

Hochweber, Jan; Hartig, Johannes

Analyzing organizational growth in repeated cross-sectional designs using multilevel structural equation modeling

formal und inhaltlich überarbeitete Version der Originalveröffentlichung in:

formally and content revised edition of the original source in:

Methodology 13 (2017) 3, S. 83-97



Bitte verwenden Sie beim Zitieren folgende URN /

Please use the following URN for citation:

urn:nbn:de:0111-pedocs-158678

<http://nbn-resolving.de/urn:nbn:de:0111-pedocs-158678>

Nutzungsbedingungen

Gewährt wird ein nicht exklusives, nicht übertragbares, persönliches und beschränktes Recht auf Nutzung dieses Dokuments. Dieses Dokument ist ausschließlich für den persönlichen, nicht-kommerziellen Gebrauch bestimmt. Die Nutzung stellt keine Übertragung des Eigentumsrechts an diesem Dokument dar und gilt vorbehaltlich der folgenden Einschränkungen: Auf sämtlichen Kopien dieses Dokuments müssen alle Urheberrechtshinweise und sonstigen Hinweise auf gesetzlichen Schutz beibehalten werden. Sie dürfen dieses Dokument nicht in irgendeiner Weise abändern, noch dürfen Sie dieses Dokument für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, aufführen, vertreiben oder anderweitig nutzen.

Mit der Verwendung dieses Dokuments erkennen Sie die Nutzungsbedingungen an.

Terms of use

We grant a non-exclusive, non-transferable, individual and limited right to using this document.

This document is solely intended for your personal, non-commercial use. Use of this document does not include any transfer of property rights and it is conditional to the following limitations: All of the copies of this documents must retain all copyright information and other information regarding legal protection. You are not allowed to alter this document in any way, to copy it for public or commercial purposes, to exhibit the document in public, to perform, distribute or otherwise use the document in public.

By using this particular document, you accept the above-stated conditions of use.

Kontakt / Contact:

peDOCS

Deutsches Institut für Internationale Pädagogische Forschung (DIPF)

Informationszentrum (IZ) Bildung

E-Mail: pedocs@dipf.de

Internet: www.pedocs.de

Mitglied der


Leibniz-Gemeinschaft

Accepted manuscript version (after peer review) of the following article:

Hochweber, J., & Hartig, J. (2017). Analyzing organizational growth in repeated cross-sectional designs. Using multilevel structural equation modeling. *Methodology*, 13(3), 83–97. <https://doi.org/10.1027/1614-2241/a000133>

© 2017 Hogrefe Publishing

This version of the article may not completely replicate the final version published in the journal. It is not the version of record and is therefore not suitable for citation.

The accepted manuscript is subject to the Creative Commons licence CC-BY-NC.

Analyzing Organizational Growth in Repeated Cross-Sectional Designs
Using Multilevel Structural Equation Modeling

Jan Hochweber

University of Teacher Education St. Gallen

German Institute for International Educational Research

Johannes Hartig

German Institute for International Educational Research

Summary

In repeated cross-sections of organizations, different individuals are sampled from the same set of organizations at each time point of measurement. As a result, common longitudinal data analysis methods (e.g., latent growth curve models) cannot be applied in the usual way. In this contribution, a multilevel structural equation modeling (MSEM) approach to analyze data from repeated cross-sections is presented. Results from a simulation study are reported which aimed at obtaining guidelines on appropriate sample sizes. We focused on a situation where linear growth occurs at the organizational level, and organizational growth is predicted by a single organizational level variable. The power to identify an effect of this organizational level variable was moderately to strongly positively related to number of measurement occasions, number of groups, group size, intraclass correlation, effect size, and growth curve reliability. The Type I error rate was close to the nominal alpha level under all conditions.

Keywords: Cluster randomization, multilevel modeling, repeated cross-section, statistical power, structural equation modeling, sample size

Analyzing Organizational Growth in Repeated Cross-Sectional Designs Using Multilevel Structural Equation Modeling

Longitudinal assessments of organizations are indispensable for research on organizational trends, stability and change of organizational constructs, and preconditions of successful organizational development. In school effectiveness research, for example, schools' effects on students' attainment were found to be relatively stable, and changes in school effectiveness to be related to, among others, changes in schools' entry policies, student composition and quality of teaching practice (e.g., Creemers & Kyriakides, 2010; Thomas, Sammons, Mortimore, & Smees, 1997). Evidently, such findings can be highly relevant when planning measures of organizational change.

Even though longitudinal studies of organizations can be done exclusively with variables measured directly at the organizational level (Level 2), many use variables originally measured at the individual level (Level 1). Samples of individuals (e.g., students in schools, patients in hospitals) are assessed at several time points, and their data is used to capture change of organizations. Such studies may be based on repeatedly measuring the same individuals at each time point. Alternatively, within the same sample of organizations, different individuals may participate at each time point. These studies, which we will refer to as "organizational longitudinal studies", have been recognized as a special case of repeated cross-sectional studies, when data are collected repeatedly in a multi-stage sampling design. Feldman and McKinlay (1994), for example, distinguished „cross-sectional designs“ from „cohort designs“, the difference being that „samples are selected independently within each cluster at each time point“ (ibid., p. 62) rather than measured longitudinally.

The resulting datasets have a structure with huge amounts of missing data (see Figure 1 for an illustration). Data for each individual is observed at only one time point, meaning that empirical information on covariances across time points is completely missing at the

individual level. This rules out the application of standard techniques for longitudinal data analysis (e.g., latent growth curve models). Even “state-of-the-art” methods for dealing with missing data—in particular, full information maximum likelihood (FIML) estimation or missing data imputation methods—do not provide a remedy in this regard, since they can be applied only if partial information on the covariances between time points is available (e.g., Duncan, Duncan, & Li, 1998). Yet, appropriate methods to analyze data from organizational longitudinal studies have been developed in different research traditions.

Econometric analyses typically evaluate the impact of individual- or organization-level variables (e.g., the availability of computers at schools; Sprietsma, 2012) on repeatedly measured outcomes (e.g., test scores), applying the pseudo-panel technique introduced by Deaton (1985) and further developed by Verbeek and Nijman (1993), and others. Its basic idea is to group individuals into „pseudo-cohorts“ based on time-invariant observable characteristics, and to aggregate the data from each time point to the cohort level. Estimation is based on fixed effects by the inclusion of cohort dummies into the model. Since the observed cohort means are error-prone measures of the true cohort means, a correction is applied to the observed cohort covariance matrix. Although well-established, this approach is different, in particular, from methods common in the social sciences, where differences between organizations (e.g., schools) are usually captured by random effects.

In health-related research, organizational longitudinal data typically arise in cluster-randomized trials, where organizations (e.g., hospitals) are randomly assigned to treatment and control conditions. Pretest-posttest designs are arguably most common, although more complex designs are increasingly considered (see Hooper & Bourke, 2015, for an overview). To estimate the effect of the treatment(s), several analysis techniques have been proposed, based, among others, on generalized estimating equations (GEE) and meta-analysis methods (see Donner & Klar, 2000; Ukoumunne & Thompson, 2001). Another approach, multilevel

regression, has also been considered in the social sciences, the models being closely related to models developed in school effectiveness research (see below).

In educational and psychological research, analyses of organizational longitudinal data have been rare to date, but the necessary datasets are increasingly available. An example using data from the Programme for International Student Assessment (PISA) will be presented below. In the more common analysis of individual longitudinal (cohort) data, two analysis approaches based on latent variables have come to widespread use, one of which is rooted in multilevel regression modeling (MRM) and another in structural equation modeling (SEM). Specific multilevel regression models to analyze organizational longitudinal data have been developed both in health-related research (Donner & Klar, 2000; Ukoumunne & Thompson, 2001) and school effectiveness research (Gray, Jesson, Goldstein, Hedger, & Rasbash, 1995; Willms & Raudenbush, 1989). In these models, the measurements at different time points (Level 1) are treated as nested in organizations (Level 2), or alternatively, the measurements (Level 1) are treated as nested in organizations at different time points (Level 2) which are nested in organizations (Level 3).

SEM offers a great variety of modeling techniques to analyze change (e.g., Little, 2013; McArdle & Nesselrode, 2014). Multilevel structural equation modeling techniques (MSEM; e.g., Kaplan & Elliott, 1997; Mehta & Neale, 2005) also allow to study change simultaneously at the individual and organizational level. However, common structural equation models to study change require individual longitudinal data. In contrast to MRM, the application of SEM to data from organizational longitudinal studies has, to our knowledge, not been discussed yet. Given the flexible and powerful modeling options SEM offers, it seems desirable to develop and explore structural equation models that are suitable to describe and explain change based on data from organizational longitudinal studies.

The aim of this paper is threefold. First, we present a multilevel structural equation model suitable for analyzing data from organizational longitudinal studies. Second, we will illustrate the presented model by means of an empirical application, using data from the PISA study. Finally, we will report results from a Monte Carlo simulation study aimed at generating guidelines on appropriate sample sizes for applications of this approach.

A Multilevel Structural Equation Model for Organizational Longitudinal Studies

In presenting the structural equation model for organizational longitudinal studies, we draw on one widely popular SEM approach to analyze individual longitudinal data, latent growth curve modeling (LGM; Bollen & Curran, 2006; Meredith & Tisak, 1990; Preacher, Wichman, MacCallum, & Briggs, 2008). In LGM, growth factors are specified to capture variation in the initial status and change of persons. “Time” is entered into the model by specifying the time-point-specific measurements as indicator variables of the growth factors. It is common to fix the initial status factor loadings at 1, and to fix the change factor loadings at values representing the time passed since the initial measurement. Initial status and change can be predicted by specifying directed paths from the predictor variables to the growth factors.

In the model to be discussed, a latent growth curve model will be used to capture organizations’ initial status and change. A standard linear growth model will be applied, since this type of model is well-known and has been frequently applied to non-hierarchical data (e.g., Shevlin & Millar, 2006; Simons-Morton, Chen, Abrams, & Haynie, 2004). It is arguably most often used in practice, though other and more complex growth models may be specified as well. Organizations’ initial status and change are assumed to be related to a single predictor variable which is directly measured at the organizational level. It can be thought of, for example, as coding different treatments implemented at the level of organizations.

In the following, we consider measurements of the same outcome variable Y at three time points, under the condition that Y_1 through Y_3 were measured in different samples of individuals nested in the same sample of organizations. For ease of illustration only three measurements are considered, but the extension to more time points is straightforward.

The resulting multilevel structural equation model is shown in Figure 2. While the growth model specified at the organizational level is of the common type described above, the individual level model is adapted to reflect the data structure in organizational longitudinal studies. For our example, we obtain the following Level 1 equations:

$$Y_{1ig} = \alpha_{1g} + r_{1ig} \quad (1)$$

$$Y_{2ig} = \alpha_{2g} + r_{2ig} \quad (2)$$

$$Y_{3ig} = \alpha_{3g} + r_{3ig} \quad (3)$$

Each equation represents the time-point specific intercept of organizations g ($\alpha_1, \alpha_2, \alpha_3$) and the deviation of individuals i from the respective intercept (r_1, r_2, r_3). Hence, no model is specified to explain growth at the individual level. In contrast, if MSEM is used to analyze individual longitudinal data, the variances and covariances are modeled at both the individual and organizational level. Obviously, this is not appropriate for organizational longitudinal studies, in which different individuals are sampled at each time point. At the individual level, the variance in Y can be estimated at each time point, but it is neither possible nor necessary to estimate covariances between time points: Since different individuals participate at each time point, no systematic relationships between their outcome have to be expected at the individual level. Relationships over time may be merely introduced at the organizational level, due to the individuals being members of organizations which are assessed repeatedly. Consequently, the Level 1 covariances between time points can be fixed to zero:

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_{11}^2 & & \\ 0 & \sigma_{22}^2 & \\ 0 & 0 & \sigma_{33}^2 \end{bmatrix} \quad (4)$$

For the organizational growth model, we obtain the Level 2 equations:

$$\alpha_{1g} = \eta_{0g} + \varepsilon_{1g} \quad (5)$$

$$\alpha_{2g} = \eta_{0g} + \eta_{1g} + \varepsilon_{2g} \quad (6)$$

$$\alpha_{3g} = \eta_{0g} + 2\eta_{1g} + \varepsilon_{3g} \quad (7)$$

α_{1g} , α_{2g} , and α_{3g} are the (latent) intercepts or means of the Level 1 outcome variables Y_1 , Y_2 , and Y_3 (cf. Kaplan & Elliott, 1997; Muthén & Asparouhov, 2011). η_{0g} and η_{1g} are latent variables capturing the organizations' initial status and change, respectively. The residuals ε_1 through ε_3 represent intercept variance not accounted for by the growth model. They are assumed to be multivariate normally distributed with covariance matrix:

$$\mathbf{\Theta} = \begin{bmatrix} \theta_{11} & & \\ 0 & \theta_{22} & \\ 0 & 0 & \theta_{33} \end{bmatrix} \quad (8)$$

Regressing the organizations' initial status and change on predictor variable Z leads to:

$$\eta_{0g} = \beta_{00} + \beta_{01}Z_g + \zeta_{0g} \quad (9)$$

$$\eta_{1g} = \beta_{10} + \beta_{11}Z_g + \zeta_{1g} \quad (10)$$

The residuals ζ_{0g} and ζ_{1g} are assumed to be multivariate normally distributed with covariance matrix:

$$\mathbf{\Psi} = \begin{bmatrix} \psi_{00} & \\ \psi_{01} & \psi_{11} \end{bmatrix} \quad (11)$$

In the Appendix, we present the *Mplus* syntax for specifying the model shown in Figure 2. It can be easily adapted to different and more complex situations.

An Application to PISA Data

The usefulness of the presented model to analyze data from organizational longitudinal studies will be illustrated using data from the Programme for International Student Assessment (PISA). PISA is a triennial international survey which evaluates education systems by testing the competencies of 15-year-olds. Schools are sampled randomly proportional to size from each participating education system, and students are sampled randomly from these schools. Student competencies in reading, mathematics, and science are assessed, and questionnaires are administered to students, parents, and school principals. The resulting datasets are commonly regarded as cross-sectional in nature, given that data for individual students is available from only one time point. However, each time a new set of schools is sampled for PISA, it may occur by chance that some schools are sampled for the second (or third, fourth, etc.) time. As a consequence, PISA may provide organizational longitudinal data with independent samples of 15-year-old students from the same set of schools.

In Germany, 502 schools that first participated in PISA 2000, and then again in PISA 2003 and/or PISA 2006 were identified, and a dataset was created to allow for longitudinal analyses at the school level. On average, each of these schools provided data from 26.8 students per time point; 306 schools participated in PISA 2000 and 2003, 134 schools participated in PISA 2000 and 2006, and 62 schools participated in all three assessments.

Our analysis proceeded in two steps. In the first step, the model illustrated in Figure 2, but without school-level predictor Z , was separately fit to students' math and science scores (i.e., Plausible Values; e.g., OECD, 2009) from PISA 2000, 2003, and 2006. We did not use the reading scores, since preliminary analyses had clearly indicated that growth did not follow a linear function. According to model fit indices, an acceptable fit of the linear growth model was obtained both for the math test scores ($\chi^2 = 83.758$, $df = 8$, $CFI = 0.943$, $RMSEA = 0.020$,

SRMR Level 1 = 0.000, SRMR Level 2 = 0.004) and the science test scores ($\chi^2 = 123.308$, $df = 8$, CFI = 0.911, RMSEA = 0.025, SRMR Level 1 = 0.000, SRMR Level 2 = 0.006).

In the second step, school-level variables were entered into the models to predict schools' initial status (path β_{01} ; cf. Figure 2) and, of primary interest, change (path β_{11}) in math and science achievement. We decided to use a set of variables from the German version of the school principal questionnaire administered in PISA 2000 which focused on four types of problems potentially encountered at school level: (1) lack of resources for teaching and learning (8-item scale, e.g., lack of teaching materials; Cronbach's $\alpha = .88$); (2) lack of teaching personnel (item with a 4-point Likert scale); (3) lack of teacher engagement (4-item scale, e.g., resistance to change; Cronbach's $\alpha = .67$); (4) lack of student discipline (6-item scale, e.g., disruptions in class; Cronbach's $\alpha = .83$). We conceived these variables to be relatively stable at the school level over several years, and to be likely to exert a negative long-term effect on a school's achievement development (e.g., Greenwald, Hedges, & Laine, 1996; Ma & Willms, 2004). For math and science achievement, respectively, each variable was entered separately into a model as predictor of schools' initial status and change (corresponding to predictor Z in Figure 2). School type (high, intermediate, low track, represented by two dummy variables), average socioeconomic status (HISEI; Ganzeboom, de Graaf, Treiman, & de Leeuw, 1992), and the proportion of students with migration background (i.e., at least one parent born abroad) were entered additionally as predictors at school level, serving as control variables.

Results are presented in Table 1. Students' discipline was related to schools' initial achievement level in both domains. A lack of resources for teaching and learning was negatively associated with schools' change in math achievement, and a lack of teacher personnel negatively predicted schools' change in science achievement. Hence, we found that

these predictors contributed to explain schools' growth in math and science achievement over a period of six years, based on organizational longitudinal data from the PISA study.

Appropriate Sample Sizes: A Monte Carlo Simulation Study

We conducted a Monte Carlo simulation to provide guidelines regarding appropriate sample sizes for organizational longitudinal studies. A number of simulation studies exists for multilevel regression models (see McNeish & Stapleton, 2014, for a review) as well as for latent growth curve models (e.g. Fan, 2003; Fan & Fan, 2005). However, simulation results are only generalizable to models and data structures similar to the simulated conditions, and to our knowledge there are no studies on the analysis of data from organizational longitudinal studies using growth curve models. Specifically, the simulation aimed at answering two research questions, both concerning the effect of organizational variable Z on organizational change η_1 (path β_{11} , cf. Figure 2): 1) Given Z 's true effect is zero, under which conditions is the proportion of Type I errors reasonably close to the nominal alpha level? 2) Given Z 's true effect is different from zero, under which conditions is there sufficient power to obtain statistical significance? We decided to focus on the linear growth model presented above, since models of this—or a closely related—type are argueably most often applied in a great diversity of settings.

The simulation design factors included Level 1 sample size, Level 2 sample size, intraclass correlation, growth curve reliability, number of measurement occasions, and effect size. Level 1 sample size (number of individuals per organization), Level 2 sample size (number of organizations), and intraclass correlation (ICC)—the proportion of variance of an outcome variable Y located at Level 2—are typical design factors in simulation studies on multilevel regression (e.g., Lüdtke et al., 2008; Maas & Hox, 2005). Each of these factors was considered relevant, in particular to the power analysis. First, power for detecting an effect on organizational growth should depend on Level 2 sample size. Power for detecting differences

in growth in single-level LGM increases with sample size (Fan, 2003; Muthén & Curran, 1997). The units in our growth model are groups instead of individuals, but still power should increase with sample size, that is, the number of organizations. Second, power should be influenced by Level 1 sample size and ICC. The indicator variables in our growth model are the time-point-specific group means, and the reliability of (observed) group means is captured by the ICC(2), which is determined by the number of individuals per organization (n) and the proportion of variance located at the organizational level (ICC [1]) (e.g., Bliese, 2000):

$$ICC(2) = \frac{n \cdot ICC(1)}{1 + (n - 1) \cdot ICC(1)} \quad (12)$$

Power for detecting an effect on organizational growth might be compromised if reliability of the group means is too low. Even if both the number of observations and the ICC are clearly different from zero (say, $n = 10$, $ICC = .10$), group-mean reliability can be unsatisfactory judged by common psychometric standards.

In single-level LGM, growth curve reliability—the proportion of variance explained in the indicator variables by the growth factors—is positively related to the power to detect individual differences in slopes (Hertzog, Oertzen, Ghisletta, & Lindenberger, 2008) and covariances between slopes (Hertzog, Lindenberger, Ghisletta, & Oertzen, 2006). This suggests that power in our model also increases with growth curve reliability, that is, with the proportion of variance in the group means explained by the organizational growth factors. Furthermore, the number of measurement occasions is positively related to the power to detect individual differences in slopes (Hertzog et al., 2008) and covariances between slopes (Hertzog et al., 2006). Even more important, it has a positive impact on LGM's power to detect mean growth rate differences (Muthén & Curran, 1997), suggesting that power to detect differences in organizational growth also increases with the number of measurements. Finally, power for detecting an effect is obviously strongly related to effect size (see Muthén

& Curran, 1997, for an example using single-level LGM). Since effect sizes of a wide range occur in organizational research, we included effect size as another design factor.

In contrast, we expected the design factors' relationship to Type I error rate to be at most moderate, and Type I error rates to be overall acceptable. In a recent overview of simulation studies on multilevel regression, McNeish and Stapleton (2014) considered 30 groups to be sufficient for obtaining accurate estimates of fixed effects standard errors for ICC values between .10 and .30 and group sizes between 5 and 30. Hox, Maas, and Brinkhuis (2010) investigated standard error bias in multilevel confirmatory factor analysis and found the coverage of the 95% confidence interval to be acceptable for Level 2 loadings (though not variances) when the sample size was small (50 groups, group size 5) and maximum likelihood estimation was used. Hence, obtaining an appropriate Type I error rate for the fixed effect of organizational variable Z should be possible under most or all studied conditions.¹

Method

Design

In this study, a 4 (number of groups; NG) \times 3 (group size; GS) \times 3 (intraclass correlation; ICC) \times 3 (growth curve reliability; GR) \times 2 (number of measurement occasions; NM) \times 3 (effect size; ES) design was used. The number of groups (organizations) was specified as 30, 50, 100, and 150, consistent with previous simulation studies on multilevel modeling (LaHuis & Ferguson, 2009; Maas & Hox, 2005). In applied research, the number of groups varies widely but small to moderately sized group samples are very common (see LaHuis & Ferguson, 2009, for several examples). We confined our analysis to balanced designs, with group size specified as 5, 10, and 25. Thus, certain emphasis was placed on small groups (such as work groups and classrooms) since problems with insufficient power should be more pronounced with small group sizes. Intraclass correlation was specified as .10,

.20, and .30. ICCs reported in educational and organizational research are typically not larger than .30, and often substantially smaller (Bliese, 2000; Hedges & Hedberg, 2007).²

Two conditions for the number of measurement occasions were implemented, three and five repeated measurements. Three measurements is the minimum in LGM, and is arguably the most observed in practice. Considering previous simulation studies (e.g., Fan & Fan, 2005; Muthén & Curran, 1997), five measurements is an intermediate number that can be reasonably expected in diverse fields of applied research. Effect size in terms of Cohen's d was specified as 0, 0.5, and 0.8, the latter two representing moderate and large effects, respectively, as suggested by Cohen (1988). We decided to not examine small effects since we primarily conceived predictor Z to be a treatment variable. Usually, implementing a treatment at the organizational level is costly and would be considered only if substantial effects were to be expected. Finally, growth curve reliability was specified as 0.5, 0.7, and 0.9, comparable to previous simulation studies on latent growth modeling (Hertzog et al., 2006, 2008).

Data Generation

Data was generated for the model depicted in Figure 2. For each combination of the design factors, 1,000 samples were simulated for the $d = 0.5$ and $d = 0.8$ conditions, and 3,000 samples were simulated for the $d = 0$ condition using R 3.01 (R Core Team, 2013). We decided to analyze a larger number of replications for the $d = 0$ condition to obtain more precise estimates of the actual alpha levels, given that relatively small random deviations from the expected proportion of Type I errors might (erroneously) suggest that the estimated standard errors are biased. The *mvtnorm* package (Genz et al., 2014) was used to generate the required draws from multivariate normal distributions.

In all models, the intercepts of the initial status factor, β_{00} , and the change factor, β_{10} , were specified as 0. That is, we assumed no change on average across groups if organizational

variable Z was zero (e.g., if no treatment was given). The total (unconditional) variance of the initial status factor was specified as 1 ($SD = 1$), and the total variance of the change factor was specified as 0.25 ($SD = 0.5$), leading to a 4:1 ratio of total intercept over change variance.

Hertzog et al. (2006), drawing on longitudinal studies of adult cognitive development, found that “variance in change is small to moderate relative to variance in initial level” (Hertzog et al., 2006, p. 245) and arrived at total intercept over change variance ratios of 2:1 and 4:1, respectively. Since a relatively smaller change variance seems more realistic in many research settings, we decided for a slightly more conservative 4:1 ratio.

The binary organizational variable Z was drawn from a Bernoulli distribution with probability 0.5. Z was conceived to have no effect on organizations’ initial status (i.e., $\beta_{01} = 0$), which seems most realistic in a group-randomized trial where the relationship between treatment status and outcome is minimized due to randomization. Z ’s effect on organizational growth was specified to arrive at $\beta_{11} / \sqrt{\psi_{11}} = 0$, $\beta_{11} / \sqrt{\psi_{11}} = 0.5$, and $\beta_{11} / \sqrt{\psi_{11}} = 0.8$, respectively, in line with Cohen’s effect size classification.³ The residual correlation between the growth factors was set to a small positive value ($r = .1$), such that organizations with a higher initial status tended to show somewhat larger growth.

The loadings of the growth factors were specified as shown in Figure 2, which is the most common specification in LGM. The residual variance of α_1 , θ_{11} , was specified to match the selected GR. Since the residual variances were assumed to be homogeneous, the residual variances of α_2 and α_3 , θ_{22} and θ_{33} , were set equal to θ_{11} . In line with previous research (Hertzog et al., 2006, 2008), the selected GR condition thus referred to the first measurement occasion, and GR was allowed to vary as a function of time.

Finally, the Level 1 variance of Y_1 , σ_{11}^2 , was specified to match the selected ICC. The Level 1 variances of Y_2 and Y_3 , σ_{22}^2 and σ_{33}^2 , were set equal to σ_{11}^2 . Although this allowed the

ICC to differ between time points, this homogeneity assumption seemed appropriate from a substantive perspective, since we focused on organizational change and did not assume any processes that might influence Level 1 variability.

Analysis

The simulated samples were analyzed one by one in *Mplus* 6.11 (Muthén & Muthén, 1998-2010), using robust full information maximum likelihood estimation (“MLR” estimator). SPSS was used to compute summary statistics from the results. The analysis model in *Mplus* was correctly specified, that is, analogous to the data generating model. Specifically, the variances at Level 1 (σ_{11}^2 through σ_{33}^2) and the random intercept residual variances (θ_{11} through θ_{33}) were fixed to be equal, respectively. The other parameters in the model were estimated freely. Type I error rate and statistical power regarding the effect of organizational variable Z were estimated using the Wald significance test of $\hat{\beta}_{11}$. They were approximated by the proportion of replications in which the null hypothesis was incorrectly (Type I error rate; $\beta_{11} = 0$) or correctly (Power; $\beta_{11} > 0$) rejected.

Results

Model Convergence

Using standard specifications in *Mplus* regarding maximum number of iterations and convergence criteria, 74 (< 0.1%) of a total of 1,080,000 samples failed to converge. 143,163 (13.3%) samples converged but *Mplus* issued a warning indicating inadmissible solutions (e.g., Heywood cases). We calculated the percentage of replications with either failed convergence or inadmissible solution for each of the 648 simulated conditions (i.e., each combination of the design factors). Then, we calculated the Pearson correlation between these percentages and each design factor across the simulated conditions. Higher percentages occurred when the number of measurement occasions was smaller ($r = \cdot .52$), number of

groups was smaller ($r = \cdot .36$), growth curve reliability was smaller ($r = \cdot .33$), group size was smaller ($r = \cdot .31$), and intraclass correlation was smaller ($r = \cdot .29$), but they were not noticeably related to effect size ($r = .03$).

Some conditions showed high percentages of failed convergence or inadmissible solutions, reaching a maximum of 63.3%. Similar problems are not uncommon in applications of MSEM, and have been previously found to be related to a small number of groups and a low ICC (Hox & Maas, 2001; Li & Beretvas, 2013). Nevertheless, the large majority of samples provided admissible solutions. The average percentage of replications with failed convergence or inadmissible solutions was 13.5% across the 648 conditions. In 56.8% of the conditions, less than 10% of the replications showed failed convergence or inadmissible solutions, while in 18.4% of the replications this percentage was 30% or higher. For the subsequent analyses, we discarded all samples which did not converge or produced inadmissible solutions, presupposing that in practice a data analyst should not proceed with interpreting the results of these models.

Proportion of Type I Errors

Next, we considered the proportion of Type I errors when the effect of predictor Z on organizational growth was zero. The minimum percentage of significant results ($p < .05$) across the simulated conditions was 3.90%, the maximum percentage was 6.91%, that is, the percentage of Type I errors differed by no more than 2% from the nominal level.

A strict criterion for judging these results is based on the standard error for the percentage of Type I errors. For example, given 3000 replications with admissible solutions for a condition with $d = 0$, the standard error for the occurrence of an event with $p = 0.05$ is $\sqrt{p(1-p)/n} = \sqrt{.05(.95)/3000} = 0.00398$, leading to an expected 95%-interval between 4.22% and 5.78% for the number of Type I errors. In determining the standard error and confidence interval, we used the actual number of replications with admissible solutions for

each simulated condition. Of all 216 conditions with $d = 0$, the 95% confidence interval did not include the expected value of 5% in 62 (28.7%) of the conditions.

On the other hand, considering the actual size of these deviations from the nominal alpha level ($< 2\%$), comparable differences have been regarded as negligible in previous simulation studies (Fan & Fan, 2005; Maas & Hox, 2005). According to a more formal criterion suggested by Bradley (1978), values within one-half of the nominal Type I error rate are acceptable for 95% non-coverage rates, that is, proportions of significant results between 2.5% and 7.5%. Hence, Bradley's criterion was fulfilled under all conditions studied.

Power Analysis

Finally, we considered the power to detect a non-zero effect of predictor Z on organizational growth. Pearson correlations calculated across the simulated conditions between the percentage of statistically significant results and each design factor indicated an increase in power with number of groups ($r = .63$), effect size ($r = .46$), number of measurement occasions ($r = .43$), growth curve reliability ($r = .24$), group size ($r = .19$), and intraclass correlation ($r = .17$). Detailed results are presented in Table 2.

Effect size and, in particular, number of groups were the design factors most closely related to power. A frequently cited recommendation by Cohen (1988) is that power should exceed .80. If this rule is applied, 30 or 50 groups were under none of the conditions sufficient to obtain adequate power. Even moderate power of .50 was generally reached only with five measurements and a large effect size ($ES = .8$), and, given 30 groups, only under otherwise favorable conditions concerning group size, intraclass correlation, and growth curve reliability. In contrast, power $\geq .80$ was typically obtained if the number of groups was at least 100, the effect size was large and five measurements were available.

Similarly, it was found difficult to obtain sufficient power with only three measurements. Power of .80 or above required a large effect size, a growth curve reliability of

at least .7 (150 groups) or even .9 (100 groups) as well as sufficient intraclass correlation and group size. Having three measurements and only 50 groups, even moderate power of .50 or above was difficult to achieve and required a large effect size, large growth curve reliability and moderate to large intraclass correlation (ICC \geq .20) and group size (GS \geq 10).

Furthermore, power was generally low when growth curve reliability was low (GR = .5).

However, reaching a power of .80 was still possible under most conditions if five measurements and at least 100 groups were available and the effect size was large.

Under appropriate conditions, a power of .80 was reached with all three intraclass correlations studied. On average, the difference in power related to intraclass correlation was rather modest, .11 between ICC = .10 and ICC = .30 across all conditions. Generally, an increase from .1 to .2 had a noticeably larger impact than an increase from .2 to .3. Strong effects of intraclass correlation, however, were found only under specific conditions.

Finally, given otherwise favorable conditions, power \geq .80 was obtained with all three group sizes examined. Still, an average power increase of .06 was associated with an increase in group size from 5 to 10, and from 10 to 25, respectively. Comparing group sizes 5 and 25, the increase in power was smallest when intraclass correlation was large (ICC = .30) and largest when intraclass correlation was small (ICC = .10). Under some conditions, the difference in power reached around .4 or above, and sampling 25 instead of 5 individuals per group would raise the power from low (e.g., .40) to acceptable (e.g., .79). For specific conditions, an increase from 5 to 10 already implied a noticeable increase in power.

Discussion

We presented a multilevel structural equation model to analyze data from “organizational longitudinal studies”, that is, repeated cross-sectional studies where different individuals are sampled at each time point from the same set of organizations. Although approaches to analyze this data type have been developed in various research disciplines,

structural equation models have some clear advantages which make them worthwhile to consider when choosing from available models. First, while the linear growth model we explored is very common in analyses of change, it can be easily replaced by more complex growth models (e.g., polynomial growth curves models, unspecified trajectory models, multiple-group models; Bollen & Curran, 2006; Preacher et al., 2008) or entirely different models to capture change (e.g., cross-lagged panel models, latent difference models; Little, 2013; McArdle, 2009). All aspects of the model can be flexibly extended, including, for instance, models to capture mediation and moderation over time (e.g., Little, 2013; Preacher, 2015). In practice, the model presented in this article (including the *Mplus* syntax in the Appendix) may serve as a starting point when specifying these alternative models. Second, SEM allows to represent the repeatedly measured variables as well as further predictor or outcome variables as latent variables, allowing to explicitly model unreliability of the observed variables. It also facilitates the evaluation of measurement invariance over time. Third, SEM allows to judge the goodness of fit of a specified longitudinal model based on various fit indices, including new methods to determine level-specific model fit (Ryu, 2014). In several regards, SEM thus appears more flexible than another well-known latent variable approach to analyze change, that is, multilevel regression.

While organizational longitudinal studies appear less common in some research areas, in particular, educational and psychological research, it might be beneficial to consider their potential in these disciplines more often. Specifically, repeatedly sampling individuals from organizations seems appropriate if the research interest lies on organizational change, and individual change appears more as a nuisance than interesting in itself. For example, when evaluating the impact of school policies on students' achievement trajectories, it may make sense to repeatedly sample students from the same age group, in order to not confound organizational change and individual development. Also, as has been discussed in the health-

related and econometric literature (e.g., Feldman & McKinlay, 1994; Ukoumunne & Thompson, 2001; Verbeek, 2008), organizational longitudinal studies may have certain advantages over individual longitudinal studies, among others, generally lower rates of attrition and nonresponse, a higher representativeness of the sample if the population changes substantially, and a higher robustness against measurement effects on participants' behavior ("Hawthorne effect").

Implications for Study Design

Since structural equation models for organizational longitudinal studies have not been widely discussed, we decided to focus on a relatively simple situation where linear growth occurs at the organizational level, and growth is predicted by a single variable measured at the organizational level. As indicated by our empirical example, this model can be usefully applied to describe and explain actual organizational change. In a simulation study, we investigated two questions concerning the prediction of linear growth at the organizational level: If the true effect on linear growth equals zero, under which conditions is the proportion of Type I errors reasonably close to the nominal alpha level? If the true effect is different from zero, under which conditions is there sufficient power to obtain statistical significance?

Results of the simulation showed that the proportion of Type I errors was significantly different from the nominal alpha level for roughly 30% of the conditions, but even in these cases, the deviations may be considered as comparatively small in size. Thus, if no statistical model assumptions are violated, the accuracy of the significance test appears not as a major concern when predicting linear growth in organizational longitudinal studies.

Power depends moderately to strongly on number of groups, effect size, number of measurement occasions and growth curve reliability, and still to some extent on group size and intraclass correlation. To researchers planning an organizational longitudinal study, the following outcomes might be most important. First, no condition with 30 or 50 groups

provided power $e .80$. Moderate, though less than optimal power around $.50$ was reached with 50 groups but only if the effect size was large and five measurements were available. Having 100 groups appeared as much more favorable, though there were still conditions under which having 150 groups provided a decisive advantage. Group size was found to play a much smaller role, and even a group size of 5 did not preclude the possibility of obtaining power $e .80$. Nevertheless, increasing group size, especially from small ($GS = 5$) to moderate ($GS = 10$), may make sense. One situation when researchers might consider increasing group size is when the intraclass correlation is expected to be small.

Furthermore, researchers are typically well advised to collect data at more than the minimum three measurement occasions. Under several conditions five instead of three measurements more than doubled the available power. With 30 or 50 groups, obtaining at least moderate power around $.50$ was difficult anyway, but rarely possible with only three measurements. Finally, although not under control of the researcher, it makes particular sense to consider the expected effect size when making design choices. Most importantly, power $e .80$ given a moderately sized effect required 150 groups *plus* five measurements *plus* otherwise favorable conditions. If expecting a moderate effect size, researchers might particularly consider sampling more than 150 organizations.

If researchers are interested in the statistical properties of conditions that were not included in this study (e.g., small effect sizes), they are encouraged to conduct a simulation study themselves using the specific parameter values they consider as suitable (see Muthén & Muthén, 2002, for an example in *Mplus*). In some situations, however, it may be unclear to the researcher which value of a design parameter seems most realistic. Then, a range of values for the parameter might be selected, and the expectable power might be compared. If the resulting power for a “worst-case scenario” (e.g., a very low ICC) seems inacceptably low,

our simulation results (or researchers' own) might be used to infer more appropriate values for parameters that are, in principle, under researchers' control (e.g., the number of groups).

Limitations

As has been pointed out, the present article did not aim at fully exploring the potential of applying SEM to data from organizational longitudinal studies. We focused on a model of linear organizational growth, which was predicted by a single variable measured at the organizational level. This model may be modified or extended, but the results of our simulation study should not be overgeneralized to judge other models' statistical performance. Furthermore, our simulation was limited to a selected range of values concerning the studied design factors. Most importantly, we restricted our analysis for conceptual reasons to moderate and large effects. However, for many conditions the power was hardly satisfying even with moderate effect sizes. Reliably identifying small effects on organizational growth should be even more demanding; in particular, researchers might have to collect data from much more than 150 organizations, and possibly at more than five measurement occasions.

Another limitation of this study was its focus on balanced group sizes. Few studies on multilevel modeling directly address unbalanced designs, and even less examine the impact of unequal versus equal group sizes (McNeish & Stapleton, 2014). However, a theoretical analysis by Konstantopoulos (2010) on two-level unbalanced designs gives some indication that power estimates obtained assuming equal group sizes are reasonably close to power estimates obtained for unequal group sizes in cases of mild or moderate imbalance. Finally, we did not consider effects of violations of statistical assumptions. Power in structural equations models can be considerably affected by statistical assumption violations (Kaplan, 1995). To date, it remains unclear how far the results of this study can be generalized to situations where statistical assumptions are not fully met.

References

- Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In S. W. J. Kozlowski & K. J. Klein (Eds.), *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (pp. 349-381). San Francisco, CA: Jossey-Bass.
- Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective*. Hoboken, NJ: John Wiley & Sons.
- Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*, 144-152. doi:10.1111/j.2044-8317.1978.tb00581.x
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Creemers, B. P. M., & Kyriakides, L. (2010). Explaining stability and changes in school effectiveness by looking at changes in the functioning of school factors. *School Effectiveness and School Improvement*, *21*, 409-427.
doi:10.1080/09243453.2010.512795
- Deaton, A. (1985). Panel data from time series of cross-sections. *Journal of Econometrics*, *30*, 109-126. doi:10.1016/0304-4076(85)90134-4
- Donner, A. & Klar, N. (2000). *Design and analysis of cluster randomization trials in health research*. London, UK: Arnold.
- Duncan, T. E., Duncan, S. C., & Li, F. (1998). A comparison of model- and multiple imputation-based approaches to longitudinal analyses with partial missingness. *Structural Equation Modeling*, *5*, 1-21. doi:10.1080/10705519809540086
- Fan, X. (2003). Power of latent growth modeling for detecting group differences in linear growth trajectory parameters. *Structural Equation Modeling*, *10*, 380-400.
doi:10.1207/S15328007SEM1003_3

- Fan, X., & Fan, X. (2005). Power of latent growth modeling for detecting linear growth: Number of measurements and comparison with other analytic approaches. *The Journal of Experimental Education, 73*, 121-139. doi:10.3200/JEXE.73.2.121-139
- Feldman, H. A., & McKinlay, S. M. (1994). Cohort versus cross-sectional design in large field trials: Precision, sample size, and a unifying model. *Statistics in Medicine, 13*, 61-78. doi:10.1002/sim.4780130108
- Ganzeboom, H. B. G., de Graaf, P. M., Treiman, D. J., & de Leeuw, J. (1992). A standard international socio-economic index of occupational status. *Social Science Research, 21*, 1-56. doi:10.1016/0049-089X(92)90017-B
- Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., & Hothorn, T. (2014). mvtnorm: Multivariate Normal and t Distributions. R package version 1.0-0. Retrieved from <http://CRAN.R-project.org/package=mvtnorm>
- Gray, J., Jesson, D., Goldstein, H., Hedger, K., & Rasbash, J. (1995). A multi-level analysis of school improvement: Changes in schools' performance over time. *School Effectiveness and School Improvement, 6*, 97-114. doi:10.1080/0924345950060201
- Greenwald, R., Hedges, L. V., & Laine, R. D. (1996). The effect of school resources on student achievement. *Review of Educational Research, 66*, 361-396. doi:10.3102/00346543066003361.
- Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis, 29*, 60-87. doi:10.3102/0162373707299706
- Hertzog, C., Lindenberger, U., Ghisletta, P., & Oertzen, T. von (2006). On the power of multivariate latent growth curve models to detect correlated change. *Psychological Methods, 11*, 244-252. doi:10.1037/1082-989X.11.3.244

- Hertzog, C., Oertzen, T. von, Ghisletta, P., & Lindenberger, U. (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
- Hooper, R., & Bourke, L. (2015). Cluster randomised trials with repeated cross sections: Alternatives to parallel group designs. *British Medical Journal, 350*, h2925. doi:10.1136/bmj.h2925
- Hox, J. J., & Maas, C. J. M. (2001). The accuracy of multilevel structural equation modeling with pseudobalanced groups and small samples. *Structural Equation Modeling, 8*, 157-174. doi:10.1207/S15328007SEM0802_1
- Hox, J. J., Maas, C. J. M., & Brinkhuis, M. J. S. (2010). The effect of estimation method and sample size in multilevel structural equation modeling. *Statistica Neerlandica, 64*, 157-170. doi:10.1111/j.1467-9574.2009.00445.x
- Kaplan, D. (1995). Statistical power in structural equation modeling. In R. H. Hoyle (Ed.), *Structural equation modeling. Concepts, issues, and applications* (pp. 100-117). Thousand Oaks, CA: Sage Publications.
- Kaplan, D., & Elliott, P. R. (1997). A didactic example of multilevel structural equation modeling applicable to the study of organizations. *Structural Equation Modeling, 4*, 1-24. doi:10.1080/10705519709540056
- Konstantopoulos, S. (2010). Power analysis in two-level unbalanced designs. *The Journal of Experimental Education, 78*, 291-317. doi:10.1080/00220970903292876
- LaHuis, D. M., & Ferguson, M. W. (2009). The accuracy of significance tests for slope variance components in multilevel random coefficient models. *Organizational Research Methods, 12*, 418-435. doi:10.1177/1094428107308984

- Li, X., & Beretvas, S. N. (2013). Sample size limits for estimating upper level mediation models using multilevel SEM. *Structural Equation Modeling, 20*, 241-264.
doi:10.1080/10705511.2013.769391
- Little, T. D. (2013). *Longitudinal structural equation modeling*. New York, NY: The Guilford Press.
- Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. O. (2008). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods, 13*, 203-229.
doi:10.1037/a0012869
- Ma, X., & Willms, J. D. (2004). School disciplinary climate: Characteristics and effects on eighth grade achievement. *Alberta Journal of Educational Research, 50*, 169-188.
- Maas, C., & Hox, J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology, 1*, 85-91. doi:10.1027/1614-2241.1.3.86
- McArdle, J. J. (2009). Latent variable modeling of differences and changes with longitudinal data. *Annual Review of Psychology, 60*, 577-605.
doi:10.1146/annurev.psych.60.110707.163612
- McArdle, J. J., & Nesselroade, J. R. (2014). *Longitudinal data analysis using structural equation models*. Washington, DC: American Psychological Association.
- McNeish, D. M., & Stapleton, L. M. (2014). The effect of small sample size on two-level model estimates: A review and illustration. *Educational Psychology Review*. Advance online publication. doi:10.1007/s10648-014-9287-x
- Mehta, P. D., & Neale, M. C. (2005). People are variables too: Multilevel structural equations modeling. *Psychological Methods, 10*, 259-284. doi:10.1037/1082-989X.10.3.259
- Meredith, W., & Tisak, J. (1990). Latent curve analysis. *Psychometrika, 55*, 107-122.
doi:10.1007/BF02294746

- Muthén, B. O. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement, 28*, 338-354. doi:10.1111/j.1745-3984.1991.tb00363.x
- Muthén, B. O., & Asparouhov, T. (2011). Beyond multilevel regression modeling: Multilevel analysis in a general latent variable framework. In J. J. Hox & J. K. Roberts (Eds.), *Handbook of advanced multilevel analysis* (pp. 15-40). New York, NY: Routledge.
- Muthén, B. O., & Curran, P. J. (1997). General longitudinal modeling of individual differences in experimental designs: A latent variable framework for analysis and power estimation. *Psychological Methods, 2*, 371-402. doi:10.1037/1082-989X.2.4.371
- Muthén, L. K., & Muthén, B. O. (1998-2010). *Mplus* – Statistical analysis with latent variables [Computer software]. Los Angeles, CA: Muthén & Muthén.
- 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
- Organisation for Economic Cooperation and Development (2009). *PISA 2006 technical report*. Paris, France: OECD.
- Preacher, K. J. (2015). Advances in mediation analysis: A survey and synthesis of new developments. *Annual Review of Psychology, 66*, 825-852. doi:0.1146/annurev-psych-010814-015258
- Preacher, K. J., Wichman, A. L., MacCallum, R. C., & Briggs, N. E. (2008). *Latent growth curve modeling*. Thousand Oaks, CA: Sage Publications.
- R Core Team (2013). R: A language and environment for statistical computing [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, URL <http://www.R-project.org/>

- Ryu, E. (2014). Model fit evaluation in multilevel structural equation models. *Frontiers in Psychology*, 5:81. doi:10.3389/fpsyg.2014.00081
- Shevlin, M., & Millar, R. (2006). Career education: An application of latent growth curve modelling to career information-seeking behaviour of school pupils. *British Journal of Educational Psychology*, 76, 141-153. doi:10.1348/000709904x22386
- Simons-Morton, B., Chen, R., Abrams, L., & Haynie, D. L. (2004). Latent growth curve analyses of peer and parent influences on smoking progression among early adolescents. *Health Psychology*, 23, 612-621. doi:10.1037/0278-6133.23.6.612
- Sprietsma, M. (2012). Computers as pedagogical tools in Brazil: A pseudo-panel analysis. *Education Economics*, 20, 19-32. doi:10.1080/09645290903546496
- Thomas, S., Sammons, P., Mortimore, P., & Smees, R. (1997). Stability and consistency in secondary schools' effects on students' GCSE outcomes over three years. *School Effectiveness and School Improvement*, 8, 169-197. doi:10.1080/0924345970080201
- Ukoumunne, O. C., & Thompson, S. G. (2001). Analysis of cluster randomized trials with repeated cross-sectional binary measurements. *Statistics in Medicine*, 20, 417-433. doi:10.1002/1097-0258(20010215)20:3<417::AID-SIM802>3.0.CO;2-G
- Verbeek, M. (2008). Pseudo-panels and repeated cross-sections. In L. Mátyás & P. Sevestre (Eds.), *The econometrics of panel data* (pp. 369-383). Berlin, Germany: Springer.
- Verbeek, M., & Nijman, T. (1993). Minimum MSE estimation of a regression model with fixed effects from a series of cross-sections. *Journal of Econometrics*, 59, 125-136. doi:10.1016/0304-4076(93)90042-4
- Willms, J. D., & Raudenbush, S. W. (1989). A longitudinal hierarchical linear model for estimating school effects and their stability. *Journal of Educational Measurement*, 26, 209-232. doi:10.1111/j.1745-3984.1989.tb00329.x

Footnotes

¹ It should be noted that given a small number of groups, Hox et al. (2010) found the performance to be less satisfying if robust maximum likelihood estimation was used (as in the present study).

² We simulated ICCs for observed scores and did not distinguish between true variation and measurement error which may be confounded in observed scores, resulting in biased estimates of the ICC (Muthén, 1991).

³ The Cohen's d formula is: $(\mu_1 - \mu_2)/\sigma$. In our case, the difference between group means $\mu_1 - \mu_2$ is equal to β_{11} (i.e., the difference in change between $Z = 0$ and $Z = 1$), and σ is equal to $\sqrt{\psi_{11}}$ (i.e., the variance within groups defined by Z). Since σ is assumed to be homogeneous across groups, no estimation of a common (pooled) standard deviation is required.

Table 1

School-Level Predictors' Effects on Schools' Initial Status and Change in PISA Mathematics and Science Test Scores

Predictors	Mathematics			Science		
	<i>B</i>	<i>SE</i>	²	<i>B</i>	<i>SE</i>	²
Initial Status						
Lack of teaching/learning resources	2.849	2.426	0.030	2.759	2.458	0.029
Lack of teaching personnel	2.327	2.223	0.028	3.250	2.220	0.038
Lack of teacher engagement	2.504	4.007	0.015	-3.952	4.004	-0.023
Lack of student discipline	-13.433***	3.783	-0.109	-14.878***	3.979	-0.119
Change						
Lack of teaching/learning resources	-2.962 [†]	1.605	-0.356	-2.506	1.596	-0.368
Lack of teaching personnel ^a	-1.806	1.445	-0.302	-3.008*	1.469	-0.608
Lack of teacher engagement ^a	-1.593	2.724	-0.114	2.722	2.821	0.308
Lack of student discipline ^a	-3.857	2.573	-0.432	-0.772	2.573	-0.127

Note. A separate model was specified for each predictor. In each model, school type, average socioeconomic status (HISEI), and proportion of students with a migration background were entered as control variables (coefficients not reported). Unstandardized coefficients (*B*) refer to the original metric of the PISA tests, scales, and items.

^a In the science models, the slope factor showed a slightly negative residual variance which was fixed to zero.

[†] $p < .10$, * $p < .05$, *** $p < .001$ (two-sided).

Table 2

Power of the Multilevel Structural Equation Model to Detect an Effect on Linear Organizational Growth

	GS = 5			GS = 10			GS = 25		
	ICC = .1	ICC = .2	ICC = .3	ICC = .1	ICC = .2	ICC = .3	ICC = .1	ICC = .2	ICC = .3
NG = 30									
ES = .5									
NM = 3									
GR = .5	.06	.09	.10	.10	.09	.10	.09	.09	.12
GR = .7	.08	.12	.12	.09	.13	.14	.15	.16	.18
GR = .9	.06	.13	.16	.11	.17	.20	.17	.23	.25
NM = 5									
GR = .5	.11	.18	.18	.15	.21	.23	.18	.23	.23
GR = .7	.15	.20	.23	.19	.23	.24	.25	.27	.25
GR = .9	.17	.25	.26	.23	.28	.29	.26	.29	.30
ES = .8									
NM = 3									
GR = .5	.13	.12	.14	.11	.15	.17	.15	.19	.21
GR = .7	.11	.18	.22	.18	.20	.29	.21	.31	.30
GR = .9	.16	.20	.29	.23	.33	.39	.35	.43	.45
NM = 5									
GR = .5	.24	.30	.36	.33	.39	.43	.40	.43	.44
GR = .7	.30	.40	.47	.37	.48	.49	.47	.50	.52
GR = .9	.32	.49	.52	.47	.53	.55	.50	.58	.58
NG = 50									
ES = .5									
NM = 3									
GR = .5	.09	.11	.12	.07	.11	.15	.14	.14	.16
GR = .7	.10	.14	.17	.13	.15	.23	.17	.23	.25
GR = .9	.12	.18	.22	.19	.22	.30	.23	.30	.32
NM = 5									
GR = .5	.17	.24	.27	.25	.27	.29	.28	.31	.30
GR = .7	.22	.30	.35	.30	.35	.35	.33	.40	.39
GR = .9	.22	.31	.38	.33	.36	.42	.39	.39	.43
ES = .8									
NM = 3									
GR = .5	.12	.17	.22	.18	.24	.24	.22	.27	.33
GR = .7	.16	.28	.31	.22	.36	.38	.35	.41	.46
GR = .9	.23	.33	.45	.29	.48	.58	.48	.62	.67
NM = 5									
GR = .5	.34	.48	.55	.50	.56	.59	.59	.61	.64
GR = .7	.44	.57	.65	.56	.66	.70	.66	.70	.74
GR = .9	.50	.67	.70	.66	.72	.76	.72	.77	.79

Table 2 (continued)

	GS = 5			GS = 10			GS = 25		
	ICC = .1	ICC = .2	ICC = .3	ICC = .1	ICC = .2	ICC = .3	ICC = .1	ICC = .2	ICC = .3
NG = 100									
ES = .5									
NM = 3									
GR = .5	.11	.17	.18	.14	.21	.24	.21	.23	.24
GR = .7	.14	.23	.30	.19	.27	.34	.30	.37	.43
GR = .9	.20	.28	.39	.30	.43	.50	.39	.51	.54
NM = 5									
GR = .5	.27	.41	.44	.39	.48	.51	.47	.54	.55
GR = .7	.39	.51	.53	.49	.58	.62	.57	.59	.64
GR = .9	.47	.58	.62	.55	.61	.65	.64	.63	.65
ES = .8									
NM = 3									
GR = .5	.21	.31	.43	.30	.40	.46	.44	.53	.52
GR = .7	.26	.45	.56	.41	.58	.70	.60	.73	.76
GR = .9	.36	.57	.73	.56	.76	.84	.78	.86	.90
NM = 5									
GR = .5	.64	.78	.83	.78	.85	.88	.85	.89	.91
GR = .7	.73	.85	.91	.87	.93	.93	.90	.95	.95
GR = .9	.82	.92	.94	.90	.94	.95	.94	.97	.97
NG = 150									
ES = .5									
NM = 3									
GR = .5	.13	.22	.25	.19	.28	.34	.28	.35	.36
GR = .7	.19	.29	.38	.28	.44	.46	.46	.52	.55
GR = .9	.27	.43	.54	.40	.52	.62	.58	.67	.73
NM = 5									
GR = .5	.44	.58	.60	.55	.64	.68	.65	.71	.70
GR = .7	.54	.66	.72	.64	.71	.75	.74	.78	.79
GR = .9	.65	.71	.74	.71	.78	.80	.79	.83	.85
ES = .8									
NM = 3									
GR = .5	.28	.42	.57	.41	.56	.63	.59	.66	.70
GR = .7	.40	.62	.74	.60	.78	.83	.79	.87	.89
GR = .9	.52	.76	.90	.74	.90	.95	.91	.96	.97
NM = 5									
GR = .5	.77	.92	.96	.89	.95	.97	.97	.98	.97
GR = .7	.90	.96	.97	.96	.99	.99	.98	.99	.99
GR = .9	.93	.97	.99	.98	.99	1.00	.99	1.00	1.00

Note. Darker color indicates higher power, brighter color indicates lower power. NG = number of groups; ES = effect size; NM = number of measurements; GR = growth curve reliability; GS = group size; ICC = intraclass correlation.

Organization ID	Individual ID	Time Point 1	Time Point 2	Time Point 3
1	1	X		
1	⋮	X		
1	5	X		
1	6		X	
1	⋮		X	
1	10		X	
1	11			X
1	⋮			X
1	15			X
2	16	X		
2	⋮	X		
2	20	X		
2	21		X	
2	⋮		X	
2	25		X	
2	26			X
2	⋮			X
2	30			X
⋮	⋮			
⋮	⋮			

Figure 1. Illustration of the data structure in organizational longitudinal studies. „X“ indicates that a measurement is available for the individual at this time point. Since all individuals are assessed at only one time point, no individual longitudinal data is available. However, since at any time point individuals are assessed from each organization (in this example, 5 individuals per organization at each time point), organizational longitudinal data is available.

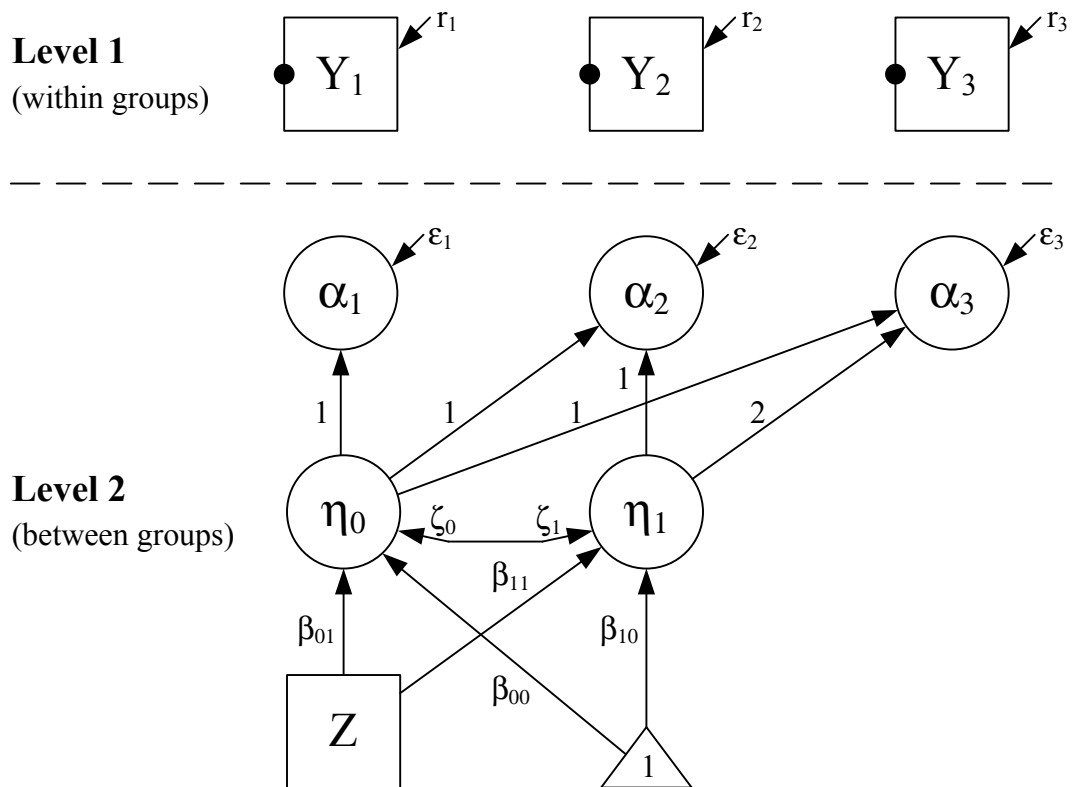


Figure 2. Model of linear organizational growth based on a multilevel structural equation modeling approach.

Appendix

Mplus Syntax for Specifying the Multilevel Structural Equation Model for Linear Organizational Growth

DATA:

```
file = ...\dataset.dat;      !data source
```

VARIABLE:

```
names   = orgid y1 y2 y3 z; !variables in dataset
usevar  = y1 y2 y3 z;      !variables in analysis
cluster = orgid;          !cluster id variable
between = z;              !level 2 variable(s)
missing = ...;           !missing data code(s)
```

ANALYSIS:

```
type = twolevel;          !twolevel model requested
```

MODEL:

```
%within%                  !level 1 model specification
y1 with y2@0;             !covariances fixed to 0
y1 with y3@0;
y2 with y3@0;

y1 y2 y3 (1);            !variances fixed to equality
                          !(may be relaxed)

%between%                 !level 2 model specification
i s | y1@0 y2@1 y3@2;    !linear growth model

i s on z;                !intercept/slope predicted by z

i with s;                !estimate intercept/slope residual covariance

y1 y2 y3 (2);           !residual variances fixed to equality
                          !(may be relaxed)
```