Published in final edited form as:

Psychol Aging. 2012 September ; 27(3): 560–576. doi:10.1037/a0026669.

A Monte Carlo Simulation Study of the Reliability of Intraindividual Variability

Ryne Estabrook¹, Kevin J. Grimm², and Ryan P. Bowles³

¹Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University

²University of California, Davis

³Michigan State University

Abstract

Recent research has seen intraindividual variability (IIV) become a useful technique to incorporate trial-to-trial variability into many types of psychological studies. IIV as measured by individual standard deviations (ISDs) has shown unique prediction to several types of positive and negative outcomes (Ram, Rabbit, Stollery, & Nesselroade, 2005). One unanswered question regarding measuring intraindividual variability is its reliability and the conditions under which optimal reliability is achieved. Monte Carlo simulation studies were conducted to determine the reliability of the ISD compared to the intraindividual mean. The results indicate that ISDs generally have poor reliability and are sensitive to insufficient measurement occasions, poor test reliability, and unfavorable amounts and distributions of variability in the population. Secondary analysis of psychological data shows that use of individual standard deviations in unfavorable conditions leads to a marked reduction in statistical power, although careful adherence to underlying statistical assumptions allows their use as a basic research tool.

Identifying individual differences in intraindividual variability, that is, understanding why individuals change over time and why individuals differ in how they change, is one of the principle goals of developmental research (Baltes & Nesselroade, 1979). Historically, research in psychology has focused on structured intraindividual variability, individual change that proceeds in a predictable manner over time. For example, techniques such as latent growth curve modeling have been used to understand and predict cognitive changes associated with adult aging (e.g., McArdle, Ferrer-Caja, Hamagami, & Woodcock, 2002). Recently, research has begun to investigate unstructured intraindividual variability, commonly referred to simply as intraindividual variability or IIV, in which change is unpredictable around a mean or structured trend over time.

Fiske and Rice (1955) provided one of the first comprehensive discussions on intraindividual variability, delineating among several types. They defined what is now termed IIV (their term was Type I IIV) as “pure” intraindividual variability, reflecting “the

Correspondence concerning this article should be addressed to Ryne Estabrook, Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University, P.O. Box 980126, Richmond, Va, 23298-0126. crestabrook@vcu.edu.

A version of this work was presented at the annual meeting of the Gerontological Society of America, Dallas, TX in November of 2006.

The following manuscript is the final accepted manuscript. It has not been subjected to the final copyediting, fact-checking, and proofreading required for formal publication. It is not the definitive, publisher-authenticated version. The American Psychological Association and its Council of Editors disclaim any responsibility or liabilities for errors or omissions of this manuscript version, any version derived from this manuscript by NIH, or other third parties. The published version is available at www.apa.org/pubs/journals/PAG

difference between the two responses of an individual at two points in time” (p. 217) under the conditions of identical or objectively indistinguishable stimuli and situations. Under this definition, the order and timing of responses is assumed immaterial such that later measurements are not impacted by the passage of time or previous measurements.¹ Key to the importance of IIV in psychological research is that individuals can differ in their typical difference between responses; that is, there can be individual differences in the magnitude of IIV. Under this conceptualization, IIV can be considered a trait specific to an individual on a particular task or measure, rather than a property of the task itself (e.g., measurement error). As a trait, IIV has been thought of as an individual’s level of instability, error, wobble, lability, inconsistency, noise (Ram, Rabbit, Stollery, & Nesselroade, 2005), or ‘hum’ (Nesselroade, 1991).

IIV has been incorporated into studies across a number of areas of psychology, particularly in cognition, personality, and lifespan development (for reviews, see Anstey, 2004; MacDonald, Nyberg & Bäckman, 2006; Nesselroade & Ram, 2004; and a recent special issue, Blanchard-Fields, 2009). To illustrate, perhaps the most common use of IIV in psychological research is in studies of reaction time (RT). Jensen (1992) showed that IIV in simple RT is an attribute of an individual distinct from his or her mean RT and demonstrated that both an individual’s mean and IIV independently contribute to the prediction of general intelligence. Lecert, Ghisletta, and Jouffray (2004) demonstrated a similar finding, showing that level and IIV contributed independently to the same working memory factor. Furthermore, IIV in RT tends to be positively associated with age, such that older adults tend to have more variability in RT (Anstey, 1999; Deary & Der, 2005; Hale, Myerson, & Smith, 1988; Salthouse & Berish, 2005; Wilkinson & Allison, 1989). Increased IIV with age has been shown to be predictive of a number of neurological disorders and conditions (Duchek, Balota, Tse, Holtzman, Fagan & Goate, 2009; Fuentes, Hunter, Strauss, & Hultsch, 2001; Hultsch, MacDonald, Hunter, Levy-Bencheton, & Strauss, 2000; MacDonald, Hultsch, & Dixon, 2003; Rentrop, Rodewald, Roth, Simon, Walther, Fielder, Weisbrod & Kaiser, 2010; Schretlen, Munro, Anthony, & Pearson, 2003), risk of cognitive impairments (Bielak, Hultsch, Strauss, MacDonald & Hunter, 2010a; Bielak, Hultsch, Strauss, MacDonald & Hunter, 2010b), and brain activity (MacDonald, Nyberg, Sandblom, Fischer, & Bäckman, 2008).

Several other studies have considered IIV with other variables in the context of developmental change. For example, IIV in perceived control predicts five-year mortality (Eizenman, Nesselroade, Featherman, & Rowe, 1997). Conversely, greater daily IIV within task predicts increased gains related to practice in an older sample (Allaire & Marsiske, 2005). Heart rate variability has been found to predict temperament in infants (Fox & Porges, 2006) and major depressive disorder in adults (Licht, de Geus, Zitman, Hoogendijk, van Dyck, & Penninx, 2009). Siegler (1994, 2002) in a review of studies of developmental processes has argued that intraindividual variability is a crucial part of studying children, as increased variability tends to indicate transitions between developmental stages.

Measuring IIV with the individual standard deviation

The measurement of IIV presents a unique challenge, as it reflects a magnitude of variability rather than an individual’s level of a static trait at a specific point in time. The most common way to measure IIV is with the individual standard deviation (ISD). The ISD is calculated by first removing any time-dependent structured trend in the data (at its simplest, an individual’s mean across time, but more often a linear trend), then calculating each

¹Variation affected by the passage of time is referred to as type II IIV, and variation affected by the measurement process itself is referred to as type III IIV.

individual's standard deviation on a single variable across multiple measurement occasions. The standard conceptualization of IIV measured with the ISD is based on the equation

$$X_{nt} = S_{nt} + U_{nt} \quad (1)$$

where X_{nt} is the observed score for person n at time t , which is composed of two parts, a structured part, S_{nt} , that is predictable from time, and an unstructured part, U_{nt} , that is independent of S_{nt} , and unpredictable from time. If there is no time-ordered trend, then S_{nt} is an individual's mean across time ($S_{nt} = S_n = \mu_n$). Alternatively, S_{nt} may show some form of time-dependence, ranging in complexity from a simple linear trend to combinations of curvilinear and oscillatory patterns of change, and S_{nt} can be detrended by subtracting the expected value of the appropriate pattern of change at each time t . Therefore, without loss of generality, we consider a conceptualization of IIV in which S_{nt} is either stable or detrended and thus not time-dependent, so that $S_{nt} = S_n = \mu_n$. Consistent with common terminology, we refer to $S_n = \mu_n$ as the intraindividual mean, recognizing that S_n can be more broadly defined. We further assume that U_{nt} follows a probability distribution within individuals (typically a normal distribution) with a standard deviation, σ_n , which is specific to individual n . An individual's value of σ_n is the ISD and is the most typical measure of IIV. Finally, the ISD is assumed to follow a probability distribution across individuals (often normal despite a necessary lower bound of zero) with a population mean, μ_{ISD} , and standard deviation, σ_{ISD} .

ISDs are the most utilized measure of IIV because they (a) are theoretically superior to indices that depend on extreme observations (e.g., range, local maxima, and minima), (b) directly reflect the quantity of unstructured IIV because they ignore the order of individual responses, (c) are simple and straightforward to calculate, and (d) are readily interpretable by researchers familiar with standard deviations. They are also independent of the number of observations, unlike some entropy-based measures of IIV that have strong relationships with the number of occasions (Richman & Moorman, 2000). However, the ISD is often correlated with the mean level such that the ISD tends to be higher for high scoring individuals (e.g., higher variability in RT for slower participants). This has been seen as a shortcoming by some researchers who advocate for the coefficient of variation (ISD divided by the mean; Haldane, 1955). In this article, we focus on the ISD because of its common usage, but note that our results are likely to generalize to the closely related coefficient of variation.

Familiarity with standard deviations may have the unintended consequence of a neglect of reliability checks. Existing empirical discussions of the reliability of variability are limited to several studies reporting test-retest reliabilities of ISDs ranging from 0.00 to 0.80 (Allaire & Marsiske, 2005; Li, et al, 2001). Other discussions of the reliability of variability shows similarly varying results, with comparisons of the standard errors of individual means and standard deviations suggesting that ISDs are very reliable (Schmiedek, 2006; Schmiedek, et al, 2009) and other analytic discussions showing ISDs to be far less reliable than individual means (Wang, under review). ISDs are implicitly treated as simple transformations of the data, comparable in complexity to individual means. This may not be a reasonable assumption as intraindividual variation presents both theoretical and practical concerns for reliability. Theoretically, ISDs differ from other types of measures by quantifying stochasticity rather than stability or systematic change. Individual differences in IIV inherently violate core assumptions of classical test theory (CTT), which assumes that any variability between repetitions of the same test is simply error variance and thus not meaningful. This violation has two important implications: (a) reliability of a measure of stochasticity is not explicitly defined within CTT, and (b) the meaning of CTT-defined

reliability for the intraindividual means is problematic (see below). Practically, unstructured variability may be harder to quantify reliably, as sufficient measurement occasions are required to describe the underlying distribution. Furthermore, tests that are reliable for measuring a score at a single point in time or for assessing the intraindividual mean may not be useful measures for the examination of intraindividual variability.

Differences between the reliabilities of intraindividual means and ISDs could have important impacts on psychological research. If ISDs are relatively reliable, then they represent an underutilized tool for incorporating variability into many types of models. However, if ISDs are generally unreliable, then the predictive validity of these measures comes into question. The lack of statistical power associated with poor reliability suggests that modest relationships involving intraindividual variability would be difficult to detect. By extension, significant relationships found involving ISDs in previous research may be substantially over- or underestimated.

Definition of Reliability within the Context of IIV

Reliability is defined conceptually as the consistency or replicability of a score under identical experimental conditions. The mathematical definition of reliability, bound in CTT for an arbitrary measure X_n , is:

$$X_n = T_n + E_n \quad (2)$$

$$E(E_n) = 0 \text{ and } E(T_n, E_n) = 0, \quad (3)$$

where X_n is individual n 's observed score on some measure, T_n is the individual's latent 'true' score, and E_n is an error component with a known (usually normal) probability distribution with standard deviation σ_E . This definition holds whether X_n is a measure at a single measurement occasion or a mean or other function of multiple occasions of measurement. Note that CTT is similar to our definition of IIV (Equation 1) with the primary difference being that unstructured portion of X_{nt} (i.e., U_{nt} in equation 1) consists of both true variation and error, both of which are subsumed in the error term in the above definition. Under CTT, the classical definition of reliability is the correlation between parallel forms of an instrument, which can be shown to equal the proportion of observed score variance attributable to true score variance:

$$\rho_{XX'} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2} = \frac{\sigma_T^2}{\sigma_X^2} \quad (4)$$

where $\rho_{XX'}$ is the reliability and σ_T^2 and σ_X^2 are variances of T (true scores) and X (observed scores), respectively.

When we move into an IIV context, calculating reliability of the intraindividual mean becomes problematic because the error variance, and therefore the observed score variance, varies across individuals. For example, an individual with high IIV will have greater sampling variability in their observed intraindividual mean (over, say, 5 occasions) than another individual with low IIV. The inability to disentangle IIV from error variance means that all variance around the mean, regardless of source, is treated as error variance. However, this variance term is not homoskedastic across individuals due to the interindividual differences in IIV. Reliability would thus vary across individuals, and sample measures of reliability would misstate the actual reliability of a measure.²

The reliability of the ISD also has challenges, but of a more practical nature. Although the CTT derivation of reliability can apply because the ISD is assumed to follow a population distribution with no individually varying parameters, there is no clear method for estimating the reliability of a measure of stochasticity. Reliability cannot be directly assessed because the variance of the true score is not observable, it is estimated with closely related concepts, such as internal consistency with coefficient alpha, split halves with a Spearman-Brown adjustment, or the coefficient of stability (i.e., test-retest reliability). To our knowledge, there are currently no versions of coefficient alpha or the Spearman-Brown prophecy for measures of variability such as the ISD. Both are based on the simple formula for the variance of a linear combination of observed scores, whereas the ISD is a nonlinear combination. The coefficient of stability is generally impractical because it requires (at least) two measures of the ISD, which itself requires a number of repeated occasions; that is, calculating the coefficient of stability would require twice as many occasions, which is cost-prohibitive (Fleeson, 2001; Marcotte et al., 2003).

To address these issues with the reliability of both the intraindividual mean and ISD and to maintain comparable conceptualizations of reliability for both, we turn to the basic definition of the *reliability index* - the correlation between an observed score and its underlying true score. Squaring this yields the proportion of observed score variance attributable to true score variance, one of the most common interpretations of reliability. Calculating the reliability index for the mean and ISD requires that the true score is known, which is impossible in empirical studies. In this article, we take a simulation approach, under which true scores are known for hypothetical samples. This allows us to examine the reliability of both the intraindividual mean and ISD under typical experimental conditions. These simulations can also be applied to empirical studies by including empirically generated parameter values into the simulation; we provide R code for the simulations so that other researchers can calculate the reliability of the intraindividual mean and ISD within their own studies (see Appendix A). Our results are then applied to an empirical study with data from the MacArthur Study of Successful Aging (Eizenman, et al, 1997) as an illustration of assessing reliability of the ISD using the simulation approach.

Simulation Study

The goal of the simulation study is to address practical concerns about the reliability of the intraindividual mean and ISD in the context of IIV. We derive reliability estimates using a Monte Carlo simulation, allowing us to investigate the reliability of ISDs in comparison to the intraindividual means. We examine the effects of two experimental conditions and two population characteristics that may impact the reliability of ISDs and intraindividual means: number of measurement occasions, reliability of the measure, mean ISD, and the magnitude of between-person differences in ISD in the population. In these simulations, we focus on two related concerns regarding the reliability of intraindividual variability: (1) identifying experimental conditions required for reliable and meaningful ISDs, and (2) examining the extent to which the validity of studies of intraindividual variability depend on characteristics of the study population. The values for these parameters are derived from the studies referenced in Table 1, which details the ranges of these experimental characteristics in studies that both (1) employ measures of IIV, and (2) published distributional characteristics of their measure of IIV (mean, variance, etc).

²A similar issue has been identified in measurement methods such as Rasch measurement and Item Response Theory that yield individual differences in measurement precision. See, e.g., person separation reliability in Wright and Stone (1979).

Model Parameters

Each simulated individual was assigned an intraindividual mean level (μ_n) and an individual standard deviation (σ_n), where both individual parameters are indexed by n . The intraindividual means were drawn from a normal distribution with an arbitrary mean of zero ($\mu_{\text{MEAN}}=0$) and an arbitrary non-zero population standard deviation ($\sigma_{\text{MEAN}}=1$), both of which were fixed across all simulation runs. These two parameters, ($\mu_{\text{MEAN}}, \sigma_{\text{MEAN}}$), define the between-subjects distribution in intraindividual means: $\mu_n \sim N(\mu_{\text{MEAN}}, \sigma_{\text{MEAN}}^2)$. Each individual was then assigned an individual standard deviation that describes how much the individual varies across time. ISDs were drawn from a normal distribution with a population mean (μ_{ISD}) and population standard deviation (σ_{ISD}). These two parameters describe the extent to which the typical individual in the population fluctuates across time (μ_{ISD}) and between-subjects differences in the amount of fluctuation across time (σ_{ISD}): $\sigma_n \sim N(\mu_{\text{ISD}}, \sigma_{\text{ISD}}^2)$.

Four factors were varied across simulation conditions, as summarized in Table 2. The first factor varied was the average amount of individual variability, μ_{ISD} , with values between 0.2 and 2.0 in steps of 0.2. These values can be compared to the between-person differences in intraindividual means ($\sigma_{\text{MEAN}} = 1$) to indicate how much the average person varies across time compared to how much people differ from one another. The ratios of within-individual variability to between-person variability were chosen to be in line with empirical studies (see Table 1). The second factor was the ratio of the standard deviation of the individual standard deviations (σ_{ISD}) to the mean individual standard deviation (μ_{ISD}), with values of .08, .16, .24, .32, .40 and .48, which were in the range of typical empirical values (see Table 1). This parameter can be thought of more simply as the standard deviation of the distribution of ISDs, scaled in terms of the mean.

By definition, ISDs cannot fall below 0, but the normal distribution from which ISDs were sampled in our simulation is unbounded, which occasionally led to negative ISDs being sampled. When negative ISDs occurred, the simulee was removed and replaced with a new draw. Thus, the actual distribution of ISDs was slightly skewed with a slightly larger mean than the true simulated value, but the departure was considered to be negligible as fewer than 0.5% of cases were affected by this procedure. Alternative versions of this simulation drew ISDs from the chi-square and gamma distributions, which showed no substantive differences in results.³

The third and fourth factors were number of occasions and test reliability. The number of occasions ranged from 3 to 10 in steps of 1 and then from 10 to 100 in steps of 5. The fewest number of occasions included in the study was three, because that is the smallest number of occasions on which ISDs can be calculated (providing more information than the range) and for which results have been reported (Salthouse, 2007). Although among studies of IIV of which we are aware, only three studies exceeded 25 measurement occasions using a non-physiological measure (Eid & Diener, 1999, 52 occasions; Allaire & Marsiske, 2005, 120 occasions; MacDonald, et al, 2008, 160 occasions), we included up to 100 occasions to more fully understand the effect of the number of occasions. The reliability of the measure, that is,

³While this bounding problem could be avoided by sampling ISDs from a distribution bounded at zero, distributional information reported in the literature cites only the means and standard deviations of the distributions of ISDs. Selecting an alternative sampling distribution would lead to a disconnect with the existing literature. Alternate versions of this simulation draw from chi-square and gamma distributions, and showed very minor differences in ISD reliability, which is expected as the family non-central chi-square distributions can be well approximated by the normal distribution (Sankaran, 1963). Reliabilities for each simulation cell correlate 0.998 between the normal and gamma distributions, with a mean difference of 0.012 ($t_{4498}=47.40$, $p<.001$, Cohen's $d=0.04$). Similar results are found for the chi-square version of the analysis, with simulation cells correlating 0.998 and a mean difference of 0.013 ($t_{4498}=47.68$, $p<.001$, Cohen's $d=0.04$).

the reliability of a score at a single point in time, ranged from .60 to 1.00 in increments of .10, reflecting a range of typical values from barely adequate to perfect.

Data Generation—The statistical parameters described above were used in the data generation for the simulation. For each simulation intraindividual means and ISDs for 100 individuals were generated (sample size did not have an appreciable impact on the results). We then simulated an individual's true scores at each occasion t as

$$TrueScore[t]_n = \mu_n + ISD_n \cdot \varepsilon[t]_n. \quad (5)$$

where $\varepsilon[t]_n \sim N(0,1)$ is a standard normally distributed random variable. In keeping with the Fiske and Rice (1955) definition of Type I variability, true scores at any occasion were unpredictable from time and independent of true scores at any other occasion. Observed scores were calculated by adding error to the true score,

$$ObservedScore_{nt} = \sqrt{\rho} * TrueScore_{nt} + \sqrt{1 - \rho} * ErrorScore_{nt} \quad (6)$$

where ρ is the reliability of the test, and $ErrorScore_{nt}$ is normally distributed with a mean of zero and a variance of $\sigma_{TrueScore}^2$, where $\sigma_{TrueScore}^2$ is the variance of the true score across persons and occasions. Specifying the observed score this way keeps the variance and scale of the observed scores constant regardless of the reliability of the measure.

The data generation process yielded a 26 (number of occasions) by 5 (levels of test reliability) by 10 (levels of mean ISD) by 6 (levels of the ratio of mean ISD to the standard deviation of ISDs) simulation design, as summarized in Table 2. Each cell in the simulation design was repeated 100 times in a Monte Carlo simulation approach, for a total of 780,000 repeated measures datasets. Observed means and ISDs were calculated for each dataset, and reliabilities of means and ISDs were defined as the squared correlation between the observed measure and the true mean and ISD used for data generation. These reliabilities were then regressed on the simulation parameters; these regressions are presented in Appendix B. All data generation and analyses were carried out in R (R Development Core Team, 2007). Sample simulation code is contained in Appendix A and the full code is available by request.

Results

The reliability of intraindividual means and ISDs are presented in Figures 1 and 2, respectively, with the median reliability across the 100 repeated simulations as the outcome variable on the y -axis and the number of occasions used to calculate the intraindividual means and ISDs on the x -axis. Each figure consists of a grid of plots, indexed by the reliability of the measure (denoted as “Scale Reliability” and presented vertically) and the amount of interindividual variability in the ISDs (denoted by $SD_{ISD}/Mean_{ISD}$ & presented horizontally). Within each plot, five curves are presented to reflect a selection of five out of the ten simulated levels of mean IIV: 0.2, 0.4, 0.6, 1.0, and 2.0 times the standard deviation of individual means. The only difference in presentation between the plots for intraindividual means and ISDs is the order of the five levels of mean ISD. That is, the results for intraindividual means (Figure 1) show that the highest levels of mean reliability (topmost curves) were associated with the *lowest* levels of mean IIV (0.2, represented by a solid line in both figures), whereas the results for ISDs (Figure 2) show that the highest levels of ISD reliability were associated with the *highest* levels of mean IIV (2.0, represented by a dotted line in both figures).

The figures give an overview of the effects of number of measurement occasions and within-to-between variability on the reliabilities of intraindividual means and ISDs. As shown in Figure 1, intraindividual means were relatively reliable, asymptotically approaching perfect reliability in all conditions. While differences between the plots (indexed by measure reliability & variance in IIV) can be perceived, most of the differences were related to the number of occasions and the mean amount of IIV. As shown in Figure 2, ISDs were far less reliable than individual means, with more distinct differences between the plots than observed for the individual means. In particular, ISDs showed more pronounced effects of measure reliability and the variance in ISDs, as well as an interaction between mean ISD and measure reliability.

In order to numerically assess the importance of each of the manipulated factors in the simulation and to simplify and summarize the presentation of results, regression analyses were performed predicting Fisher's z -transformed reliability from the four simulation parameters (number of occasions, test reliability, μ_{ISD} , & the ratios of σ_{ISD} to μ_{ISD}), a log transform of these variables, and two-way interactions.⁴ The regressions accounted for almost all of the variability in reliability ($R^2 = .998$ for intraindividual means, .988 for ISDs). To simplify interpretation, results in this section are presented as the proportion of variance accounted for (R^2) for each of the four factors (i.e., the total R^2 for the linear and log-transformed factor combined, representing the total contribution of each effect) and six interactions. Details about the regressions and full regression results are included in Appendix B.

The reliability of intraindividual means was considered as a reference to compare to the reliability of ISDs. The reliability of intraindividual means was negatively impacted by the mean amount of IIV and positively by the number of occasions. These two factors accounted for 45.8% and 44.6% of the variance in the reliability of means, respectively. The reliability of the measure and the interactions between mean IIV and measure reliability accounted for an additional 9.2% of the reliability of intraindividual means. Reliability of intraindividual means was improved with higher test reliability, but the effect was attenuated at higher levels of mean ISD. The variability in IIV and all other interactions, some of which were statistically significant, combined for only 0.2% of the variance in mean reliability, and were not important factors in estimating reliability of the mean.

Although the reliability of intraindividual means was predominantly accounted for by only two manipulated factors, the reliability of the ISD was significantly and meaningfully related to all four simulation factors. The reliability of ISDs increased with more measurement occasions ($R^2 = .313$), higher scale reliability ($R^2 = .163$), higher mean IIV ($R^2 = .109$), and greater variability in IIV ($R^2 = .316$). In addition, all sets of two-way interactions accounted for at least 0.8% of the variance in ISDs, with a maximum of 2.4%. The strongest interaction involved mean IIV and measure reliability, indicating that as measure reliability approached 1.00, the mean IIV had a smaller impact on the reliability of ISDs.

Summary of simulation results

In summary, the simulation study indicated the reliability of the intraindividual means was generally acceptable, and affected primarily by the number of occasions and the average

⁴Non-linear transformations of the variables are required, as linear regression provides an incomplete description of the data ($R^2 = .898$ for means, .870 for ISDs). Log transformations provide good fit for the diminishing effects of increasing measurement occasions, for the correlational nature of measure reliability, and the dependence of the two ISD ratios on variances, which are bounded at zero. Unfamiliar readers should look to introductory regression texts (for example Cohen, Cohen, Aiken & West, 2003, section 6.4) for more information on this non-linear transformation. All regression effects were estimated simultaneously in a single multiple regression model.

amount of IIV in the sample. ISDs, on the other hand, were less reliable, especially with few occasions and few interindividual differences in IIV. While both the intraindividual mean and the ISD benefited from increasing the number of occasions and scale reliability, other factors affected the reliability of the ISD without having an important impact on the reliability of the intraindividual mean. Based on these results, ISD reliability appears to be poor in most psychological studies.

Example: MacArthur Study

In light of the theoretical issues with defining and estimating reliability for the intraindividual mean and ISD within CTT, we demonstrate how these simulations can be applied to assess reliability in an empirical study. Eizenman et al. (1997) used data from 30 individuals on 25 occasions of measurement, collected as part of the MacArthur Foundation Research Network on Successful Aging, to study intraindividual variability in perceived control. Eizenman et al. reported that higher levels of intraindividual variability in perceived control predicted five-year mortality in a logistic regression model, which we've replicated in Table 3. In this example, we first demonstrate how the simulations can be used to estimate the reliability of the ISD in an empirical study. We then conduct a simulation parallel to the first study, using parameter values equal to the empirical values from the MacArthur Studies, to yield a more precise estimate of reliability for all possible numbers of occasions from 3 to 100. We also performed a bootstrap power analysis, resampling from the observed data to create samples with the same distributional characteristics but varying numbers of observations. In doing so, we demonstrate how varying the number of occasions affects the reliability of IIV and its resulting impact on power in the ensuing analysis. We focus on the number of occasions because it was the most crucial factor associated with ISD reliability that can be experimentally manipulated.

Reliability of ISD

We first demonstrate the use of the previous simulations to estimate reliability in the MacArthur study. To do so, we determined the values for several parameters: number of occasions, reliability of the measure, standard deviation of the intraindividual mean, and mean and standard deviation of the ISD. In this study, the number of occasions was 25 and the reliability of the perceived control measure was .702 based on coefficient alpha taken at the first occasion. Reanalyzing the MacArthur dataset showed that the between-person differences in intraindividual means was 2.679 (i.e., $\sigma_{\text{MEAN}}=2.679$), the average amount of within-person variability was $\mu_{\text{ISD}} = 1.453$, and the between-person differences in intraindividual variability was $\sigma_{\text{ISD}} = .586$ for the perceived control measure.⁵ These values yield the distributional parameters listed in Equations 7 and 8.

$$\frac{\mu_{\text{ISD}}}{\sigma_{\text{Mean}}} = \frac{1.453}{2.679} = .542 \quad (7)$$

$$\frac{\sigma_{\text{ISD}}}{\mu_{\text{ISD}}} = \frac{0.586}{1.453} = .403 \quad (8)$$

We then repeated the simulation with the above parameters to yield a more precise and tailored estimate of reliability of the ISD. We produced 1000 datasets for numbers of observations from 3 to 1000⁶ using the same reliability of the measurement instrument (.

⁵There is a distinction to be made between the true and observed ISDs, as observed IIV consists of a combination of true IIV and residual error. The stated reliability of the measure (.702) is used to add error variance to the simulation as shown in Equation 8.

702), distributional characteristics (see Equations 7 and 8) and sample size (30) as the original study. This procedure takes minimal computational time, making it an effective method for researchers to estimate ISD reliability for studies on which these parameters can be calculated, either from pilot data, meta-analysis, or literature review. The results of this simulation are presented in Figure 3, and show that the reliability of the ISD for this population and scale ranges from .054 with three occasions to .760 with 100 occasions and .947 with 1000. The estimated reliability of the ISD in the 25 occasion MacArthur study was .449.

Effect of number of occasions and reliability on power

Finally, we considered how the reliability of ISDs might affect the power to detect effects using a parametric bootstrap power analysis (see Efron & Tibshirani, 1993 for review of bootstrap methods). The original data was resampled, such that new data for any individual in the original study was randomly drawn from a normal distribution with the same mean and ISD that was observed for that individual in the original data collection. This procedure was then repeated just as the reliability simulation was, creating 1000 datasets of 30 individuals each with numbers of occasions from 3 to 1000. Logistic regression analysis from Eizenman, et al (1997) were then repeated for each dataset. Power for a given number of observations can be estimated as the proportion of bootstrapped datasets in which the effect was replicated.

The results of the power analysis as well as the previous reliability simulation are presented in Figure 3, showing the power for the 25 occasion version of the study to be .402. Both ISD reliability and power increase strongly with the number of observations, though these two values increase at differing rates. The relationship between estimated reliability and power is more strongly shown in Figure 4, where we see a near perfect non-linear relationship between the two.

The above power analysis was contingent on the original study to estimate the correct effect size, which was somewhat tenuous given the relatively low reliability of the ISD and low power of the study. This manifests itself in the confidence interval for the logistic regression effect (1.700, CI = [0.040, 3.361]) and its associated odds ratio (5.476, CI = [1.041, 28.897]), showing that increasing the ISD by one unit could increase odds of mortality by as little as 4% and as much as a 28-fold increase. We can also get a rough sense of the effect of IIV in perceived control on 5-year mortality if the ISD were measured perfectly by adjusting the point-biserial correlation between mortality and IIV of perceived self-control for unreliability (Thorndike, 1949). This changes the correlation from .422 to .653, changing the R^2 for the relationship between these two variables from .178 to .426.

Discussion

The results of this study showed that the reliability of individual standard deviations was considerably lower than the reliability of intraindividual means in the range of experimental conditions presented in Tables 1 and 2. The reliability of intraindividual means was found to be generally high and so the use of this technique in psychological research is psychometrically appropriate. On the other hand, individual standard deviations were found to have low reliability, especially in experiments with few occasions, poor test reliability, or suboptimal distributions of variability in the population being studied.

⁶The range of observations includes all possible values from 3 to 100 and values from 100 to 1000 in steps of 10.

Reliability of the intraindividual mean

The simulation showed intraindividual means to be reliable in virtually all simulation conditions. The reliability of means approaches perfect reliability after only a few occasions. Even in the worst of circumstances for the reliability of intraindividual means (minimum measure reliability and maximum mean and variance of IIV), .80 reliability was achieved after thirty-five measurement occasions. Intraindividual means were affected primarily by the number of occasions and the amount of IIV, and to a lesser extent, the reliability of the measurement instrument. The results suggest the intraindividual mean was affected by all forms of variability, whether meaningful (IIV) or not (error), in a similar manner. It appears that intraindividual means were most reliable when variability from all sources was minimized. Similar relationships are likely to be found with more complex longitudinal trends; detecting the correct longitudinal trend depends on unstructured variation around that trend being minimized regardless of the attribution of that variability to error or the individual.

Another important finding was that the amount of interindividual differences in intraindividual variation (σ_{ISD}) had small if any effect on the reliability of intraindividual means (R^2 for the main effect = 0.001, R^2 for all interactions = 0.000). When individual differences in intraindividual variability exist, assumptions of homogeneity of variance that form the basis of reliability calculations are violated, making the interpretation of reliability challenging. Conceptually, individuals with low intraindividual variability are more accurately described by an intraindividual mean or trend, or even the score from a single occasion, while individuals with high intraindividual variability are more poorly measured by a mean or trend. However, we found that the variability in IIV had little effect on the reliability of the intraindividual mean, indicating that we can interpret reliability in any empirical study as approximately equal to the reliability under a condition of no variability in IIV. The reliability of the intraindividual mean can therefore be considered the ‘typical’ or ‘average’ reliability for an individual, albeit with some amount of interindividual variation around that value. Reliability is thus not merely a summary or descriptive statistic, but has interpretational value at the individual level, despite the theoretical issues highlighted earlier.

Reliability of the Intraindividual Standard Deviation

While the reliability of intraindividual means was acceptable under most circumstances, the reliability of ISDs was much lower than typically acceptable. This was especially true with relatively few measurement occasions; even in the best circumstances, the reliability of individual standard deviations did not surpass .80 in fewer than 15 measurement occasions or .90 in fewer than 35. Aggregating across all simulation values for scale reliability and distributional parameters for IIV in the population, the median reliability of the ISD surpassed .80 with a minimum of 90 measurement occasions. All four simulation parameters affected the results of individual standard deviations. ISDs were more reliable when the average IIV was high relative to interindividual differences in intraindividual means, and when there were greater individual differences in the ISD. Mirroring the effects for intraindividual means, ISDs were more reliable with more measurement occasions and more reliable measures. However, ISDs required more occasions and scale reliability to be useful, and this may simply be too unreliable for use in typical psychological research without many measurement occasions.

These results become understandable and expected when considered within typical sampling theory. Measuring intraindividual variability essentially treats each individual as a population of responses with each measurement occasion being a sample from a time-invariant distribution. Once time dependence is removed through detrending, there is no

conceptual difference between describing a single individual's population of repeated responses and summarizing a sample's distribution of responses at a single measurement occasion. ISDs from a number of occasions should be as reliable as sample standard deviations from the same number of individuals.

The possibility of needing dozens or hundreds of measurement occasions to reliably calculate ISDs should not preclude studying intraindividual variability. Despite the reliability problems demonstrated in this study, many studies have found statistically significant effects of IIV, such as the results of the MacArthur Study described in our example, indicating that these effects may be large enough to overcome low reliability, and therefore may have strong scientific importance. Reliability of a scale represents only one component in determining the statistical power of a study. Design factors, including the sample size and magnitude of the effect in question, are also important to the power of a study. It is clear, however, that beyond statistical significance, a large number of occasions are needed to assess the *size* of the effect with reasonable precision in the populations and constructs examined in the context of IIV thus far. For example, our results suggest that the magnitude of the effect of IIV in perceived control on 5-year mortality may be significantly underestimated, as evidenced by the correction for unreliability. Alternatively, it is possible that IIV may be reliably measured in experimental conditions where the distribution of IIV is far more favorable than in existing research cited in Table 1. It is vital for researchers interested in the impact of IIV to conduct appropriate literature reviews and pilot studies to ascertain the reliability of IIV in their planned studies and the resulting impact on power.

Novel approaches must be explored to obviate for the need for a large number of measurement occasions in studies of IIV. ISDs from related constructs, subscales or individual items may be factor analyzed, and the resulting factors treated as measures of IIV on the construct being measured (Gerstorf, et al, 2009). Bowles' (2009) work on intratask change provides an alternative method for assessing IIV within the context of a single test.

Importance of measurement characteristics

Given the low reliability of the ISD, researchers should be sure to use measurement instruments and scoring techniques that do not further adversely affect reliability. We highlighted a few aspects of the measurement process that may have important effects on the reliability of the ISD. First, it is vital that the scores are at least interval scaled, where a single unit has the same meaning across the range of scores. Without this characteristic, the meaning of the ISDs may change with even slight differences in mean level. This may introduce an additional source of error in measuring the true underlying level of IIV, and further degrade any measure of IIV, including the ISD.

Second, the precision of the scores may also impact the reliability of the ISD. We simulated data with as many decimal places as generated using a random number generator, yielding near perfect precision of the scores. This level of precision is well beyond the capability of most psychological measurement. Less than perfect precision adds additional error into the scores, and by extension, the ISDs. Researchers should be cautious when investigating intraindividual variability using scales with few possible responses.

Third, the scores should be absent of notable ceiling or floor effects. In these cases, relationships may exist between the mean and variability that would inhibit reliable measurement of IIV. For instance, if a measurement instrument is bounded at zero with many responses at or near zero and no ceiling effect, individuals with means very close to zero would be unable to express IIV due to the scale's floor. Any response much below an individual's mean would count as zero, reducing the reliability of the ISD. Furthermore, individuals with higher mean levels would be able to demonstrate higher levels of IIV, likely

resulting in higher mean IIV than individuals with low mean performance regardless of the true relationship between mean level and intraindividual variability.

Individual homoscedasticity

One important assumption of IIV research deserves particular attention: meaningful measures of IIV require homoscedasticity within individuals; that is, individuals are assumed to have the same magnitude of IIV over the course of the study. When homoscedasticity assumptions are violated, each individual's magnitude of IIV changes, rendering results based on the ISD of questionable meaningfulness. Violations of homoscedasticity may simply indicate misspecification of the longitudinal trend, as any calculation of variance is dependent on the mean or trend used to calculate it. Violations may also indicate an incorrect scaling of the measurement instrument that could be corrected by an appropriate data transformation akin to typical practice in cases of across-individual heteroscedasticity in regression analysis. A third possibility is that changes in IIV are meaningful and an avenue for further scientific study (Ram, Rabbit, Stollery, & Nesselroade, 2005).

Conclusion

The shortcomings of the individual standard deviation as highlighted by the simulation study seem at odds with the significant relationship between intraindividual variability in perceived control and mortality, as well as other findings of significant relations between IIV and other outcomes. Using a typical correction for attenuation suggests that the actual correlation between IIV and mortality in the provided example was .653 rather than .422, yielding an increase in the R^2 from .178 to .426. This strikingly large difference suggests that documented relationships involving IIV have been underestimated and that IIV may be a much more important aspect of human behavior than typically appreciated. The resulting impact of poor reliability on power further suggests that many null findings involving IIV may mask important effects. Alternatively, it may be that some of the empirical results attributed to IIV were in fact relationships involving structured intraindividual variation too complex to be captured by the intraindividual mean or simple detrending procedures. Constructs may show oscillations based on various biological and social factors, including biorhythms, seasonal effects, and other short-and long-term oscillatory factors, which should be treated not as unstructured IIV but patterned longitudinal change and modeled with appropriate methods (see reviews in Boker, Deboeck, Edler & Keel, 2010; Chow, Hamaker, Fujita & Boker, 2009; Deboeck, 2010). When ISDs are calculated using data governed by a dynamical system, the ISD may approximate otherwise undetected structured intraindividual variability.

Although the results of the simulation indicate ISDs should only be used with caution, the *methodological* shortcomings shown here do not diminish the importance of intraindividual variation within psychological *theory*. These limitations are not a problem with ISDs as a technique any more than poor statistical power with few subjects is a problem with regression analysis or ANOVA. Instead, these results highlight the need for careful consideration of research design before data collection to ensure sufficient reliability and power. Only when researchers take appropriate care to consider IIV in study design will ISDs provide a reliable, informative, and useful measure of intraindividual variability.

Acknowledgments

The authors wish to thank John Nesselroade and Steven Boker for their help with and review of this work, as well as Jason Allaire and Michael Marsiske for sharing their data for use in this document. Ryne Estabrook is supported by grant R25 DA026119-03 from the National Institute on Drug Abuse, and was previously supported by grant T32

AG20500-06 from the National Institute on Aging during the completion of this work; Kevin J. Grimm was supported by National Science Foundation REECE Program Grant (DRL-0815787).

References

- Allaire JC, Marsiske M. Intraindividual Variability May Not Always Indicate Vulnerability in Elders' Cognitive Performance. *Psychology & Aging*. 2005; 20(3):390–401. [PubMed: 16248699]
- Anstey KJ. Sensorimotor variables and forced expiratory volume as correlates of speed, accuracy and variability in reaction time performance in late adulthood. *Aging, Neuropsychology, and Cognition*. 1999; 6:84–95.
- Anstey KJ. Within-person variability as a dynamics measure of late-life development: new methodologies and future directions. *Gerontology*. 2004; 50(4):255–258. [PubMed: 15258432]
- Baltes, PB.; Nesselroade, JR. History and rationale of longitudinal research. In: Nesselroade, JR.; Baltes, PB., editors. *Longitudinal research in the study of behavior and development*. New York: Academic Press; 1979. p. 1-39.
- Bielak AAM, Hultsch DF, Strauss E, MacDonald SWS, Hunter M. Intraindividual variability is related to cognitive change in older adults: Evidence of within-person coupling. *Psychology and Aging*. 2010a; 25(3):575–586. [PubMed: 20853965]
- Bielak AAM, Hultsch DF, Strauss E, MacDonald SWS, Hunter M. Intraindividual variability in reaction time predicts cognitive outcomes 5 years later. *Neuropsychology*. 2010b September; Online first publication.
- Blanchard-Fields, Freda. Intraindividual variability and aging. *Psychology and Aging*. 2009; 24(Special Issue)(4)
- Boker, SM. Generalized local linear approximation of derivatives from time series. In: Chow, S-M.; Ferrer, E.; Hsieh, F., editors. *Statistical Methods for Modeling Human Dynamics: An Interdisciplinary Dialogue*. New Jersey: Lawrence Erlbaum Associates; 2010.
- Bowles, RP. Using intratask change item response models for the assessment of intraindividual variability. In: Chow, SM.; Ferrer, E.; Hsieh, F., editors. *Statistical methods for modeling human dynamics: An interdisciplinary dialogue*. Mahwah, NJ: Erlbaum; 2009.
- Butler A, Hokanson J, Flynn H. A comparison of self-esteem lability and low trait self-esteem as vulnerability factors for depression. *Journal of Personality and Social Psychology*. 1994; 66:166–177. [PubMed: 8126646]
- Chow SM, Hamaker EL, Fujita F, Boker SM. Repreasing time-varying cyclic dynamics using multiple-subject state-space models. *British Journal of Mathematical and Statistical Psychology*. 2009; 62:683–712. [PubMed: 19200409]
- Cohen, J.; Cohen, P.; West, SG.; Aiken, L. *Applied Multiple Regression/Correlation for the Behavioral Sciences*. 3. Mahwah, NJ: Erlbaum; 2003.
- Deary IJ, Der G. Reaction time, age, and cognitive ability: Longitudinal findings from age 16 to 63 years in representative population samples. *Aging, Neuropsychology and Cognition*. 2005; 12:187–215.
- Deboeck PR. Estimating dynamical systems: estimation hints from Sir Ronald A. Fisher. *Multivariate Behavioral Research*. 2010; 45(4):725–745.
- Duchek JM, Balota DA, Tse CS, Holtzman DM, Fagan AM, Goate AM. The utility of intraindividual variability in selective attention tasks as an early marker for Alzheimer's disease. *Neuropsychology*. 2009; 23(6):746–758. [PubMed: 19899833]
- Eid M, Diener E. Intraindividual variability in affect: Reliability, validity, and personality correlates. *Journal of Personality and Social Psychology*. 1999; 76:662–676.
- Eizenman D, Nesselroade JR, Featherman D, Rowe J. Intraindividual variability in perceived control in an older sample: The MacArthur successful aging studies. *Psychology and Aging*. 1997; 12:489–502. [PubMed: 9308096]
- Efron, B.; Tibshirani, R. *An Introduction to the Bootstrap*. Boca Raton, FL: Chapman & Hall; 1993.
- Fiske DW, Rice L. Intra-individual response variability. *Psychological Bulletin*. 1955; 57:217–250. [PubMed: 14371891]

- Fleeson W. Toward a Structure- and Process-Integrated View of Personality: Traits as Density Distributions of States. *Journal of Personality and Social Psychology*. 2001; 80:1011–1027. [PubMed: 11414368]
- Fox N, Porges S. The relation between neonatal heart period patterns and developmental outcome. *Child Development*. 1985; 56:28–37. [PubMed: 3987407]
- Fuentes K, Hunter MA, Strauss E, Hultsch DF. Intraindividual variability in cognitive performance in persons with chronic fatigue syndrome. *The Clinical Neuropsychologist*. 2001; 15:210–227. [PubMed: 11528543]
- Gerstorf D, Siedlecki KL, Tucker-Drob EM, Salthouse TA. Within-person variability in state anxiety across adulthood: Magnitude and associations with between-person characteristics. *International Journal of Behavioral Development*. 2009; 33(1):55–64.
- Haldane JBS. The measurement of variation. *Evolution*. 1955; 9:484.
- Hale S, Myerson J, Smith GA. Age, variability, and speed: Between-subjects diversity. *Psychology and Aging*. 1988; 3:407–410. [PubMed: 3268286]
- Hultsch DF, McDonald SWS, Hunter MA, Levy-Bencheton J, Strauss E. Intraindividual variability in cognitive performance in older adults: Comparison of adults with mild dementia, adults with arthritis, and healthy adults. *Neuropsychology*. 2000; 14:588–598. [PubMed: 11055261]
- Jensen AR. The importance of intraindividual variation in reaction time. *Personality and Individual Differences*. 1992; 13:869–881.
- Lecert T, Ghisletta P, Jouffray C. Intraindividual variability and level of performance in four visuo-spatial working memory tasks. *Swiss Journal of Psychology*. 2004; 63:261–272.
- Li SC, Aggen SH, Nesselroade JR, Baltes PB. Short-term fluctuations in elderly people's sensorimotor functioning predict text and spatial memory performance: The MacArthur successful aging studies. *Gerontology*. 2001; 47:100–116. [PubMed: 11287736]
- Licht CMM, de Geus EJC, Zitman FG, Hoogendijk WJG, van Dyck R, Penninx BWJH. Association between major depressive disorder and heart rate variability in the Netherlands study of depression and anxiety (NESDA). *Archives of General Psychiatry*. 2009; 65(12):1358–1367.
- MacDonald SWS, Hultsch DF, Dixon RA. Performance variability is related to change in cognition: Evidence from the Victoria Longitudinal Study. *Psychology and Aging*. 2003; 18:510–523. [PubMed: 14518812]
- MacDonalds SWS, Nyberg L, Bäckman L. Intra-individual variability in behavior: links to brain structure, neurotransmission and neuronal activity. *Trends in Neurosciences*. 2006; 29(8):474–480. [PubMed: 16820224]
- MacDonald SWS, Nyberg L, Sandblom J, Fischer H, Bäckman L. Increased response-time variability is associated with reduced inferior parietal activation during episodic recognition in aging. *Journal of Cognitive Neuroscience*. 2008; 20(5):779–786. [PubMed: 18201127]
- Marcotte, TD.; Roberts, D.; Rosenthal, TJ.; Heaton, RK.; Bentley, H.; Grant, I. the HRNC Group. Test-retest reliability of standard deviation of lane position as assessed on a PC-based driving simulator [Abstract]. *Proceedings of the Second International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design*; Iowa City, IA: Human Factors Research Program, University of Iowa; 2003. p. 199-200.
- McArdle JJ, Ferrer-Caja E, Hamagami F, Woodcock RW. Comparative longitudinal structural analyses of the growth and decline of multiple intellectual abilities over the life-span. *Developmental Psychology*. 2002; 38:115–142. [PubMed: 11806695]
- Nesselroade, JR. *Visions of aesthetics, the environment & development: The legacy of Joachim F Wohlwill*. Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc; 1991. The warp and the woof of the developmental fabric; p. 213-240.
- Nesselroade JR, Ram N. Studying intraindividual variability: what we have learned that will help us understand lives in context. *Research in Human Development*. 2004; 1(1 & 2):9–29.
- Penner JA, Shiffman S, Paty JA, Fritzsche BA. Individual differences in intraperson variability in mood. *Journal of Personality and Social Psychology*. 1994; 66(4):712–721. [PubMed: 8189348]
- R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing; Vienna, Austria: 2006. URL <http://www.R-project.org>

- Ram N, Rabbitt P, Stollery B, Nesselroade JR. Cognitive performance inconsistency: Intraindividual change and variability. *Psychology & Aging*. 2005; 20:623–633. [PubMed: 16420137]
- Rentrop M, Rodewald K, Roth A, Simon J, Walther S, Fielder P, Weisbrod M, Kaiser S. Intraindividual variability in high-functioning patients with schizophrenia. *Psychiatry Research*. 2010; 178(1):27–32. [PubMed: 20447695]
- Richman JS, Moorman JR. Physiological time-series analysis using approximate entropy and sample entropy. *Am J Physiol Heart Circ Physiol*. 2000; 278:H2039–H2049. [PubMed: 10843903]
- Salthouse TA. Implications of within-person variability in cognitive and neuropsychological functioning for the interpretation of change. *Neuropsychology*. 2007; 21:401–411. [PubMed: 17605573]
- Salthouse TA, Berish DE. Correlates of within-person (across-occasion) variability in reaction time. *Neuropsychology*. 2005; 19:77–87. [PubMed: 15656765]
- Sankaran M. Approximations to the non-central chi-square distribution. *Biometrika*. 1963; 50(1–2): 199–204.
- Schmiedek F. The dark side of the mean. Paper presented at the 11th Cognitive Aging Conference; Atlanta, GA. 2006 Apr.
- Schmiedek F, Lovden M, Lindenberger U. On the Relation of Mean Reaction Time and Intraindividual Reaction Time Variability. *Psychology and Aging*. 2009; 24(4):841–857. [PubMed: 20025400]
- Schretlen DJ, Munro CA, Anthony JC, Pearlson GD. Examining the range of normal intraindividual variability in neuropsychological test performance. *Journal of the International Neuropsychological Society*. 2003; 9:864–870. [PubMed: 14632245]
- Siegler RS. Cognitive variability: A key to understanding cognitive development. *Current Directions in Psychological Science*. 1994; 3:1–5.
- Siegler RS. Variability and infant development. *Infant Behavior & Development*. 2002; 25:550–557.
- Thorndike, RL. *Personnel selection: Test and measurement techniques*. New York: Wiley; 1949.
- Vickery CD, Sepehri A, Evans CC, Lee JE. The association of level and stability of self-esteem and depressive symptoms in the acute inpatient stroke rehabilitation setting. *Rehabilitation Psychology*. 2008; 53(2):171–179.
- Wang, L. Investigating reliabilities of intra-individual variability indicators. (under review) Manuscript submitted for publication
- Wilkinson RT, Allison S. Age and simple reaction time: Decade differences for 5,325 subjects. *Journals of Gerontology*. 1989; 44:29–35.
- Wright, BD.; Stone, MH. *Best test design*. Chicago: MESA Press; 1979.

Appendix A: R Script for Simulation

```
#ISD simulation
#Author: Ryne Estabrook
#Institution: University of Virginia, Department of Psychology
start.time<-proc.time()
#set working directory (folder where you want files saved)
setwd("C:/Users/Ryne/Documents/ISD")
#define values for four sim parameters
#sample size (scalar)
N<-100
#number of datasets per set of simulation values (scalar)
X<-100
#number of observations (vector)
T<-c(3:9,seq(10,100,5))
#measure reliabilities (vector)
M<-c(.6, .7, .8, .9, 1)
```

```

#Mean ISD (vector)
C<-c(.2, .4, .6, .8, 1, 1.2, 1.4, 1.6,1.8,2)
#SD of ISD/Mean ISD (vector)
#going above .33 runs increased risk of non-normality due to
#resampling negative ISDs
K<-c(.08, .16, .24, .32,.4, .48)
###end required user input###
#define vectors related to user parameters
R<-sqrt(M)
meas.error<-sqrt(1-M)
results<-matrix(rep(NA, 15*length(T)*length(R)*length(C)*length(K)*X),
ncol=15)
index<-0
m.index<-0
#create ids
id<-1:N
#start looping here
for(t in 1:length(T)){
  for(r in 1:length(R)){
    for(c in 1:length(C)){
      for(k in 1:length(K)){
        m.index<-m.index+1
        for(x in 1:X){
          index<-index+1
          #define individual true parameters, with resampling of negative ISDs
          true.mean<-rnorm(N, 0, 1)
          true.sd<-rnorm(N, C[c], K[k]*C[c])
          for(a in 1:N){
            while(true.sd[a]<=0){true.sd[a]<-rnorm(N, C[c], K[k]*C[c])}
          }
          # create blank data matrix
          true<-matrix(rep(NA, N*T[t]), nrow=N)
          obs<-matrix(rep(NA, N*T[t]), nrow=N)
          error<-matrix(rep(NA, N*T[t]), nrow=N)
          #create true and observed data
          for(i in 1:N){
            for(o in 1:T[t]){
              true[i,o]<-rnorm(1,true.mean[i], true.sd[i])
            }
          }
          error<-matrix(rnorm(N*T[t], 0, sqrt(1+C[c]^2)), nrow=N)
          obs<-R[r]*true+meas.error[r]*error
          #calculate ISDs
          ind.mean<-rep(NA,N)
          ISD<-rep(NA, N)
          for (i in 1:N){
            ind.mean[i]<-mean(obs[i, 1:T[t]])
            ISD[i]<-sd(obs[i, 1:T[t]])
          }
          results[index, 1]<-T[t]

```

```

results[index, 2]<-R[r]
results[index, 3]<-C[c]
results[index, 4]<-K[k]
results[index, 5]<-x
results[index, 6]<-cor(true.mean, ind.mean)
results[index, 7]<-cor(true.sd, ISD)
results[index, 8]<-mean(true.mean)
results[index, 9]<-mean(true.sd)
results[index, 10]<-mean(ind.mean)
results[index, 11]<-mean(ISD)
results[index, 12]<-sd(true.mean)
results[index, 13]<-sd(true.sd)
results[index, 14]<-sd(ind.mean)
results[index, 15]<-sd(ISD)
} #X
status<-paste("Condition ",m.index," of
",length(T
)*length(R)*length(C)*length(K),
"TRCK",t,r,c,k)
print(status)
} #T
} #R
} #C
} #K
#create data frame for results
t<-results[,1]
r<-results[,2]
c<-results[,3]
k<-results[,4]
x<-results[,5]
m<-results[,6]
i<-results[,7]
m.true.mean<-results[,8]
m.true.sd<-results[,9]
m.ind.mean<-results[,10]
m.isd<-results[,11]
s.true.mean<-results[,12]
s.true.sd<-results[,13]
s.ind.mean<-results[,14]
s.isd<-results[,15]
results<-data.frame(t,r,c,k,x,m,i,m.true.mean, m.true.sd, m.ind.mean, m.isd,
s.true.mean, s.true.sd, s.ind.mean, s.isd)
#write to file (adjust working directory)
#write.table(results, file= "ISDresults.txt", row.names=FALSE)
end.time<-proc.time()

```

Appendix B: Regression Results for Study 1

Table B1

Reliability of Individual Means: Main Effects

Effect	Estimate	SE	p	95% CI	
				Upper	Lower
Intercept	29.950	2.381	>.001	25.283	34.617
Timepoints ($R^2=.489$)					
Linear Effect (t)	-0.003	0.011	0.807	-0.024	0.018
Log Transformation ($\log(t)$)	0.744	0.312	0.017	0.133	1.355
Test Reliability ($R^2=.058$)					
Linear Effect (ρ)	-29.210	2.389	>.001	-33.892	-24.528
Log Transformation ($\log(\rho)$)	25.510	2.108	>.001	21.378	29.642
Mean ISD ($R^2=.422$)					
Linear Effect (μ_{ISD})	-34.560	0.701	>.001	-35.934	-33.186
Log Transformation ($\log(\mu_{\text{ISD}})$)	44.660	0.579	>.001	43.525	45.795
ISD Variability ($R^2=.001$)					
Linear Effect (σ_{ISD})	9.946	3.582	0.006	2.925	16.967
Log Transformation ($\log(\sigma_{\text{ISD}})$)	-1.320	0.809	0.103	-2.906	0.266

Note: Fisher-Z transformed reliability of individual means regressed on linear and natural logarithm transformations of the four simulation parameters and their interactions. All parameters and effects are untransformed, and R^2 are assigned first to main effects and then to interactions. Data are a single dataset of median reliability for all 100 datasets used in study 1. $R^2=.998$ for model including both main effects and two-way interactions. Inclusion of additional non-linear forms of the above variables (i.e. quadratic, square roots, etc) provides significant but negligible improvements in model fit ($\Delta R^2=.0005$) with no changes in substantive interpretation.

Table B2

Reliability of Individual Means: Interactions

Effect	Estimate	SE	p	95% CI	
				Upper	Lower
Timepoints * Test Reliability (R ² =.000)					
t* ρ	0.002	0.011	0.847	-0.019	0.023
t*log(ρ)	-0.004	0.009	0.693	-0.022	0.015
log(t)* ρ	-0.186	0.312	0.552	-0.798	0.426
log(t)*log(ρ)	0.292	0.276	0.290	-0.249	0.832
Timepoints * Mean ISD (R ² =.001)					
t* μ _{ISD}	0.001	0.000	>.001	0.001	0.002
t*log(μ _{ISD})	0.000	0.000	0.171	-0.001	0.000
log(t)* μ _{ISD}	-0.086	0.005	>.001	-0.096	-0.076
log(t)*log(μ _{ISD})	0.021	0.004	>.001	0.012	0.029
Timepoints * ISD Variability (R ² =.000)					
t* σ _{ISD}	0.000	0.001	0.919	-0.002	0.002
t*log(σ _{ISD})	0.000	0.000	0.832	-0.000	0.000
log(t)* σ _{ISD}	-0.018	0.027	0.516	-0.071	0.036
log(t)*log(σ _{ISD})	0.000	0.006	0.985	-0.012	0.012

				95% CI	
Effect	Estimate	SE	p	Upper	Lower
Test Reliability * Mean ISD (R ² =.027)					
ρ * μ _{ISD}	34.760	0.703	>.001	33.383	36.137
ρ *log(μ _{ISD})	-45.620	0.580	>.001	-46.757	-44.483
log(ρ)* μ _{ISD}	-28.860	0.620	>.001	-30.075	-27.645
log(ρ)*log(μ _{ISD})	37.080	0.512	>.001	36.076	38.084
Test Reliability * ISD Variability (R ² =.000)					
ρ * σ _{ISD}	-10.460	3.594	0.004	-17.504	-3.416
ρ *log(σ _{ISD})	1.377	0.812	0.090	-0.214	2.968
log(ρ)* σ _{ISD}	8.216	3.171	0.010	2.001	14.431
log(ρ)*log(σ _{ISD})	-1.109	0.716	0.122	-2.513	0.295
Mean ISD * ISD Variability (R ² =.000)					
μ _{ISD} * σ _{ISD}	0.103	0.061	0.094	-0.018	0.223
μ _{ISD} *log(σ _{ISD})	-0.011	0.014	0.438	-0.038	0.016
log(μ _{ISD})* σ _{ISD}	-0.153	0.051	0.002	-0.253	-0.054
log(μ _{ISD})*log(σ _{ISD})	0.015	0.011	0.181	-0.007	0.038

Table B3

Reliability of Individual Standard Deviations: Main Effects

Effect	Estimate	SE	p	95% CI	
				Upper	Lower
Intercept	35.480	(4.653)	<.001	26.360	44.600
Timepoints ($R^2=.357$)					
Linear Effect (t)	0.033	(0.021)	0.113	-0.008	0.074
Log Transformation ($\log(t)$)	-3.584	(0.609)	<.001	-4.778	-2.390
Test Reliability ($R^2=.151$)					
Linear Effect (ρ)	-35.580	(4.669)	<.001	-44.731	-26.429
Log Transformation ($\log(\rho)$)	33.690	(4.120)	<.001	25.615	41.765
Mean ISD ($R^2=.100$)					
Linear Effect (μ_{ISD})	-39.020	(1.370)	<.001	-41.705	-36.335
Log Transformation ($\log(\mu_{\text{ISD}})$)	57.940	(1.132)	<.001	55.721	60.159
Variability in ISD ($R^2=.290$)					
Linear Effect (σ_{ISD})	-1.970	(7.000)	0.778	-15.690	11.750
Log Transformation ($\log(\sigma_{\text{ISD}})$)	-6.254	(1.581)	<.001	-9.353	-3.155

Note: Fisher-Z transformed reliability of individual standard deviations regressed on linear and natural logarithm transformations of the four simulation parameters and their interactions. All parameters and effects are untransformed, and R^2 are assigned first to main effects and then to interactions. Data are a single dataset of median reliability for all 100 datasets used in study 1. $R^2=.988$ for model including both main effects and two-way interactions. Inclusion of additional non-linear forms of the above variables (i.e. quadratic, square roots, etc) provides significant but negligible improvements in model fit ($\Delta R^2 < .0001$) with no changes in substantive interpretation.

Table B4

Reliability of Individual Standard Deviations: Interactions

Effect	Estimate	SE	p	95% CI	
				Upper	Lower
Timepoints * Test Reliability (R ² =.012)					
t* ρ	-0.025	(0.021)	0.230	-0.066	0.016
t*log(ρ)	0.020	(0.019)	0.283	-0.016	0.056
log(t)* ρ	4.218	(0.611)	<.001	3.021	5.415
log(t)*log(ρ)	-3.018	(0.539)	<.001	-4.074	-1.962
Timepoints * Mean ISD (R ² =.012)					
t* μ _{ISD}	-0.001	(0.000)	0.102	-0.001	0.000
t*log(μ _{ISD})	0.000	(0.000)	0.134	0.000	0.001
log(t)* μ _{ISD}	-0.073	(0.010)	<.001	-0.093	-0.052
log(t)*log(μ _{ISD})	0.138	(0.009)	<.001	0.121	0.155
Timepoints * ISD Variability (R ² =.024)					
t* σ _{ISD}	-0.012	(0.002)	<.001	-0.016	-0.009
t*log(σ _{ISD})	0.002	(0.000)	<.001	0.001	0.003
log(t)* σ _{ISD}	0.103	(0.053)	0.054	-0.002	0.207
log(t)*log(σ _{ISD})	0.140	(0.012)	<.001	0.117	0.164
Test Reliability * Mean ISD (R ² =.024)					
ρ * μ _{ISD}	39.130	(1.373)	<.001	36.439	41.821
ρ *log(μ _{ISD})	-57.980	(1.134)	<.001	-60.203	-55.757
log(ρ)* μ _{ISD}	-34.310	(1.212)	<.001	-36.686	-31.934
log(ρ)*log(μ _{ISD})	49.880	(1.001)	<.001	47.918	51.842
Test Reliability * ISD Variability (R ² =.010)					
ρ * σ _{ISD}	2.654	(7.024)	0.706	-11.113	16.421
ρ *log(σ _{ISD})	6.470	(1.586)	<.001	3.361	9.579
log(ρ)* σ _{ISD}	-3.839	(6.198)	0.536	-15.987	8.309
log(ρ)*log(σ _{ISD})	-4.339	(1.400)	0.002	-7.083	-1.595
Mean ISD * ISD Variability (R ² =.008)					
μ _{ISD} * σ _{ISD}	-0.193	(0.120)	0.106	-0.428	0.041
μ _{ISD} *log(σ _{ISD})	-0.101	(0.027)	<.001	-0.154	-0.048
log(μ _{ISD})* σ _{ISD}	-0.004	(0.099)	0.967	-0.198	0.190
log(μ _{ISD})*log(σ _{ISD})	0.238	(0.022)	<.001	0.194	0.281

Figure 1.

Median reliability of intraindividual means for $\mu_{ISD} = [.2, .4, .6, 1.0, 2.0] * \sigma_{MEAN}$. For all curves, maximum reliability occurs in the minimum intraindividual variability condition (for $\mu_{ISD} = .2 * \sigma_{MEAN}$) and is indicated by a solid line. Number of occasions from 0 to 100; reliability from 0 to 1.

Figure 2.

Median reliability of intraindividual standard deviations for $\mu_{ISD}=[.2, .4, .6, 1.0, 2.0] * \sigma_{MEAN}$. For all curves, maximum reliability occurs in the maximum intraindividual variability condition (for $\mu_{ISD}=2.0 * \sigma_{MEAN}$), and is indicated by the highest number of line breaks. Number of occasions from 0 to 100; reliability from 0 to 1.

Figure 3. Reliability of ISDs and estimated power in the Eizenman, et al (1997) logistic regression presented as functions of sample size.

Figure 4. Estimated logistic regression power for the Eizenman, et al (1997) study as a function of the reliability of ISDs.

Table 1

Distributional Characteristics of ISDs in Previous Research

Study	T	IIV	μ_{ISD}/σ_{MEAN}	σ_{ISD}/μ_{ISD}	Notes
Salthouse, 2007	3	ISD	.28 – .78	.50 – .84	16 cognitive measures spanning vocabulary, reasoning, spatial abilities and memory.
Vickery, et al, 2008	8	ISD	.49–.62	.51–.73	Twice-daily measures of self-esteem
Lecerf, Ghisletta & Jouffray, 2004	15	ISD, CV, CIP	1.06 – 2.06	.15 – .30	4 visio-spatial working memory tasks.
Bertie, 1969	20	ISD	.47 – 1.30	.73 – 1.03	6 Repetitive Psychometric Measures (Moran & Mefferd, 1959).
Deary & Der, 2005	20	ISD	.80 – 1.23	.58 – .71	Simple reaction time.
Deary & Der, 2005	20	ISD	1.30 – 1.72	.25 – .29	Reaction time for a four-choice reaction time measure.
Eizenman, et al, 1997	25	ISD	.51	.39	Variability in perceived control.
Eid & Diener, 1999	52	ISD		.29–.38	Variability in Affect.
Allaire & Marsiske, 2005	120	ISD	.31 – .59	.19 – .27	Measures of inductive reasoning, memory and perceptual speed.
MacDonald, et al, (2008)	160	ISD	.02 – .03	.16	Response time for word recognition task.
Fox & Porges, 1985	>100	ISD		.34 – .43	Variability in infant heart rate over a three minute period.

Note. Number of occasions (T), measures used (IIV) and distributional characteristics for studies which provide some distributional information. μ_{ISD}/σ_{MEAN} denotes the ratio of the mean ISD to the standard deviation of individual means, indicating the magnitude of intraindividual differences relative to variance in interindividual differences. σ_{ISD}/μ_{ISD} denotes the ratio of intraindividual differences in IIV relative to mean IIV. ISD=Individual standard deviation; CV=coefficient of variation; CIP=Coefficient of Inconsistency of Performance.

Table 2

Summary of Simulation Parameters

Parameter	Levels	Values
μ_{ISD}	10	0.2, 0.4, 0.6, 0.8, 1.0, 1.2, 1.4, 1.6, 1.8, 2.0
σ_{ISD}	6	0.08, 0.16, 0.24, 0.32, 0.40, 0.48
Occasions	26	3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100
Reliability	5	0.60, 0.70, 0.80, 0.90, 1.00

Note. Summary of parameters in simulation study. μ_{ISD} parameters are in σ_{MEAN} units, such that $\mu_{\text{ISD}} = 0.2$ indicates that $\mu_{\text{ISD}} = 0.2 * \sigma_{\text{MEAN}}$. σ_{ISD} is in μ_{ISD} units, such that $\sigma_{\text{ISD}} = 0.08$ indicates that $\sigma_{\text{ISD}} = 0.08 * \mu_{\text{ISD}}$.

Table 3

Logistic regression of 5-year mortality on IIV in perceived control

Predictor	Estimate	SE	p	Odds	95% CI
Intercept	-3.880	(1.4987)	0.010		
IIV	1.700	(0.847)	0.045	5.476	1.041 28.807

Note. Logistic regression results from predicting 5-year mortality from ISD using data from Eizenman, et al (1997).