frontiers in Psychology

METHODS
published: 17 April 2018
doi: 10.3389/fpsyg.2018.00294

Precision, Reliability, and Effect Size of Slope Variance in Latent Growth Curve Models: Implications for Statistical Power Analysis

Andreas M. Brandmaier [1,2]*, ***Timo von Oertzen*** [1,3,4], ***Paolo Ghisletta*** [5,6,7], ***Ulman Lindenberger*** [1,2,8] ***and Christopher Hertzog*** [9]

[1] Center for Lifespan Psychology, Max Planck Institute for Human Development, Berlin, Germany, [2] Max Planck UCL Centre for Computational Psychiatry and Ageing Research, Berlin, Germany, [3] Department of Psychology, University of Virginia, Charlottesville, VA, United States, [4] Department für Psychologie, Universität der Bundeswehr München, Neubiberg, Germany, [5] Faculty of Psychology and Educational Sciences, University of Geneva, Geneva, Switzerland, [6] Faculty of Psychology (French), Swiss Distance Learning University, Brig, Switzerland, [7] Swiss National Center of Competences in Research LIVES-Overcoming Vulnerability: Life Course Perspectives - University of Geneva, Geneva, Switzerland, [8] Department of Political and Social Sciences, European University Institute, Fiesole, Italy, [9] School of Psychology, Georgia Institute of Technology, Atlanta, GA, United States

OPEN ACCESS

Edited by:
Holmes Finch,
Ball State University, United States

Reviewed by:
Michael C. Neale,
Virginia Commonwealth University, United States
Gregory R. Hancock,
University of Maryland, College Park, United States

****Correspondence:***
Andreas M. Brandmaier
brandmaier@mpib-berlin.mpg.de

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *04 September 2017*
Accepted: *21 February 2018*
Published: *17 April 2018*

Citation:
Brandmaier AM, von Oertzen T, Ghisletta P, Lindenberger U and Hertzog C (2018) Precision, Reliability, and Effect Size of Slope Variance in Latent Growth Curve Models: Implications for Statistical Power Analysis. Front. Psychol. 9:294. doi: 10.3389/fpsyg.2018.00294

Latent Growth Curve Models (LGCM) have become a standard technique to model change over time. Prediction and explanation of inter-individual differences in change are major goals in lifespan research. The major determinants of statistical power to detect individual differences in change are the magnitude of true inter-individual differences in linear change (LGCM slope variance), design precision, alpha level, and sample size. Here, we show that design precision can be expressed as the inverse of *effective error.* Effective error is determined by instrument reliability and the temporal arrangement of measurement occasions. However, it also depends on another central LGCM component, the variance of the latent intercept and its covariance with the latent slope. We derive a new reliability index for LGCM slope variance—effective curve reliability (ECR)—by scaling slope variance against effective error. ECR is interpretable as a standardized effect size index. We demonstrate how effective error, ECR, and statistical power for a likelihood ratio test of zero slope variance formally relate to each other and how they function as indices of statistical power. We also provide a computational approach to derive ECR for arbitrary intercept-slope covariance. With practical use cases, we argue for the complementary utility of the proposed indices of a study's sensitivity to detect slope variance when making *a priori* longitudinal design decisions or communicating study designs.

Keywords: linear latent growth curve model, statistical power, effect size, effective error, structural equation modeling, reliability, longitudinal study design

INTRODUCTION

People differ in rates of change in many functional domains, both at behavioral and neural levels of analysis (e.g., Lindenberger, 2014). Describing, explaining, and modifying between-person differences in change are the central goals of lifespan research (Baltes and Nesselroade, 1979; Baltes et al., 2006; Ferrer and McArdle, 2010; McArdle and Nesselroade, 2014). Successful detection of inter-individual differences in change requires longitudinal (within-person) study designs with adequate statistical power.

Statistical power, the likelihood of rejecting a false null hypothesis, is a function of sample size, test size (that is, alpha level), population effect size, and the precision of measurement. In their considerations on optimal power in longitudinal designs, Rast and Hofer (2014) illustrated that Willett's (1989) reliability of the growth rate—as captured by Growth Rate Reliability (GRR)—plays an important role in determining statistical power in longitudinal designs using linear latent growth curve models. Latent growth curve models (LGCM) have become a commonly used analysis technique to capture change in longitudinal data (e.g., Meredith and Tisak, 1990; Ferrer and McArdle, 2003, 2010; Duncan et al., 2006). Here we formally show why GRR is a useful index of statistical power to detect slope variance in LGCM. However, we also reveal that GRR may provoke misleading interpretations because it remains constant over conditions that alter statistical power to detect individual differences in LGCM slopes using likelihood ratio tests. Specifically, the magnitude of stable individual differences (intercept variance) influences the power of 1-*df* likelihood ratio (LR) tests on slope variance. However, this is not reflected by GRR. Furthermore, the existence of intercept-slope covariance implies the existence of slope variance, yet GRR does not change as a function of intercept-slope covariance. These observations motivate and justify the development of a new, more comprehensive reliability measure that is coherent with the statistical power of LR tests.

Central factors affecting statistical power to detect individual differences in linear change include (a) the time elapsing from the beginning to the end of a study (henceforth referred to as total study duration); (b) the number of measurement occasions and their distribution over total study duration; (c) for each measured construct, the precision of the measurement instruments administered at each occasion; and (d) the number of participants, including any partial sampling designs, such as withholding assessments to estimate practice effects (Baltes et al., 1988), or other forms of planned missingness (McArdle, 1994; Wu et al., 2016). Simulations of statistical power to detect LGCM slope variances have already taken at least some of these factors into account when evaluating questions about statistical power (e.g., Hertzog et al., 2008; Wänström, 2009; Rast and Hofer, 2014; Ke and Wang, 2015).

In this article, we present a formal derivation of measures of precision, reliability, and statistical power that conform to LR tests of slope variance. All three types of measures can be regarded as gauging a design's sensitivity to detect individual differences in linear slopes (henceforth simply referred to as *change sensitivity*). We derive a new measure of reliability, which we term *effective curve reliability* (ECR). ECR can be regarded as a standardized effect size measure that is coherent with power to detect individual differences in linear change using LR tests. We formally outline the conditions under which GRR and ECR are identical, as well as the conditions under which they diverge. We also discuss potential applications of such measures in the service of *a priori* longitudinal design decisions.

EFFECTIVE ERROR AND EFFECT SIZE FOR SLOPE VARIANCE

Linear Latent Growth Curve Models

A linear LGCM for balanced designs assumes linear growth in a variable x_{ij} over time points t_j with $j = 1, \dots, M$ on $i = 1, \dots, N$ persons with inter-individual differences in the intercept and the linear slope, which are represented by latent variable means and (co)variances. In a linear LGCM, the mean vector, μ, and the covariance matrix, Σ, of the observed variables are a function of factor loadings, Λ, latent variables' intercepts, ν, a latent variance-covariance matrix, Ψ, and a residual variance-covariance matrix, Θ (e.g., Bollen, 1989):

$$\begin{aligned} \Sigma &= \Lambda \Psi \Lambda' + \Theta \\ \mu &= \Lambda \nu \end{aligned} \tag{1}$$

Under the assumption of homoscedastic and uncorrelated residual errors, we define

$$\begin{aligned} \Lambda &= \begin{bmatrix} 1 & t_1 \\ 1 & t_2 \\ 1 & \vdots \\ 1 & t_M \end{bmatrix} \\ \Psi &= \begin{bmatrix} \sigma_I^2 & \sigma_{IS} \\ \sigma_{IS} & \sigma_S^2 \end{bmatrix} \\ \nu &= \begin{bmatrix} \mu_I \\ \mu_S \end{bmatrix} \\ \Theta &= \begin{bmatrix} \sigma_\epsilon^2 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma_\epsilon^2 \end{bmatrix} \end{aligned} \tag{2}$$

with the number of measurement occasions M at times t_1 to t_M, the homogeneous residual error σ_ϵ^2, the mean μ_I, and variance σ_I^2 of the latent intercept, and the mean μ_S, and the variance of the latent slope σ_S^2, and the latent intercept-slope covariance σ_{IS}. One of the t_j, $j \in \{1, \dots, M\}$ can be conveniently fixed to zero to identify the slope and intercept. Often t_1 is fixed to 0 to define the intercept at the first occasion of measurement ($j = 1$), but other choices are possible.

In principle, SEM, multi-level model (e.g., Goldstein, 1986), and random-effects model (e.g., Laird and Ware, 1982) representations of the LGCM are empirically and analytically identical (e.g., Rovine and Molenaar, 2000; Curran, 2003). For the sake of clarity, we restrict our perspective to a treatment from a SEM approach but our results and conclusions equally apply to algebraically equivalent growth curve modeling approaches.

Specific and Generalized Variance Tests

LGCM captures individual differences in rates of change in the between-person variance in linear slopes (e.g., Singer and Willett, 2003). In the univariate case, a formal statistical test of the null hypothesis of zero slope variance can be obtained by two types of

LR χ^2-tests. Each of the tests is based on the difference in -2LL (log-likelihood) goodness of fit of two nested models (meaning that the parameters of the simpler model represent a subset of the parameters of the larger model):

(a) A restricted model specifying that Ψ has a free intercept variance (σ_I^2) but fixing both the slope variance ($\sigma_S^2 = 0$) and the intercept-slope covariance ($\sigma_{IS} = 0$), and a second model freely estimating all three parameters in Ψ. This 2-*df* LR test relies on information in both the slope variance and intercept-slope covariance to test the null hypothesis of zero population slope variance.

(b) An alternative 1-*df* LR test ignores the information in the intercept-slope covariance (by assuming it to be zero in the less restricted model) and tests the loss of information by fixing only the variance (that is, $\sigma_S^2 = 0$) in the more restricted model.

In practice, a third test is more often employed, a 1-*df* Wald test (Bollen, 1989), generated from an estimated LGCM solution: the estimated slope variance divided by its estimated standard error of estimate. There are principled reasons in actual practice to prefer the 2-*df* LR test over the 1-*df* LR test or the Wald test (see section Discussion), but the 1-*df* LR test is particularly useful for change-sensitivity analysis in longitudinal research design when intercept-slope covariance is not considered substantial.

Prelude on Change Sensitivity and Statistical Power

How can one best examine the sensitivity of linear change measurements? von Oertzen (2010) and von Oertzen and Brandmaier (2013) developed the concept of *effective error*, which captures the precision with which properties of the latent slope in a LGCM can be measured. In a single measurement, measurement error of an instrument quantifies the amount of unsystematic variance (and its inverse quantifies the precision of measurement). In a longitudinal design, effective error variance quantifies the magnitude of unsystematic variance in the outcome of interest over repeated measures (and its inverse is proportional to the precision of the repeated measures design to measure a given outcome). Effective error is thus a measure of sensitivity to detect a hypothesized effect. It can be computed as a weighted composite of all variance sources in a LGCM that potentially distort the measurement of the effect of interest. As we will show below, major components of effective error for LGCM linear slope variance are the temporal arrangement of measurement occasions and instrument reliability. Effective error is also scaled by intercept variance expressing the stable between-person differences. As such, effective error is primarily an index of change sensitivity that requires no assumption about the true absolute (or, unstandardized) effect size, sample size, or test size (significance level of the statistical test). Conceptually, effective error can be construed as the error a researcher would have experienced if one were able to measure the latent construct of interest (here, variance in linear change) with a single measurement. Despite the fact that this one-shot measurement of change can never be attained in reality, it can be shown that the corresponding minimal study design has equivalent statistical power to the original study design (von Oertzen, 2010).

How is the concept of effective error relevant for understanding statistical power? Recall that classic treatments of power consider two main parameters that govern the power of a study to reject the null hypothesis about a parameter value when it is false (e.g., Cohen, 1988): effect size (as it relates to the non-centrality parameter of the sampling distribution), and dispersion of the sampling distribution, with two main constituents: sample size and "error variance." Consider the dependency graph in **Figure 1**, which we use as a scheme to develop a hierarchical conceptualization of change sensitivity. We start with effective error and arrive at statistical power. From top to bottom, the measures require an increasing number of assumptions to be made about design factors or population values (represented as white boxes): measurement precision, true effect size, and test size. At the top of the hierarchy, effective error is a measure of precision of a repeated-measures design, construed broadly to imply all sources of variance that contribute to the dispersion in the sampling distribution of the test statistic. If an unstandardized population effect size (i.e., the absolute value of the population slope variance parameter) is known or assumed, one can derive a standardized effect size that rescales the parameter against effective error to generate a metric-free standardized effect size. Put differently, standardized effect size can be conceptualized as the reliability of the population effect, for instance, slope variance, in relation to the precision with which it can be measured (e.g., Willett, 1989; Kelley and Preacher, 2012). When scaling the standardized effect size with sample size, we obtain *evidence* (the magnitude of evidence against the null hypothesis) as a measure of change sensitivity. Once one is willing to determine criteria of the statistical test (i.e., the Type I error rate), statistical power itself is possibly the ultimate index of change sensitivity as it determines the probability of correctly rejecting the null hypothesis when it is false. We stress that no single measure of change sensitivity is generally superior to any other; instead, they complement each other and convey different perspectives to change sensitivity, and each is based on different population values or design factors (see **Figure 1**).

Effective Error of the Specific Variance Test

The analytical derivations of effective error, standardized effect size, and statistical power in this paper are based on the assumption of no intercept-slope covariance, which renders them coherent with the assumptions of the specific 1-*df* LR test of the null hypothesis of zero slope variance. When the true intercept-slope covariance parameter is different from zero, the 2-*df* test should be used because (a) it has greater power to reject the null hypothesis (Pinheiro and Bates, 2000; Hertzog et al., 2008), and (b) the 1-*df* specific variance test is misspecified, as it represents a LR test of two restricted models both incorrectly specifying zero intercept-slope covariance. Currently, there is no known closed-form solution for ECR in the general case. However, in section Generalized ECR, we will provide an algorithm to compute the ECR numerically for cases in which the intercept-slope covariance is non-zero. This yields a standardized effect size

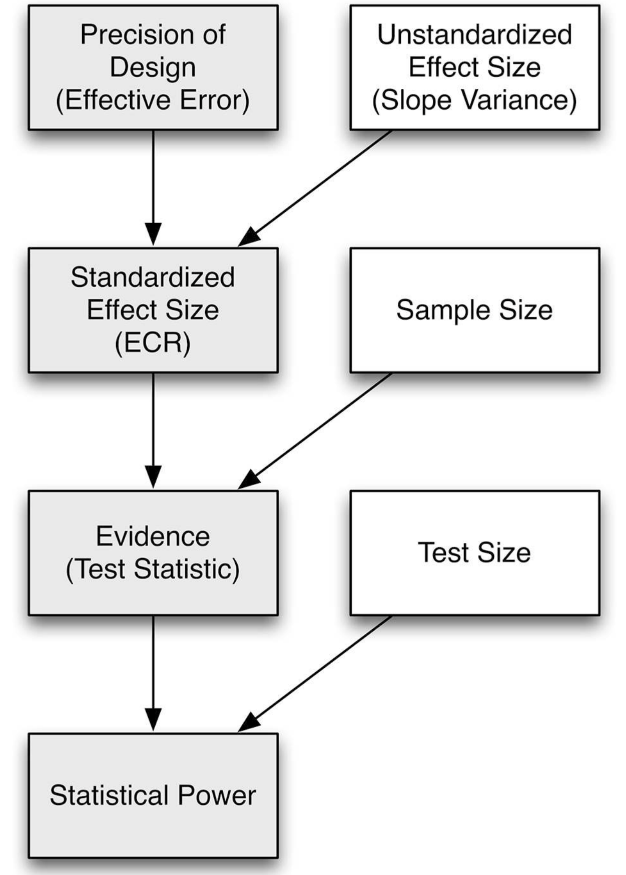
Precision of Design (Effective Error)
Unstandardized Effect Size (Slope Variance)
Standardized Effect Size (ECR)
Sample Size
Evidence (Test Statistic)
Test Size
Statistical Power

FIGURE 1 | The relation of effective error (as the inverse of precision of a study design), reliability, standardized effect size, evidence, and statistical power is shown as a dependency graph. Gray boxes indicate measures, and white boxes indicate design factors or population values. Each measure conveys change sensitivity, with the number of assumptions to be made about design factors and population values increasing from top to bottom.

metric coherent with the statistical power of the 2-*df* generalized variance test. We proceed by presenting the analytical solutions of precision, reliability, and power for the 1-*df* test. This test is not only practically relevant but enables us to set out a comprehensive perspective on the roles of GRR and ECR in determining statistical power to detect individual differences in linear change.

We now review a formal solution that allows one to derive the proposed measures of change sensitivity, starting with the variance of the effective error, for tests of linear slope variance in a longitudinal study. We use power-equivalent[1] transformations, defined as transformations that structurally change the model but preserve statistical power for a given hypothesis (see MacCallum et al., 2010; von Oertzen, 2010), to transform the original complex design into a power-equivalent model with minimal parameterization (see **Figure 2**). The resulting condensed model specifies direct observation of the slope variance (which is, of course, not possible in reality) and its corresponding residual, effective error, which captures the aggregate effects of all variables on precision of measurement.

Effective error serves as a basis for comparison of change sensitivity across multiple studies that differ in specification (von Oertzen and Brandmaier, 2013). The major practical value of an expression for effective error is that models with equal effective error have equal power to reject the null hypothesis of non-zero slope variance, even if they are structurally different (under the assumption that all parameters but the slope variance are assumed fixed; see von Oertzen, 2010).

Here, we argue in favor of measures of change sensitivity that are coherent with power for the specific variance test, that is, a test that assumes no intercept-slope covariance (see *null model* in Section Specific and generalized variance test). Extending previous work (von Oertzen and Brandmaier, 2013), we derive a general-case, analytical solution of effective error for known and fixed non-zero intercept-slope covariance (see section 1 in Supplementary Material). In the simplifying case of no intercept-slope covariance, we obtain:

$$\sigma^2_{eff} = \frac{\sigma^2_\epsilon}{\sum_{j=1}^{M} t_j^2 - \eta \left(\sum_{j=1}^{M} t_j \right)^2} \tag{3}$$

with

$$\eta = \frac{1}{M + \frac{\sigma^2_\epsilon}{\sigma^2_I}} = \frac{\sigma^2_I}{M\left(\sigma^2_I + \sigma^2_\epsilon / M\right)} = \frac{ICC_2}{M} \tag{4}$$

and σ^2_ϵ being the LGCM residual variance, σ^2_I being the intercept variance, t_j being the time point of measurement occasion j (varying from 1 to M) and ICC_2 being the intra-class correlation coefficient for M repeated measures.

As can be seen from Equations (3) and (4), the effective error depends directly on ICC$_2$, and, thus, also on intercept variance σ^2_I. Effective error increases with increasing ICC$_2$ such that change sensitivity decreases when intercept variance is large in comparison to residual variance, or in other words, change sensitivity increases with the amount of shrinkage captured as ICC$_2$. In summary, effective error increases with residual error variance σ^2_ϵ, increases with intercept variance σ^2_I, decreases with the number of measurement occasions M, and quadratically decreases with the measurement time points t_j, corresponding to an asymptotic decrease proportional to the variance of the distribution of measurement occasions.

RELIABILITY OF RANDOM SLOPES IN LGCM

Reliability is typically conceived as a measure of consistency, precision, or repeatability of measurement in a population

[1] We are not referring here to equivalent models, that is, cases in which models with different parameter specifications generate identical derived mean and covariance structures (Stelzl, 1986; Lee and Hershberger, 1990; MacCallum et al., 1993). Instead, we are referring to models that, even if qualitatively dissimilar in terms of model parameter specification, generate identical effective error for the effect of interest.

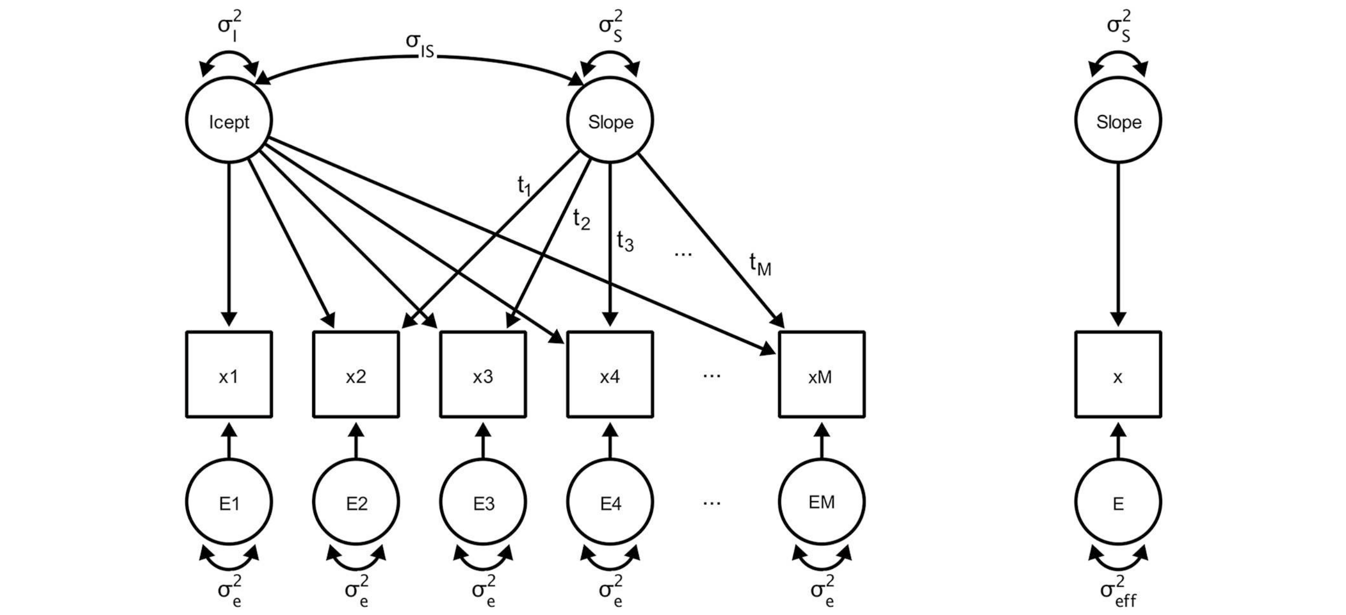

FIGURE 2 | Left: A LGCM with a latent intercept "Icept" with variance σ_I^2 and a latent slope "Slope" with variance σ_S^2, and intercept-slope covariance σ_{IS}. A single variable is observed for a total of *M* occasions of measurement (labeled x1 to xM) and an error term with variance σ_e^2. **Right**: A hypothetical minimal and power-equivalent model representing a design in which the slope is measured directly with a single source of effective error.

(Kline, 1998). In practice, a reliable measurement instrument reproduces the same or similar scores when measuring the same construct from one assessment to another. This was the principal justification used by Willett (1989) in deriving GRR, which he treated as an extension of the reliability of a simple difference score to cover the reliability of slope variance across larger numbers of measurement occasions. We take a similar approach here, but directly link the concept of reliability of the slope variance to effective error, grounding that concept in the basis for statistical power of the LR test of slope variances. In a hypothetical model, in which slope variance could be directly measured, total observed slope variance would be due to true score variation, that is, individual differences in linear slope and measurement error variance.

We depart from Willett (1989) in the error term used to scale reliability of measurement for LGCM slopes, using the effective error σ^2_{eff} instead, to rescale the true effect of inter-individual differences σ^2_S in order to yield an alternative reliability index for slopes, *effective curve reliability* (ECR). ECR scales slope variance as a proportion of the sum of slope variance and slope measurement error:

$$ECR = \frac{\sigma^2_S}{\sigma^2_S + \sigma^2_{eff}} \tag{5}$$

ECR is thus an index that ranges between zero and one, and becomes larger as population slope variance increases, or as the precision of the longitudinal study design reduces effective error. Conceptualized as reliability, it can also be regarded as a standardized effect size statistic (Kelley and Preacher, 2012).

When intercept-slope covariance is negligible, ECR is obviously similar to the aforementioned GRR. In fact, GRR is related to ECR and equal to it in a few limiting cases. We illustrate the relations of ECR and GRR using the equations above that assume that $\sigma_{IS} = 0$. One noteworthy limiting case is when $\text{ICC}_2 = 1$, where individual differences account for all of the reliable sources of variance in the dependent variable over time; that is, each measurement within a person strongly depends on the others and the information gained about intercept differences by any given measurement within one person is largely redundant:

$$\tilde{\sigma}^2_{eff} = \frac{\sigma^2_\epsilon}{SST} \tag{6}$$

with *SST* being the squared deviations of t_j about their mean. Formally:

$$SST = \sum_{j=1}^{M} \left(t_j - \bar{t}\right)^2 \text{ and } \bar{t} = \sum_{j=1}^{M} \frac{t_j}{M} \tag{7}$$

Inserting this special-case effective error, $\tilde{\sigma}^2_{eff}$, into the definition of ECR, we obtain GRR:

$$GRR = \frac{\sigma^2_S}{\sigma^2_S + \sigma^2_\epsilon / SST} \tag{8}$$

A second limiting case is when $\text{ICC}_2 = 0$ renders $\eta = 0$ in Equation (3), and produces a similar simplification on effective error. It renders ECR = GRR because there is no intercept variance. Otherwise, ECR and GRR diverge by the degree to which ICC_2 is smaller than one, that is, to the extent that intercept variance is relatively small or residual error variance large. As a final special case, GRR = ECR when the sum across time points is zero, $\sum t_i = 0$, which is the case in models in which the intercept is centered with respect to t:

$$\sigma^2_{eff} = \frac{\sigma^2_{\epsilon}}{\sum_{j=1}^{M} t_j^2 - \eta \cdot \left(\sum_{j=1}^{M} t_j\right)^2} = \frac{\sigma^2_{\epsilon}}{\sum_{j=1}^{M} t_j^2} = \frac{\sigma^2_{\epsilon}}{SST} = \tilde{\sigma}^2_{eff} \quad (9)$$

Otherwise, ECR will differ from GRR. As we show next, ECR's direct connection to intercept variance ensures that it, and not GRR, is coherent with power of the 1-*df* LR test for slope variance.

Monte Carlo Simulation of ECR, GRR, and Power of the 1-*df* Slope Variance Test

To illustrate that the statistical power of the 1-*df* LR test is dependent on intercept variance, as expected based on the definition of effective error, we ran selected Monte Carlo simulations manipulating intercept variance. We hypothesized that the effect of intercept variance on the power of the LR test is captured by ECR but not by GRR. We also simulated the Wald test as defined in section Specific and Generalized Variance Tests. The baseline model for the simulations had three measurement occasions with equally spaced intervals and a sample size of 100 (cf. **Figure 3**). For illustration, slope variance was set to 2, residual variance was varied between 10, 20, and 30 (as indicated by the different lines) and intercept variance was varied between 0.1 and 200 to cover a wide range of possible ICC_2 values. The intercept was defined at the first occasion of measurement by fixing the basis vector loading t_1 to zero. The critical purpose of the simulation is to show the effect of ICC_2 on power of the 1-*df* LR test and the Wald test.

Figure 3 depicts the simulation results based on 1,000 Monte Carlo replications of each condition. The left-hand panel plots power as a function of varying intercept variance scaled as ICC_2. Clearly power of the specific variance LR test (solid lines) is strongly affected by the proportion of intercept variance, as predicted. However, this is not the case for the Wald test (dashed lines). The right-hand panel plots ECR and GRR as a function of the same ICC_2. GRR, plotted in dashed lines, is unaffected by ICC_2. In contrast, ECR (solid lines of right-hand panel) tracks the effects of ICC_2 on estimated power. ECR does not ignore intercept variance and thus conforms to the effective error and power of the 1-*df* LR test, whereas GRR is in line with the Wald test, which ignores intercept variance. The simulation shown in **Figure 3** is grounded in a particular set of LGCM parameter values, but reflects a general principle of coherence of ECR with power under the given assumptions.

In **Figure 4**, we extended this simulation to show how statistical power and ECR change as a function of either shifting measurement time points in time (including centering the design) and prolonging or shortening total study duration by a multiplicative factor. The baseline model is a linear change model with three measurement occasions, baseline, 2 years after, and 4 years after beginning of the study. Intercept, slope, and residual variance were modeled after the values reported by Rast and Hofer (2014) for the *Memory-in-Reality* measure from the OCTO-Twin study (Johansson et al., 1999, 2004). **Figure 4** shows that both ECR and statistical power increase when total study duration is increased as measurement occasions are added. Importantly (but not shown in the figure), GRR does not change as a function of time shift (because *SST* is constant under time shift).

In summary, both ECR and GRR connect closely to statistical power to detect non-zero slope variance in LGCMs because they are standardized effect size measures for the slope variance, which makes them independent of the original metric in the dependent variable. Under the assumption of no intercept-slope correlation and if one wanted to use a Wald test, GRR is the appropriate effect size statistic. However, ECR is superior to GRR when computing effect sizes for the 1-*df* specific variance test because it traces statistical power more accurately.

If one assumes for the sake of argument that the selection of studies by Rast and Hofer (2014) is representative for a particular future longitudinal study of individual differences in psychological aging, one can compare expected differences between ECR and GRR in the range of plausible parameter values from their reported data sets. We used the reported data from Rast and Hofer (2014) and observed ICC_2 between 0.62 and 0.98 (median: 0.89), which may—depending on the density of measurements—lead to substantial differences in the predictions about change sensitivity as captured by ECR vs. GRR. That is, within the reported range of the studies used in their meta-analysis, *post-hoc* GRR estimates range from 0.02 to 0.72 (median = 0.36) and ECR estimates range from 0.03 to 0.76 (median = 0.40).

Below, we illustrate that this magnitude of difference can have a serious impact on design decisions. Before doing so, we corroborate our perspective on the usefulness of ECR as a proxy for statistical power by formally showing how ECR relates to statistical power in the specific variance test.

FORMALIZING THE CONNECTION OF ECR TO STATISTICAL POWER OF THE 1-*df* SLOPE VARIANCE TEST

In the typical fourfold table capturing hybrid Neyman–Pearson inference, the probability of Type I errors is denoted α and that of Type II errors is denoted as β. The power of a LR test (1 − β) to correctly reject a (false) null hypothesis is defined as the probability of obtaining a value of the test statistic that is larger than a critical value derived from the sampling distribution of that test statistic under the null hypothesis:

$$1 - \beta = \int_{\chi^2_{\nu,0}(\alpha)}^{\infty} \chi^2{}_{\nu,\lambda}\,(x)\,dx \quad (10)$$

with $\chi^2_{\nu,\lambda}$ being a non-central χ^2-distribution with ν degrees of freedom and non-centrality parameter λ, and $\chi^2_{\nu,0}(\alpha)$ denoting the *critical value* for test size α . Note that the start of the integral is the critical value for the null distribution, which is usually a

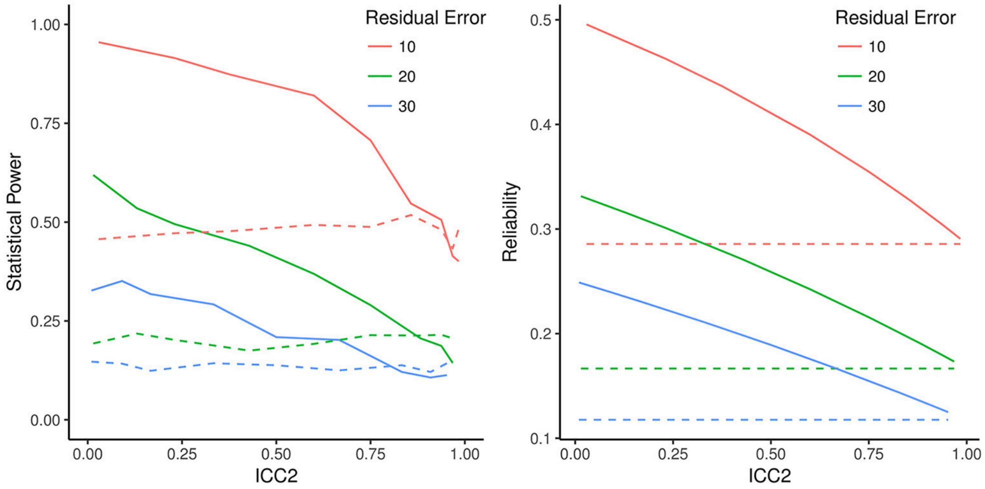

FIGURE 3 | Statistical power for specific variance test and Wald test, and effective curve reliability (ECR) and Growth Rate Reliability (GRR), as an effect of varying intercept variance scaled as the intra-class correlation coefficient (ICC$_2$). **Left**: Lines are Monte Carlo estimates of statistical power of the specific variance test (solid) and Wald test (dashed) with each line representing a model with different residual error variance. **Right**: Corresponding ECR (solid lines) and GRR (dashed lines). GRR is constant for all intercept variance values (scaled as ICC$_2$) whereas ECR traces the shape of change in statistical power as seen in the left panel. GRR and ECR converge to the same value as ICC$_2$ approaches 1.

central χ^2-distribution with ν degrees of freedom. However, if the variance is at a boundary (i.e., in our situation, if the slope variance is restricted to be positive), this distribution may be different, e.g., a mixture of χ^2-distributions (see Self and Liang, 1987; Stoel et al., 2006). Ignoring the mixture distribution may lead to lower power than one could achieve when attending to it (but see Kolenikov and Bollen, 2012). However this pertains equally across all possible designs and, thus, can be ignored at the level of effective error and reliability.

For a test of a single parameter, in particular, a restriction on the slope variance, we set $\nu = 1$. The left integral bound $\chi^2(\nu, 0)$ is then determined by the critical value α as chosen by the investigator. The area to be integrated over depends on the non-centrality parameter λ, which again depends on the true slope variance, and effective error here. Using the approximation for λ by Satorra and Saris (1985), we obtain its value from the likelihood ratio of the minimal model with the slope variance restricted to zero over the minimal model without any restriction[2]. The model-predicted covariance of the unrestricted model is the sum of true slope variance and effective error, $\Sigma_{min} = \left[\sigma_S^2 + \sigma_{eff}^2\right]$. The restricted minimal model assumes zero slope variance, which yields $\Sigma_{res} = \left[\sigma_{eff}^2\right]$. Under the assumption of an unrestricted mean structure, the log-likelihood ratio (see section 2 in Supplementary Material) as the estimate of the non-centrality parameter is

$$\lambda = N\left[\frac{1}{1-ECR} - \ln\left(\frac{1}{1-ECR}\right) - 1\right] \tag{11}$$

Omitting third-order terms, an approximation of this Equation is given by $\lambda \approx N\left(\frac{ECR^2}{1-ECR}\right)$ which reveals that the relation of ECR and λ is dominated by a quadratic term of reliability and a linear term of sample size.

To summarize, by calculating the effective error σ_{eff}^2 from the structural design parameters of a given LGCM (residual variance, time points of occasions of measurement, and intercept variance) and assuming a true slope variance σ_S^2, a sample size N, and a test size α, we can derive λ, the non-centrality parameter of the χ^2-distribution under the hypothesis H$_1$. By numerically integrating this χ^2-distribution between a left bound depending on α, and infinity, we obtain the statistical power for the specific variance test under the assumption of all parameters but slope variance known and fixed. Note that under this assumption, the analytical power value is typically overestimated as freely estimating the remaining parameters will cost power. Still, it can be obtained that ECR and sample size are the primary determinants of statistical power of the LR test. The minimal power-equivalent metrics, effective error and ECR, are particularly useful as they constitute projections of the multidimensional design parameter space of a LGCM to a univariate index. They can both serve as comprehensive indices of change sensitivity in a longitudinal design exercise. ECR, in particular, is suitable as a general-purpose measure of slope effect size that is comparable across different studies because it does not depend on the units of measurement of the dependent variable.

Generalized ECR

When intercept-slope covariance is non-zero, the 1-*df* specific variance test is mis-specified because it assumes a zero covariance $\sigma_{IS}^2 = 0$, and the analytical solution of ECR as presented above, which does not capture this misspecification, is no longer fully

[2] Note that this likelihood ratio is equal to the likelihood ratio of the original model from which the power-equivalent minimal model was derived (see **Figure 2** for illustration), with a zero-slope restriction on the original model (see von Oertzen, 2010).

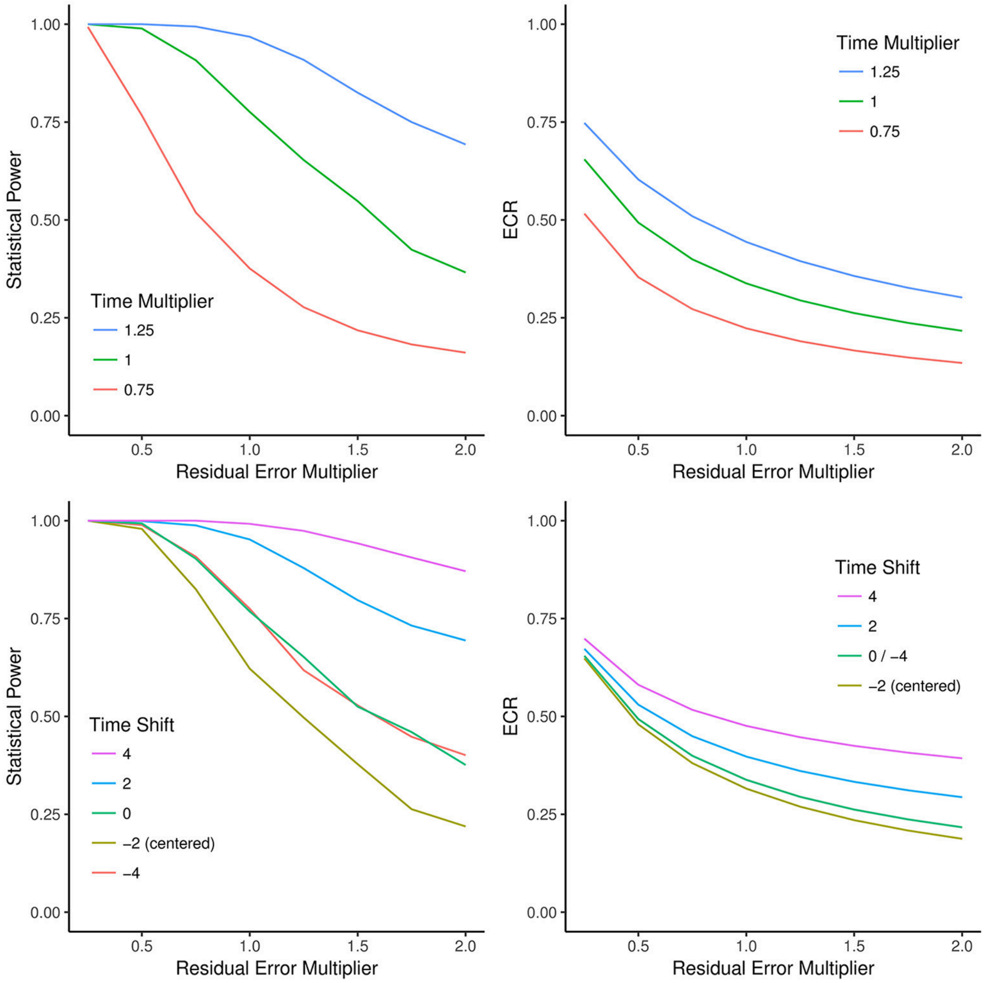

FIGURE 4 | The effects of shifting, prolonging, or compressing the study duration on statistical power and effective curve reliability (ECR). **Top left**: Monte Carlo simulation of the effect of decreasing or increasing residual error variance (x-axis) or shortening/prolonging total study duration (blue, green, and red lines) on statistical power (y-axis). **Top right**: The effect of decreasing or increasing residual error variance (x-axis) or shortening/prolonging total study duration (blue, green, and red lines) on ECR (y-axis). ECR traces the same shapes of decreases in statistical power as observed in the **top left** panel. **Bottom left**: Monte Carlo simulation of the effect of shifting the measurement time points by factor (−4, −2, 0, 2, 4; colored lines) and changing residual error variance (x-axis) on statistical power (y-axis). Note that anchoring the intercept at the first or last measurement has identical power. Lowest power is achieved with a centered design. Bottom right: The effect of shifting the measurement time points by factors (−4, −2, 0, 2, 4; colored lines) and changing residual error variance (x-axis) on ECR (y-axis). Note that the green line represents both the original design (shift: 0) and a design with intercept variance anchored at the last measurement (shift: −4), both of which empirically have identical power whereas the centered design has lowest ECR and power. As shown in Equation (11), for all variations in time sampling, GRR yields the values corresponding to the ECR of the time-centered design (yellow line; $t = -2$), and thus differential power because time shifts are not reflected in GRR.

coherent with its power. Currently, there is no analytical solution to compute ECR from a set of LGCM parameters for arbitrary intercept-slope covariance, which is then coherent with the 2-*df* generalized variance test. However, we provide an algorithm that can be used to compute ECR in the general case (see section 3 in Supplementary Material). This 2-*df* ECR can be interpreted as a standardized effect size of the total latent information about individual differences in linear slope. Under the assumption that $\sigma^2_{IS} = 0$, the analytical and computational solutions are identical.

To illustrate the coherence of this generalized index of ECR with power to detect slope variance, we ran an extension for one condition of our previous simulation considering the effect of intercept-slope covariance. **Figure 5** shows a slice of the simulation where residual error variance was set to 20, As before, we varied ICC$_2$ and also varied intercept-slope correlation, σ_{IS}, between −0.5 and +0.5. The left-hand panel of **Figure 5** shows the power of this test as a function of ICC$_2$ (as in **Figure 3**), but also as a function of σ_{IS}. The right-hand panel also displays the computed ECR as a function of ICC$_2$ and σ_{IS}. Clearly power of the 2-*df* generalized variance test is strongly affected by the proportion of intercept variance and the intercept-slope correlation. Most noteworthy, comparison of the two panels shows that ECR completely tracks these variations in power. In contrast, GRR (dashed line) is identical across all

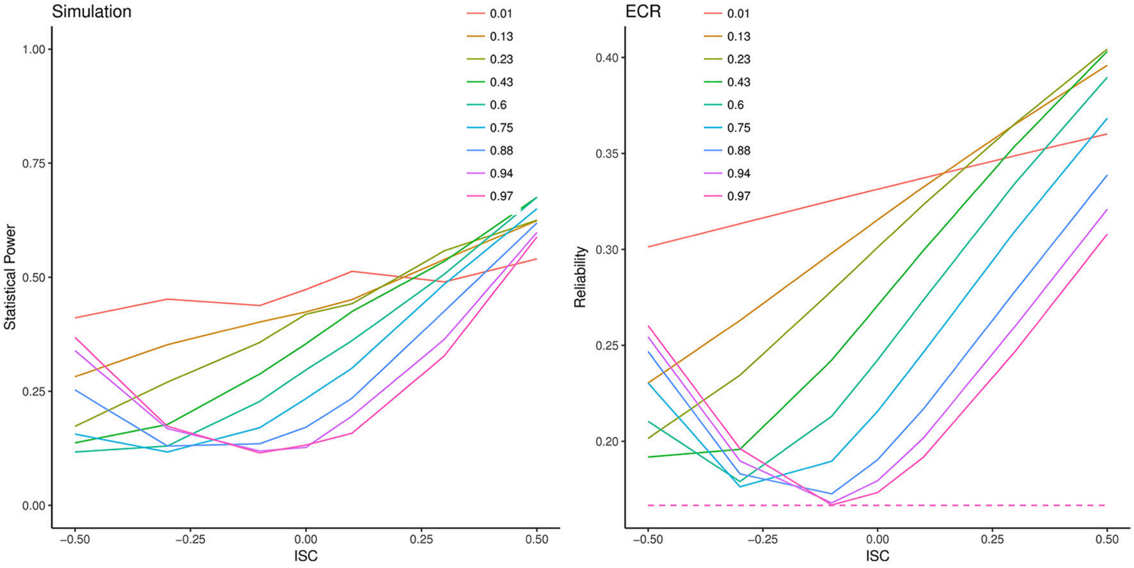

FIGURE 5 | Statistical power and effective curve reliability (ECR) values for the 2-df generalized variance test as a function of intercept-slope correlation (ISC). **Left**: Lines are Monte Carlo estimates of statistical power of the generalized variance test, with each line representing a model with different intercept variance scaled as the intra-class correlation coefficient (ICC_2). **Right**: Corresponding ECR (solid lines). GRR is constant for all conditions shown here.

conditions shown, despite the fact that the observed power ranges between 10% and 70%. Clearly ECR, not GRR, is the appropriate standardized effect size index for the 2-*df* test.

Considerations When Gauging Change Sensitivity

We now consider the practical implementation of the concepts introduced above for the purpose of *a priori* power analysis when designing longitudinal studies. To illustrate our analytical derivations, we continue to assume no intercept-slope correlation, $\sigma_{IS} = 0$, but the same considerations can be made using the computational approach allowing for non-zero intercept-slope correlations. Let us assume that a developmental researcher is engaged in design decisions when planning a longitudinal study. In practice, one cannot necessarily use existing studies to fully determine the configuration of LGCM parameters that are needed to specify the statistical power of a given design precisely. One reason for this claim is that phenomena do not always generalize across different constructs and measures, but another is that published papers often do not report all the relevant parameter estimates that are needed to compute ECR. Instead, partial or indirect evidence regarding parameters is often reported, such as an estimate of slope variance but none of the remaining parameters. How should one proceed on the basis of whatever minimal information is available in advance about the magnitude of individual differences in change so as to inform the study design process and its sensitivity to detecting said change? Although there are no incontrovertible general answers to this question, there are a number of practical guidelines and solutions available depending on what is considered to be known about the phenomena of interest. In the following, we present several cases showing how one can make use of minimal information about measures and their change to generate feasible study designs.

The parameters that determine ECR and hence power to detect LGCM slope variances can be divided into two categories: those that describe the process under investigation and those that describe the study design. The former parameters determine the structural model (Ψ; see Equation 2) of the LGCM, that is, the slope variance, intercept variance, and intercept-slope covariance. The latter parameters comprise the measurement model of the LGCM (Λ and Θ; see Equation 2), and consist of measurement precision, number and temporal distribution of measurement occasions, and total study duration. For the most part, the parameters of the measurement model can be directly influenced by design decisions. Hence variations in those parameter values can and should be selected in ways that are appropriate to design pragmatics and inevitable external constraints, such as budget and time (see Brandmaier et al., 2015), while optimizing power to detect the effect of interest. Making informed choices about population values is typically more difficult but equally important, as the choice of these values will also influence the outcome of a power analysis for a given design problem. In particular, deriving a plausible estimate of the population slope variance for a given measure or set of measures is crucial because it fundamentally determines the ECR and hence statistical power. If estimates of unstandardized slope variance or values of ECR from previous studies are available, they likely provide the basis for a good first guess. We would therefore like to emphasize the pressing need to report such indices of change sensitivity. If necessary, a conservative best-guess must be made from whatever other information is available. According to Willett (1989), the most logical strategy is probably to adopt a specific target reliability, such as GRR or ECR, when planning a study. We endorse this idea, although we demonstrate some of its limitations below. But how does one arrive at reasonable LGCM parameters needed to compute an index like ECR when no estimates from existing studies are available? In the next section, we offer some heuristics on how to go about solving this problem, given the limited information that is typically available at the time of study design.

Some heuristics for estimation of ECR need to include the separate contributions of different sources of error variance to effective error. The total error variance in a univariate LGCM is a sum of slope regression residual variance and indicator error variance. Slope regression residuals are errors of prediction in the usual sense of a regression equation, and their variance summarizes dispersion of occasion-specific latent factor scores around the best-fit line estimated by the LGCM (that is, they are occasion-specific shocks or disturbances in the latent variable space). Indicator error captures stochastic measurement error variance in the observed variable at each occasion. The distinction between these sources of error variance should be borne in mind when attempting to set parameter values in *a priori* power analysis. In practice, one usually has a static reliability estimate for a given measure that can be used to estimate measurement error variance, even though assumptions about measurement equivalence across time may not strictly hold in a given data set (Meredith and Horn, 2001). However, one also needs an estimate of regression residual variance. Our current approach to this problem invokes the assumption of homogeneous residual error variances as a heuristic that can be relaxed in more advanced applications. The following simple illustration of the underlying more complex logic exemplifies this.

Assume that information is available on the stability of individual differences of the target variable in a comparable population over a particular epoch of time. Given this information, provisional estimates of parameter values for the LGCM power analysis can be based on what is known about relations between stability of individual differences and LGCM parameters (Rogosa et al., 1982; Rogosa and Willett, 1983). In particular, there is a close correspondence between stability over time and ICC. In the following we show how the estimates of indicator reliability (such as parallel-forms reliability) and estimates of stability over time can be leveraged to estimate error variance in the LGCM for power analysis purposes.

Consider classical test theory approaches to alternate-forms reliability (e.g., Jöreskog, 1971). Under the assumption of homogeneous measurement error and no change between measurements, the correlation of these two measurements estimates their reliability (and corresponds to the intra-class correlation, ICC):

$$\rho_{reliability} = \frac{\sigma_I^2}{\sigma_I^2 + \sigma_\epsilon^2} \tag{12}$$

This allows us to relate the alternate-forms reliability to the individual differences in intercept and the residual error variance, which can be considered invariant over time for our purposes. If we are willing to assume a linear change process with true individual differences in change, then we can leverage a stability coefficient (a correlation of a variable with itself over time in a given sample (e.g., Bloom, 1964), sometimes labeled a test-retest correlation), to achieve estimates of intercept variance and regression residual variance. This approach is based on simplifying the following equation for the stability coefficient:

$$\rho_{stability} = \frac{\sigma_I^2 + T\sigma_{IS}}{\sqrt{\sigma_I^2 + \sigma_\epsilon^2}\sqrt{\sigma_I^2 + \sigma_\epsilon^2 + T^2\sigma_S^2 + 2T\sigma_{IS}}} \tag{13}$$

Equations (12) and (13) allow us to translate stability and reliability estimates into growth curve parameter estimates and thus aid in study design planning by reducing degrees of freedom in the choice of parameter values. In the following, we present illustrative cases to show how these transformations can be applied, and more broadly, how design decision can be informed by using effective error and ECR.

ILLUSTRATED APPLICATIONS

Case 1: Deriving LGCM Weights

Here we will illustrate how the relationships identified in Equations (12) and (13) using minimal pre-existing information about the phenomena under investigation can generate useful parameter values for the evaluation of statistical power. Suppose a researcher is interested in longitudinally measuring episodic memory performance in a group of older adults. If a similar study has already been conducted via LGCM, we would generally be willing to use its estimates to generate a power analysis. For this purpose, we would assume generalizability in measures, procedures, occasions, and populations. But what if a comparable study with LGCM results is lacking, that is, there is neither a report of existing effect sizes nor of parameter values that could serve as a starting point for our analysis of change sensitivity? In a literature review, the researcher does find (a) a cross-sectional study reporting an alternate-forms reliability of 0.9 for the memory task, and (b) longitudinal data on a similar task reporting a 3-year stability coefficient of 0.85. These values reduce uncertainty about possible LGCM parameter values and can be leveraged to start the power analysis process.

In our example, we assume a scaling in T-score units, such that the total observed variance is 100 (given that $SD = 10$). If there were no individual differences in change (i.e., $\sigma_S^2 = 0$), there would be perfect stability of individual differences and hence zero regression residual error. In this limiting case, which we assume to hold with a parallel-forms reliability estimate, we find that $\sigma_I^2 + \sigma_E^2 = 100$. Taking this information together with the other given, $\rho_{reliability} = 0.9$, and solving Equation (12), we see that $\sigma_I^2 = 90$ and $\sigma_E^2 = 10$. Of course, our focus is on variance in change, and if we assume true individual differences in change, then the 3-year test-retest stability estimate must be lower than the parallel-forms estimate because these systematic differences in rates of change contribute to the observed score variances.

We can apply Equation (13) to generate a crude estimate of σ_S^2. A practical constraint is that it is usually impossible to specify a point value for the intercept-slope covariance term σ_{IS}

in advance. However, one solution to the problem is to assume that $\sigma_{IS} = 0$, at least as a starting point. With this assumption we can solve Equation (13) for σ_S^2 by inserting $T = 3$ and using our previous estimates of σ_I^2 and σ_E^2, yielding an estimate of $\sigma_S^2 = 1.35$. These parameter values can then be used to conduct statistical power analysis for a longitudinal design based on the specific parameters.

Brandmaier et al. (2015) developed a power analysis utility for LGCM slope variances called LIFESPAN. Using LIFESPAN (or the R code in the Supplementary Material) we can fill in the derived values for a design that is supposed to measure participants, for example, 3 times over 5 years, and we obtain estimated $\sigma_{eff}^2 = 0.76$, ECR $= 0.64$, and GRR $= 0.63$. The effective error gauges the precision of the design in units of unstandardized effect size without relying on a specific unstandardized effect size. The standardized effect size metrics indicate that a substantial amount of change-related variance in the variable of interest is associated with systematic individual differences in rates of change.

We might want to be more conservative when estimating true slope variance based on point estimates of residual variability and indicator reliabilities, given that these may be based on minimal information from the literature. For example, we might be targeting a slightly different population for which we believe change to be more difficult to detect (e.g., a sample that is slightly younger, slightly better educated, or that includes a lower proportion of participants with dementia) than the population generating available information on indicator reliability and stability. **Table 1** presents a range of slope variance estimates for different test-retest stabilities crossed with different alternative-forms reliabilities. The smaller the test-retest stability, the smaller the proportion of stable individual differences and the larger the estimate of variance in change is.

Table 1 highlights that there can be no true variance in slopes when both estimates of reliability and stability are identical. As noted by Rogosa et al. (1982), this is the degenerate case in which there are no individual differences in change, and in which the reliability of the difference score must be zero. In our terms, there is no opportunity for change sensitivity because there are no individual differences in change that the design must be able to detect.

In practice, engaging in power analysis means that we can entertain a range of possible slope variance effect sizes. Use of a power analysis program like LIFESPAN facilitates generation of a range of possible power values by systematically varying values of error variance components, but the same thing can also be done with values of ECR and unstandardized effect size (see Brandmaier et al., 2015).

TABLE 1 | Slope variance as a function of reliability and stability.

Reliability	Stability				
	0.750	0.800	0.850	0.875	0.890
0.890	4.54	2.64	1.07	0.38	0.00
0.900	4.89	2.95	1.35	0.64	0.25
0.925	5.79	3.74	2.05	1.31	0.89

Slope variance values are derived under the assumption of no intercept-slope covariance and no slope regression residuals.

Case 2: Evaluating and Comparing Study Designs

Evaluating and comparing existing study designs is often a useful first step when planning a longitudinal study. When *post-hoc* power estimates are available from a set of pertinent studies (e.g., Rast and Hofer, 2014), one can call on these existing designs as a starting point for a future design. As argued earlier, statistical power can be decomposed into several components influencing change sensitivity. In the following example, we show how different indicators of change sensitivity may inform us in complementary and useful ways when comparing longitudinal studies with respect to their change sensitivity.

Assume we have singled out two existing studies from the literature targeting the same or a similar outcome of interest. In each, participants were measured every 3 months for 1 year and it happens that the estimated intercept variances were similar, suggesting an approximate $\sigma_I^2 = 90$. The studies fundamentally differ in the precision of measurement instruments used (5.9 in Study A and 10 in Study B), in the magnitude of estimated change variance (3.0 in Study A and 5.0 in Study B), and in sample size (200 participants in Study A and 300 in Study B). Now, a *post-hoc* power analysis reveals that Study A has a power of roughly 70% to reject the null hypothesis, $\sigma_S^2 = 0$, whereas Study B has a corresponding power of 85%. If presented with these power values only, we might be inclined to start our own study design by building upon the original design of the more powerful study, namely Study B. However, the difference in sample sizes between the studies must also be considered in evaluating the space of possible designs that are made up of alternative sample sizes, different temporal spacing of observations, and so on. As we noted earlier (see **Figure 1**), statistical power is determined by standardized effect size and sample size. Since we may not yet have decided on a sample size for our new study, it seems reasonable to compare the reliabilities of the studies given that reliabilities are independent of sample size. Using the LIFESPAN program, we find that for both studies, ECR $= 0.246$. Thus, one possible inference, based only on reliability, is that the difference in observed statistical power is solely due to the larger sample size in Study B. Upon closer inspection of the differences between the studies, we are reminded that the dependent variables also had different indicator reliabilities. Willett (1989) already noted that reliability *confounds* both individual differences in change and measurement precision, which may present hidden problems for the unwary investigator. If we wanted to plan our future study design independently of the absolute magnitude of individual differences, we can turn to effective error as a metric that is in units of (estimated) true effect size but independent of its magnitude. Doing so, we obtain $\sigma_{eff}^2 = 9.20$ for Study A and $\sigma_{eff}^2 = 15.33$ for Study B. By comparing effective error (which only works when variables are scaled in the same metric), we consider the measurement model and factor out sample size and true effect size. Then we find that Study A is the more sensitive measurement model (with more favorable precision and temporal spacing of measurements) to detect change variance despite its lower estimated *post-hoc* power even though Study B is the overall more powerful study when respective unstandardized effect size (magnitude of population slope variance) and sample sizes are also taken into account. This result may seem trivial as the difference in effective errors in this illustration—for the sake of simplicity—is solely due to apparent differences in indicator reliability. In actual applications, study differences in ECR may often reflect differences arising from constellations of total study duration, measurement density, and indicator reliability. Although we advocate using a standardized effect size measure like ECR in design planning, this example shows that no single measure of change sensitivity is inherently superior to the others; instead, they convey complementary summaries of study design properties.

Case 3: Finding Optimal Measurement Density

Suppose a group of researchers is about to start a longitudinal study. The initial design calls for five occasions of measurement, each 1 year apart, thus spanning a total time of 4 years. Based on a literature review, the group assumes $\sigma_I^2 = 10$, $\sigma_S^2 = 1$, and $\sigma_\epsilon^2 = 50$. Using the LIFESPAN program, the researchers compute an effective error $\sigma_{eff}^2 = 2.50$ and a relatively low reliability of ECR = 0.29. When discussing the design, the question comes up whether a different spacing of measurements in time would be better to detect individual differences in change. The initial design will be denoted as a set of time points at which measurement takes place relative to study onset, {0, 1, 2, 3, 4}. The following alternative spacings of measurement occasions are discussed: for example, for organizational reasons, either the first three or last three measurement occasions could be moved closer together with time differences between measurements of 6 months each: {0, 0.5, 1, 3, 4}, {0, 1, 3, 3.5, 4}. **Table 2** lists ECR, GRR, and Monte-Carlo-simulated power values for all three designs. Note that ECR values are scaled proportionally to statistical power whereas the two alternative designs have equal GRR values but different statistical power. The difference arises because the $\text{ICC}_2 = 10/(10 + 50/5) = 0.5$, which is different from the asymptotic case of $\text{ICC}_2 = 1$, which would render both effect size indices identical. ECR, not GRR, adequately quantifies differences in effective error due to the different spacing of measurement occasions. Given that sample size and unstandardized magnitude of individual differences in change are constant across all alternative designs considered, deciding upon the best (that is, the most change-sensitive) design can actually be based on effective error (coherent with ECR) alone and need not necessarily rely on simulated power values.

TABLE 2 | Measures of study design quality for different distributions of five measurement occasions over 4 years.

Design	Dispersion	GRR	ECR	Statistical power (N = 100)
{0, 1, 2, 3, 4}	2.50	0.17	0.29	0.51
{0, 0.5, 1, 3, 4}	2.95	0.19	0.28	0.51
{0, 1, 3, 3.5, 4}	2.95	0.19	0.33	0.63

Statistical power refers to the 1-df variance test and is based on a Monte Carlo estimate with 2,500 repetitions. Dispersion is given as the variance of the measurement time points. Note that the second and the third design have equal GRR values despite different ECR values and statistical power.

DISCUSSION

This paper has demonstrated how measures of sensitivity to detect individual differences in linear change including statistical power can be leveraged to help researchers communicate and make decisions about longitudinal designs. Building on earlier results (von Oertzen, 2010; von Oertzen and Brandmaier, 2013), we have described how the central concept of effective error relates to concepts of reliability and standardized effect size for the LGCM slope variance parameter, captured either as GRR or ECR. The direct connection among effective error, standardized effect size, and statistical power of the 1-*df* LR test in a maximum likelihood framework provides a comprehensive perspective on sensitivity to detect individual differences in change that can be used as a benchmark for evaluating alternative longitudinal designs in terms of change sensitivity.

The developments presented here allow for a much broader and more comprehensive approach to analyzing, diagnosing, and planning longitudinal studies than is common practice in the field. In our view, both a lack of a formal system for comparison and contrast of influences of design decisions on statistical power, and the absence of a practically implementable approach to testing a wide range of alternative design configurations have led researchers to neglect *a priori* power analysis of change sensitivity at the planning stages of longitudinal investigations.

The approach illustrated here provides a means by which researchers can systematically and comprehensively consider trade-offs of design features that generate equivalent power. They can also examine how change sensitivity can be improved by attending to specific design features, such as adding occasions of measurement or prolonging the total study duration. Furthermore, this approach has been implemented in a power analysis utility program (LIFESPAN; Brandmaier et al., 2015) that facilitates iterative consideration of how power changes as a function of manipulating different design features. In the Supplementary Materials, we also provide R code to compute the discussed indices of precision and reliability. This paper provides a formal treatment of the elements that are essential to make this approach both statistically coherent and practically feasible. Of course, the LIFESPAN approach is a first pass at a utility aiding in the formalization of the optimal-design problem, and we fully expect that future efforts can greatly improve upon how the *a priori* design process can be supported by alternative optimization approaches.

Considerations of statistical power are inevitably a function of the specific hypotheses being tested. Researchers typically focus on more specific questions than detecting variance in change. Hence a generic assessment of change sensitivity is no substitute for an evaluation of a longitudinal design's power with respect to a more specific hypothesis critical to the viability of the respective study. Given that most longitudinal studies include multiple outcome variables, one would need to consider the possibility of a change sensitivity analysis within each relevant construct domain, an aspect of the problem that we do not consider further in this paper.

Building on previous work by Willett (1989) and Rast and Hofer (2014), we have shown that both ECR and GRR can be regarded as a metric-free index of slope variance effect size. We also showed that, when $\sigma_{IS} = 0$, GRR is a special case of ECR, and that only ECR is coherent with the effective error term for the specific and generalized LR test of zero slope variance that governs its power. Although GRR and ECR are closely related, we have shown here that ECR provides a better basis for capturing statistical power of the 1-*df* LR test (assuming no intercept-slope covariance) or the 2-*df* LR test (assuming an arbitrary intercept-slope covariance), because it does not ignore the contribution of intercept variance (scaled as ICC_2) and intercept-slope covariance to the power of the LR test. On the other hand, GRR may be better suited for indexing an effect size coherent with the statistical power of the Wald test as it ignores intercept variance and intercept-slope covariance like the Wald test does. Along with others in the literature, we continue to argue (Hertzog et al., 2008) that the generalized 2-*df* test of zero slope variance is superior to the other tests, and generally preferable, because a non-zero intercept-slope covariance is also evidence of individual differences in change. Therefore there is a pressing need to extend the line of research presented in this paper to focus on how the more complex effective error for a 2-*df* LR test of slope variance—and an effect size index based upon it—may be analytically derived to achieve a better understanding about how design factors, individually and in interaction, determine statistical power to detect variance in change.

Previously, Rast and Hofer (2014) criticized simulations of LR test power attending to indices of error that were influenced by intercept variance (e.g., Hertzog et al., 2008) and advocated the merits of GRR for power analysis based on the fact that it ignored intercept variance. We have shown, to the contrary, that the power of the specific variance LR test is influenced by intercept variance (as scaled by ICC_2) and that this influence on power is reflected in effective error and ECR, but not by GRR. The interconnection of intercept variance and slope variance can also be shown via the asymptotic non-independence of both components; the fact that the Fisher information matrix in an LGCM has non-zero entries for the covariance of the intercept variance parameter and slope variance parameter (even if intercept-slope-covariance is zero) violates the asymptotic independence assumption and has implications on bias in parameter estimates, standard errors, and statistical power (Kaplan and Wenger, 1993).

In general, we advocate that studies of LGCM begin to report estimates of ECR and GRR to provide a means to scale change sensitivity across different studies with different outcome variables[3]. One benefit of reliability indices such as ECR and GRR is that they allow for meta-analysis of effect sizes based on existing findings in the literature. Rast and Hofer (2014) used GRR to characterize existing studies in the field of normal cognitive aging. Our findings demonstrate another benefit: The availability of ECR as an effect size statistic will further improve the field's ability to conduct *a priori* power analyses that attend specifically to change sensitivity, often by leveraging information from other studies. Given that studies often do not report the full set of unconditional LGCM variance parameter estimates that can be used to generate estimates of ECR, we recommend that studies using linear LGCM begin reporting an estimate of ECR as a measure of slope effect size that can be useful for communication of results and can facilitate longitudinal design decisions by other researchers.

Although the ultimate target for evaluating change sensitivity is the statistical power to detect slope variance, statistical power is not the only index of change sensitivity to consider or report. Rather, our illustrative cases show that any of the three measures—effective error, ECR, or statistical power—can provide helpful information about aspects of change sensitivity. When examining alternative design possibilities within a single study, it seems reasonable to operate on the most basic component of change sensitivity proposed here, that is, the effective error of the slope variance test, as it already allows for the analysis of trade-offs of study design properties against each other. In particular, effective error enables researchers to analyze the effect of adding or removing occasions of measurement, prolonging or shortening the study's duration, or switching between differentially reliable measurement instruments (von Oertzen, 2010; von Oertzen et al., 2010; von Oertzen and Brandmaier, 2013). Reliability, as indexed by ECR, scales a candidate slope variance parameter by its effective error in a given design. As such, it provides a metric for standardized effect size that can be used to compare different study designs against each other directly and efficiently. Furthermore, one can consider absolute precision of the design in terms of effective error as well (see Case 2). Effective error is expressed in absolute units of unstandardized effect size but is independent of both sample size and true effect size.

In the study design phase, guesses about effect sizes may draw upon reports of published studies. If no plausible estimates of absolute magnitude of slope variance can be made when planning a longitudinal study, effective error is still applicable as a metric for comparisons of change sensitivity when deciding between different proposed study designs for the same outcome (i.e., when change is measured in the same units across designs). If longitudinal outcomes of different studies can legitimately be scaled to a common metric (e.g., T-scores), then effective error provides a metric for comparisons of relative precision even across different scales.

[3] An estimate of ECR from a LGCM using ML estimation is obtained by substituting estimates of LGCM parameters in Equation (7). Currently, this can be accomplished by inputting LGCM parameter estimates into the LIFESPAN program, which returns values for effective error, ECR, and GRR.

Another commonly reported measure of reliability in latent variable models is ICC$_2$. ICC$_2$ is closely connected to ECR but targets the reliability of intercept variance. It is usually defined as a reliability measure in which variance due to differences between persons in a repeated-measures design is expressed as a proportion of the total variance (Raudenbush and Bryk, 2002). ICC$_2$ can be identically obtained by following the logic proposed here to derive reliability (ECR) in an intercept-only LGCM, in which no change of the construct is assumed. Then, we obtain ICC$_2$ as the reliability of the intercept variance (representing stable between-person differences).

Limitations

The limitations of using any of the indices developed in this paper inherit the well-known limitations of statistical power analysis in general. If all parameters of the LGCM can be specified *a priori* (i.e., they can be considered known at the time of planning), existing software such as Mplus (Muthen and Muthen, 2007), PinT (Snijders et al., 2007), OpenMx (Boker et al., 2011), simsem (Pornprasertmanit et al., 2016), or SIMR (Green and MacLeod, 2016) can be used to evaluate the power to reject candidate values of the parameter of interest—in our case, slope variance. For practical purposes, the need to specify values of parameters in advance can be an impediment to further efforts to use *a priori* power analysis to inform design decisions. At one extreme, this problem can be avoided by simply calculating the power to reject a particular magnitude of overall model fit using a relative goodness-of-fit index, as suggested by MacCallum et al. (2006). One can thereby make statements about the power of the entire model to detect consequential violations in overall model fit. However, this is not equivalent to evaluating power to reject a critical null hypothesis about a specific parameter. Our claim is that change sensitivity in a longitudinal study is often best captured by the ability to detect LGCM slope variance.

If *post-hoc* estimates of population values are used to determine indices of change sensitivity, the uncertainty about the empirical point estimates of the parameters induces uncertainty in the derived indices. That is, effective error, effective growth curve reliability, and statistical power reflect true change sensitivity only to the degree to which *post-hoc* parameter estimates involved in their computation (e.g., residual error, slope variance, or intercept variance) reflect the corresponding population values. The same is necessarily true for *a priori* guesses that can only be as accurate as the researcher's *a priori* knowledge or intuition in picking parameter values. Therefore, it is typically recommended that a designer choose conservative values during study design planning (von Oertzen and Brandmaier, 2013). In the light of uncertainty of best guesses, it is also advisable to calculate reliability of change or statistical power for a range of possible values and either to integrate over this range or to select the most conservative value among them. Alternatively, it may be useful to treat uncertainty in estimates formally (Kelley and Rausch, 2011; Lai and Kelley, 2011; Gribbin et al., 2013) and to derive confidence intervals for effect size estimates in accordance with best practices proposed by Kelley and Preacher (2012). Future work should focus on capturing the uncertainty around single-point estimates of ECR and GRR.

In a similar vein, effective error, reliability, and statistical power are only valid under the assumption that the growth process is adequately modeled, that is, there is no misspecification in the underlying LGCM. The perfect interchangeability of power-equivalent operations—for instance between adding more occasions within a given total study duration or increasing total study duration—is based on the assumption that linearity of change is strictly true, and not just a useful approximation for shorter stretches of time. Also note that our considerations are only valid under the assumption of homogeneous residual error variances over time. The present manuscript is a central building block for planned future work addressing more complex models, such as non-linear growth curve models or (dual) change score models, and other types of hypotheses (e.g., the generalized variance test). We recommend to use Monte-Carlo-based simulation approaches (Muthén and Muthén, 2002) for assessing more general designs currently not covered by our approach. However, simulation-based approaches will always be slower and their results will always be more difficult to generalize than an analytical solution that allows a fuller overall theoretical understanding of what is being studied. For example, solving the problem of finding alternative, power-equivalent designs to an initial study design (Brandmaier et al., 2015) may be time-consuming, at best, if not infeasible when purely relying on simulation-based approaches.

Finally, missing data, and specifically attrition in longitudinal data, decrease the change sensitivity of longitudinal designs. von Oertzen and Brandmaier (2013; Appendix B, Theorem 4) proved that under a missing-completely-at-random (MCAR) assumption, an effective error can be derived for each group that has a unique attrition pattern, and the resulting effective error is a weighted harmonic mean of the effective errors aggregated over attrition groups. Given that attrition from longitudinal studies is decidedly not MCAR, further work on this problem is needed. With the same approach, spacing between measurements varying both within and across people can be accounted for by conceiving of such a study design as a multiple-group design, in which each person is his or her own group, and in which the overall effective error is a function of the person-specific effective errors.

Future work should also address additional statistical issues. The power equivalence theory that is the basis for the present work assumes that all parameters other than the values subject to a hypothesis test are known and fixed. Power-equivalent operations are performed under this assumption; however, in practice, all LGCMs are typically freely estimated from data. The same assumption is made in other analytical approximations of statistical power, such as that of Satorra and Saris (1985). When researchers are exploring a design space for variations in change sensitivity, they need to be aware that the LR test in question may have lower power than is estimated by the power equivalence approach because it actually freely estimates all parameters. Thus, after a candidate design has been chosen using power-equivalence trade-offs, a final Monte

Carlo simulation to generate a more accurate estimate of power with the candidate design should be conducted. The LIFESPAN program can generate the Monte Carlo simulation for a specified set of hypothetical model parameters generated by power-equivalence searches of the design-trade-off space (Brandmaier et al., 2015).

To conclude, we believe that the notions of precision (scaled as effective error), reliability, and statistical power help to promote the insight that increasing sample size is not the only, and not necessarily even the best way to optimize change sensitivity of a longitudinal design. Instead, all design decisions influence change sensitivity, and research designs can be optimized accordingly even before the size of the sample has been specified. Greater sample size increases statistical power of the LR test, to be sure, but one can consider other means to achieve this that may be more practically feasible in a given study context (e.g., by administering instruments with higher reliability). The present approach enables researchers to evaluate, optimize, and communicate longitudinal designs comprehensively by considering how design features interactively influence change sensitivity in LGCM.

AUTHOR CONTRIBUTIONS

All authors contributed to this manuscript. AB and TvO derived the mathematical proofs in the Supplementary Material.

ACKNOWLEDGMENTS

We are grateful to Julia Delius for her helpful assistance in language and style editing. The authors acknowledge support from the Max Planck Society.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2018.00294/full#supplementary-material

REFERENCES

Baltes, P. B., Lindenberger, U., and Staudinger, U. M. (2006). "Lifespan theory in developmental psychology," in *Handbook of Child Psychology,* eds W. Damon and R. M. Lerner (New York, NY: Wiley), 569–664.

Baltes, P. B., and Nesselroade, J. R. (1979). "History and rationale of longitudinal research," in *Longitudinal Research in the Study of Behavior and Development,* eds J. R. Nesselroade and P. B. Baltes (New York, NY: Academic Press), 1–39.

Baltes, P. B., Reese, H. W., and Nesselroade, J. R. (1988). *Life-Span Developmental Psychology: Introduction to Research Methods*. Hillsdale, NJ: Erlbaum.

Bloom, B. S. (1964). *Stability and Change in Human Characteristics*. New York, NY: John Wiley & Sons.

Boker, S. M., Neale, M., Maes, H., Wilde, M., Spiegel, M., Brick, T. R., et al. (2011). OpenMx: an open source extended structural equation modeling framework. *Psychometrika* 76, 306–317. doi: 10.1007/s11336-010-9200-6

Bollen, K. A. (1989). *Structural Equations with Latent Variables.* Oxford, UK: John Wiley.

Brandmaier, A. M., von Oertzen, T., Ghisletta, P., Hertzog, C., and Lindenberger, U. (2015). LIFESPAN: a tool for the computer-aided design of longitudinal studies. *Front. Psychol.* 6:272. doi: 10.3389/fpsyg.2015.00272

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences, 2nd Edn.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Curran, P. J. (2003). Have multilevel models been structural equation models all along? *Multivariate Behav. Res.* 38, 529–569. doi: 10.1207/s15327906mbr3804_5

Duncan, T. E., Duncan, S. C., and Strycker, L. A. (2006). *An Introduction to Latent Variable Growth Curve Modeling: Concepts, Issues, and Application.* Mahwah, NJ: Lawrence Erlbaum Associates.

Ferrer, E., and McArdle, J. J. (2003). Alternative structural models for multivariate longitudinal data analysis. *Struct. Equat. Model.* 10, 493–524. doi: 10.1207/S15328007SEM1004_1

Ferrer, E., and McArdle, J. J. (2010). Longitudinal modeling of developmental changes in psychological research. *Curr. Dir. Psychol. Sci.* 19, 149–154. doi: 10.1177/0963721410370300

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika* 73, 43–56. doi: 10.1093/biomet/73.1.43

Green, P., and MacLeod, C. J. (2016). SIMR: an R package for power analysis of generalized linear mixed models by simulation. *Methods Ecol. Evol.* 7, 493–498. doi: 10.1111/2041-210X.12504

Gribbin, M. J., Chi, Y. Y., Stewart, P. W., and Muller, K. E. (2013). Confidence regions for repeated measures ANOVA power curves based on estimated covariance. *BMC Med. Res. Methodol.* 13:57. doi: 10.1186/1471-2288-13-57

Hertzog, C., von Oertzen, T., Ghisletta, P., and Lindenberger, U. (2008). Evaluating the power of latent growth curve models to detect individual differences in change. *Struct. Equat. Model.* 15, 541–563. doi: 10.1080/10705510802338983

Johansson, B., Hofer, S. M., Allaire, J. C., Maldonado-Molina, M. M., Piccinin, A. M., Berg, S., et al. (2004). Change in cognitive capabilities in the oldest old: the effects of proximity to death in genetically related individuals over a 6-year period. *Psychol. Aging* 19, 145–156. doi: 10.1037/0882-7974.19.1.145

Johansson, B., Whitfield, K., Pedersen, N. L., Hofer, S. M., Ahern, F., and McClearn, G. E. (1999). Origins of individual differences in episodic memory in the oldest-old: a population-based study of identical and same-sex fraternal twins aged 80 and older. *J. Gerontol. B Psychol. Sci. Soc. Sci.* 54, P173–P179. doi: 10.1093/geronb/54B.3.P173

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika* 36, 109–133. doi: 10.1007/BF02291393

Kaplan, D., and Wenger, R. N. (1993). Asymptomatic independence and separability in convariance structure models: implications for specification error, power, and model modification. *Multivariate Behav. Res.* 28, 467–482. doi: 10.1207/s15327906mbr2804_4

Ke, Z. J., and Wang, L. J. (2015). Detecting individual differences in change: methods and comparisons. *Struct. Equat. Model.* 22, 382–400. doi: 10.1080/10705511.2014.936096

Kelley, K., and Preacher, K. J. (2012). On effect size. *Psychol. Methods* 17, 137–152. doi: 10.1037/a0028086

Kelley, K., and Rausch, J. R. (2011). Sample size planning for longitudinal models: accuracy in parameter estimation for polynomial change parameters. *Psychol. Methods* 16, 391–405. doi: 10.1037/a0023352

Kline, R. B. (1998). *Principles and Practice of Structural Equation Modeling.* New York, NY: Guilford Press.

Kolenikov, S., and Bollen, K. A. (2012). Testing negative error variances:is a Heywood case a symptom of misspecification? *Sociol. Methods Res.* 41, 124–167. doi: 10.1177/0049124112442138

Lai, K. K., and Kelley, K. (2011). Accuracy in parameter estimation for targeted effects in structural equation modeling: sample size planning for narrow confidence intervals. *Psychol. Methods* 16, 127–148. doi: 10.1037/a0021764

Laird, N. M., and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* 38, 963–974. doi: 10.2307/2529876

Lee, S., and Hershberger, S. (1990). A simple rule for generating equivalent models in covariance structure modeling. *Multivariate Behav. Res.* 25, 313–334. doi: 10.1207/s15327906mbr2503_4

Lindenberger, U. (2014). Human cognitive aging: corriger la fortune? *Science* 346, 572–578. doi: 10.1126/science.1254403

MacCallum, R. C., Browne, M. W., and Cai, L. (2006). Testing differences between nested covariance structure models: power analysis and null hypotheses. *Psychol. Methods* 11, 19–35. doi: 10.1037/1082-989X.11.1.19

MacCallum, R. C., Lee, T., and Browne, M. W. (2010). The issue of isopower in power analysis for tests of structural equation models. *Struct. Equat. Model.* 17, 23–41. doi: 10.1080/10705510903438906

MacCallum, R. C., Wegener, D. T., Uchino, B. N., and Fabrigar, L. R. (1993). The problem of equivalent models in applications of covariance structure-analysis. *Psychol. Bull.* 114, 185–199. doi: 10.1037/0033-2909.114.1.185

McArdle, J. J. (1994). Structural factor analysis experiments with incomplete data. *Multivariate Behav. Res.* 29, 409–454. doi: 10.1207/s15327906mbr2904_5

McArdle, J. J., and Nesselroade, J. R. (2014). *Longitudinal Data Analysis Using Structural Equation Models*. Washington, DC: American Psychological Association.

Meredith, W., and Horn, J. (2001). "The role of factorial invariance in modeling growth and change," in *Decade of Behavior. New Methods for the Analysis of Change,* eds L. M. Collins and A. G. Sayer (Washington, DC: American Psychological Association), 203–240.

Meredith, W., and Tisak, J. (1990). Latent curve analysis. *Psychometrika* 55, 107–122. doi: 10.1007/BF02294746

Muthen, L. K., and Muthen, B. O. (2007). *Mplus: Statistical Analysis with Latent Variables: User's guide.* Los Angeles, CA: Muthén & Muthén.

Muthén, L. K., and Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Struct. Equat. Model.* 9, 599–620. doi: 10.1207/S15328007SEM0904_8

Pinheiro, J. C., and Bates, D. M. (2000). "Linear mixed-effects models: basic concepts and examples," in *Mixed-Effects Models in S and S-PLUS. Statistics and Computing* (New York, NY: Springer).

Pornprasertmanit, S., Miller, P., and Schoemann, A. (2016). *simsem: SIMulated Structural Equation Modeling*. Available online at: https://CRAN.R-project.org/package=simsem

Rast, P., and Hofer, S. M. (2014). Longitudinal design considerations to optimize power to detect variances and covariances among rates of change: simulation results based on actual longitudinal studies. *Psychol. Methods* 19, 133–154. doi: 10.1037/a0034524

Raudenbush, S. W., and Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods.* Thousand Oaks, CA: Sage.

Rogosa, D. R., Brandt, D., and Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychol. Bull.* 92, 726–748. doi: 10.1037/0033-2909.92.3.726

Rogosa, D. R., and Willett, J. B. (1983). Demonstrating the reliability of the difference score in the measurement of change. *J. Educ. Meas.* 20, 335–343. doi: 10.1111/j.1745-3984.1983.tb00211.x

Rovine, M. J., and Molenaar, P. C. (2000). A structural modeling approach to a multilevel random coefficients model. *Multivariate Behav. Res.* 35, 51–88. doi: 10.1207/S15327906MBR3501_3

Satorra, A., and Saris, W. E. (1985). Power of the likelihood ratio test in covariance structure analysis. *Psychometrika* 50, 83–90. doi: 10.1007/BF02294150

Self, S. G., and Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. Am. Stat. Assoc.* 82, 605–610. doi: 10.1080/01621459.1987.10478472

Singer, J. D., and Willett, J. B. (2003). *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. New York, NY: Oxford university press.

Snijders, T., Bosker, R., and Guldemond, H. (2007). *PinT: Power analysis in Two-level designs.* Available online at: https://www.stats.ox.ac.uk/~snijders/multilevel.htm - progPINT

Stelzl, I. (1986). Changing a causal hypothesis without changing the fit - some rules for generating equivalent path models. *Multivariate Behav. Res.* 21, 309–331. doi: 10.1207/s15327906mbr2103_3

Stoel, R. D., Garre, F. G., Dolan, C., and van den Wittenboer, G. (2006). On the likelihood ratio test in structural equation modeling when parameters are subject to boundary constraints. *Psychol. Methods* 11, 439–455. doi: 10.1037/1082-989X.11.4.439

von Oertzen, T. (2010). Power equivalence in structural equation modelling. *Br. J. Math. Stat. Psychol.* 63, 257–272. doi: 10.1348/000711009X441021

von Oertzen, T., and Brandmaier, A. M. (2013). Optimal study design with identical power: an application of power equivalence to latent growth curve models. *Psychol. Aging* 28, 414–428. doi: 10.1037/a0031844

von Oertzen, T., Hertzog, C., Lindenberger, U., and Ghisletta, P. (2010). The effect of multiple indicators on the power to detect inter-individual differences in change. *Br. J. Math. Stat. Psychol.* 63, 627–646. doi: 10.1348/000711010X486633

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

Willett, J. B. (1989). Some results on reliability for the longitudinal measurement of change: implications for the design of studies of individual growth. *Educ. Psychol. Meas.* 49, 587–602. doi: 10.1177/001316448904900309

Wu, W., Jia, F., Rhemtulla, M., and Little, T. (2016). Search for efficient complete and planned missing data designs for analysis of change. *Behav. Res.* 48, 1047–1061. doi: 10.3758/s13428-015-0629-5

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.