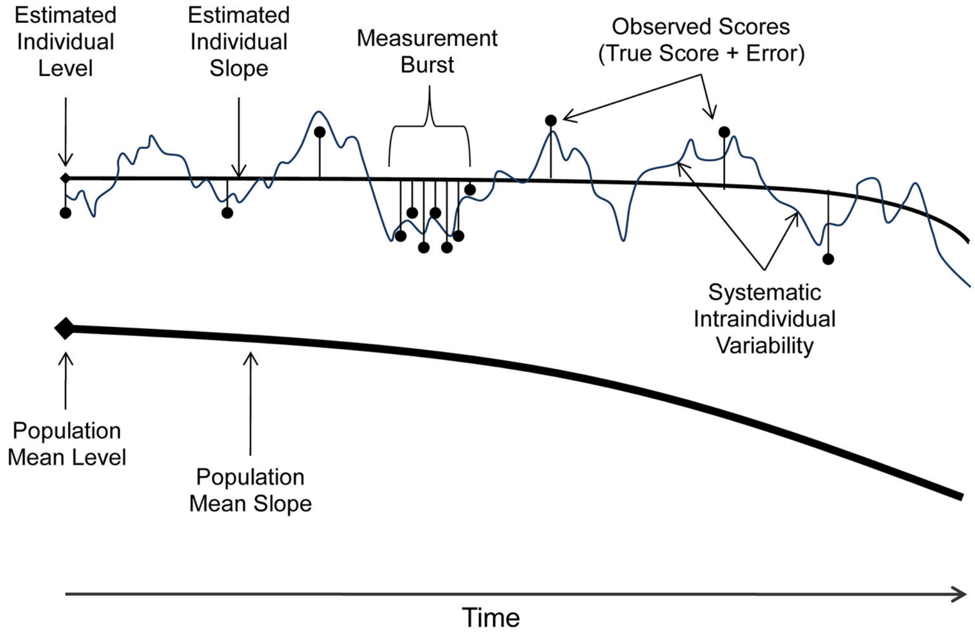
Estimated Individual Level
Estimated Individual Slope
Measurement Burst
Observed Scores (True Score + Error)
Systematic Intraindividual Variability
Population Mean Level
Population Mean Slope
Time

Figure 1.

Theoretical decomposition of an individual's observed scores into population mean level and slope, individual deviation from level and slope, and systematic intraindividual deviations from the individual slope as distinct from measurement error.

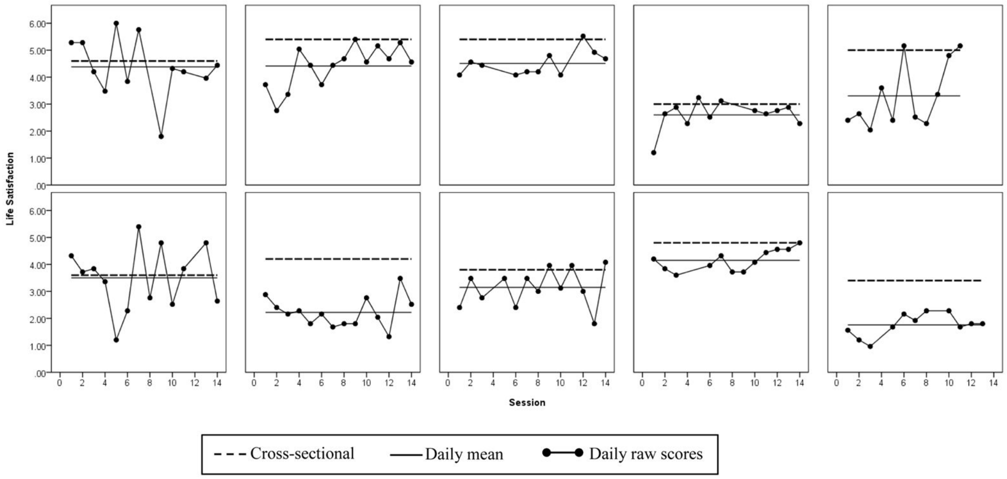

Figure 2.

Reported life satisfaction values from ten random participants displaying differences in cross-sectional, aggregated daily mean, and daily raw score measures.

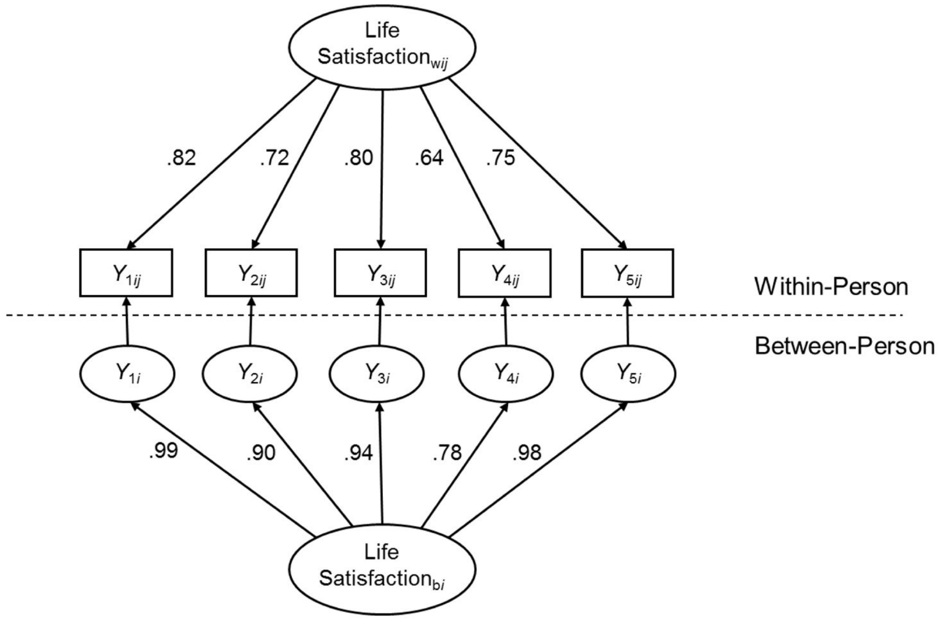

Figure 3.

Multilevel confirmatory factor model of life satisfaction with one within-person factor and one between-person factor.

Table 1

Reliability estimates (ω) comparing cross-sectional measures and daily measures of well-being.

	Cross-Sectional Measure			Daily Measure					
Variable	***M***	**SD**	**Omega(*BP*)**	***M***	**SD**	**Omega (*BP*)**	**Omega (*WP*)**	**ICC**	***r***
Life Satisfaction	4.33	0.82	0.789	3.56	0.91	0.970	0.855	0.53	.58
Positive Affect	3.53	0.65	0.862	2.77	0.55	0.956	0.844	0.45	.51
Negative Affect	1.98	0.65	0.804	1.61	0.48	0.966	0.823	0.48	.54

Note. BP = between-person; WP = within-person; ICC = intraclass correlation coefficient; *r* = correlation between cross-sectional and daily measure.

Table 2

Standardized factor loadings from multilevel confirmatory factor analyses of the daily measures of life satisfaction and competence.

		Factor Loadings	
Multilevel Factor Models	**ICC**	**WP**	**BP**
Life Satisfaction[a]			
LS1	.51	.82	.99
LS2	.51	.72	.90
LS3	.50	.80	.94
LS4	.29	.64	.78
LS5	.48	.75	.98
Factor variance	.61	1.766	2.751
Competence[b]			
Comp1	.33	.63	.88
Comp2	.29	.34	.16
Comp3	.36	.73	.96
Comp4	.38	.63	.89
Comp5	.30	.33	−.05
Comp6	.34	.54	.33
Comp7	.41	.70	.96
Factor variance	.49	1.107	1.052

Note. LS=Life Satisfaction. Comp=competence. ICC = intraclass correlation coefficient. WP = within-person. BP = between-person.

[a] $\chi^2(10) = 15.83$, CFI = .997, SRMR(WP) = 0.01, SRMR(BP) = 0.02, RMSEA = 0.02.

[b] $\chi^2(28) = 256.44$, CFI = .89, SRMR(WP) = 0.04, SRMR(BP) = 0.21, RMSEA = 0.07.