Method Note

American Journal of Evaluation
2014, Vol. 35(4) 543-561

DOI: 10.1177/1098214014523823
aje.sagepub.com
SAGE

Multilevel Growth Modeling: An Introductory Approach to Analyzing Longitudinal Data for Evaluators

Kevin A. Gee[1]

Abstract

The growth in the availability of longitudinal data—data collected over time on the same individuals—as part of program evaluations has opened up exciting possibilities for evaluators to ask more nuanced questions about how individuals' outcomes change over time. However, in order to leverage longitudinal data to glean these important insights, evaluators responsible for analyzing longitudinal data face a new set of concepts and analytic techniques that may not be part of their current methodological tool kit. In this article, I provide an applied introduction to one method of longitudinal data analysis known as multilevel growth modeling. I ground the introductory concepts and illustrate the method of multilevel growth modeling in the context of a well-known longitudinal evaluation of an early childhood care program, the Carolina Abecedarian Project.

Keywords

multilevel growth modeling, longitudinal data, impact evaluation, longitudinal methods

Introduction

The growth in the availability of longitudinal data—data collected over time on the same individuals—as part of program evaluations has opened up exciting possibilities to ask more nuanced questions about program impact. One key question of interest to impact evaluators is how changes over time on some outcome differ based on program participation. However, in order to leverage longitudinal data to understand differences over time on outcomes between program participants and nonparticipants, evaluation analysts responsible for analyzing longitudinal data face a new set of concepts and analytic techniques that may not be part of their current methodological tool kit. In this article, I provide an applied introduction to one method of longitudinal data analysis known as multilevel growth modeling.[1] The content and focus of this article is inspired by and synthesizes prior work that provides detailed and comprehensive explanations of the methodological and analytical approaches that underlie multilevel growth modeling,

[1] School of Education, University of California, Davis, CA, USA

Corresponding Author:
Kevin A. Gee, School of Education, University of California Davis, One Shields Ave., Davis, CA 95616, USA.
Email: kagee@ucdavis.edu

including the influential works of Singer and Willett (2003; Applied Longitudinal Data Analysis) and Raudenbush and Bryk (2002; Hierarchical Linear Models: Applications and Data Analysis).

Given a wealth of prior works devoted to explaining multilevel growth modeling, what is the impetus behind this methodological primer? The purpose and rationale for this introductory piece is threefold. First, although the technique of multilevel growth modeling has been in existence for over 20 years with seminal texts devoted to the method (Fitzmaurice, Laird, & Ware, 2004; Hedeker & Gibbons, 2006; Raudenbush & Bryk, 2002; Singer & Willett, 2003; Snijders & Bosker, 1999), currently, the published evaluation literature is limited in providing a concise and accessible ''how-to'' approach to multilevel growth modeling that speaks to the needs of both evaluation analysts and consumers of the evaluation literature.[2] Second, evaluators who are educated through degree and certificate granting academic programs in evaluation, particularly at the master's level,[3] often do not receive training in longitudinal methods. Thus, although evaluators may have the statistical foundations to conduct basic descriptive analyses of longitudinal data, they may lack knowledge of how to analyze data using more sophisticated longitudinal methods. Finally, given the interdisciplinary nature of the field of evaluation—spanning fields as diverse as education, psychology, management studies, public health, and economics—methodological traditions vary in evaluation. Therefore, evaluators may not have been exposed to the method of multilevel growth modeling as it may lie outside the methodological conventions of their own disciplinary backgrounds. Accordingly, three potential audiences may find this introduction to multilevel growth modeling of particular interest and value: (1) evaluation analysts who are interested in carrying out longitudinal analyses but need a starting point to do so; (2) consumers of the evaluation literature who want to understand studies that employ such methods; and (3) evaluators who are trained in longitudinal methods from other disciplinary areas, such as econometrics, and want to understand evaluation studies and conduct analyses using multilevel growth modeling.

Throughout this article, I ground the introductory concepts and illustrate the method of multilevel growth modeling in the context of a well-known longitudinal evaluation of an early childhood care program, the Carolina Abecedarian Project (hereafter referred to as the Abecedarian Project). The multilevel growth modeling concepts and procedures I introduce through this example can be applied more generally to the analysis of evaluations that are designed to track participants on some prespecified continuous measure over time. There are numerous examples of such evaluations from a variety of substantive fields including social policy, education, and public health. For example, there are job training programs that track wage information; educational interventions that capture educational test scores; smoking cessation programs that document reduction in smoking behaviors (the number of cigarettes smoked); and obesity prevention programs that track body mass index.

I begin this article with a brief overview of the Abecedarian Project, including its overall evaluation design and data. Then, I guide the reader through the following topics:

1. data requirements for multilevel growth models;
2. visualizing change over time;
3. specifying multilevel growth models;
4. interpreting and displaying the results of multilevel growth models.

Readers should note that the last two topics do require familiarity with statistical concepts underlying ordinary least squares (OLS) regression analysis.

The Carolina Abecedarian Project and Data Set

The Carolina Abecedarian Project was a longitudinal study that began in 1972 and tracked the developmental outcomes of 111 children in North Carolina who, at infancy, were randomly assigned to

Table 1. Data for Four Sample Infants From the Abecedarian Project Displayed in Cross-Sectional Format.

IDNO	TREAT	MDI6	MDI12	MDI18
5110	0	100	115	88
5111	0	116	134	132
5401	1	94	100	100
5402	1	110	134	123

Note. IDNO = identification number; MDI = mental development index. Two infants (IDNO = 5110 and IDNO = 5111) were randomly assigned into the control group (TREAT = 0), while the remaining two (IDNO = 5401 and IDNO = 5402) were randomized into the treatment group (TREAT = 1).

either a child-centered preschool program (treatment group) or no preschool (control group). A majority of the participating children were African American and all were from economically disadvantaged households. Children participating in the project have been followed into adulthood to determine the long-term consequences of providing high-quality early childhood care. Readers further interested in the substantive background and details of the study design, including the design of the intervention as well as key findings, should refer to Campbell and Ramey (1995).

In this article, I illustrate how multilevel growth modeling can be applied to the Abecedarian Project data[4] to examine how infants' growth in cognitive and linguistic functioning differs by treatment status. I use infants' Mental Development Index (MDI) scores on the Bayley Scales of Infant Development (BSID) as a measure of their cognitive and linguistic functioning. I compare MDI scores between control and treatment group infants across three distinct time points—at 6, 12, and 18 months of age.

Data for Multilevel Growth Modeling

Data Collection and Organization

To carry out multilevel growth modeling, you first need data on a continuous outcome[5] that is measured and collected at multiple time points for a sample of individuals. Also, the outcome should capture the same underlying construct (e.g., cognitive and linguistic functioning) on a consistent scale across each time point.[6] Once data are collected, they can be arranged in cross-sectional format with one row of data per individual in the data set. Data can also be organized in a panel format so that each individual has multiple rows, one for each measurement occasion. Often, analysts enter, clean, and store data in cross-sectional format. They then arrange their data into panel format in preparation for analysis. Several widely used statistical software programs such as Stata, SPSS, and SAS have data manipulation routines and commands that automate the process of transforming data from cross-sectional to panel formats (and vice versa).[7]

Tables 1 and 2 display data for four sample infants from the Abecedarian Project in cross-sectional and panel formats, respectively. In Table 1, there are several variables to note: (1) the identification number (IDNO) uniquely identifying each infant (e.g., 5110); (2) a dichotomous variable (TREAT) indicating whether or not the infant was assigned to the treatment (TREAT = 1) or control (TREAT = 0) group; and (3) the outcome variables (MDI6, MDI12, and MDI18) recording infants' MDI scores at 6, 12, and 18 months of age. In Table 2, there are three features to note for the data in panel format: (1) each infant has three rows of data, one for each measurement occasion; (2) the variable MDI (column 3) is a single variable, having lost its numerical suffixes (6, 12, or 18) that existed in a cross-sectional format. These suffixes are now recorded in a new variable labeled MONTH (column 4) and document the age, in months, at which each infant's MDI score was measured; (3) finally, the

Table 2. Data for Four Sample Infants From the Abecedarian Project Displayed in Panel Format.

IDNO	TREAT	MDI	MONTH
5110	0	100	6
5110	0	115	12
5110	0	88	18
5111	0	116	6
5111	0	134	12
5111	0	132	18
5401	1	94	6
5401	1	100	12
5401	1	100	18
5402	1	110	6
5402	1	134	12
5402	1	123	18

Note. IDNO = identification number; MDI = mental development index.

Table 3. Descriptive Statistics of Mental Development Index (MDI) Scores for Infants From the Abecedarian Project.

		Control Group		Treatment Group	
Variable	Description	*N*	Mean	*N*	Mean
MDI6	Scores on the Mental Development Index (MDI) of the	53	101.34 (14.70)	53	107.43 (15.48)
MDI12	*Bayley Scales of Infant Development* (the numbered suffix	53	105.83 (14.47)	51	111.59 (14.33)
MDI18	denotes month of measurement)	49	89.98 (11.70)	51	107.51 (13.98)

Note. Standard deviation in parentheses. $n = 111$.

arrangement of data in panel format illustrates the multilevel nature of the data—there are three MDI scores that are "nested" within each individual infant. Table 3 presents average MDI scores at each measurement occasion for the full sample ($n = 111$) of infants by treatment status.

Visualizing Change Over Time

With data collected and organized in either cross-sectional or panel format, you might be inclined to immediately begin fitting multilevel growth models. Rather, you should consider visualizing your data to determine how your outcomes are changing over time. One powerful way to visualize your data is to create and display a set of empirical growth trajectories (Singer & Willett, 2003, p. 28). These trajectories consist of a set of OLS regression lines fit through your outcomes for each individual.[8] Importantly, creating and displaying empirical growth trajectories for separate groups (e.g., program participants versus nonprogram participants) helps you gauge whether there are potential differences, on average, in growth trajectories.[9]

Figure 1 displays a set of empirical growth trajectories for infants in the Abecedarian Project study sample. The upper panel of Figure 1 is for all infants, irrespective of their treatment status. The lower panel displays two side-by-side plots. Plot A is for treatment group infants, while Plot B is for control group infants. Note that a linear relationship is fit through these points, given only three time points. More complex nonlinear trajectories could be specified with additional time points. In each of these panels, the thickest line represents the average of those trajectories.

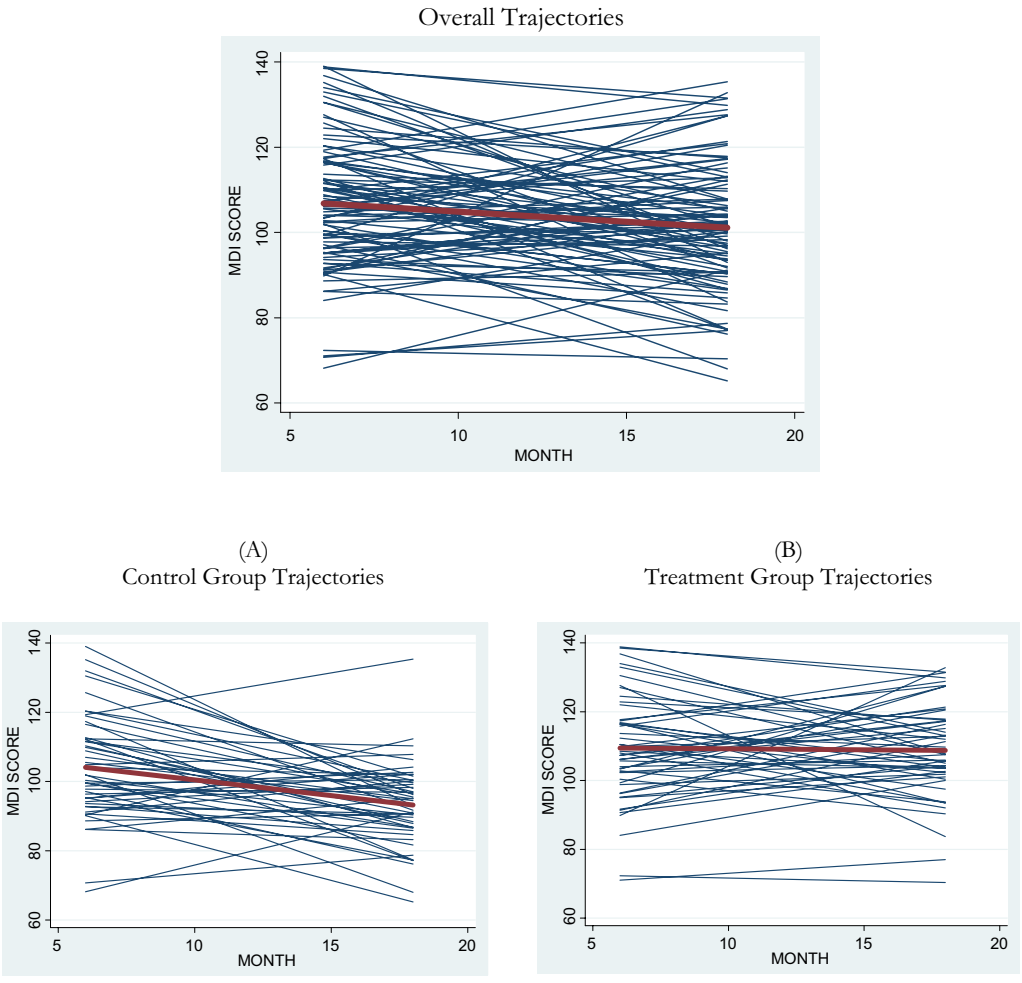

Figure 1. Empirical growth trajectories of Mental Development Index (MDI) scores for infants from the Abecedarian Project. The thickest line represents the average of the trajectories ($n = 111$).

The upper panel of Figure 1 shows substantial variation across infants in their MDI scores. Each infant starts[10]out with a different MDI score at 6 months of age and has a different MDI trajectory. For example, some infants start out relatively low and rapidly advance while others start out high and decline. In the lower panel of Figure 1, the side-by-side plots of these trajectories provide important visual cues foreshadowing how MDI scores might differ over time by treatment status. In Plot A, the average trajectory of MDI scores among treatment group infants remains relatively flat. Yet, in Plot B, the average trajectory of MDI scores for control group infants declines. These plots potentially signal that growth in MDI scores differ by treatment group. Given these visual guideposts, the next section describes how multilevel growth models are formally specified.

Specifying Multilevel Growth Models

To evaluate the impact of the Abecedarian child care program on differences in infants' MDI scores over time, we specify two models: (1) a level-1 model that addresses how each infant changes over

Table 4. A Two-level Multilevel Growth Model Describing the Hypothesized Change in MDI Scores over Time for Infants in the Abecedarian Project by Treatment Status.

Model	Equation	
Level-1 model (within-person)	$MDI_{ij} = \alpha_{0i} + \alpha_{1i}(MONTH_{ij}) + \varepsilon_{ij}$	(1)
Level-2 model (person-level)	$\alpha_{0i} = \beta_{00} + \beta_{01}(TREAT_i) + \mu_{0i}$	(2a)
	$\alpha_{1i} = \beta_{10} + \beta_{11}(TREAT_i) + \mu_{1i}$	(2b)

Note. The level-1 growth parameters (α_{0i} and α_{1i}) are outcomes at level 2.

time and (2) a level-2 model that addresses how growth in MDI scores differs between individual infants by treatment group. The level-1 model is often referred to as a *within-person* model since it models changes over time in the outcome *within* each individual. The level-2 model is often referred to as a *person-level* (or *between-person*) model since it examines how these individual changes over time vary *between* individuals. These models are specified in Table 4.

The Level-1 Model

The level-1 model posits that a given infant *i*'s MDI score on each occasion *j* (MDI_{ij}) is a linear function of the month ($MONTH_{ij}$) at which infant *i* was assessed plus individual error (ε_{ij}). In this model, there are two *individual growth parameters* (Singer & Willett, 2003, p. 51). The first growth parameter (α_{0i}) represents a given infant's initial MDI score at baseline when MONTH = 0.[11] The second growth parameter (α_{1i}) represents the monthly rate of change in MDI scores for infant *i*.[12] Finally, the error term ε_{ij} represents all factors other than time influencing infant *i*'s MDI score on occasion *j*. By convention, we assume these errors are normally distributed and have a mean of zero with a constant variance σ^2_{ε}. Constant variance, or *homoscedasticity*, means that the variability of this error term remains constant at each level of the predictor in the model. Figure 2 illustrates the level-1 error term for a given infant *i*.

Given that this level-1 model focuses on individual growth, it provides only one part of the story about growth in infants' cognitive and linguistic development. Looking across infants, we know that infants will have different individual initial statuses (α_{0i}) as well as individual growth rates (α_{1i}). Thus, we want to model how each of these individual growth parameters (1) vary across individual infants and (2) are predicted by characteristics, such as their treatment status, that vary from infant to infant but are constant across time. To accomplish this, we need to specify a level-2 model.

The Level-2 Model

For the level-2 model, we specify as many equations as there are level-1 individual growth parameters. Given two individual growth parameters (α_{0i} and α_{1i}) at level 1, we posit two separate equations for the level-2 model. Importantly, as shown by the boxes and arrows overlaying Equations 1, 2a, and 2b in Table 4, the level-1 individual growth parameters, α_{0i} and α_{1i}, serve as outcome variables in the level-2 model.

The first Equation 2a of the level-2 model expresses the relationship between the first growth parameter (initial status, α_{0i}) and treatment status. Parameters β_{00} and β_{01} (commonly referred to as *fixed effects*) represent the mean initial MDI score for control group infants (when $TREAT_i$ = 0) and the effect of $TREAT_i$ on mean initial MDI score, respectively. Finally, the error term in

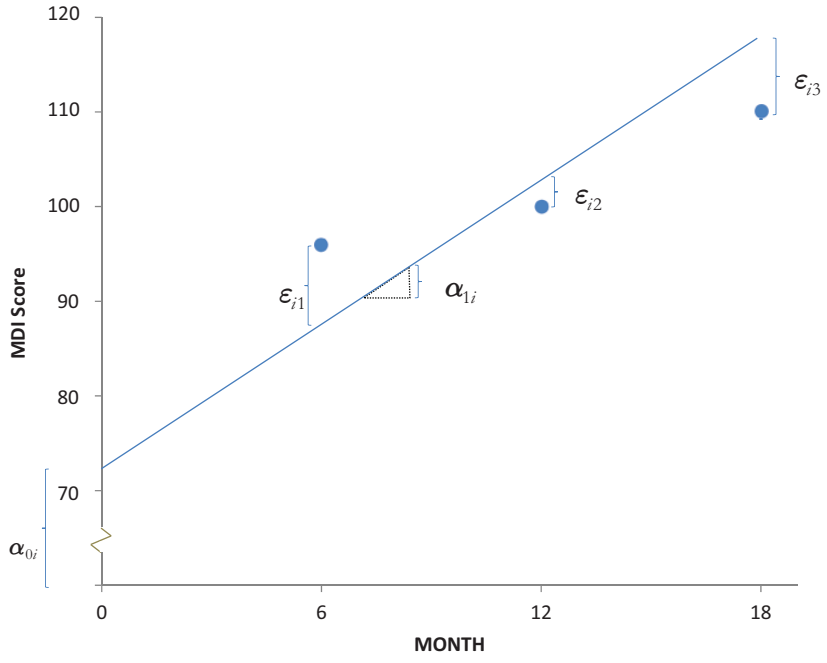

Figure 2. Illustration of the level-1 model error term, ε_{ij}. ε_{ij} is the vertical deviation between infant *i*'s observed Mental Development Index (MDI) score on each occasion *j* and his or her MDI growth trajectory. α_{0i} is the infant's initial status and α_{1i} is the growth rate.

Equation 2, μ_{0i} (known as a *random effect*) represents infant *i*'s deviation from β_{00} after controlling for assignment into the Abecedarian Program. Figure 3 shows μ_{0i} for a given infant *i* in the control group as an example. Technically, we want to estimate the *variance* of μ_{0i} (often denoted as τ_{00}). The variance of μ_{0i} captures the spread of these initial statuses and indicates how much each individual infant varies from each other on his or her initial statuses.

The second equation of the level-2 model (2b) is of primary importance as it determines whether the second growth parameter (growth rate, α_{1i}) varies by treatment status. In this equation, β_{10} is the average rate of change in MDI scores for control group infants. Importantly, β_{11} captures the relationship, on average, between an infant's growth rate (α_{1i}) and treatment status (indicated by $TREAT_i$)—this parameter will help answer our main impact evaluation question since it distinguishes how growth rates differ, on average, by treatment group. Visually, recall the empirical growth plots in the 2 lower panels of Figure 1 (Panels A and B) for control and treatment groups. You can think of β_{11} as capturing the difference, on average, in the slopes of these trajectories by treatment status.

Finally, μ_{1i}, captures infant *i*'s deviation from his or her population average growth trajectory, controlling for random assignment into child-centered care. Figure 3 shows, μ_{1i} for a given infant *i* in the control group. When fitting this model to data, we are interested in estimating the variance of, μ_{1i}, often denoted as τ_{11}. The variance tells us the extent of the dispersion of growth trajectories around an average population growth trajectory, conditional on assignment into the Abecedarian Program.

We also assume that both level-2 error terms, μ_{0i} and μ_{1i}, are bivariate normally distributed with a mean of 0 with constant variance. Importantly, we allow these two terms to covary. The covariance, often denoted as τ_{01}, allows us to determine the relationship between initial statuses and growth

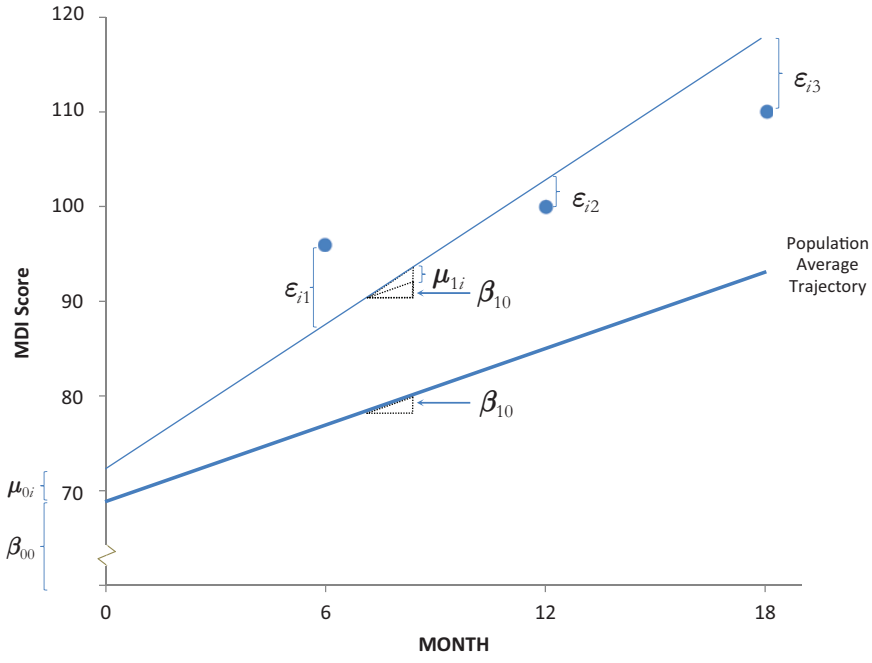

Figure 3. Illustration of the level-2 model error terms (μ_{0i} and μ_{1i}) for a given control group infant *i* (TREAT = 0). The thicker line is the population average trajectory for all control group infants while the thinner line is the trajectory for infant *i*. β_{00} and β_{10} are the population average initial status and population average growth trajectory for all control group infants, respectively. μ_{0i} is infant *i*'s deviation from β_{00} and μ_{1i} and is the deviation from β_{10}. ε_{ij} represents the vertical deviations between observed Mental Development Index (MDI) scores on each occasion *j* and infant *i*'s growth trajectory.

rates, conditional on assignment into the Abecedarian child care program. For example, we may think that infants who have relatively low initial MDI scores have higher rates of growth versus infants with higher initial MDI scores. A positive τ_{01} suggests that infants with relatively high initial MDI scores have more rapid growth rates while a negative τ_{01} suggests that infants starting with lower MDI scores have more rapid growth rates. τ_{01} can also be expressed as a correlation: $\tau_{01}/\sqrt{\tau_{00} \times \tau_{11}}$.

Model Assumptions

There are two key assumptions of the multilevel model error terms mentioned earlier that are important to highlight. First, we assume that the level-1 and level-2 error terms (ε_{ij}, μ_{0i}, and μ_{1i}) are normally distributed. Second, we assume that the variances of these error terms are homoscedastic. Violations of these assumptions may cast doubt on the validity of our findings. For example, if the level-1 error term is not normally distributed, standard errors can be biased at both levels (Raudenbush & Bryk, 2002, p. 266). This could lead to incorrect inferences about the statistical significance of program impact. As addressed in West, Welch, and Galecki (2007) and Singer and Willett (2003, pp. 127–132), there are ways to visually determine whether these assumptions hold true by producing plots that display estimates of these errors terms known as residuals.

Table 5. Explanation and Interpretation of Population Parameters in a Two-level Multilevel Growth Model (*i* indexes infant; *j* indexes occasion).

Model	Parameter	Explanation and Interpretation
Level-1 model (within-person)		
$MDI_{ij} = \alpha_{0i} + \alpha_{1i}(MONTH_{ij}) + \varepsilon_{ij}$	α_{0i}	The first individual growth parameter representing a given infant *i*'s initial MDI score at baseline when MONTH = 0.
	α_{1i}	The second individual growth parameter representing the monthly rate of change in MDI scores for infant *i*.
	ε_{ij}	All factors besides the effect of time that influence infant *i*'s MDI score on occasion *j*. The variance of ε_{ij} is often denoted as σ^2_{ε}.
Level-2 model (person-level)		
$\alpha_{0i} = \beta_{00} + \beta_{01}(TREAT_i) + \mu_{0i}$	β_{00}	The mean initial MDI score for control group infants (when $TREAT_i = 0$). Often referred to as a "fixed effect."
	β_{01}	The effect of $TREAT_i$ on mean initial MDI score. Often referred to as a "fixed effect."
	μ_{0i}	The unexplained portion of infants' initial statuses that remains after accounting for their assignment into the Abecedarian Program. Often referred to as a "random effect". The variance of μ_{0i} is often denoted as τ_{00}.
$\alpha_{1i} = \beta_{10} + \beta_{11}(TREAT_i) + \mu_{1i}$	β_{10}	The average rate of change in MDI scores for control group infants. Often referred to as a "fixed effect."
	β_{11}	The relationship, on average, between an infant's growth rate (α_{1i}) and treatment status (indicated by $TREAT_i$). Often referred to as a "fixed effect."
	μ_{1i}	The portion of the individual growth rate unique to each individual infant that remains unexplained after accounting for random assignment into the Abecedarian Program. Often referred to as a "random effect". The variance of this error term is often denoted as τ_{11}.

Summary: Specifying Multilevel Growth Models

To recap, establishing a basic linear multilevel growth model with repeated observations (three or more) on some continuous measure collected over time on individuals requires you to posit two models: (1) a level-1 within-person model; and (2) a level-2 person-level model.[13]

The Level-1 Model (Within Person). The level-1 model (Table 4; Equation 1) hypothesizes, for an individual, the relationship between each individual's outcome on each measurement occasion and the passage of time plus unobserved error. This relationship is specified using an equation comprising two individual growth parameters (1) initial status (the value of an individual's outcome at baseline) and (2) the effect of time on the outcome, representing an individual's growth over time. Finally, we include an unobserved individual-level error term that varies across each measurement occasion.

The Level-2 Model (Person Level). Given that there may be variability across individuals in their initial statuses and growth rates and, most importantly, we want to evaluate what determines that variability, we posit a level-2 model. This model consists of as many equations as there are level-1 growth parameters. If growth is linear, there are two growth parameters at level 1: initial status and rate of growth. Thus, we posit two equations each having one of the growth parameters as an outcome (Table 4;

Table 6. A Two-level Unconditional Means Model.

Model	Equation	
Level-1 model (within-person)	$MDI_{ij} = \alpha_{0i} + \varepsilon_{ij}$	(1)
Level-2 model (person-level)	$\alpha_{0i} = \beta_{00} + \mu_{0i}$	(2a)

Equations 2a and 2b). To assess how each growth parameter differs by particular characteristics, such as program participation, we include key predictors in each equation. Specifying the second equation of the level-2 model is of primary importance as it allows us to model how average growth trajectories differ according to characteristics such as treatment group assignment. Table 5 summarizes the key parameters of the level-1 and -2 models.

Specifying Two Diagnostic Models

Before fitting a full multilevel growth model to data, you should specify and fit two preliminary models: (1) an *unconditional means model* (also referred to as a *null model*; Garson, 2013) and (2) an *unconditional growth model* (Singer & Willett, 2003, p. 92). Results of these unconditional models provide important quantitative information that serve as guideposts signaling whether it is viable to proceed further in your analyses.

Unconditional Means Model

The unconditional means model allows you to understand (1) the extent to which there is significant variation within individuals (do their outcomes actually change over time?) and (2) whether there is significant variation between individuals in outcomes over time, thus establishing the feasibility of including predictors to help explain this variation.

As shown in Table 6, the model is specified without any predictors, only intercept terms at each level. In this model, we are mainly interested in the fixed-effect parameter β_{00} and the variance parameters for μ_{0i} and ε_{ij} (τ_{00} and σ^2_ε, respectively). β_{00} represents the overall mean MDI score across all measurement occasions and infants. τ_{00} represents the degree to which there is variation in MDI scores between infants, while σ^2_ε captures the degree to which there is variation in MDI scores within infants. Together, the between and within variation is the total variation ($\tau_{00} + \sigma^2_\varepsilon$) in the outcome. If estimates of both τ_{00} and σ^2_ε statistically significantly differ from zero, we can conclude that there is significant variability between and within infants in their MDI scores. This variability is important because it establishes the feasibility of including predictors in our subsequent level-1 and level-2 models. For example, if there was no variability in MDI scores within infants—hypothetically imagine that infants did not differ in their MDI scores on each occasion—then it would be impossible to include any additional predictors in the level-1 equation since there would be no variation to predict.

Estimates of τ_{00} and σ^2_ε from the unconditional means model can also be used to calculate the proportion of the total variation in MDI scores that exists between infants. This magnitude is known as the intraclass correlation coefficient (ICC)[14] and is often denoted by the Greek symbol ρ (Killip, Mahfoud, & Pearce, 2004). The ICC is the ratio of the variation in MDI scores that lies between infants to the total variation in MDI scores :$\tau_{00}/\tau_{00} + \sigma^2_\varepsilon$. The ICC ranges from 0 to 1, with values closer to 1 indicating that a higher proportion of the total variation in the outcome is attributable to differences between individuals. The ICC can also be interpreted in percentage form. For example,

Table 7. A Two-Level Unconditional Growth Model.

Model	Equation	
Level-1 model (within-person)	$MDI_{ij} = \alpha_{0i} + \alpha_{1i}(MONTH_{ij}) + \varepsilon_{ij}$	(1)
Level-2 model (person-level)	$\alpha_{0i} = \beta_{00} + \mu_{0i}$	(2a)
	$\alpha_{1i} = \beta_{10} + \mu_{1i}$	(2b)

an ICC of .25 means that 25% of the total variation in a particular outcome lies between individuals while the remaining 75% lies within individuals.

Unconditional Growth Model

A subsequent diagnostic model, the unconditional growth model, establishes whether MONTH is related to infants' MDI scores. This model is specified without any predictors at level 2 as shown in Table 7. In this model, β_{00} is the average MDI score for all infants at the beginning of the study (when MONTH = 0), while β_{10} is the average monthly rate of change in MDI across all infants. Importantly, we are interested in the estimates and statistical significance of the variance parameters for μ_{0i} and μ_{1i} (τ_{00} and τ_{11}, respectively). Estimates of τ_{00} and τ_{11} indicate whether there is variability in both infants' overall initial status in MDI scores and, most importantly, change over time in MDI scores. If there is significant variability especially in MDI growth rates, then including predictors at level 2 may help predict this variation.

Results of Multilevel Growth Models

Software for Fitting Multilevel Growth Models

There are several statistical software packages that can fit multilevel growth models including HLM, MLWiN, Mplus, R, SAS, SPSS, and Stata. Each software program has its own unique syntax as well as data handling and preparation requirements. In addition, there are numerous guides and reference books written specifically for each software package. A selected list of references is given in Table 8.

Diagnostic Models Results

Unconditional means model (Column 1, Table 9)

The fixed effect estimate.[15] The overall mean MDI score across all three measurement occasions and infants is $\hat{\beta}_{00} = 104.2$ points, and we can conclude that infants' cognitive and linguistic functioning from 6 to 18 months differs from zero.

The random effect estimates. Estimates of the variance parameters for the infant-specific residual ($\hat{\sigma}^2_{\varepsilon} = 137.90$) and mean MDI score ($\hat{\tau}_{00} = 104.86$) both differ from zero.[16] This tells us that infants do have different MDI scores across time and that infants differ from each other. The ICC, calculated as $104.86/(104.86 + 137.90) \approx 0.431$, indicates that approximately 43% of the variation in MDI scores is due to between-infant differences. This also suggests that including predictors in our level-2 model that differentiate one infant to another infant might be helpful in predicting this variation.

Table 8. Software and Recommended References for Fitting Multilevel Growth Models.

Software Package	Reference
HLM	Raudenbush, S. W. (2004). *HLM 6: Hierarchical linear and nonlinear modeling*. Lincolnwood, IL: Scientific Software International.
MLwiN	Rasbash, J., Browne, W., Goldstein, H., Yang, M., Plewis, I., Healy, M., ... Lewis, T. (2000). *A user's guide to MLwiN*. London, England: University of London, Institute of Education, Centre for Multilevel Modelling.
Mplus	Muthén, L. K., & Muthén, B. O. (1998–2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.
R	Bliese, P. (2013). *Multilevel modeling in R (2.5)*. Retrieved from http://cran.r-project.org/doc/contrib/Bliese_Multilevel.pdf
SAS	Singer, J. D. (2002). Fitting individual growth models using SAS PROC MIXED. In D. S. Moskowitz & S. L. Hershberger (Eds.), *Modeling intraindividual variability with repeated measures data: Methods and applications* (pp. 122–153). Mahwah, NJ: L. Erlbaum Associates.
SPSS	Heck, R. H., Thomas, S. L., & Tabata, L. N. (2013). *Multilevel and longitudinal modeling with IBM SPSS*. New York, NY: Routledge.
Stata	Rabe-Hesketh, S., & Skrondal, A. (2012). *Multilevel and longitudinal modeling using Stata*. College Station, TX: Stata Press.

Table 9. Results of Fitting Multilevel Growth Models to the Abecedarian Project Data.

	(1) Unconditional Means	(2) Unconditional Growth	(3) Full Model Results
Fixed effects			
Initial status ($\hat{\beta}_{00}$)	104.02*** (1.20)	109.20*** (2.04)	109.99*** (2.90)
Treatment ($\hat{\beta}_{01}$) Effect of treatment on initial status			−1.44 (4.09)
Month (rate of change) ($\hat{\beta}_{10}$)		−0.44** (0.14)	−0.89*** (0.19)
Treatment ($\hat{\beta}_{11}$) Effect of treatment on rate of change			0.90*** (0.27)
Random effects			
Level 1 (within infant)			
Temporal variation ($\hat{\sigma}^2_{\varepsilon}$)	137.90 (13.62)	117.74 (16.51)	119.80 (16.84)
Level 2 (between infant)			
Infant mean initial status ($\hat{\tau}_{00}$)	104.86 (21.27)	160.95 (71.34)	160.08 (72.21)
Infant mean rate of change ($\hat{\tau}_{11}$)		0.37 (0.36)	0.15 (0.35)
Covariance ($\hat{\tau}_{01}$)		−4.26 (4.63)	−3.95 (4.58)

Note. Models fit using Stata's xtmixed command. Statistical significance of the random effects not reported because the standard tests of significance (e.g., Wald statistics) for such effects are not entirely reliable (see Hedeker & Gibbons, 2006, p. 52). $\alpha = .05$ ($n = 111$).
$\sim p < .10$. *$p < .05$. **$p < .01$. *** $p < .001$.

Unconditional growth model (Column 2, Table 9)

The fixed effect estimates. $\hat{\beta}_{00}$ tells us that infants initially have a mean MDI score of 109.20. $\hat{\beta}_{10}$ is −0.44 indicating that infants' MDI scores significantly decline by about 0.44 points per month irrespective of their treatment status. Figure 4 displays this fitted linear growth trajectory.

The random effect estimates. The variance in mean initial status ($\hat{\tau}_{00} = 160.95$) tells us that infants significantly vary on their initial MDI scores. However, the variance in the rate of change

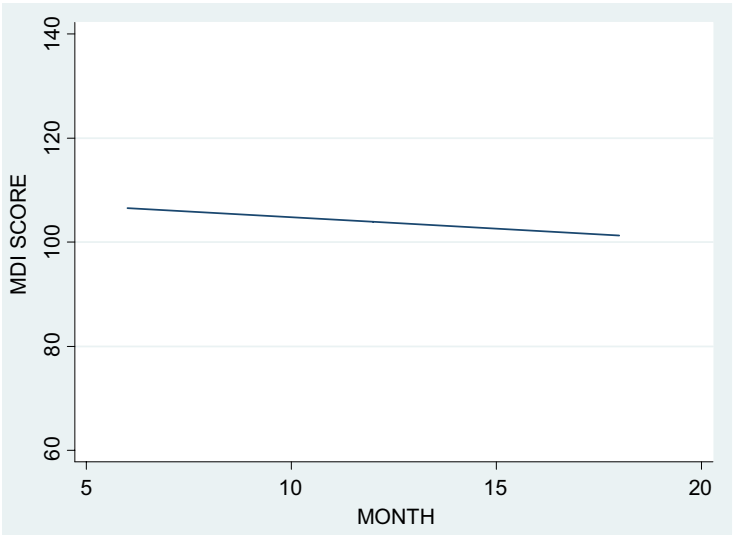

Figure 4. Predicted mean growth rate in Mental Development Index (MDI) scores for infants from the Abecedarian Project ($n = 111$).

($\hat{\tau}_{11} = 0.37$) indicates that infants' growth rates do not significantly vary. Despite this, adding in a predictor at level 2 (e.g. random assignment) could still be helpful in predicting infants' MDI growth rates.[17] Finally, the estimated covariance $\hat{\tau}_{10} = -4.26$ (expressed as a correlation, $-4.26/\sqrt{160.95 \times 0.37} \approx -0.55$) is negative indicating that children with low initial MDI scores tend to have faster growth rates; however, it is not statistically significant.

Full model results (Column 3, Table 9)

The fixed effect estimates. The mean initial MDI score ($\hat{\beta}_{00}$) is 110 points and a nonsignificant $\hat{\beta}_{01}$ tells us that these initial scores do not differ by treatment status. This is what we would expect, given that random assignment has created a treatment and a control group that are initially equivalent in expectation. $\hat{\beta}_{10}$ is $-.89$ indicating that MDI scores for control group infants declined by 0.89 points per month. Finally, the primary estimate of interest in the full model, $\hat{\beta}_{11}$, tells us that the MDI growth rate between the treatment and the control group significantly differs, on average, by 0.90 points per month.

The random effect estimates. Note that the estimated variance of the mean rate of change ($\hat{\tau}_{11}$) is lower in the full model versus the unconditional growth model (0.15 vs. 0.37, a decline of approximately 60%). Although we cannot reject the null that $\hat{\tau}_{11}$ is zero, this decline is what we would expect, given that TREAT_i helps predict variation in MDI growth trajectories.

Interpreting and displaying the full model results

To interpret the full model, we can derive equations describing the control and treatment group growth trajectories. First, we substitute estimates from column 3 of Table 4 into the level-2 equations:

$$\hat{\alpha}_{0i} = 109.99 - 1.44(\text{TREAT}_i) \tag{1}$$

$$\hat{\alpha}_{1i} = -0.89 + 0.90(\text{TREAT}_i) \tag{2}$$

Table 10. Predicted Mental Development Index (MDI) Scores by Month for Control Group and Treatment Group Infants From the Abecedarian Project.

Month	Control	Treatment
6	104.65	108.61
12	99.31	108.67
18	93.97	108.73

Note. Predicted Scores are Based on a Fitted Multilevel (2-level) Growth Model (column 3 of Table 9). $n = 111$.

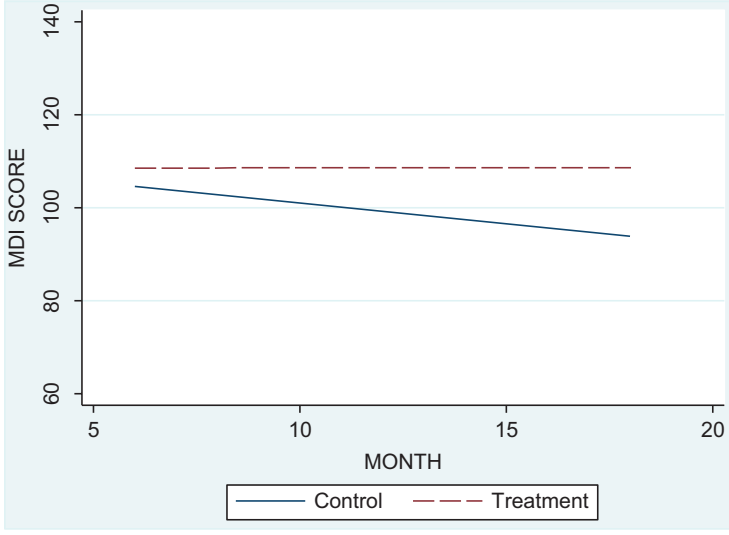

Figure 5. Predicted mean growth rates in Mental Development Index (MDI) scores for infants from the Abecedarian Project by control and treatment status ($n = 111$).

Given that the individual growth parameters, α_{0i} and α_{1i}, are the intercept and slope of the level-1 model, we substitute Equations 1 and 2 into those terms, respectively, and simplify:

$$\hat{\text{MDI}}_{ij} = 109.99 - 1.44(\text{TREAT}_i) - .89(\text{MONTH}_{ij}) + .90(\text{TREAT}_i \times \text{MONTH}_{ij}) \quad (3)$$

Then, we develop a fitted model for the treatment group by setting $\text{TREAT}_i = 1$ in Equation 3 which simplifies to:

$$\hat{\text{MDI}}_{ij} = 108.55 + .01(\text{MONTH}_{ij}) \quad (4)$$

Similarly, we derive a fitted model for control group infants by setting $\text{TREAT}_i = 0$ in Equation 3:

$$\hat{\text{MDI}}_{ij} = 109.99 - .89(\text{MONTH}_{ij}) \quad (5)$$

In Equations 4 and 5, the slope coefficients on the variable MONTH_{ij}, .01 and −.89, are the estimated changes in MDI scores per month for treatment and control group infants, respectively. Treatment group infants gain .01 points per month, while control group infants decline .89 points per month. The difference (.01 − (−.89)) is .90 points per month. To display predicted growth

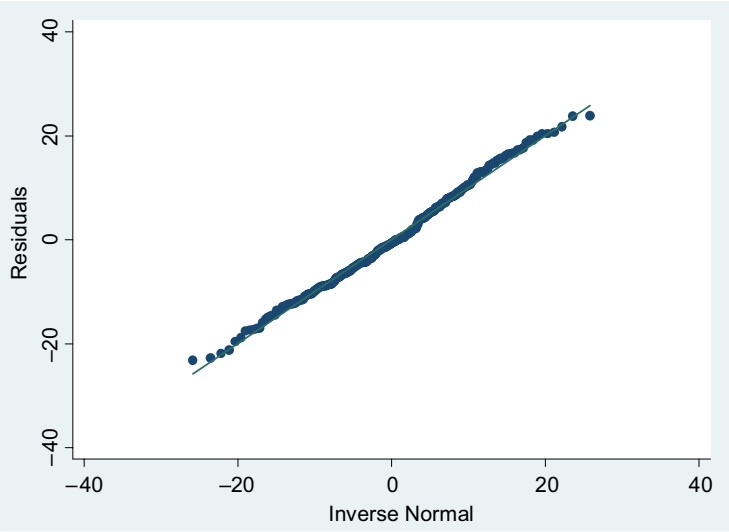

Figure 6. Quantile–Quantile (Q-Q) plot for the level-1 residuals.

trajectories for the control and the treatment groups, you can insert values for the months at which MDI scores were measured (6, 12, and 18) into Equations 4 and 5 and then plot these values in a program such as Microsoft Excel. Table 10 summarizes the predicted average MDI scores at months 6, 12, and 18 separately for control and treatment group infants. Figure 5 displays these two trajectories. Note the clear distinction between MDI growth rates in the control group (a noticeable decline) versus in the treated group (a slightly upward tilting trajectory).

Thus, we now have an answer to our primary impact evaluation question. By fitting a multilevel growth model to the Abecedarian Project data, we find that random assignment into child-centered care causes cognitive and linguistic growth rates to differ, on average, by .90 points per month.[18] From age 6 to 18 months, infants randomly assigned into child-centered care gained .01 points per month in their cognitive and linguistic functioning (as measured by MDI scores) while control group infants declined by .89 points every month.

Testing the model assumptions: Normality and homoscedasticity

After fitting the full model, it is important to check for normality and homoscedasticity in the level-1 and -2 error terms. We do this by visually inspecting the estimates of these error terms, or residuals, which are derived from the final fitted model.

Level-1 residuals. To test for normality of our level-1 residuals (the observed values of our outcome minus the predicted values), we can produce a Quantile–Quantile (Q-Q) as shown in Figure 6. In these plots, we determine how well the points, which represent the model's residuals, line up against a normal distribution line. If the points deviate from the normal distribution line, this suggests a departure from the normality assumption. Figure 6 shows that the points closely line up with the normal distribution line. This visually signals that the assumption of normality is satisfied. To visually test for the assumption of homoscedasticity, we can plot the model's residuals versus the model's predicted values of the outcome and examine whether the vertical spread of the residuals at each predicted value remains consistent. This plot, shown in Figure 7, shows no discernible pattern; it resembles a random point cloud and therefore we can conclude that the level-1 residuals are homoscedastic.

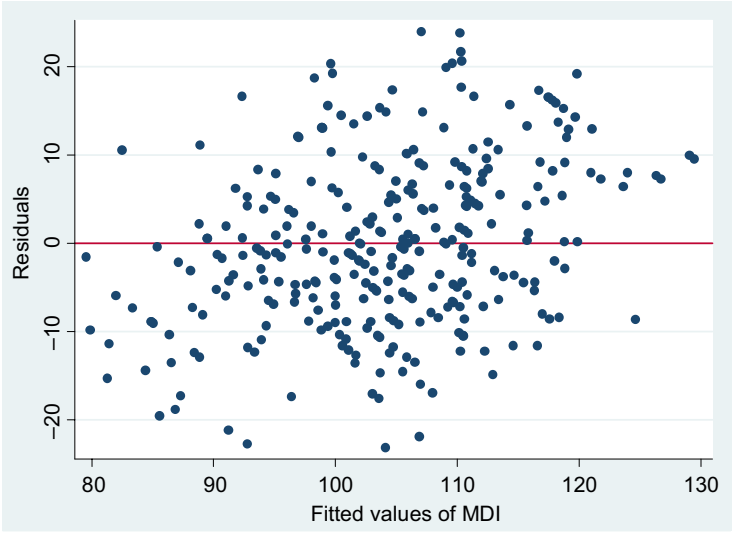

Figure 7. Plot of the level-1 residuals against the model's predicted values of the outcome (fitted Mental Development Index [MDI] scores).

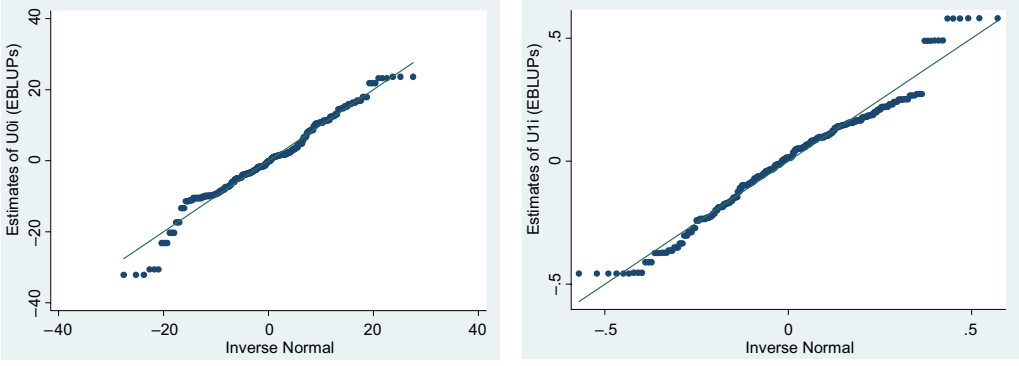

Figure 8. Quantile–Quantile (Q-Q) plots for the level-2 residuals (estimates of μ_{0i} appear in the left panel, while estimates of μ_{1i} appear in the right panel).

Level-2 residuals. Figure 8 shows the Q-Q plots for each of the level-2 residuals.[19] As shown, although some of the residuals align fairly closely with the normal distribution line, the points begin to trail off the normal distribution line, particularly at the extreme ends.[20] Thus, there is some evidence that the normality assumption for both these errors may not necessarily hold. However, as noted by Raudenbush and Bryk (2002, p. 274), departures from normality for the level-2 residuals will not necessarily bias the fixed effects—importantly, this will not bias our main impact estimate, $\hat{\beta}_{11}$. To test for homoscedasticity, we can examine the vertical spread of each of the level-2 residuals against values of the level-2 predictor in our model, TREAT_i. Figure 9 shows the plot of each level-2 residual versus TREAT_i. As shown, there is reasonable similarity in the distribution of the residuals at each value of TREAT_i with a slightly wider range when TREAT_i takes on the value of 1 versus 0. Based on this visual evidence, we do conclude that the level-2 error terms are homoscedastic.

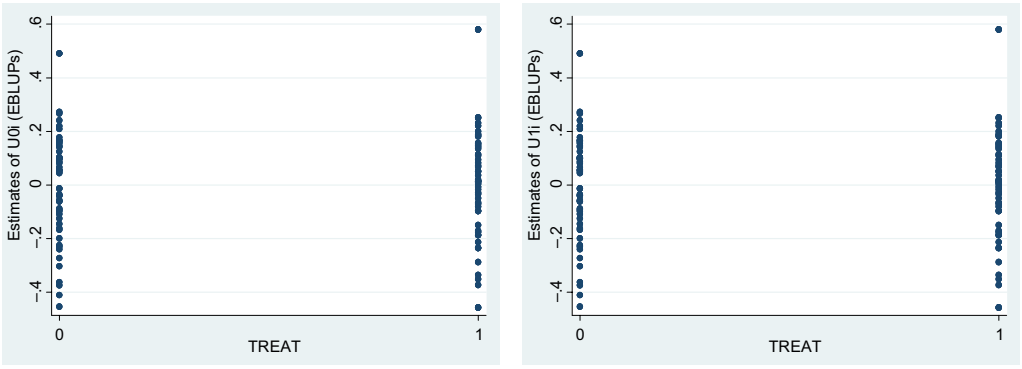

Figure 9. Plots of the level-2 residuals against values of the level-2 predictor, TREAT (estimates of μ_{0i} appear in the left panel while estimates of μ_{1i} appear in the right panel).

Conclusion

The increasing popularity of longitudinal evaluations that are designed to track outcomes over time presents program evaluators with new opportunities to move beyond evaluations that focus solely on a single end point in time. Longitudinal designs enhance evaluators' ability to understand how outcomes change over time as well as how this change can differ for particular groups (e.g., those who participated in a program versus those who did not). However, the methodological tools to analyze longitudinal data may not be part of an evaluator's current methodological tool kit. To help evaluators make use of longitudinal data to understand changes over time and to help readers of the evaluation literature to become more critical consumers of longitudinal studies, I have provided an applied introduction to one increasingly popular method of analyzing longitudinal data known as *multilevel growth modeling*. Specifically, I have illustrated how to organize and visually display data on growth over time; how to specify multilevel growth models; and finally, how to interpret and display the results. Readers seeking a deeper understanding of multilevel methods beyond the introductory material covered in this article should seek out the works listed in references of this article to more thoroughly understand and realize the power and potential of multilevel growth modeling in their own evaluation work.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. Readers should note that *multilevel growth models* may also be referred to as mixed-effects regression models; growth curve models; hierarchical linear models; growth models; linear mixed-effects models; random coefficient models, random-effects models, and mixed models.
2. A search was conducted of key peer-reviewed evaluation journals including the *American Journal of Evaluation*; *Evaluation Review: A Journal of Applied Social Research*; *Evaluation: The International Journal of Applied Social Research*; and *Evaluation and Program Planning*.

3. A comprehensive list can be found on the American Evaluation Association website: http://www.eval.org/p/cm/ld/fid=43
4. I use the public use version of data from the Abecedarian Project available through the Interuniversity Consortium for Political and Social Research (ICPSR; Ramey et al., 2004).
5. Outcomes can also be dichotomous, ordinal, nominal, or count data (Hedeker & Gibbons, 2006); however, the multilevel modeling procedures are more complex and beyond the scope of this article. Readers should consult the work of Hedeker and Gibbons (2006) for further details on multilevel growth models with non-continuous outcomes.
6. For this example, I have used only the observations from the Abecedarian Project data set that were collected on a consistent and regularly spaced (6 months) schedule. Multilevel growth modeling can also accommodate irregularly spaced collection schedules for each individual (see Singer and Willet, 2003, pp. 139–46).
7. The Institute for Digital Research and Education (IDRE) at the University of California, Los Angeles has excellent step-by-step tutorials on how to transform data from cross-sectional to panel format. Readers should refer to the following commands and accompanying online tutorials for their selected statistical software package (the links provided are applicable at the time of the writing of this article): For Stata users, the **reshape** command: http://www.ats.ucla.edu/stat/stata/modules/reshapel.htm. For SPSS users, the **vartocases** command: http://www.ats.ucla.edu/stat/spss/modules/reshapel115.htm. For SAS users, the **proc transpose** statement: http://www.ats.ucla.edu/stat/sas/modules/wtol_transpose.htm
8. These are also referred to as *time plots* (Fitzmaurice et al., 2004, p. 62).
9. Producing these empirical growth plots with longitudinal data can be accomplished in several popular software packages including Stata, SPSS, and SAS. Examples of how to create these plots can be found on the Statistical Computing site of IDRE website: http://www.ats.ucla.edu/stat/
10. In the actual Abecedarian Project, MDI measurements were initially taken when infants were 3 months old, but for illustrative purposes, I have chosen to use infants' Mental Development Index (MDI) scores at 6 months as a starting point.
11. To more easily interpret the first growth parameter, initial status, the values of the predictor MONTH could be centered so that initial status is an individual's true MDI score at 6 months old rather that at 0 months old. To center the predictor MONTH, 6 is subtracted from the values of the variable MONTH (thus, MONTH would be recorded as 0, 6, and 12); however, for illustrative purposes, I have chosen to leave the value of MONTH in its original uncentered form. For more information about centering, see Raudenbush and Bryk (2002) pp. 33–35.
12. The subscripts 0 and 1 index each of the parameters.
13. If individuals were nested in higher order units such as patients in hospitals, employees in office sites, or students in classrooms, we would need to posit a three-level model. For an example of a three-level multilevel growth model, see Raudenbush and Bryk (2002, pp. 237–245).
14. Technically, this is known as an *unconditional* intraclass correlation (Rabe-Hesketh & Skrondal, 2008, p. 97) since there are no predictors included in the model.
15. By convention, the symbol "^" (called a *hat*) is placed above the model parameter to indicate that the parameter has been estimated from sample data.
16. The statistical significance of the random effects was assessed using a single parameter test (Raudenbush & Bryk, 2002, p. 63; Singer and Willett, 2003, p. 73) by dividing the estimate by its standard error to obtain a *z*-statistic. However, given general disagreement over the appropriateness of this test, readers should also consider using deviance statistics to examine the statistical significance of the random effects as illustrated in Singer and Willett (2003, pp. 116–122).
17. Technically, we could specify a model with a nonrandomly varying slope (Raudenbush & Bryk, 2002, p. 28) since we would omit μ_{1i} in the second equation of the level-2 model. Also, despite finding no slope variation, adding in level-2 predictors can enhance the statistical power to detect differences in the slope variation (Muthén, 2013).

18. Note that technically we are estimating what is known as the intent to treat effect of child-centered care which is the impact of a *randomized offer* to participate in the program, not actual participation in program itself.
19. Technically, these are the empirical best linear unbiased predictors for each of the level-2 random effects.
20. In addition, also note that residuals that deviate from the ends of the normal distribution line indicate that we should check for the possible influence of outliers in our data and consider conducting sensitivity analyses by refitting our models with these outliers removed.

References

Campbell, F. A., & Ramey, C. T. (1995). Cognitive and school outcomes for high-risk African-American students at middle adolescence: Positive effects of early intervention. *American Educational Research Journal*, *32*, 743–772.

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2004). *Applied longitudinal analysis*. Hoboken, NJ: John Wiley & Sons.

Garson, G. D. (2013). *Hierarchical linear modeling: Guide and applications*. Thousand Oaks, CA: Sage.

Hedeker, D., & Gibbons, R. D. (2006). *Longitudinal data analysis*. Hoboken, NJ: John Wiley & Sons.

Killip, S., Mahfoud, Z., & Pearce, K. (2004). What is an intracluster correlation coefficient? Crucial concepts for primary care researchers. *The Annals of Family Medicine*, *2*, 204–208.

Muthén, B. O. (2013). *Mplus discussion*. Retrieved from http://www.statmodel.com/discussion/messages/14/228.html?1378745081

Rabe-Hesketh, S., & Skrondal, A. (2008). *Multilevel and longitudinal modelling using Stata*. College Station, TX: Stata Press.

Ramey, C. T., Gallagher, J. J., Campbell, F. A., Wasik, B. H., & Sparling, J. J. (2004). *Carolina Abecedarian Project and the Carolina Approach to Responsive Education (CARE)*, *1972–1992*. Retrieved from: http://doi.org/10.3886/ICPSR04091.v1

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

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford, England: Oxford University Press.

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

West, B., Welch, K. B., & Galecki, A. T. (2007). *Linear mixed models: A practical guide using statistical software*. Boca Raton, FL: Chapman & Hall/CRC Press.