European Journal of
Oral Sciences

Eur J Oral Sci 2009; 117: 343–350
Printed in Singapore. All rights reserved

European Journal of
Oral Sciences

Focus

An introduction to latent growth curve modelling for longitudinal continuous data in dental research

Yu-Kang Tu[1,2], Francesco D'Aiuto[3], Vibeke Baelum[4], Mark S. Gilthorpe[2]

[1]Department of Periodontology, Leeds Dental Institute, Leeds; [2]Biostatistics Unit, Centre for Epidemiology and Biostatistics, Leeds Institute of Genetics, Health and Therapeutics, University of Leeds, Leeds; [3]Periodontology Unit, UCL Eastman Dental Institute, London, UK; [4]School of Dentistry, Faculty of Health Sciences, University of Aarhus, Aarhus, Denmark

Tu Y-K, D'Aiuto F, Baelum V, Gilthorpe MS. An introduction to latent growth curve modelling for longitudinal continuous data in dental research. Eur J Oral Sci 2009; 117: 343–350. © 2009 The Authors. Journal compilation © 2009 Eur J Oral Sci

Many studies in dental research are based on repeated measurements of several continuous variables. Statistical analyses of such data require advanced methods to explore the complexity of information within the data. Currently, the most frequently adopted approach is to undertake multiple univariate tests. Occasionally, more advanced and sophisticated statistical methodologies, such as multilevel modelling and generalized estimating equation, have been used. In the last decade, a novel statistical methodology known as latent growth curve modelling has been developed in the social sciences. Latent growth curve modelling can be considered a special application of structural equation modelling and is generally conducted using structural equation modelling software. Recent development of statistical theory shows that latent growth curve modelling is equivalent to multilevel modelling, and both approaches yield identical results. However, in some study designs latent growth curve modelling can provide a more flexible framework of statistical modelling than multilevel modelling and generalized estimating equation for longitudinal data. The aim of this article was to present a non-technical introduction to latent growth curve modelling for dental researchers. The emphasis was on conceptual understanding, rather than mathematical rigor, so path diagrams were used for visual presentations of various statistical models. When properly applied, latent growth curve modelling has great potential to give new directions for future longitudinal dental research.

Dr Yu-Kang Tu, Biostatistics Unit, Centre for Epidemiology and Biostatistics, University of Leeds, Room 8.01, Level 8, Worsley Building, Leeds LS2 9JT, UK.

Telefax: +44–113–3434877
E-mail: y.k.tu@leeds.ac.uk

Key words: latent growth curve modelling; longitudinal analysis; multilevel modelling; repeated measurements; structural equation modelling

Accepted for publication March 2009

Many studies in dental research require repeated measurements of continuous variables during the period of observation. For instance, in randomized controlled clinical trials on the treatment of periodontal disease, gingival probing pocket depth and clinical attachment level on the same teeth and sites are often measured repeatedly. Similarly, the study of the growth of facial profiles among boys and girls is based on repeated tracings of cephalometric landmarks over a period of several years.

Faced with such longitudinal data, dental researchers often resort to the use of univariate statistical methods of data analysis, such as multiple *t*-tests (or analysis of variance when there are more than two groups), to compare differences in the chosen outcomes between groups at different time-points (1–3). This common approach poses a number of potential problems, such as inflated type-1 error rates as a result of multiple testing. Perhaps more importantly, univariate statistical methods can only test limited research hypotheses. For example, dental researchers might find differences in cephalometric landmarks between boys and girls when using a simple *t*-test to compare changes in the data obtained from the children between 8 and 14 yr of age. However, this approach does not provide answers to more complex and vital research questions, such as whether or not there are different patterns and trajectories of facial growth for girls and boys between 8 and 14 yr of age; whether or not children with smaller values at baseline may undergo catch-up growth in later years; and whether the manifestation of this catch-up growth is different for girls and boys. All these research questions require more advanced statistical methods for careful analysis of the longitudinal data. Repeated-measures analysis of variance and multivariate analysis of variance have therefore been popular for the analysis of longitudinal continuous data. However, those methods are not flexible when dealing with missing values and require certain strict assumptions, such as constant variances and covariances between repeated measures (known as sphericity) (4–6). Therefore, they have been gradually replaced by more recently developed statistical methodologies, such as multilevel modelling (5–12) and generalized estimating equations (5, 13).

In parallel with the development of multilevel modelling in the social sciences (also known as random-effects

modelling in biostatistics), another statistical methodology, latent growth curve modelling, has been developed (14–16) within the framework of structural equation modelling (17–20). In the last decade, statisticians have come to the conclusion that structural equation modelling yields answers equivalent to those of multilevel modelling with regards to longitudinal data analysis (12, 21, 22), and therefore software for structural equation modelling can be used to analyze multilevel (random effects) models. This discovery may impact on the practical issues of data analysis because the structural equation modelling framework offers greater flexibility in statistical modelling by allowing the incorporation of latent variables (21). The aim of this article was to show to dental investigators the possible benefits they might gain from using these methodologies when analysing longitudinal continuous data. The emphasis of this article is on conceptual understanding, rather than mathematical rigor, and therefore path diagrams will be extensively used for visual explanations of the statistical models with illustrations of practical examples.

Path diagrams

Structural equation modelling can be considered as a general theoretical framework for all univariate and multivariate linear statistical models (i.e. correlation, linear regression, analysis of variance, multivariate analysis of variance, canonical correlation, and factor analysis) (21–23). The statistical theory of structural equation modelling is complex, and the equations are usually written using matrix algebra. An alternative way for non-statisticians to appreciate the structural equation modelling concepts is through understanding the path diagrams of these statistical models. Indeed, some software packages for structural equation modelling, such as AMOS (Amos Developmental Corporation, Spring House, PA, USA), EQS (Multivariate Software, Encino, CA, USA), and LISREL (Scientific Software International, Lincolnwood, IL, USA), provide a graphical interface that allows users to draw path diagrams for their models on the computer, and the software then performs the analyses specified in the path diagrams.

The path diagram for simple linear regression

To illustrate what a path diagram is, we begin with the well-known example of linear regression with one outcome variable (known as the dependent variable) and one explanatory variable (known as the independent variable or covariate). Figure 1 is the path diagram for a simple linear regression model given as:

$$y = b_0 + b_1 x + e, \quad (1)$$

where y is the outcome variable, x is the explanatory variable, e is the residual error term, b_0 is the intercept, and b_1 is the regression coefficient for x. The interpretation of Eqn (1) is that when x is zero, y is b_0, and when x increases by one unit, y is expected to increase by the amount of b_1. The residual error term is the difference between the observed values of the outcome and the predicted (linear model) values of the outcome. In path diagrams, observed variables such as x and y are within squares, whereas latent (i.e. not directly measured and to be estimated in the model) variables such as residual errors (e in Eqn 1) are within circles. An arrow from variable x to variable y in a path diagram means that x affects y in the specified statistical model, but y does not affect x. By contrast, a double arrow connecting x and y means that these two variables are correlated without specific causal directions. For instance, suppose x is children's age, and y is their body height. It is reasonable to draw an arrow from x to y, as age causally determines body height. However, when x is body height and y is body weight, it is sensible to draw a double arrow, because both variables are determined by other factors, such as age. When there is no arrowed line (single or double) between x and y, this means that x and y are assumed to be causally independent (i.e. the underlying population correlation between them is assumed to be zero in the specified model). In the linear regression model, x and e are assumed to be uncorrelated, which is one of the assumptions behind regression analysis: explanatory variables and residual errors are independent. The arrow from one variable to another is called a 'path' in the diagram. In Fig. 1, there are two paths that specify the relationships between variables in the model: one from x to y, and another from e to y. As a result, two parameters associated with those two paths may be estimated. The parameter for the path from x to y is b_1, which is unknown but can be estimated, whereas the parameter for the path from e to y is fixed to be 1. Thereby, only one free (i.e. not already known or not fixed to a certain value) parameter for the relationship between x and y requires estimation.

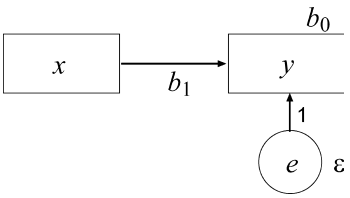

Fig. 1. Path diagram of simple linear regression in Eqn (1). For endogenous (dependent) variable y, the intercept b_0 is estimated. The regression coefficient b_1 in the path diagram is equivalent to the regression coefficient b_1 in Eqn (1). For the residual error e, its mean is fixed to zero but its variance (ϵ) is estimated. The regression coefficient for e is always fixed to unity. x is the explanatory variable.

Regression weights, path coefficients, and factor loadings

In linear regression, b_1 is usually called the regression coefficient, but in structural equation models, the parameters for the paths are sometimes called path coefficients or factor loadings. Despite the confusing

jargon, all these terms can be interpreted as regression coefficients, and here we simply call them regression coefficients. One exception is when a double arrow connects two variables (not in Fig. 1), in which case the estimated path coefficient is the covariance between the two variables.

Exogenous and endogenous variables

In a path diagram, like that shown in Fig. 1, variables such as x are known as exogenous variables because there is no arrow from other variables in the model directed towards them. By contrast, variables such as y are known as endogenous variables because there is at least one arrow from other variables (x in this model) in the model directed towards it. Endogenous variables are accompanied by residual errors, such as e in our model, because it is unlikely that the variations in y can be completely explained by x. Structural equation modelling estimates the means and variances for exogenous variables whilst estimating the intercepts for the endogenous variables. This is because the variances of endogenous variables are derived from exogenous variables and the associated residual errors. For example, in the linear regression given as Eqn (1), the intercept for y will be estimated and is equivalent to b_0. Both the mean and the variance of x will be estimated, although they are not explicitly expressed in Eqn (1). The mean of the residual errors (e) is fixed to zero (as it is in regression analysis) and the path coefficient from it to the associated endogenous variable is fixed to be 1 (as it is in Eqn 1). Therefore, the only parameter to be estimated is its variance. The mean and variance of y can then be derived from Eqn (1). Note that observed and unobserved variables can be exogenous or endogenous variables.

The path diagram for multiple linear regression

Multiple linear regression tests the relationship between one outcome variable, y, and more than one explanatory variable, x_1, x_2,… x_n. Figure 2 is the path diagram for a multiple linear regression with three explanatory variables denoted:

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + e, \qquad (2)$$

where y is the outcome variable, x_1 to x_3 are the explanatory variables, e is the residual error term, b_0 is the intercept, and b_1 to b_3 are the regression coefficients for x_1 to x_3, respectively. The interpretation of Eqn (2) is that when x_1 to x_3 is zero, y is b_0, and when x_1 increases by one unit and x_2 and x_3 are held constant, y is expected to increase by the amount of b_1.

In Fig. 2, the three paths from each of the three explanatory variables to y are equivalent to the regression coefficients given by Eqn (2), and the interpretations of these paths are the same as those for the regression coefficients. Note that there are three double arrows in Fig. 2 that connect x_1, x_2, and x_3, and these represent the covariance amongst the three explanatory variables (whose means and variances will also be estimated). This indicates that the relationship between y and each x is determined whilst also taking into account the correlations amongst the three explanatory variables. Note that when multiple regression analysis is undertaken using standard software packages, the explanatory variables are always assumed to be correlated, regardless of whether or not subsequent interpretation of the regression coefficients recognizes this.

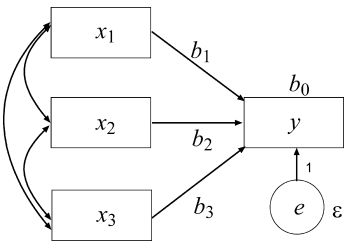

Fig. 2. The path diagram for multiple regression. For endogenous (dependent) variable y, the intercept b_0 is estimated. The three regression coefficients (b_1, b_2, and b_3) in the path diagram are equivalent to those in Eqn (2). For the residual error e, its mean is fixed to zero but its variance (ϵ) is estimated. The regression coefficient for e is always fixed to unity. x_1 to x_3 are the explanatory variables.

Two-level growth models for longitudinal data analysis

For illustration, we used the data of four repeated measures of the distance between sella and the pterygomaxillary fissure (*Dist*), obtained from children assessed at 8, 10, 12, and 14 yr of age, published by Potthoff & Roy (24). Figure 3 shows that there seems to be a difference in growth trajectories of *Dist* between 11 girls and 16 boys. Boys have a greater value of *Dist* at baseline and gain a greater increase in the distance than girls. We therefore wanted to investigate the following research questions.

A. Is there a difference between boys and girls in the changes in *Dist* between ages 8 and 14 yr?
B. Are the growth patterns/ trajectories linear or nonlinear?
C. When does the growth spurt take place?
D. Do boys and girls follow similar growth trajectories?

The path diagram in Fig. 4 shows the general concept of the latent growth curve model for the analysis of growth in the distance between sella and the pterygomaxillary fissure (*Dist*) for these children. As explained previously, observed and measured variables are in squares. In this model, the observed variables are the four measurements of *Dist* made at 8 (*Distance*8), 10 (*Distance*10), 12 (*Distance*12), and 14 (*Distance*14) yr of age. Another variable observed in the model was *Sex*, which is a binary variable (girls are coded 0 and boys are coded 1). The

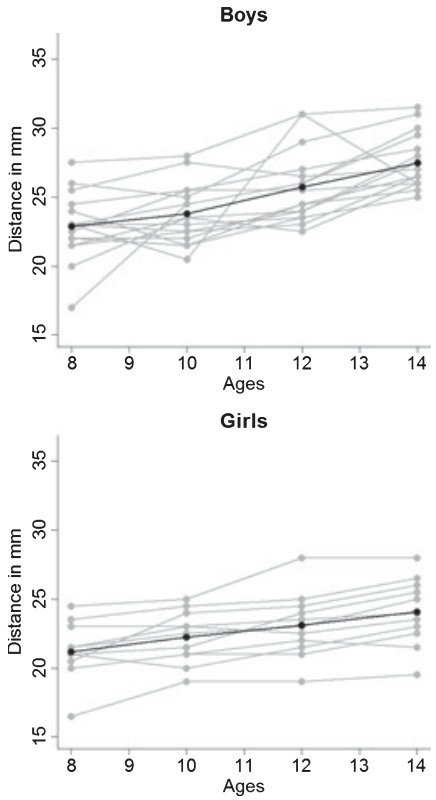

Fig. 3. Observed growth patterns of 16 boys and 11 girls.

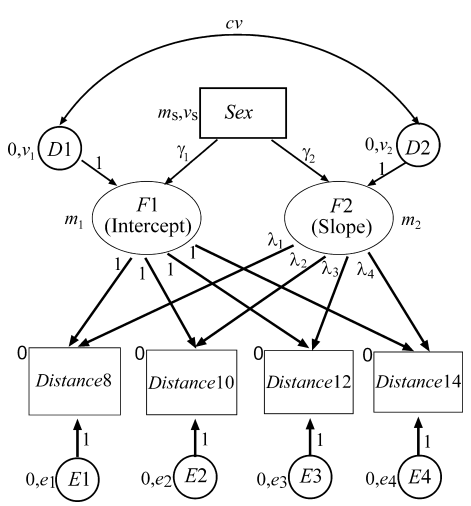

Fig. 4. Path diagram for the random effects model. *cv* – covariance between *D*1 and *D*2; m_s – mean of *Sex*; v_s – variance of *Sex*; v_1 – variance of *D*1; v_2 – variance of *D*2; m_1 – intercept for *F*1; m_2 – intercept for *F*2; γ_1 – regression coefficient for the path from *Sex* to *F*1; γ_2 – regression coefficient for the path from *F*2 to *Sex*; λ_1 to λ_4 – regression coefficients for the paths from *F*2 to *Distance*8 to *Distance*14; *e*1 to *e*4 – variances for *E*1 to *E*4. *Sex* is a dummy variable in which boys are coded 1 and girls coded 0. Please see the texts for explanation of other variables in the path diagram.

parameters m_s and v_s are the mean and variance of *Sex*, respectively.

To estimate a linear growth model for the changes in *Dist*, two latent variables are required (shown in circles in Fig. 4): *F*1 is the estimated baseline *Dist*, and *F*2 is the estimated slope for the growth curve. The parameters m_1 and m_2 are the intercepts for *F*1 and *F*2; and *D*1 and *D*2 are residual error terms for *F*1 and *F*2. Recall that for endogenous variables (i.e. *F*1, *F*2, and the four measurements of *Dist* in this model) only the intercepts are estimated because they are affected and 'explained' by exogenous variables (i.e. *Sex* in this model) and their associated residual errors (*D* and *E*). Similarly to residual error terms in regression analysis, the means of *D*1 and *D*2 are fixed to be zero, and v_1 and v_2 are their variances, respectively. *E*1 to *E*4 are the error terms for each observed variable; *E*1 to *E*4 are assumed to be uncorrelated and to have a mean of zero. By contrast, the two latent variables *F*1 and *F*2 are assumed to be correlated (there is a double arrow between *D*1 and *D*2, and their covariance is estimated in Fig. 4).

Note that *F*1 and *F*2 are unobserved (latent) variables, which means that unlike *Dist*, they are not directly measured but are estimated by extracting information from the observed variables. Therefore, the meaning of *F*1 and *F*2 depends upon how this information is extracted (i.e. it depends upon how the relationships between them and *Dist* are defined in the model by specifying the parameters for the arrows from *F*1 and *F*2 to *Dist*). Each arrow represents a path from one variable to another, and as explained previously, these paths can be interpreted as regression coefficients. For instance, all the regression coefficients for the arrows from *F*1 to the four measurements of *Dist* are 1, and those for the arrows from the residual errors (*D* and *E*) are also fixed to be 1. So, a latent growth curve model or structural equation model can be viewed as an attempt to use multiple equations to define the relationships amongst observed and unobserved variables in the model.

For instance, the equation for the relationship between *Distance*8 and other variables in Fig. 4 is given as:

$$Distance8 = 1 * F1 + \lambda_1 * F2 + 1 * E1. \quad (3)$$

Similarly, the equations for *Distance*10, *Distance*12, and *Distance*14 in Fig. 4 are given as:

$$Distance10 = 1 * F1 + \lambda_2 * F2 + 1 * E2, \quad (4)$$

$$Distance12 = 1 * F1 + \lambda_3 * F2 + 1 * E3, \quad (5)$$

$$Distance14 = 1 * F1 + \lambda_4 * F2 + 1 * E3, \quad (6)$$

$$F1 = m_1 + \gamma_1 * Sex + 1 * D1, \text{and} \quad (7)$$

$$F2 = m_2 + \gamma_2 * Sex + 1 * D2, \quad (8)$$

where m_1 and m_2 are the intercepts in Eqn (7) and Eqn (8) for *F*1 and *F*2, respectively; and λ_1, λ_2, λ_3, λ_4, γ_1, and γ_2 are regression coefficients. Recall the simple linear regression given by Eqn (1): $y = b_0 + b_1x + e$, where b_0 and b_1 are two unknown parameters that need to be estimated. The regression coefficient for *e* is actually fixed to be unity just like those for *E*1 to *D*2 in Eqns (3–8). To estimate a linear curve, the paths from *F*2 to *Distance*8, *Distance*10, *Distance*12, and *Distance*14 (λ_1, λ_2, λ_3, and λ_4) are fixed to be 0, 2, 4, and 6, respectively, and Eqns (3–6) are now given as:

$$Distance8 = 1 * F1 + *E1, \quad (3')$$

$$Distance10 = 1 * F1 + 2 * F2 + 1 * E2, \quad (4')$$

$$Distance12 = 1 * F1 + 4 * F2 + 1 * E3, \text{and} \quad (5')$$

$$Distance14 = 1 * F1 + 6 * F2 + 1 * E3. \quad (6')$$

As a result, $F1$ is the estimated baseline *Dist* at 8 yr of age, and $F2$ is the estimated growth in *Dist* per year. If λ_1, λ_2, λ_3, and λ_4 were fixed to be 8, 10, 12, and 14 (i.e. the actual ages when these measurements were taken), $F1$ would be the estimated *Dist* at birth (age 0), but $F2$ would still be the average growth in *Dist* per year. Therefore, we can view latent growth curve modelling (or structural equation modelling in general) as a system of multiple equations for the relationships amongst the observed and latent variables and their relationships can be identified by solving these equations simultaneously. This is in contrast to linear regression where we only need to solve one equation to identify the relationship between one dependent variable and other explanatory variables. In Eqns (3–8), some of the parameters are already given (such as the regression coefficients fixed to be unity), and the unknown parameters, such as m_i, and γ_i, can be estimated. In Fig. 4, all the means of residual errors are fixed to be zero, just like the means of residual errors in linear regression models. It is noted that the intercepts of the observed outcome variables *Distance*8, *Distance*10, *Distance*12, and *Distance*14 in Eqns (3–6) are also fixed to be zero, because the means of these variables will be estimated via the means or the intercepts (i.e. m_1 and m_2) of latent variables $F1$ and $F2$.

In summary, for this longitudinal data analysis, latent growth curve modelling estimates a linear growth trajectory (i.e. a change pattern) for each patient, just as a multilevel model does. Variations in the intercepts and slopes of these trajectories, regarded as fixed and random effects in multilevel models, are explicitly specified in latent growth curve models as latent variables, because unlike *Dist*, which is directly observed and measured, these trajectories and their variations (random effects) are unknown and need to be estimated.

Results from analyses using linear latent growth curve modelling

Research question A: is there a difference between boys and girls in the changes in *Dist* between 8 and 14 yr of age?

We first fitted linear growth curves for the changes in *Dist*. The linear latent growth curve model in Fig. 4 was analysed using the statistical software package MPLUS (25). The results from MPLUS, using maximum likelihood estimation, are shown in Table 1. The baseline average *Dist* for girls at 8 yr of age was 21.21 mm, and the increase (growth) in *Dist* was 0.48 mm per year. The baseline average *Dist* for boys at 8 yr of age was 21.21 + 1.41 = 22.62 mm, and the average growth in *Dist* per year was 0.48 + 0.30 = 0.78 mm. The variances of the intercepts and slopes for all the 27 growth trajectories were 2.91 and 0.02, respectively, and their covariance was −0.01.

MPLUS (and other structural equation modelling software) provides several indices for assessing overall model fit. The most commonly used are the Chi-square test and the root mean square error of approximation (RMSEA). The rationale behind the model fit assessment is as follows. Each structural equation model provides a framework of causal relationships amongst observed variables. This framework entails a certain structure of correlational relationships among the observed variables, which is usually given as a correlation/covariance matrix Σ. The difference between Σ and the observed correlation/covariance matrix S is formulated by a likelihood function, which is then minimized using iterative procedures. The Chi-square test is then used to evaluate the difference between these two matrices by taking into account the number of the estimated parameters in the proposed model. In theory, if the relationships amongst the observed and latent variables specified in the model are correct, there should be no difference between Σ and S, although we may find some small differences attributed to random sampling errors. When the Chi-square value is large (i.e. the overall difference between the two matrices is large) relative to the model's degree of freedom, the proposed model is rejected (i.e. something in the relationships specified by the proposed model is not correct and requires further consideration). When the Chi-square value is small, we fail to reject the model or tentatively accept the model as adequate.

The Chi-square value for the linear growth model is 11.297 with 10 degrees of freedom (d.f.) ($P = 0.335$) and the RMSEA is 0.069. In structural equation modelling, the null hypothesis for the Chi-square test is that

Table 1

Results of the linear growth curve model from MPLUS

	Linear growth curve model		
Effect	Estimate	SE	*P*-value
Regression coefficients			
m_1	21.209	0.611	< 0.001
m_2	0.480	0.100	< 0.001
γ_1	1.407	0.794	0.089
γ_2	0.305	0.130	0.021
Variance/covariance			
v_1	2.905	1.141	0.006
cv	−0.008	0.147	0.956
v_2	0.024	0.034	0.243
e_1 to e_4	1.716	0.330	< 0.001

m_1 – intercept for $F1$, i.e. the mean baseline *DIST* for girls; m_2 – intercept for $F2$, i.e. the growth in *DIST* for girls; γ_1 – regression coefficient for the path from *Sex* to $F1$; γ_2 – regression coefficient for the path from *Sex* to $F2$; cv – covariance between $D1$ and $D2$; v_1 – variance of $D1$; v_2 – variance of $D2$; $e1$ to $e4$ – variances for $E1$ to $E4$.
SE, standard error.

there is no difference in the covariance structures between the proposed model and the data, and a *P*-value of > 0.05 means that we cannot reject the null hypothesis. A small RMSEA means that the proposed model fits the data relatively well (0.06 is usually the cut-off value) (19). Although our model seems to be acceptable according to the Chi-square test, the RMSEA indicates that it may be modified to achieve better model fit.

Results from analyses using non-linear latent growth curve modelling

Research question B: are the growth patterns/ trajectories linear or non-linear?

Above, we tested a linear growth model (i.e. the fitted growth curve in *Dist* for each child was assumed to follow a straight line). However, the observed growth trajectories that are shown in Fig. 3 may suggest a non-linear growth curve for boys because the mean trajectory for *Dist* (the black line) seems to have a steeper slope after 10 and 12 yr of age than before 10 yr of age. There are many simple and advanced approaches to model non-linear curves in the statistical literature. As only four measurements of *Dist* were taken over the 6-yr period, some advanced methods, such as fractional polynomials and splines, are not suitable. Alternatively, we can introduce a quadratic term into the latent growth curve model to capture the non-linearity in the growth by incorporating a third latent variable, *F*3, as shown in Fig. 4, to model the quadratic growth curves. The regression weights for the paths from *F*3 to *Distance*8, *Distance*10, *Distance*12, and *distance*14 are fixed to be 0, $2^2 = 4$, $4^2 = 16$, and $6^2 = 36$, respectively. There is also an arrow from *Sex* to *F*3 (γ_3) to test whether girls and boys have a different acceleration in growth speed. The results from MPLUS are shown in Table 2. The growth acceleration in *Dist* for girls (m_3) is −0.003, indicating that the growth speed decreased with age, while the regression coefficient for γ_3 is 0.056, indicating that the growth speed ($-0.003 + 0.056 = 0.053$) for boys increased with age. Even so, neither estimate was statistically significant (i.e. the quadratic growth curve model is not statistically better than the linear curve model in explaining the growth patterns).

Table 2

Results of the quadratic growth curve model from MPLUS

Quadratic growth curve model			
Effect	Estimate	SE	*P*-value
Regression coefficients			
m_1	21.198	0.641	< 0.001
m_2	0.497	0.309	0.112
γ_1	1.621	0.833	0.063
γ_2	−0.017	0.401	0.967
m_3	−0.003	0.049	0.954
γ_3	0.054	0.063	0.399
Variance/covariance			
v_1	2.940	1.140	0.005
cv	−0.016	0.147	0.916
v_2	0.026	0.034	0.220
e_1 to e_4	1.667	0.330	< 0.001

m_1 – intercept for *F*1, i.e. the mean baseline *DIST* for girls; m_2 – intercept for *F*2, i.e. the growth in *DIST* for girls; m_3 – intercept for *F*3; i.e. the acceleration in growth speed in *DIST* for girls; γ_1 – regression coefficient for the path from *Sex* to *F*1; γ_2 – regression coefficient for the path from *Sex* to *F*2; γ_3 – regression coefficient for the path from *Sex* to *F*3; *cv* – covariance between *D*1 and *D*2; v_1 – variance of *D*1; v_2 – variance of *D*2; *e*1 to *e*4 – variances for *E*1 to *E*4.
SE, standard error.

Research question C: when does the growth spurt take place?

We took advantage of the fact that latent growth curve modelling has another elegant way to model the non-linear growth curves without the need for a third latent variable. Recall that in the linear growth curve model, the paths from *F*2 to *Distance*8, *Distance*10, *Distance*12, and *Distance*14 (λ_1, λ_2, λ_3, and λ_4) were fixed to be 0, 2, 4, and 6, respectively. To capture the non-linearity, we fixed the first path (λ_1) to be 0 and the final path (λ_4) to be 6, but allowed λ_2 and λ_3 to be free parameters for estimations. In other words, we estimated the proportion of growth taking place between two consecutive measurements. When the growth speed remains constant from 8 to 14 yr of age, the estimated values for λ_2 and λ_3 should be very close to 2 and 4. The results from this approach using MPLUS showed that λ_2 and λ_3 were 1.458 and 3.815, indicating that the growth seemed to be slower between 8 and 10 yr of age and became faster after 10 yr of age. The Chi-square value for the model was 9.989 with 8 d.f. ($P = 0.266$) and an RMSEA of 0.096. The difference in the Chi-square values between the linear and non-linear models could be used to compare the two models, as the linear model is a restricted form of the non-linear model (i.e. the linear model can be derived from the non-linear model by invoking the constraints on the values of λ_2 and λ_3). In structural equation modelling, these two models are known as nested models (18–20). In the comparison of the nested models, the one with fewer d.f. (i.e. the non-linear model) will always have a smaller Chi-square value. The difference in Chi-square values is therefore used to test whether the models that yield smaller Chi-square values but use more d.f. are more parsimonious [i.e. whether the additional use of d.f. to estimate free parameters (such as λ_2 and λ_3) achieve significant improvement in the model fit] (18–20) The difference in Chi-square values between linear and non-linear models is 1.308 with 2 d.f. ($P = 0.52$), indicating that the non-linear model does not provide a significantly better model fit. However, the statistical power to reject the linear model in favour of the non-linear model is quite weak as a result of the small sample size, and the issue of sample size estimation will be discussed later.

Results from multiple-group latent growth curve modelling analyses

Research question D: do boys and girls follow similar growth trajectories?

This research question calls for a contrast of the groups of boys and girls. In the previous linear or non-linear models, we assumed that boys and girls both have the same growth patterns but have different baseline values and amount of changes. Latent growth curve modelling is able to test whether both sexes have the same growth pattern, and this is known as multiple-group analysis. We first specified a two-group model in which boys and girls are assumed to follow the same linear growth patterns and to have the same amount of growth in the distance (the baseline distance was allowed to be different). The results showed that the baseline values of *Dist* were 20.771 and 22.917 for girls and boys, respectively, and the growth in the distance was 0.66 for both groups. The Chi-square value was 40.094 (d.f. = 21) with a *P*-value of 0.007, indicating a poor model fit. This is hardly surprising because the amount of growth has been constrained to be equal in both groups. We then estimated different changes in *Dist* for boys and girls, and the results showed that the baseline values of *Dist* are 21.4 mm and 22.9 mm and the growth in *Dist* is 0.356 mm and 0.591 mm for girls and boys, respectively. The Chi-square value was 35.064 (d.f. = 20) with a *P*-value of 0.02, and the difference in the Chi-square was 5.03, suggesting that the growth in the distance was different between boys and girls. However, the overall model fit can still be improved, and this is achieved by allowing the residual variances to be different. The results showed that the residual variance was 0.454 and 2.649 for girls and boys, respectively, and that the Chi-square value was reduced to 14.403 (d.f. = 19, $P = 0.76$). This indicates that the growth trajectories are more homogeneous amongst girls than boys, which is corroborated by the data shown in Fig. 3.

We were then able to test whether or not there was a non-linear growth pattern by estimating the regression coefficients λ_2 and λ_3 for *Distance*10 and *Distance*12, respectively, as in the previous section. By plotting the growth trajectories for each group, Fig. 3 seemed to suggest that girls had a linear growth pattern and that boys may have a non-linear growth pattern. To test this hypothesis, we relaxed our assumption for linearity for both groups (i.e. we estimated λ_2 and λ_3 for each group). The results showed that $\lambda_2 = 2.128$ and $\lambda_3 = 4.108$ for girls, and that $\lambda_2 = 1.250$ and $\lambda_3 = 3.758$ for boys, suggesting that the growth speed seems to follow a linear pattern in girls but that in boys the growth speed is slower between 8 and 10 yr of age and then becomes faster. The model Chi-square was 11.902 (d.f. = 15) with a *P*-value of 0.686, and because of the small sample size, the difference in the Chi-square values was not statistically significant ($\chi^2 = 2.501$, d.f. = 4). Therefore, we failed to reject the hypothesis that both girls and boys had a linear growth pattern.

Statistical power calculation in latent growth curve modelling

The non-linear model testing and multigroup analysis in the previous sections raised the issue of sample size and statistical power calculation. As the Chi-square value for model fit is directly related to sample size, the Chi-square test will have greater statistical power when sample size increases. The estimation of statistical power is more complex in latent growth curve modelling as it deals with more than one outcome in the model. Two commonly used methods for sample size and statistical power calculation for structural equation modelling are the Satorra-Saris method (26) and Monte-Carlo simulations (25, 26). It is beyond the scope of this article to give details of these methods, so we only demonstrate how to use the former to estimate the required sample size to test whether boys and girls have different non-linear growth patterns. The first step is to request estimated covariance matrices and means from the output of structural equation modelling software for the model in which non-linear growth patterns were estimated for both groups. The second step is to use those matrices and means to run the same model, and because they are the estimated covariances and means, the model Chi-square should be very close to zero, indicating a nearly perfect fit (taking into account the rounding errors in the original means and matrices). In our example, the Chi-square value is 0.005. Then we use the same means and covariances to estimate a linear model for both groups, and the results show that the difference in Chi-square values is 1.545, which follows a non-central Chi-square distribution with d.f. = 4 (23), and the estimated statistical power is therefore 0.14. To achieve statistical power of 0.8, the required sample size for each group is estimated to be around 135 subjects (i.e. to obtain sufficient statistical power to test whether girls and boys have different non-linear growth trajectories or both groups follow linear growth trajectories, a total of 270 subjects is required). For an accessible account of statistical power estimation within structural equation modelling, we refer the readers to an excellent book by Brown (26).

Discussion

The methods introduced in the previous sections, if applied properly, can be very useful and powerful statistical tools for dental researchers. Some complex, but vital, research hypotheses can only be tested using these sophisticated methodologies. The conceptual difference in the analysis of longitudinal data between multilevel modelling and structural equation modelling is that longitudinal data, such as the example in this article, are analysed using a two-level analysis in multilevel modelling (four repeated measurements of *Dist* is level 1 and subjects is level 2), while they are analysed using a single level analysis in structural equation modelling (21, 22). As a result, the fixed and random effects at level 2 in multilevel modelling are explicitly specified as latent

variables in structural equation modelling. Latent growth curve modelling can be extended to analyze multiple longitudinal continuous outcomes to delineate the direct and indirect effects of interventions (27).

Recent developments in statistical theory and software programming have also extended the uses of this methodology for counts, and for ordinal and binary outcomes (28–31). Although it is beyond the scope of this article to describe these new developments, they can potentially change the way in which dental researchers analyse their data. For instance, in clinical trials on the efficacy of dental hygiene products, ordinal site-level plaque scores coded as 0, 1, 2, and 3 are often analysed as a continuous variable or aggregated to the subject level before the analyses. An alternative approach is to use latent growth curve modelling by assuming an underlying unobserved continuous variable for the amount of dental plaque.

Any statistical method has its limitations, and latent growth curve modelling is no exception. In our examples, the intervals between the measurements of outcomes were approximately identical for all subjects. If the intervals between the measurements were quite different for many patients, and these differences cannot be assimilated, this will not pose any problem in the analyses employing multilevel modelling software, but this may be a problem when undertaking latent growth curve modelling within some SEM software packages (although this is not a problem for MPLUS). Therefore, researchers should choose the methods (and software packages) best suited for their research questions and study design. We strongly encourage dental researchers to work with experienced biostatisticians when they wish to use these methods to analyse their longitudinal continuous data.

References

1. Apatzidou DA, Kinane DF. Quadrant root planing versus same-day full-mouth root planing: I. Clinical findings. *J Clin Periodontol* 2004; **31**: 132–140.
2. Preshaw PM, Heasman PA. Periodontal maintenance in a specialist periodontal clinic and in general dental practice. *J Clin Periodontol* 2005; **32**: 280–286.
3. Jackson MA, Kellett M, Worthington HV, Clerehugh V. Comparison of interdental cleaning methods: a randomized controlled trial. *J Periodontol* 2006; **77**: 1421–1429.
4. Gunsolley JC, Chinchllll VM, Koertge TE, Palcanis KG, Sarbln AG, Brooks CN. The use of repeated measures analysis of variance for plaque and gingival indices. *J Clin Periodontol* 1989; **16**: 156–163.
5. Twisk JWR. *Applied longitudinal data analysis for epidemiology*. Cambridge: Cambridge University Press, 2003.
6. Hox J. *Multilevel analysis*. Mahwah, NJ: Laurence Erlbaum Associates, 2002.
7. Raudenbush SW, Bryk AS. *Hierarchical linear models: applications and data analysis methods. 2nd edn*. Thousand Oaks, CA: Sage Publication, 2002.
8. Singer JB, Willett JD. *Applied longitudinal data analysis*. New York: Oxford University Press, 2003.
9. Goldstein H. *Multilevel statistical models. 2nd edn*. New York: John Wiley, 1995.
10. Gilthorpe MS, Zamzuri AT, Griffiths GS, Maddick IH, Eaton KA, Johnson NW. Unification of the "burst" and "linear" theories of periodontal disease progression: a multilevel manifestation of the same phenomenon. *J Dent Res* 2003; **82**: 200–205.
11. Twisk JWR. *Applied multilevel analysis*. Cambridge: Cambridge University Press, 2006.
12. Steele F. Multilevel models for longitudinal data. *J R Stat Soc Ser A* 2008; **171**: 1–15.
13. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; **73**: 13–22.
14. Byrne BM, Crombie G. Modeling and testing change: an introduction to the latent growth curve model. *Underst Stat* 2003; **2**: 177–203.
15. Bollen KA, Curran PJ. *Latent curve models*. Hoboken, NJ: Wiley, 2006.
16. Duncan TE, Duncan SC, Strycker LA. *An introduction to latent variable growth curve modeling. 2nd edn*. Mahwah, NJ: Laurence Erlbaum Associates, 2006.
17. Bollen KA. *Structural equations with latent variables*. New York: Wiley, 1989.
18. Diamantopoulos A, Siguaw JA. *Introducing LISREL*. Thousand Oaks, CA: Sage Publication, 2000.
19. Kline RB. *Principles and practice of structural equation modeling, 2nd edn*. New York: Guilford Press, 2005.
20. Loehlin JC. *Latent variable models: an introduction to factor, path, and structural equation analysis. 4th edn*. Mahwah, New Jersey: Laurence Erlbaum Associates, 2004.
21. Curran P. Have multilevel models been structural equation models all along? *Multivariate Behav Res* 2003; **38**: 529–569.
22. Bauer DJ. Estimating multilevel linear models as structural equation models. *J Educ Behav Stat* 2003; **28**: 135–167.
23. Tu Y-K, Baelum V, Gilthorpe MS. A structural equation modelling approach to the analysis of change. *Eur J Oral Sci* 2008; **116**: 291–296.
24. Potthoff RF, Roy SN. A generalized multivariate analysis of variance model useful especially for growth curve problem. *Biometrika* 1964; **52**: 313–326.
25. Muthén LK, Muthén B. *Mplus user's guide. 4th edn*. Los Angeles, CA: Muthén & Muthén, 2006.
26. Brown TA. *Confirmatory factor analysis for applied research*. New York: The Guilford Press, 2006.
27. Tu Y-K, Jackson M, Kellett M, Clerehugh V. Direct and indirect effects of interdental hygiene in a clinical trial. *J Dent Res* 2008; **87**: 1037–1042.
28. Muthén BO. Growth modeling with binary responses. In: Von Eye A, Clogg C, eds. *Categorical variables in developmental research: methods of analysis*. San Diego, CA: Academic Press, 1996; 37–54.
29. Muthén B. Second-generation structural equation modeling with a combination of categorical and continuous latent variables: new opportunities for latent class/latent growth modeling. In: Collins LM, Sayer A, eds. *New methods for the analysis of change*. Washington, D.C: American Psychological Association, 2001; 291–322.
30. Skrondal A, Rabe-Hesketh S. *Generalized latent variable modeling: multilevel, longitudinal, and structural equation models*. Boca Raton, FL: Chapman & Hall, 2005.
31. Little TD, Preacher KJ, Selig JP, Card NA. New developments in latent variable panel analyses of longitudinal data. *Int J Behav Dev* 2007; **31**: 357–365.