British Journal of Mathematical and Statistical Psychology (2010), 63, 627–646

The British Psychological Society

www.bpsjournals.co.uk

# The effect of multiple indicators on the power to detect inter-individual differences in change

Timo von Oertzen[1]*, Christopher Hertzog[2], Ulman Lindenberger[1] and Paolo Ghisletta[3,4]

[1]Max Planck Institute for Human Development, Berlin, Germany
[2]School of Psychology, Georgia Institute of Technology, Atlanta, USA
[3]Faculty of Psychology and Educational Sciences, University of Geneva, Switzerland
[4]Distance Learning University, Geneva, Switzerland

Hertzog *et al.* evaluated the statistical power of linear latent growth curve models (LGCMs) to detect individual differences in change, i.e., variances of latent slopes, as a function of sample size, number of longitudinal measurement occasions, and growth curve reliability. We extend this work by investigating the effect of the number of indicators per measurement occasion on power. We analytically demonstrate that the positive effect of multiple indicators on statistical power is inversely related to the relative magnitude of occasion-specific latent residual variance and is independent of the specific model that constitutes the observed variables, in particular of other parameters in the LGCM. When designing a study, researchers have to consider trade-offs of costs and benefits of different design features. We demonstrate how knowledge about power equivalent transformations between indicator measurement designs allows researchers to identify the most cost-efficient research design for detecting parameters of interest. Finally, we integrate different formal results to exhibit the trade-off between the number of measurement occasions and number of indicators per occasion for constant power in LGCMs.

## 1. Introduction

Latent growth curve models (LGCMs) are an increasingly popular method for assessing change in longitudinal data (e.g., Duncan, Duncan, Strycker, Li, & Alpert, 1999; McArdle, 1988; Meredith & Tisak, 1990; Raudenbush & Bryk, 2002). A major advantage of these models is that they allow one to directly estimate random effects (i.e., individual

*Correspondence should be addressed to Timo von Oertzen, Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany (e-mail: vonoertzen@mpib-berlin.mpg.de).

DOI:10.1348/000711010X486633

differences) in latent intercepts and slopes across time. For a given variable, the intercept and slope variances represent individual differences in initial performance and change, respectively (Rogosa & Willett, 1985; Singer & Willett, 2003), and the intercept–slope covariance reflects the extent to which individual differences in initial status correlate with subsequent change.[1]

Recently, Hertzog, Lindenberger, Ghisletta, and von Oertzen (2006, 2008) reported a series of simulations on the statistical power to detect covariances and variances of change in linear single-indicator LGCMs. Statistical power is defined as the probability that a statistical test, such as a likelihood-ratio (LR) test, rejects a null hypothesis of, for instance, no variances in slopes (Snijders & Bosker, 1999). There are a number of study design parameters that influence power in longitudinal studies, including sample size, the observation time, the observation density, and the growth curve reliability (GCR). Hertzog *et al.* (2008) found that GCR, defined as the ratio of growth curve determined variance to total variance at the first longitudinal occasion, is an important determinant of power. When GCR was high, the power to detect moderate-sized true slope variance of half the intercept variance was typically high as long as sample size was moderate (say, 500 or greater) and there were four or more occasions of measurement. When GCR was low, the power to detect the same true slope variance was often low, even with large sample sizes and many longitudinal occasions.

GCR deviates from unity to the extent that residual variance not determined by the growth curve increases from 0. The total residual variance for each measurement in an LGCM is a sum of two independent variance sources: (a) measurement error of the observed variable induced by an error of the measurement itself (Lord & Novick, 1968; McDonald, 1999); and (b) latent residuals, representing the imperfection of the model's simplified approximation to reality. In a study design using a single indicator for each observed variable that changes over time, these two sources of residual variances cannot be separately identified. Instead, both sources are absorbed into a single residual variance parameter. Nevertheless, they have distinct meanings and interpretations (McArdle, 1988).

Measurement error in observed variables is well understood in terms of classical notions of reliability or dependability (e.g., Cronbach, Gleser, Nanda, & Rajaratnam, 1972); namely, it is the extent to which respondents generate inconsistent observed scores on a given measure, given equivalent underlying scores on the latent variable. Measurement errors in latent variable models will typically also include systematic sources of errors that are unique to the measure. The latent residuals, in contrast, reflect the extent to which the functional form of growth in the LGCM fails to fit the changing variable over time. A variable can be highly reliable at a given point in time but highly fluctuant over time. For example, this pattern could emerge when measuring cyclic mood states (e.g., Hertzog & Nesselroade, 1987; Nesselroade, 1991). On the other hand, measurement error at the level of each indicator imposed on a linear latent growth process may cause substantial deviations in a given variable around its latent intra-individual regression function, even when the growth process for true scores is perfectly linear in its functional form (Rogosa, Brandt, & Zimowski, 1982; Singer & Willett, 2003).

[1] This interpretation assumes that the LGCM is scaled to have the inception of the longitudinal observations define the intercept (cf. Rovine & Molenaar, 1998).

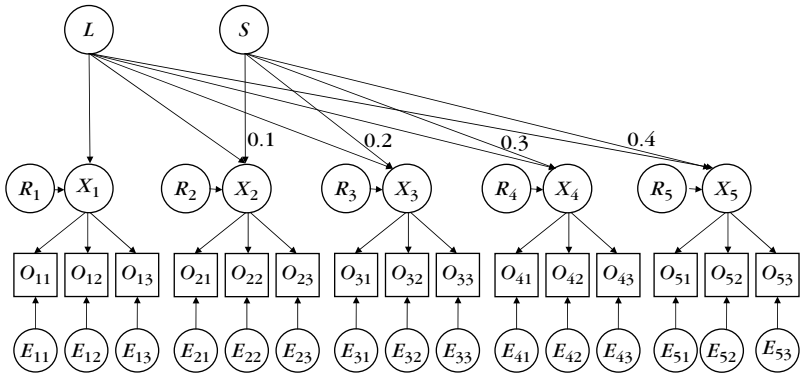


**Figure 1.** Example LGCM with five occasions of measurement and three indicators at each measurement. The latent residuals $R_1$–$R_5$ affect the latent score at each time point, while the measurement errors $E_{11}$–$E_{53}$ affect each indicator separately. *L*, level; *S*, slope; *X*, latent score; *O*, observation.

In the present paper, we investigate multiple-indicator LGCMs (MI-LGCMs), in which two or more indicators are available at each occasion as measures for each latent variable. For our purposes, we assume that observed variables are available at all measurements, although this assumption can be relaxed (e.g., McArdle & Hamagami, 1992). However, our results are not limited to LGCMs but generalize to all structural equation models (SEMs) in which multiple indicators are exogenous manifest variables of common endogenous latent variables, irrespective of the regression model linking the two. Multiple indicators can be used for all SEMs, and the results regarding the benefits of multiple indicators for statistical power given in the present paper are valid regardless of the specification of the latent regression model defining constructs with multiple indicators.

An MI-LGCM is a second-order factor model (McArdle, 1988; Sayer & Cumsille, 2001) in which the first-order factors are defined by a classical occasion-specific longitudinal factor model (Tisak & Meredith, 1990). We assume the same indicators are available as measures of the first-order factors at all occasions. Figure 1 shows a hypothetical multiple-indicator model for a construct *X* measured at five time points by three indicators $O_1$, $O_2$, and $O_3$. MI-LGCMs have been recommended by methodologists (McArdle, 1988; Sayer & Cumsille, 2001) but have not been as widely used in practice as single-indicator models.

In contrast to single-indicator LGCMs, MI-LGCMs allow one to separately estimate the two sources of residual variance, since by definition measurement error (*E* in Figure 1) acts independently on each indicator variable, whereas the latent residual (*R* in Figure 1) at each occasion of measurement influences all indicators for that occasion.[2] A second advantage of MI-LGCMs is that multiple indicators correct for attenuation due to random measurement error (e.g., Bollen, 1989), and hence the GCR to detect linear constraints in the parameters of the model increases with multiple-indicator models compared to single indicators. Because of the importance of GCR to

[2] For convenience, we ignore the mixed case in which temporally correlated, systematic measurement error is shared by more than one indicator variable, but not all available indicator variables. Mixed cases can be handled by specifying and estimating the implied measurement residual covariances.

statistical power, Hertzog *et al.* (2008) speculated that multiple indicators could be an appropriate way to improve statistical power in LGCMs. Nevertheless, use of two or more tests to gain additional indicators may be costly, for instance in terms of money and respondent burden. Hence, it is important to evaluate how much power exactly can be gained from using multiple indicators in LGCMs.

Arguably, multiple indicators should be effective in reducing the impact of measurement error, but they should not affect the impact of latent residual variance. In the present paper, we apply extant formal results (cf. Dolan, Wicherts, & Molenaar, 2004; Hancock, 2006; Penev & Raykov, 2006; von Oertzen, 2010) to provide a simple proof for this assertion. We also show some implications of the independent effects on power for both residual variance sources $R$ and $E$ for applications of MI-LGCMs.

The separable influences on power lead to larger practical questions concerning the design of studies using LGCMs to analyse individual differences in change. Specifically, to what extent could limitations in GCR be compensated by using more indicators at each occasion? Adding indicators to a design may be less expensive than adding measurement occasions to study development in a longitudinal study. Hence, developmental researchers might wish to know the answer to the following question: how might one trade adding extra indicators for fewer longitudinal occasions of measurement and still maintain equal power to detect variances in slopes? Von Oertzen (2010) provided a general methodology to approach problems like this; combining his results with formal results about multiple indicators leads us to a formal expression for the trade-off between number of measurement occasions and number of indicators at each measurement for constant power to detect slope effects in an MI-LGCM. That is, we give an exact expression for how many occasions of measurement can be saved if additional indicators are added to the design while maintaining equal statistical power.

This problem can be understood as a case of optimal design for quantitative analysis of longitudinal data (e.g., Raudenbush & Xiao-Fang, 2001). As noted by Tomarken and Waller (2005), treatment of design optimality, in terms of cost and efficiency of a given design for adequately addressing critical hypotheses, is a much needed and somewhat neglected aspect of designing studies that target the use of SEMs. Statistical power provides a means of operationally defining optimal trade-offs between different features of the research design to make efficient and effective design choices (MacCallum, Lee, & Browne, 2010; Raudenbush & Xiao-Fang, 2001; see also MacCallum, Browne, & Li, 2006).

The remainder of this paper is organized as follows. In Section 2 we introduce the formal relation between multiple-indicator and single-indicator models and discuss its implications for designing multiple-indicator studies. We then provide a realistic research example to exemplify and validate the effects described analytically.

In Section 4 we show three main results that emerge from the combination of the analytical and empirical power analysis. First, we show how the two residual variance sources, i.e., latent residual and measurement error, affect statistical power in an MI-LGCM. Second, we give an example to show how costs of a study design may be reduced using the formal comparison of multiple-indicator models. Third, we investigate how many occasions of observation can be saved when adding additional indicators for the latent variable at each occasion in the example MI-LGCM. We then close by discussing the implications of these findings for the applied researcher.

## 2. Analytical comparison of multiple-indicator models

Using multiple indicators to measure a single variable can be perceived as a method of improving the reliability of the measurement. For example, from a psychometric perspective, two indicators with low reliability could have equal or greater efficiency to answer a given research question than a model with a single but more reliable indicator. Although this fact is commonly accepted, it is difficult to concretize what precisely is meant by 'efficiency' without a quantitative conception of it.

One operationalization of efficiency is the statistical power of a given test on the data. While some applied researchers harbour a good conception about the reliability needed for a single indicator, it is much less intuitive to decide on the number of multiple indicators and their reliabilities when multiple indicators are a viable option. So, the question that arises is: what reliability would a single indicator need in order to achieve equivalent statistical power to that of a given set of indicators? In other terms, what is the 'effective' reliability of a given set of indicators?

In this paper, we make use of a fairly simple analytical equation: measuring a construct with $K$ indicators with regression weights $\lambda_1, \ldots, \lambda_K$ and normally distributed measurement error with known, fixed variances $\sigma_1^2, \ldots, \sigma_K^2$, the power of any LR test on parameters relevant to the construct is the same as that of a test on the construct measured by a single indicator with an effective error variance ($\sigma_{\text{eff}}^2$) of

$$\sigma_{\text{eff}}^2 = \frac{1}{\sum_{i=1}^{K} \frac{\lambda_i^2}{\sigma_i^2}}. \tag{1}$$

In other words, for the power of a test, it is irrelevant whether the construct is measured by multiple indicators of specified error variances or by a single indicator of accordingly lower error variance. As a side remark, equation (1) is the harmonic mean of the error variances of the normalized indicators. It is also the error variance of a weighted pooling of the indicators, with loadings $1/\lambda_i$.

Von Oertzen (2010) systematically investigated *power equivalent* models, i.e., measurement models that produce the same power for LR tests. He proved that the above equation is in fact necessary and sufficient for precise power equivalence. In the following, we will adopt the term *power equivalent* for two measurement models that produce the same power for an LR test against the same specified hypothesis.

Other papers have investigated implications of this equation for statistical power in SEMs. Dolan *et al.* (2004) showed that a special case of this equation is a necessary condition for two alternative models having equal mean of the theoretical distribution of the log-likelihood ratio. Using the fact that power can be approximated by assuming a non-central chi-square distribution (cf. Satorra, Saris, & de Pijper, 1991), Dolan *et al.* showed that the power is approximately equal for two sets of indicators if the above equation yields the same value for $\sigma_{\text{eff}}^2$. Penev and Raykov (2006) provided an alternative proof that increased the generality of this condition. Independently, Hancock (2006) gave the same equation as a condition under which two sets of indicators allow for equally powerful tests, also using the non-centrality factor of the non-central chi-square distribution.

Dolan *et al.* (2004), Hancock (2006), and von Oertzen (2010) all argued that the main advantage of equation (1) over power approximations or Monte Carlo simulations is its generality. No matter how the construct that is measured by the multiple indicators emerges, the interdependency of the indicators can be described by equation (1). For example, consider a variable $X$ that may be measured either by two indicators of

variance 1 and 7, or by two other indicators with variance 2 and 3 (in any unit). Intuitively, it is difficult to judge which set may yield higher data quality. Equation (1) shows that the effective error is 0.875 for the first set of indicators, and 1.2 for the second set of indicators. Thus, a test of a hypothesis regarding $X$ will be more powerful with the first set of indicators. This will be equally true if $X$ is measured just once in two groups (e.g., to test for equal means), or if $X$ is measured longitudinally over several years assuming a linear LGCM (e.g., to test for a covariance between initial intercept and slope), or if $X$ is measured in comparison to a second variable $Y$ in a bivariate model such as a bivariate dual change score model (cf. McArdle *et al.*, 2004).

In this paper, we apply the equation to an MI-LGCM as a frequently used model which is important for many longitudinal studies. To compare the relative effect of the latent residual and measurement error on power, we consider the total variance of one measurement, which is the effective error from equation (1) plus the latent residual (see Appendix). This sum allows us to investigate the relative effect of two sources of residual variance.

In a second step, we compare the relative influence of indicators against the number of measurement occasions in the MI-LGCM, i.e., how often the variable of interest has been measured longitudinally in a fixed time-span. To do this, we integrate equation (1) into an equation given by von Oertzen (2010) that computes the effective error of a linear LGCM; the corresponding mathematical derivations can be found in the Appendix. The result allows us to compare how many measurement occasions can be saved if each construct is measured by multiple indicators while keeping the compared models power equivalent, i.e., while keeping the power of a test against a hypothesis involving the latent variables of the MI-LGCM constant. Note that although we exemplify the power here with a specific LR test (namely, a test on variance of slope), von Oertzen (2010) showed that an analogous result can be obtained for all other LR tests in this MI-LGCM.

## 3. Simulation model

We applied equation (1) to the specific problem of estimating power to detect a variance of slope in a linear LGCM. If it can be shown analytically for a family of LGCMs that they are power equivalent, power only needs to be computed for one of the LGCMs. Specifically, two LGCMs from that family that use two different sets of indicators with the same effective error given by equation (1) will always yield the same power for an LR test against a hypothesis of zero variance.

In the following, we estimated statistical power in a linear LGCM for four instances of the set of parameter variations explored more fully by Hertzog *et al.* (2008). The model is given in Figure 1. The target variable is observed at five occasions of measurement, with factor loadings of the slope increasing from 0 to .4 in steps of .1. All means are fixed to 0. The variance of the intercept is set to 100, and the variance of the slope to 50. The covariance of intercept and slope was set either to 25.36, which corresponds to a correlation of .5, or to 0. The simulated test population consists of $N = 200$ or 500 observations.

The latent score determined by the growth process of the variable in the model is measured by multiple indicators. Hence, each measurement has two sources of variance in addition to its latent score: a normally distributed residual variance that is common to all indicators, and a measurement error that is independently

distributed and of equal variance for each indicator. Both of these sources were manipulated in some conditions to validate the accuracy of the analytical transformations (see Table 1). We verified these transformations in a Monte Carlo simulation on an MI-LGCM with 10,000 repetitions, using the engine introduced by von Oertzen, Ghisletta, and Lindenberger (2010).

For all further power values given in this paper, we made use of the fact that if power is known as a function of GCR, equation (1) allows one to compute all different variations of MI-LGCMs with the same effective error variance. Figure 2, adapted from Hertzog *et al.* (2008), provides the power to reject the null hypotheses of zero slope variance, using an LR test with two degrees of freedom and a 5% significance criterion. The test compares the full model to a nested model with the slope variance and the intercept–slope covariance fixed at zero.[3] Each function plots power as a function of GCR. Each panel has a family of power curves for different numbers of occasions of measurement, and the different panels manipulate the intercept–slope correlation and sample size. The thick lines within four of the panels of Figure 2 designate four power curves we used to evaluate the trade-offs in power as a function of the number of indicators, latent residual and measurement error. We replicated each power value with 10,000 repetitions of a Monte Carlo simulation, yielding the same power values within the precision of the simulation.

## 4. Results

### *4.1. Latent residual versus measurement error*

To compare the influence of both residual variance sources with a fixed number of indicators, and also to check on the accuracy of the power equivalence transformations, we compared results from the reformulated model to actual simulations of a set of possible models. The results are shown in Table 1. The number of indicators was fixed to three, all with a factor loading of 1.0 and identical measurement error variance. The error variance is given in the first column, and the latent residual variance is shown in the second. These two values were varied systematically to compare a variety of different residual variances and measurement error variances. The sum of these two variances is the basis for the GCR (i.e., the ratio of all variance which is not part of the growth process to the total residual variance) of a single indicator at the first measurement occasion (column 3). The effective error, i.e., the error that a single-indicator model with identical power would have, computed using equation (1), is given in column 4. The estimated power to reject a hypothesis of zero slope variance in columns 6–9 is given for two different intercept–slope correlations (0 and .5) and two sample sizes ($N = 200$ and 500).

As predicted by the analytical equation, lines with identical effective error are indeed power equivalent, i.e., generate equal power, to the level of precision afforded by the simulation (within a margin of 1.3 percentage points). For example, compare lines 4, 11, and 13, in which three models with an effective error of 20.33, but with different

[3] This LR test is at a boundary of the parameter space. Stoel, Garre, Dolan, and Wittenboer (2006) argue for a mixture distribution test in general practice for such cases; Savalei and Kolenikov (2008) give reasons against this suggestion. When following Stoel *et al.* (2006), all power values in the sequel would be slightly higher. Importantly, though, the relative contribution of the indicators would still be unchanged, and all subsequent results in this paper remain the same.

**Table 1.** Power and effective reliability for different indicator error and latent residual

| | | | | Power [%] | | | |
|---|---|---|---|---|---|---|---|
| | | | | $r = 0$ | | $r = .5$ | |
| Indicator error | Latent residual | Reliability of first indicator | Effective error | $N = 200$ | $N = 500$ | $N = 200$ | $N = 500$ |
| 0.0 | 1.0 | .991 | 0.33 | 100.0 | 100.0 | 100.0 | 100.0 |
| 5.0 | 1.0 | .943 | 5.33 | 100.0 | 100.0 | 100.0 | 100.0 |
| 10.0 | 1.0 | .901 | 10.33 | 96.2 | 100.0 | 100.0 | 100.0 |
| 20.0 | 1.0 | .826 | 20.33 | 57.5 | 92.9 | 97.0 | 100.0 |
| 0.0 | 16.0 | .862 | 5.33 | 100.0 | 100.0 | 100.0 | 100.0 |
| 5.0 | 16.0 | .826 | 10.33 | 96.6 | 100.0 | 100.0 | 100.0 |
| 10.0 | 16.0 | .794 | 15.33 | 76.6 | 99.2 | 99.4 | 100.0 |
| 20.0 | 16.0 | .735 | 25.33 | 40.4 | 80.2 | 92.6 | 100.0 |
| 0.0 | 31.0 | .763 | 10.33 | 96.0 | 100.0 | 100.0 | 100.0 |
| 5.0 | 31.0 | .735 | 15.33 | 77.5 | 99.0 | 99.5 | 100.0 |
| 10.0 | 31.0 | .709 | 20.33 | 56.3 | 93.0 | 97.1 | 100.0 |
| 20.0 | 31.0 | .662 | 30.33 | 32.3 | 66.6 | 86.9 | 99.9 |
| 0.0 | 61.0 | .621 | 20.33 | 56.1 | 93.2 | 97.2 | 100.0 |
| 5.0 | 61.0 | .602 | 25.33 | 41.7 | 80.1 | 92.8 | 100.0 |
| 10.0 | 61.0 | .585 | 30.33 | 31.6 | 66.9 | 86.7 | 99.9 |
| 20.0 | 61.0 | .552 | 40.33 | 20.1 | 46.3 | 75.0 | 98.8 |

*Note*. Systematically altering measurement error (first column) and latent residual (second column), the table gives the reliability of a single indicator (third column) and the effective error (fourth column), i.e., the error variance that a single-indicator model would need to have to provide identical power. Columns 5–8 give the power to reject the null hypothesis of zero slope variance for two different intercept–slope correlations and two different sample sizes. All power values were independently simulated using a Monte Carlo simulation with 10,000 repetitions each. All rows with identical effective error (which are lines 2 and 5; lines 3, 6, and 9; lines 7 and 10; lines 4, 11, and 13; lines 8 and 14; and lines 12 and 15) show identical power within an error margin of 1.3 percentage points, which is within the confidence interval of the simulation.

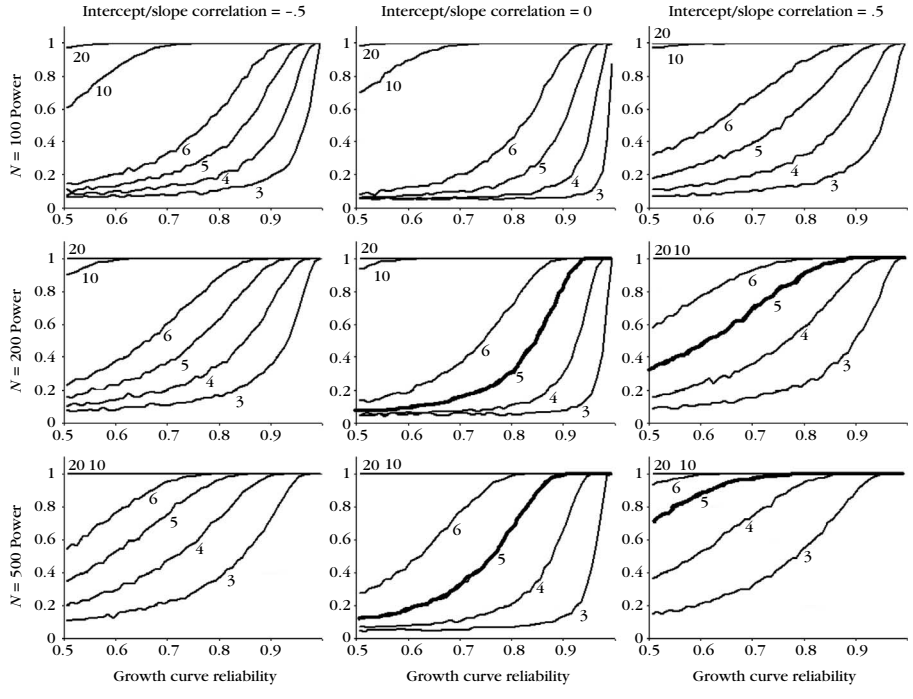


**Figure 2.** The power of a test to reject the null hypothesis of zero slope variance as a function of GCR, as reported by Hertzog *et al.* (2008). The data are plotted for three sample sizes ($N = 100$, 200, and 500) and three different intercept–slope correlations ($-.5$, 0, and .5). The curves in each panel correspond to different number paper are shown as of longitudinal occasions. The four situations used in this thick curves. (Figure adapted from Hertzog *et al.*, 2008).

constellations of measurement error variance and latent residual variance, produce approximately the same estimated power to detect the slope variance in all four conditions (power values in all four lines are close to 0.57, 0.93, 1.0, and 1.0, respectively).

To reiterate, the effective error is independent of the actual model that is used and would be identical for other models, for instance group comparisons or dual change score models, or any other model with normally distributed measurement error. For different pairs of values of latent residual and measurement error with identical effective error, the power for a test against any given hypothesis on the latent parameters of these models would have the same value. What this value is depends on the actual given model, but when comparing different combinations between latent residual and measurement error, the model and the test are irrelevant (cf. Dolan *et al.*, 2004; Hancock, 2006; von Oertzen, 2010).

Figure 3 shows the increase in power (scaled as the proportion of simulated cases when the null hypothesis is correctly rejected) for detecting slope variance with increasing number of indicators for the same reliability of a single indicator at the first measurement occasion, partitioned into different proportions of latent residual and measurement error. The four panels correspond to two different correlations $r$ of intercept and slope (0 and .5) and two different sample sizes ($N = 200$ and 500). Note that only the thick curves in Figure 2 were used (i.e., just one series of univariate LGCM

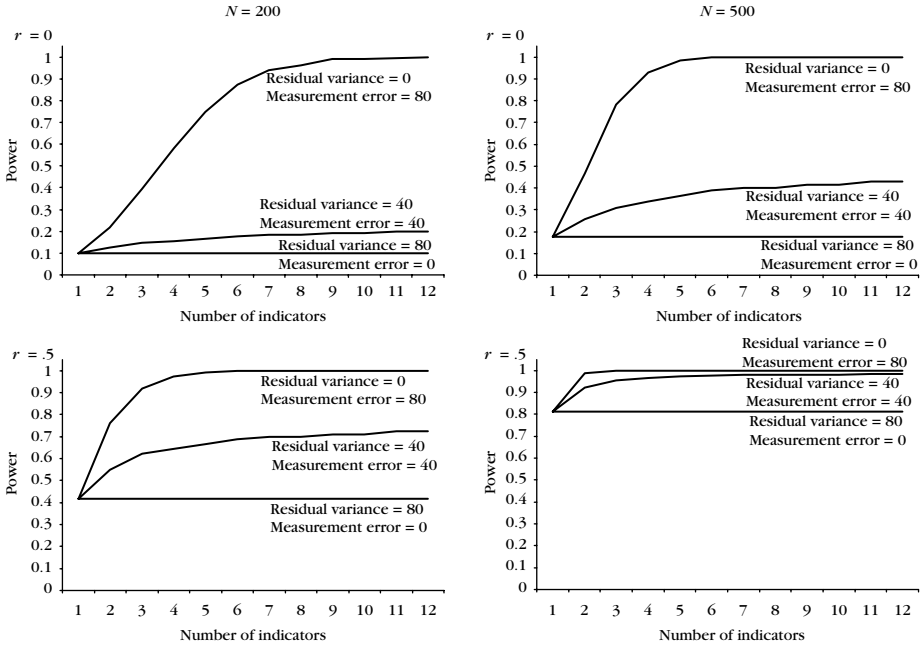


**Figure 3.** Power to detect a slope variance of 50 at an intercept variance of 100, using five measurements occasions (weights from 0.0 to 0.4, in steps of 0.1). Power is given for four different situations, crossing two values for intercept–slope correlation ($r = 0$ and .5) and two sample sizes ($N = 200$ and 500). The *x*-axis gives the number of indicators at each measurement occasion. The error of one indicator at the first measurement occasion is fixed to 80 (reliability .55) in all three curves, but differently partitioned on residual at the latent level of the model and the measurement error at each indicator. While the curves with no latent residual quickly rise to an asymptote of perfect power in all four panels, the curves with all residual variance in the latent residual are unaffected by increasing the numbers of indicators. The middle curve in all four panels corresponds to a residual variance of 40 and measurement error of 40, and rises less quickly to a lower asymptote.

simulations per panel) to create the full set of values crossing residual and measurement error values. Every possible partitioning of residual variance could be derived using equation (1).

It can be seen that if the total residual variance is comprised solely of indicator measurement error (the highest curve in each panel), power quickly rises to an asymptote of 1.0 with an increasing number of indicators per measurement. If on the other hand the total residual variance is completely comprised of the latent residual (the lowest curve in each panel), multiple indicators do not increase the power to detect the variance of slope. The middle curve in each panel shows a mixed case where half of the residual variance is latent, the other half due to measurement error. Power here also increases to an asymptote, but one that is below 1.0. To be precise, the asymptote is the power value of a model with the latent residual variance as total residual variance. Note that the same qualitative picture emerges from all four situations given, but that in situations where power was high to begin with (e.g., $N = 500$ and $r = .5$), the beneficial effect of adding more indicators is accordingly less pronounced.

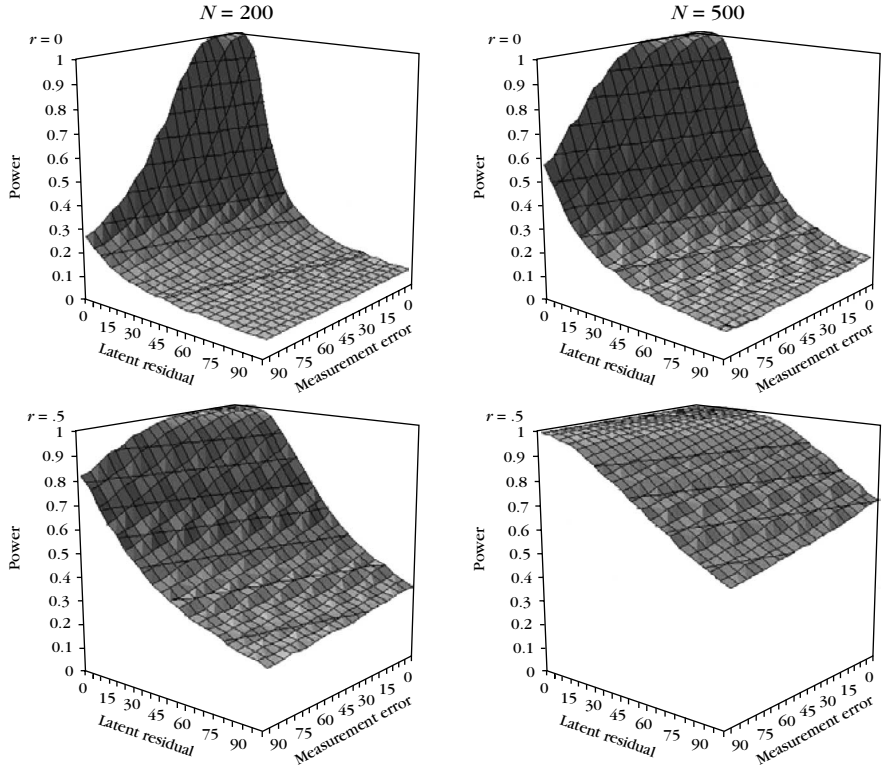


**Figure 4.** Power to detect variance of slope in an LGCM with three indicators of the same reliability, for two values of intercept–slope correlation ($r = 0$ and .5) and two sample sizes ($N = 200$ and 500). The *x*-axis in each panel shows the latent residual, i.e., the variance at the latent intercept common to all three indicators, while the *y*-axis shows the variance of the measurement error at each indicator. Note that in all four panels, the iso-powers are always straight lines with a slope of $-3$.

Figure 4 gives three-dimensional power plots for the four base curves selected from Figure 2. These plots examine a full range of jointly manipulated residual variance and measurement error variance for each base model that has been simulated, applying equation (1). The *x*- and *y*-axes show the latent residual and the measurement error specific to each indicator, respectively. The *z*-axis shows the power to detect a variance of slope in the LGCM described above. Power is at 1.0 in the upper left corner where both residual variance sources are zero and monotonically drops when either residual variance source is increased. Yet, the decrease is steeper when increasing the latent residual (right face) than when increasing measurement error (left face). The actual power values in the front left face of each panel in Figure 4 were estimated by a Monte Carlo simulation. Then, using equation (1), all other values of the three-dimensional space were computed from that.

The iso-power lines, that is, the lines between measurement error and latent residual that give the same power in all panels of Figure 4, are all linear (cf. Figure 5). While the actual power values are dependent on the parameters of the specified LGCM, the structure of the iso-power lines in Figure 5 is independent of the underlying model; they may correspond to different power values, but they will still be linear with

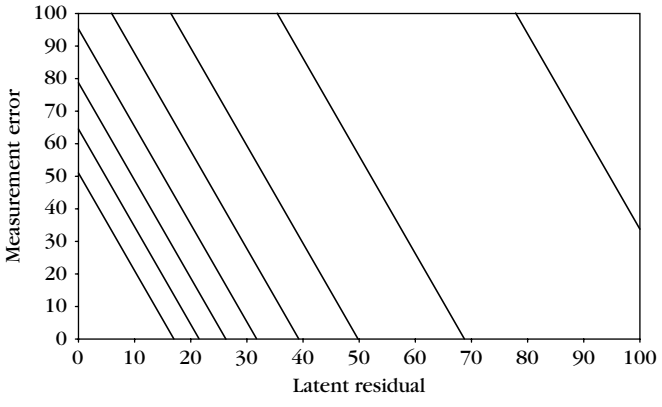


**Figure 5.** Lines of equal power to detect a parameter with three indicators of the same reliability. The *x*-axis gives the latent residual, the *y*-axis gives the measurement error. The lines shown are the iso-power lines for all four panels of Figure 4, and in fact for all SEMs using multiple indicators. We see that although the specific latent model is important for the values of power, the iso-power lines are completely independent of the particular latent model constituting the measured construct.

a slope of $-3$ (minus the number of indicators per measurement), as can be seen from equation (1). In other words, regardless of the model underlying the constructs, a decrease in residual variance is always three times better than an equal decrease in measurement error if three indicators of equal quality are used to measure the construct, which means that in order to balance an increase in residual variance, a three times greater decrease in measurement error must be achieved.

For an illustration of the effect of the number of indicators on the plots in Figure 4, the interested reader may access www.powerequivalence.com. On this website there is an interactive Excel spreadsheet where the reader may specify a number of indicators and observe how Figures 4 and 5 change accordingly.

### *4.2. Selecting indicators in a longitudinal design*

As shown, increasing the number of indicators has benefits for statistical power in LGCMs. Nevertheless, the practical issue from a design perspective is in choosing which indicators to use, considering the costs of implementing different measurement schemes. We demonstrate a simple use of equation (1) which is generic to all multiple-indicator models. This approach complements the procedures for *a priori* power calculations using goodness-of-fit indices recommended by MacCallum *et al.* (2006), while additionally addressing a specific design question.

As an example, assume a researcher wants to design a longitudinal study with five occasions of measurement to investigate a training effect using a single indicator with a measurement error variance of 10 units, at a resource cost of \$100 per participant and measurement occasion. As an alternative, four other indicators could be used in any combination to measure the same underlying construct. Among these, one has an error variance of 15 and costs of \$30 per participant and measurement, while the other three have an error variance of 30 and costs of \$20 per participant and measurement occasions.

As can be seen in Table 1, the initial design then would have a power of approximately 97% (third line, with a slight correction because of the measurement error). The total monetary costs would be $200 \times 5 \times \$100 = \$100{,}000$. Equation (1) shows that identical power can be achieved by using the three indicators with identical error variance of 30:

$$\frac{1}{\frac{1}{30} + \frac{1}{30} + \frac{1}{30}} = 10. \tag{2}$$

This corresponds to line 9 in Table 1. The costs would then reduce to $200 \times 5 \times 3 \times \$20 = \$60{,}000$. In fact, again the same power is reached by using the indicator with medium reliability together with one of the least expensive indicators:

$$\frac{1}{\frac{1}{30} + \frac{1}{15}} = 10. \tag{3}$$

The costs for this third possible study design are $200 \times 5 \times (\$20 + \$30) = \$50{,}000$.

All three possible designs have exactly the same power to detect the covariance between intercept and slope, but differ by a factor of 2 with respect to the financial costs of the study. Moreover, other factors, such as higher risk of drop-out and greater retest effects, can be considered when applying this method.

### *4.3. Multiple indicators versus measurement occasions*

An important decision when planning a longitudinal study involves finding the best compromise between the number of occasions of measurement and the number of indicators at each measurement occasion. Assume a researcher wants to investigate whether there is a non-zero variance of slope in a cognitive training programme. For example, assume that in the population, the slope is uncorrelated with the intercept, and its variance is half the size of the intercept variance (e.g., as before, 100 and 50 units). Assume further that the initial study design has six measurement occasions using four indicators at each occasion, each indicator with an error variance of 20, and a negligible latent residual. In this situation, we estimated the power to be 99%.

The Appendix demonstrates how equation (1) can be combined with results from von Oertzen (2010) to analytically describe how many indicators must be chosen for a given number of observations to get the same power as in the original study design. Figure 6 shows the graph of this equation. Note that theoretically the number of indicators and the number of observations can take any non-negative real value, but since in practice these are positive integers, the circles denote points with integer values for the number of indicators. With four indicators, we need 6 measurement occasions to get a power of 0.99, as already mentioned. This corresponds to a total of 24 tests for each participant. With three indicators, we get the same power with 9 measurement occasions; this corresponds to 27 tests in total, so this study design with the same power will presumably be more expensive. Only 2 indicators are necessary to get the same power with 15 measurement occasions (30 tests in total), and with a single-indicator design, 31 occasions would be necessary to get the same power. The design with the fewest measurements per participant uses 4 measurement occasions with 5 indicators, which gives a total of only 20 tests per participant. Note that the advantage is presumably even stronger, since administering multiple tests at a single session is usually cheaper in terms of resources than administering a single test over multiple sessions.

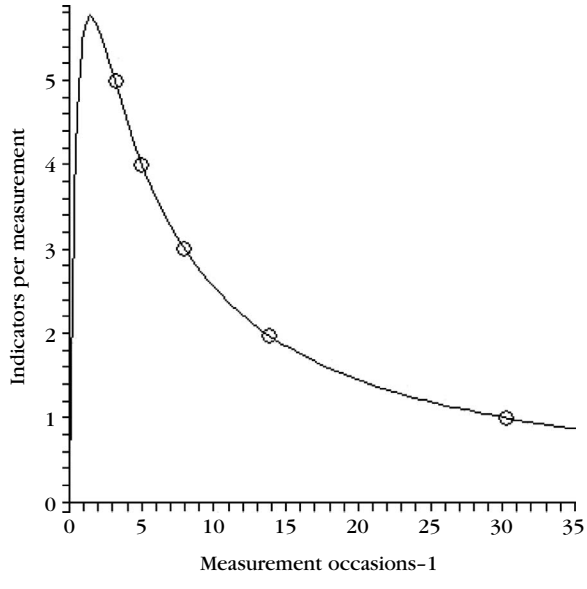


**Figure 6.** Iso-power line to detect variance of slope in an LGCM for number of indicators per occasion and number of measurement occasions. The circles mark points with an integer number of indicators. For $N = 200$ participants, intercept variance $\sigma_L^2 = 100$, slope variance (effect size) $\sigma_S^2 = 50$, indicator error $\sigma^2 = 20$, and 5 time units of observations, the power corresponds to 0.99. Note that the iso-power line itself is independent of these parameters.

Note, though, that this illustration assumes identically reliable, unbiased, and congeneric indicators, which may not be available.

Figure 7 illustrates the interaction of measuring more often and using more indicators against power. The three panels correspond to the same proportions of latent residual and measurement error as in Figure 3, again assuming a constant total residual variance for a single indicator. Figure 7a shows power assuming that the total residual variance is entirely a measurement error and the latent construct has no residual. In this case, power is strongly affected by adding more indicators. With five indicators, power starts at 70% even with only two occasions, quickly rising to maximum power. In Figure 7b the total residual variance is equally split between measurement error and latent residual. Multiple indicators still have a clearly visible effect, but much less strong than in Figure 7a; for five indicators, power is hardly increased for two measurement occasions, and still below 90% for 20 measurements. In Figure 7c, measurement error is assumed to be zero, while the latent residual is maximal. In this case, power is unaffected by adding more indicators.

## 5. Discussion

Our results confirm earlier speculations (e.g., Hertzog *et al.*, 2008) that adding multiple indicators increases statistical power to detect variance of slopes in LGCMs. Increasing the number of indicators is an important method available to the investigator to ensure maximal power in a given applied research setting. The cost is merely added complexity to the analysis, forcing the use of a multiple-indicator SEM to

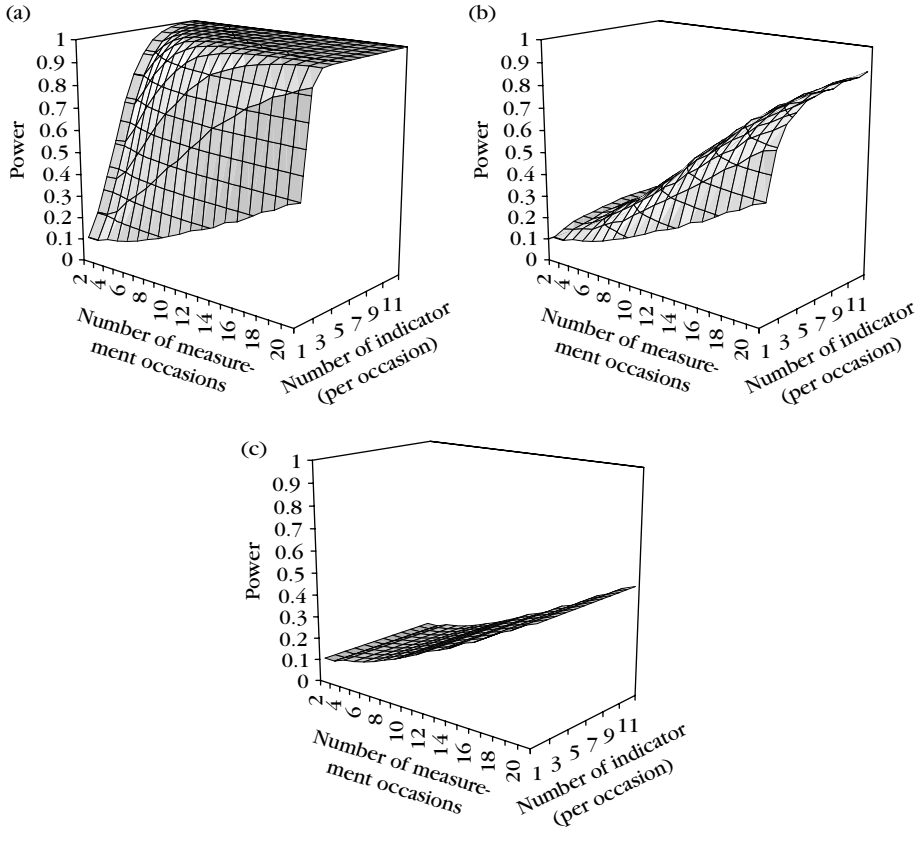


**Figure 7.** Power to detect variance of slope in an LGCM plotted against number of indicators per occasion and number of measurement occasions in a fixed time-span. Power values are given for $N = 200$ participants, intercept variance $\sigma_L^2 = 100$, slope variance (effect size) $\sigma_S^2 = 50$, no intercept–slope covariance, and total residual variance of 80. Part (a) shows the power for a measurement error of 80 and no latent residual. In (c) the indicators have no measurement error, but the residual of the latent construct is 80. Part (b) is the intermediate case with a measurement error of 40 and a latent residual of 40.

estimate the LGCM. This is obviously a small cost relative to the benefit of increased power when adding indicators of the latent variable to the model. Hence, we recommend the use of multiple-indicator designs when the model to be applied is an LGCM.

The analytical generality of equation (1) implies that the same principle applies to tests for covariances of slopes in a bivariate LGCM. We validated this by some simulations omitted from this paper. With fixed sample size, GCR, and effect size, adding multiple indicators will increase statistical power up to the maximum implied by a GCR of .99 (Hertzog *et al.*, 2006). As noted in that paper, power to detect correlated change between two variables is often disappointing when GCR is below .90, but is often excellent when GCR approaches asymptote, even for small effect sizes. If it there are substantial deviations of the measurements from the latent score of the variables, thereby producing a substantial amount of measurement error variance, then the

investigator can avoid low-power problems by adding more indicators to the model. This inference assumes, of course, that one can find multiple congeneric indicators for the latent variable.[4] Furthermore, in applied research settings other critical design aspects need consideration (e.g., drop-out, retest effects, learning).

These findings resolve the apparent inconsistency that power to detect slope correlations in single-indicator LGCMs in Hertzog *et al.* (2006, 2008) was often low when the number of occasions was three, but that studies using two occasions of measurement with multiple indicators produced significant correlations of change slopes using the latent difference-score model (McArdle & Nesselroade, 1994; for representative empirical results, see Hertzog, Dixon, Hultsch, & MacDonald, 2003; Raz *et al.*, 2005). The greater power in the latent difference score models can now be understood as a consequence of using multiple rather than single indicators to define latent variables.

The investigator interested in estimating power *a priori* for MI-LGCM applications can do so by using available statistical software to generate a simulated solution for a given set of parameters. This approach can be accomplished using Mplus (Muthén & Muthén, 2002) or other structural equation modelling software programs. The analytical technique demonstrated in this paper can be used to select an appropriate number of indicators. Methods introduced by von Oertzen (2010) can be used analogously for other study design parameters.

A strength of the method used in this paper is that it transfers directly to more complicated models that may be a better approximation to aggregate development functions, including nonlinear LGCMs (e.g., McArdle, Ferrer-Caja, Hamagami, & Woodcock, 2002). Clearly the benefits of an MI-LGCM over its univariate counterparts are a strong motivation for considering the structural equation approach to LGCMs (Curran, 2003; Rovine & Molenaar, 2000) when random effects are at the heart of the research question of interest.

A major problem in multivariate research is that researchers often design studies without sufficient attention to the power to detect conceptually important statistical parameters, as well as precision of estimation and probability of replication (Maxwell, Kelley, & Rausch, 2008). MacCallum *et al.* (2006) have shown that it is possible to use goodness-of-fit statistics to estimate *a priori* power for a model to detect a researcher-defined loss of fit that may be practically meaningful. This approach is useful and important, and is possibly used too infrequently.

Our results regarding the trade-off between density of time-based sampling (number of measurement occasions) and number of indicators help to show another dimension of how power can be increased. Large numbers of indicators help to decrease the effect of measurement error, but they are not effective in reducing the latent residual variance. Our results show that denser measurement, in contrast, can reduce the effect of the latent residual. Whether to choose more indicators or rather fewer indicators but denser measures is hence dependent on the way the residual variance is distributed between latent residual and measurement error. Appreciation of this fact, combined with the use of the techniques illustrated in this paper, can aid in optimizing the power of a longitudinal design to detect random effects in change,

[4] For other considerations regarding the selection of good multiple indicators, see Little, Lindenberger, and Nesselroade (1999).

provided that the researcher is willing to make informed guesses about the reliability of target measures and the lability of the growth processes under study.

## References

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability scores and profiles*. New York: Wiley.

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

Dolan, C. V., Wicherts, J. M., & Molenaar, P. C. M. (2004). A note on the relationship between the number of indicators and their reliability in detecting regression coefficients in latent regression analysis. *Structural Equation Modeling*, *11*(2), 210–216. doi:10.1207/s15328007sem1102_4

Duncan, T. E., Duncan, S. C., Strycker, L. A., Li, F., & Alpert, A. (1999). *An introduction to latent variable growth curve modeling: Concepts, issues, and applications*. Mahwah, NJ: Erlbaum.

Hancock, G. R. (2006). Power analysis in covariance structure modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 69–115). Greenwich, CT: Information Age Publishing.

Hertzog, C., Dixon, R. A., Hultsch, D. F., & MacDonald, S. W. S. (2003). Latent change models of adult cognition: Are changes in processing speed and working memory associated with changes in episodic memory? *Psychology and Aging*, *18*, 755–769. doi:10.1037/0882-7974.18.4.755

Hertzog, C., Lindenberger, U., Ghisletta, P., & von Oertzen, T. (2006). On the power of multivariate latent growth curve models to detect correlated change. *Psychological Methods*, *11*(3), 244–252. doi:10.1037/1082-989X.11.3.244

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

Hertzog, C., & Nesselroade, J. R. (1987). Beyond autoregressive models: Some implications of the trait–state distinction for the structural modeling of developmental change. *Child Development*, *58*, 93–109. doi:10.2307/1130294

Little, T. D., Lindenberger, U., & Nesselroade, J. R. (1999). On selecting indicators for multivariate measurement and modeling with latent variables: When 'good' indicators are bad and 'bad' indicators are good. *Psychological Methods*, *4*, 192–211. doi:10.1037/1082-989X.4.2.192

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

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

MacCallum, R., Lee, T., & Browne, M. W. (2010). The issue of isopower in power analysis for tests of structural equation models. *Structural Equation Modeling*, *17*, 23–41. doi:10.1080/10705510903438906

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for power and accuracy in parameter estimation. *Annual Review of Psychology*, *59*, 537–563. doi:10.1146/annurev.psych.59.103006.093735

McArdle, J. J. (1988). Dynamic but structural equation modeling of repeated measures data. In J. Nesselroad & R. Catell (Eds.), *Handbook of multivariate experiment psychology* (pp. 561–614). New York: Plenum Press.

McArdle, J. J., Ferrer-Caja, E., Hamagami, F., & Woodcock, R. W. (2002). Comparative longitudinal structural analysis of growth and decline of multiple intellectual abilities over the life span. *Developmental Psychology*, *38*, 115–142.

McArdle, J. J., & Hamagami, F. (1992). Modeling incomplete longitudinal and cross-sectional data using latent growth structural models. *Experimental Aging Research*, *18*, 145–166.

McArdle, J. J., Hamagami, F., Jones, K., Jolesz, F., Kikinis, R., Spiro, A., & Albert, M. S. (2004). Structural modeling of dynamic changes in memory and brain structure using longitudinal data from the Normative Aging Study. *Journal of Gerontology: Psychological Sciences*, *59B,* 294–304.

McArdle, J. J., & Nesselroade, J. (1994). Structuring data to study development and change. In S. H. Cohen & H. W. Reese (Eds.), *Life-span developmental psychology: Methodological innovations* (pp. 223–268). Hillsdale, NJ: Erlbaum.

McDonald, R. P. (1999). *Test theory: A unified theory.* Mahwah, NJ: Erlbaum.

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

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

Nesselroade, J. R. (1991). The warp and the woof of the developmental fabric. In R. Downs, L. Liben, & D. Palermo (Eds.), *Visions of aesthetics, the environment and development: The legacy of Joachim Wohlwill* (pp. 213–240). Hillsdale, NJ: Erlbaum.

Penev, S., & Raykov, T. (2006). Maximal reliability and power in covariance structure models. *British Journal of Mathematical and Statistical Psychology*, *59*, 75–87. doi:10.1348/000711005X68183

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis models* (2nd ed.). Thousand Oaks, CA: Sage.

Raudenbush, S. W., & Xiao-Fang, L. (2001). Effects of study duration, frequency of observation, and sample size on power in studies of group differences in polynomial change. *Psychological Methods*, *6*, 387–401.

Raz, N., Lindenberger, U., Rodrigue, K. M., Kennedy, K. M., Head, D., Williamson, A., … Acker, J. D. (2005). Regional brain changes in aging healthy adults: General trends, individual differences and modifiers. *Cerebral Cortex*, *15*, 1676–1689.

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

Rogosa, D., & Willett, J. (1985). Understanding correlates of change by modeling individual differences in growth. *Psychometrika*, *50*, 203–228. doi:10.1007/BF02294247

Rovine, M. J., & Molenaar, P. C. M. (1998). The covariance between level and shape in the latent growth curve model with estimated basis vector coefficients. *Methods of Psychological Research Online*, *3*, 95–107. Retrieved from http://www.pabst-publications.de/mpr/

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

Satorra, A., Saris, W. E., & de Pijper, W. M. (1991). A comparison of several approximations to the power function for the likelihood ratio test in covariance structure analysis. *Statistica Neerlandica*, *45*, 173–185. doi:10.1111/j.1467-9574.1991.tb01302.x

Savalei, V., & Kolenikov, S. (2008). Constrained versus unconstrained estimation in structural equation modeling. *Psychological Methods*, *13*, 150–170. doi:10.1037/1082-989X.13.2.150

Sayer, A. G., & Cumsille, P. E. (2001). Second-order latent growth models. In L. Collins & A. Sayer (Eds.), *New methods for the analysis of change* (pp. 177–200). Washington, DC: American Psychological Association.

Singer, J. D., & Willett, J. (2003). *Applied longitudinal data analysis*. New York: Oxford University Press.

Snijders, J. D., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London: Sage.

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

Tisak, J., & Meredith, W. (1990). Descriptive and associative developmental models. In A. von Eye (Ed.), *Statistical methods in longitudinal research* (Vol. 1, pp. 125–149). San Diego, CA: Academic Press.

Tomarken, A. J., & Waller, N. G. (2005). Structural equation modeling: Strengths, limitations, and misconceptions. *Annual Review of Clinical Psychology*, *1*, 31–65. doi:10.1146/annurev.clinpsy.1.102803.144239

von Oertzen, T. (2010). Power equivalence in structural equation modelling. *British Journal of Mathematical and Statistical Psychology*, *63*, 257–272. doi:10.1348/000711009X441021

von Oertzen, T., Ghisletta, P., & Lindenberger, U. (2010). Simulating statistical power in latent growth curve modeling: A strategy for evaluating age-based changes in cognitive resources. In M. Crocker & J. Siekmann (Eds.), *Resource adaptive cognitive processes* (chap. 2.5, pp. 95–118). Heidelberg: Springer-Verlag.

*Received 17 July 2009; revised version received 2 November 2009*

## Appendix

### *Measurement error and latent residual*

For $K$ indicators of identical reliability,[5] i.e., with regression weight 1 and variance $\sigma^2$, equation (1) simplifies to $\sigma^2/K$. If the residual variance $\sigma^2_{\text{res}}$ of the latent variable measured by these indicators is not zero, this latent residual must be added to the effective error, so that in total the effective error is given by

$$\sigma^2_{\text{eff}} = \sigma^2_{\text{res}} + \frac{\sigma^2}{K}. \tag{A1}$$

Solving this equation for $\sigma^2$ yields

$$\sigma^2 = -K\sigma^2_{\text{res}} + K\sigma^2_{\text{eff}} \tag{A2}$$

which shows that for constant power, $\sigma^2$ and $\sigma^2_{\text{res}}$ form a line of slope $-K$, as shown in Figure 5.

### *Number of indicators and number of measurement occasions*

As an example of power equivalent models, von Oertzen (2010) gives the following equation for the effective error to perform a test of the variance of slope in a linear LGCM:

$$\sigma^2_{\text{eff}} := \frac{12M\sigma^2_E\left((M+1)\sigma^2_L + \sigma^2_E\right)}{T^2(M+1)\left(2(2M+1)\sigma^2_E + (M+1)(M+2)\sigma^2_L\right)}. \tag{A3}$$

Here, $T$ is the total observation time over all measurements, $M$ is the number of measurement occasions in this time-span in addition to the baseline test, $\sigma^2_L$ is the

[5] In classical test theory (Lord & Novick, 1968), indicators with these properties are termed parallel forms.

variance of the intercept, and $\sigma_E^2$ the homogeneous error variance of each measurement occasion. Covariance between intercept and slope is assumed to be zero. Hence, this equation allows us to describe the data quality for an LR test on slope variance for different numbers of occasions. As can be seen, reducing $M$ (i.e., fewer observations in the same time interval) increases the effective error. To see how this effect can be balanced by adding more indicators of identical reliability, we substitute $\sigma_E^2$ with the right-hand side of equation (A1), assuming $\sigma_{\text{res}}^2 = 0$:

$$\sigma_{\text{eff}}^2 = \frac{12M\frac{\sigma^2}{K}\left((M+1)\sigma_L^2 + \frac{\sigma^2}{K}\right)}{T^2(M+1)\left(2(2M+1)\frac{\sigma^2}{K} + (M+1)(M+2)\sigma_L^2\right)}. \tag{A4}$$

To find out how many indicators are needed for a known number of measurement occasions, we solve this equation for $K$:

$$\sigma_{\text{eff}}^2 = \frac{12M\sigma^2\left(K(M+1)\sigma_L^2 + \sigma^2\right)}{T^2(M+1)K\left(2(2M+1)\sigma^2 + K(M+1)(M+2)\sigma_L^2\right)}, \tag{A5}$$

$$\begin{aligned}\Leftrightarrow 0 = {} & K^2\left(T^2\sigma_{\text{eff}}^2(M+1)^2(M+2)\sigma_L^2\right) \\ & + K\left(2T^2\sigma_{\text{eff}}^2(M+1)(2M+1)\sigma^2 - 12M(M+1)\sigma^2\sigma_L^2\right) - 12M\sigma^4.\end{aligned} \tag{A6}$$

$$\begin{aligned}\Leftrightarrow K = {} & \frac{\left(6M\sigma_L^2 - T^2\sigma_{\text{eff}}^2(2M+1)\right)\sigma^2}{T^2\sigma_{\text{eff}}^2\sigma_L^2(M+2)(M+1)} \\ & \pm \frac{\sigma^2\sqrt{12M\sigma_L^2(3M\sigma_L + (1-M)T^2\sigma_{\text{eff}}^2) + T^4\sigma_{\text{eff}}^4(2M+1)^2}}{T^2\sigma_{\text{eff}}^2\sigma_L^2(M+2)(M+1)}.\end{aligned} \tag{A7}$$

In our example, numerical values were $M = 5$, $T = 5$, $\sigma_L^2 = 100$, $\sigma^2 = 20$, $K = 1$. Following equation (A4), to achieve a reasonable power of .9, the effective error must be $\sigma_{\text{eff}}^2 = 121/431$ in this situation. Substituting all numerical values with the exception of $M$ into equation (A7) yields

$$K = \frac{20204M - 242 \pm \sqrt{209107304M^2 + 5123624M + 29282}}{605(M+2)(M+1)} \tag{A8}$$

This equation was used to plot the graph in Figure 6.