INTERNATIONAL REVIEW OF SOCIAL PSYCHOLOGY

Sommet, N., & Morselli, D. (2021). Keep Calm and Learn Multilevel Linear Modeling: A Three-Step Procedure Using SPSS, Stata, R, and Mplus. *International Review of Social Psychology*, 34(1): 24, 1–19. DOI: https://doi.org/10.5334/irsp.555

RESEARCH ARTICLE

Keep Calm and Learn Multilevel Linear Modeling: A Three-Step Procedure Using SPSS, Stata, R, and Mplus

Nicolas Sommet and Davide Morselli

This piece is meant to help you understand and master two-level linear modeling in an accessible, swift, and fun way (while being based on rigorous and up-to-date research). It is divided into four parts:

- PART 1 presents the three key principles of two-level linear modeling.
- PART 2 presents a three-step procedure for conducting two-level linear modeling using SPSS, Stata, R, or Mplus (from centering variables to interpreting the cross-level interactions).
- PART 3 presents the results from a series of simulations comparing the performances of SPSS, Stata, R, and Mplus.
- PART 4 gives a Q&A addressing multilevel modeling issues pertaining to statistical power, effect sizes, complex design, and nonlinear two-level regression.

The empirical example used in this tutorial is based on genuine data pertaining to '90s and post-'00s boy band member hotness and Instagram popularity. In reading this paper, you will have the opportunity to win a signed picture of Justin Timberlake.

Keywords: linear regression; multilevel modeling; grand- and cluster-mean centering; slope residual variance and covariance terms; cross-level interaction; confidence intervals; simulations; three-step procedure; Justin Timberlake

It's a freaking bad day. You've spent countless hours on the Internet trying to figure out how multilevel modeling works, but the only things you can find are academic papers filled with jargon, obscure equations, and indecipherable lines of code. 'Why can't I understand anything about stats?!' you ask yourself. Well, you've got to cool it now! Learning multilevel modeling can be a real bear, and this paper is *precisely* made for you to get the hang of it as easily as possible.

If you're here, you probably already know that the general aim of multilevel modeling is to simultaneously analyze data at a lower level (usually participants) *and* at a higher level (usually clusters of participants). In other words, multilevel modeling enables one to disentangle the effects of lower-level variables (e.g., individual effects) from the effects of higher-level variables (e.g., contextual effects) and examine how lower-level and higher-level variables interact with one another (interactions involving variables at different levels are called 'cross-level interactions').

Let us give you an example. In early 2000, a New Zealand team of scientists conducted research involving approximately 700 cats from 200 households (i.e., on average, 3.5 cats per household; Allan et al., 2000). The team treated the cats (level-1 units) as nested in households (level-2 units) and used multilevel modeling to disentangle the effects of level-1 cat variables (e.g., does the cat have long legs?) from the effects of level-2 household variables (e.g., is there a dog in the household?) in predicting cat obesity. They found that short-legged cats living in dog-free households tend to be chubbier.[1]

After reading the present paper, you will be able to handle this kind of (feline) two-level hierarchical design. Our paper is divided into four parts:

- PART 1 presents the three key principles of two-level linear modeling (two levels mean two types of residuals, predictors, and level-1 effect parameters).
- PART 2 presents a ready-to-use three-step procedure for conducting two-level linear modeling using SPSS, Stata, R, or Mplus.
- PART 3 presents the results from a series of simulations comparing the performances of the above statistical software.
- PART 4 gives a Q&A addressing multilevel modeling issues pertaining to statistical power (Q1), effect sizes (Q2), complex design (Q3), and nonlinear regression (Q4).

Swiss National Centre of Competence in Research LIVES, University of Lausanne, CH

Corresponding author: Nicolas Sommet (nicolas.sommet@unil.ch)

The empirical example used in the present tutorial is based on genuine data pertaining to '90s and post-'00s boy band member hotness and Instagram popularity.[2] We published the findings in various predatory journal using fake names (e.g., Abelkermit et al., 2021a, 2021b, 2021c). Our mock paper actually offered a good example of how to report multilevel analyses. The mock paper, the boy band dataset, and the software-specific instructions and scripts to perform our three-step procedure are available on the OSF (https://osf.io/4yhbm/ DOI: 10.17605/OSF.IO/4YHBM).

...Oh and yeah, we've hidden the names of twelve boy-band songs that reached the top quartile of the U.S. Billboard chart (including the best song ever from *NSYNC). The first reader to send the corresponding author the ten correct names will receive a signed picture of Justin Timberlake (displayed in **Figure 1**). The editor and reviewers were not allowed to take part in this competition.

PART 1. The Three Key Principles of Two-Level Linear Modeling

The Aim of This Part Is for You to Understand How Two-Level Modeling Works

A Very Brief Recap on Linear Regression

Imagine you conduct a study on the popularity of the best-selling '90s and post-'00s boy band leaders. You spend your day patiently gathering the number of Instagram followers of each of these boy band leaders and used a continuous scale ranging from 1 = *not popular* (≤100 Instagram followers) to 7 = *Beyoncé popular* (100,000,000 followers).[3]

At the end of the day, you have a dataset of $N = 50$ boy band leaders that you intend to analyze using regression. Regression can be thought of as a tool for describing data using an object named 'MODEL.' Obviously, we're only social scientists, and we can only expect our models to explain so much of the real world; in other words, our models can never be perfectly accurate, and the amount

Figure 1: The 8 × 10 hand signed photo of the great Justin Timberlake that YOU could win (upper panel), along with its numbered hologram Certificate of Authenticity (lower panel).

by which a model fails to properly represent the data is referred to as 'RESIDUALS' (Judd et al., 2017).

As such, the 'E = MC²' of data analysis is:

$$\text{DATA} = \text{MODEL} + \text{RESIDUALS} \qquad \text{(Eq. 1)}$$

Importantly, all regression equations have the same format as the above equation. In your case, the simplest regression equation you can use to describe your boy band leader data is a regression with no predictor and where the constant is the mean:

$$Y_i = B_0 + e_i \qquad \text{(Eq. 2)}$$

To make sense of Equation 2, take a look at **Figure 2**:

$Y_i \rightarrow$ Each circle in **Figure 2** represents the observed popularity score Y_i of a particular boy band leader i ('DATA' in Eq. 1). For instance, Justin Timberlake (*NSYNC) has a popularity score of $Y_1 = 6.75$.

$B_0 \rightarrow$ The horizontal thick line in **Figure 2** represents the mean popularity score B_0 ('MODEL' in Equation 1). You can see that the mean popularity score for *all* boy band leaders is $B_0 = 4.75$ (your very simple model). Note that B_0 is called the 'intercept' when the regression equation includes a predictor.

$e_i \rightarrow$ The vertical dotted lines in **Figure 2** represent the residuals e_i associated with each boy band leader i ('RESIDUALS' in Equation 1). These correspond to the distance of the observed popularity score Y_i of boy band leader i from the mean popularity score B_0 (i.e., the distance from the model). For instance, you can see that the observed score of Justin Timberlake ($Y_1 = 6.75$) is not properly described by the mean ($B_0 = 4.75$), and that the magnitude of the error is $e_1 = Y_1 - B_0 = 2.00$.

In simple ordinary least square linear regression, the aggregate of the residuals is *the variance of the residuals*, written as var(e_i). It is calculated by taking the mean of the squared residuals: $\text{var}(e_i) = (e_1^2 + e_2^2 + \ldots + e_N^2)/N$ (the mean of the squared distance of Justin Timberlake, Joey Zehr, Nick Jonas, etc., from the model). In a nutshell, the general goal of regression is to estimate whether making your model more complex by adding a particular predictor will lead to a reduction in the unexplained remaining variations of your outcome. For instance, you could include the level of boy band leader hotness X_i as an additional predictor in your equation ($Y_i = B_0 + B_1 \times X_i + e_i$) and see whether doing so leads to a significant reduction in var(e_i). In the context of null hypothesis testing, and assuming that you formulated a prediction for this variable, this reduction would entail accepting the hypothesis that boy band leader hotness is associated with boy band leader popularity.

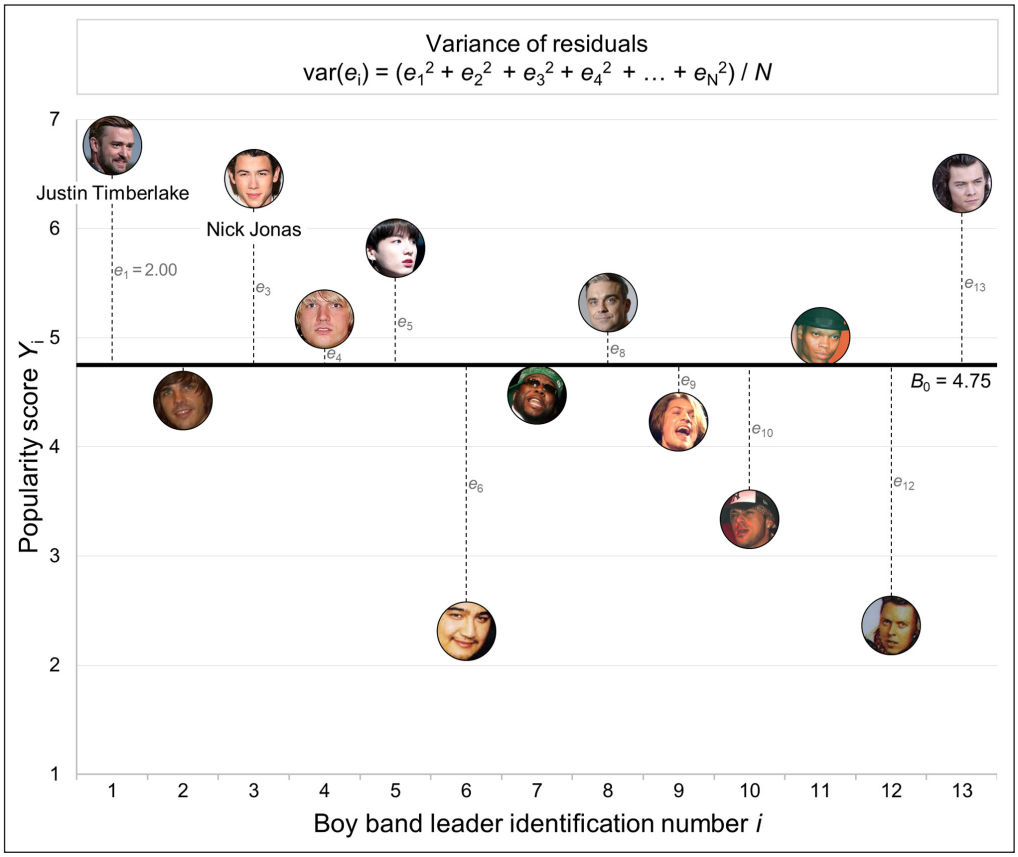


Figure 2: Graphical representation of a linear regression with no predictor (Eq. 2) in which the observed popularity score Y_i (y-axis) of a particular boy band leader i (x-axis) corresponds to the mean popularity score B_0 (horizontal thick line) plus the residuals e_i (vertical dotted lines). *Notes*: Only the first 13 observations are represented; despite the controversial fact that there is no official leader in One Direction, we treated Harry Styles as their boy band leader.

The First Principle of Two-Level Modeling: Two Levels Mean Two Types of Residuals

Now imagine you conduct a study on the popularity of *all* members of the best-selling '90s and post-'00s boy bands (not only the boy band leaders). You spend another day gathering the number of Instagram followers of each of these new members, and you end up with a dataset of $N = 175$ boy band members. The structure of your dataset is different than before. You now have a two-level hierarchically structured dataset with two types of units: $N = 175$ members (level-1 units) nested in $K = 50$ boy bands (level-2 units or clusters; i.e., a mean cluster size of $n = 3.50$ members per boy band).

In this situation you cannot use traditional regression, because it would violate the assumption of independence of the residuals, that is, the basic assumption that the residual associated with a given data point is independent of the residual of another data point (Snijder & Bosker, 1999; the two other basic assumptions are normality and homoscedasticity). Specifically, in your dataset it is easy to understand that members of similar boy bands are likely to share similar levels of popularity; thus, the residuals associated with any two members of the *same* boy band will be closer than the residuals associated with any two members of *different* boy bands. If you choose to ignore this problem and use traditional regression, you will most certainly obtain biased standard errors, which will result in false-positive or false-negative findings (depending on the nature of nonindependence; Scariano & Davenport, 1987).

In this situation, you therefore need to use two-level linear regression. Like traditional regression, two-level regression aims to describe data using an object named 'MODEL.' Similar to traditional regression, the amount by which such a model fails to properly represent the data is referred to as 'RESIDUALS.' However, this time there are *two types of residuals* (Hox, 2017): (i) the amount by which the model fails to properly represent the *between*-cluster variations is referred to as 'LEVEL-2 RESIDUALS' and (ii) the amount by which the model fails to properly represent the *within*-cluster variations is referred to as 'LEVEL-1 RESIDUALS.'

As such, the 'E = MC²' of two-level modeling is:

$$\text{DATA} = \text{MODEL} + \text{LEVEL} - 2\ \text{RESIDUALS} + \text{LEVEL} - 1\ \text{RESIDUALS} \quad \text{(Eq. 3)}$$

Importantly, all two-level regression equations have the same format as the above equation. In your case, the simplest two-level linear regression equation you can use to describe your boy band data is a regression with no predictor and where the constant is the overall mean:

$$Y_{ij} = B_{00} + u_{0j} + e_{ij} \quad \text{(Eq. 4)}$$

To make sense of Equation 4, take a look at **Figure 3**:

$Y_{ij} \rightarrow$ Each circle in **Figure 3** represents the observed popularity score Y_{ij} of a particular boy band member *i* from a particular boy band *j* ('DATA' in Equation 3). For instance, Justin Timberlake (*NSYNC) has an actual popularity score of $Y_{41} = 6.75$, his buddy Lance Bass has a score of $Y_{51} = 4.73$, and Kevin Jonas (The Jonas Brother) has a score of $Y_{13} = 6.45$.

$B_{00} \rightarrow$ The horizontal thick line in **Figure 3** represents the overall mean popularity score B_{00}, regardless of clustering ('MODEL' in Equation 3). You can see that the mean popularity score for *all* boy band members and across *all* boy bands is $B_{00} = 3.46$ (your very simple model). Note that B_{00} is also called the 'fixed intercept' when the two-level regression equation includes a predictor.

$u_{0j} \rightarrow$ The vertical thick dotted lines in **Figure 3** represent the level-2 residuals u_{0j} (also called intercept residuals or 'random intercept') associated with each boy band *j* ('LEVEL-2 RESIDUALS' in Equation 3). These correspond to the distance of the specific mean popularity score of a given boy band *j* from the overall mean popularity score B_{00}. For instance, you can see that the observed mean popularity score of *NSYNC ($\overline{NSYNC} = 4.83$) is not properly described by the overall mean ($B_{00} = 3.46$), and that the magnitude of the level-2 error is $u_{01} = \overline{NSYNC} - B_{00} = 1.37$.

$e_{ij} \rightarrow$ The vertical thin dotted lines represent the level-1 residuals e_{ij} associated with each boy band member *i* within boy band *j* ('LEVEL-1 RESIDUALS' in Equation 3). These correspond to the distance of the observed popularity score of boy band member *i* from the specific mean score of his boy band *j*. For instance, you can see that the observed popularity score of Justin Timberlake ($Y_{41} = 6.75$) is not properly described by the specific mean score of his boy band ($\overline{NSYNC} = 4.83$) and that the magnitude of the level-1 error is $e_{41} = Y_{41} - \overline{NSYNC} = 1.92$.

Therefore, in two-level linear regression there are two aggregates of residuals. First, *the variance of level-2 residuals*, written as $\text{var}(u_{0j})$, is calculated by taking the mean of the squared level-2 residuals: $\text{var}(u_{0j}) = (u_{01}^2 + u_{02}^2 + \ldots + u_{0K}^2)/K$ (the mean of the squared distance of *NSYNC, The Click Five, The Jonas Brothers, etc., from the overall mean). This captures the unexplained between-cluster variations. When $\text{var}(u_{0j})$ is larger than zero, this indicates that popularity varies between boy bands, with some bands being more popular than others.

Second, *the variance of level-1 residuals*, written as $\text{var}(e_{ij})$, is calculated by taking the mean of the squared level-1 residuals: $\text{var}(e_{ij}) = (e_{11}^2 + e_{21}^2 + \ldots + e_{nK}^2)/N$ (the mean of the squared distance of Justin Timberlake, Lance Bass, ..., Kevin Jonas, etc., from their boy band-specific means. This captures the unexplained within-cluster variations. When $\text{var}(e_{ij})$ is larger than zero, this indicates that popularity varies within boy bands, with some members being more popular than others.

Just for your general information, traditional regression and multilevel modeling use two different methods of estimation (Goldstein, 2013). Traditional regression typically uses the ordinary least squares (OLS) estimator (the coefficients and variance terms are estimated by minimizing the

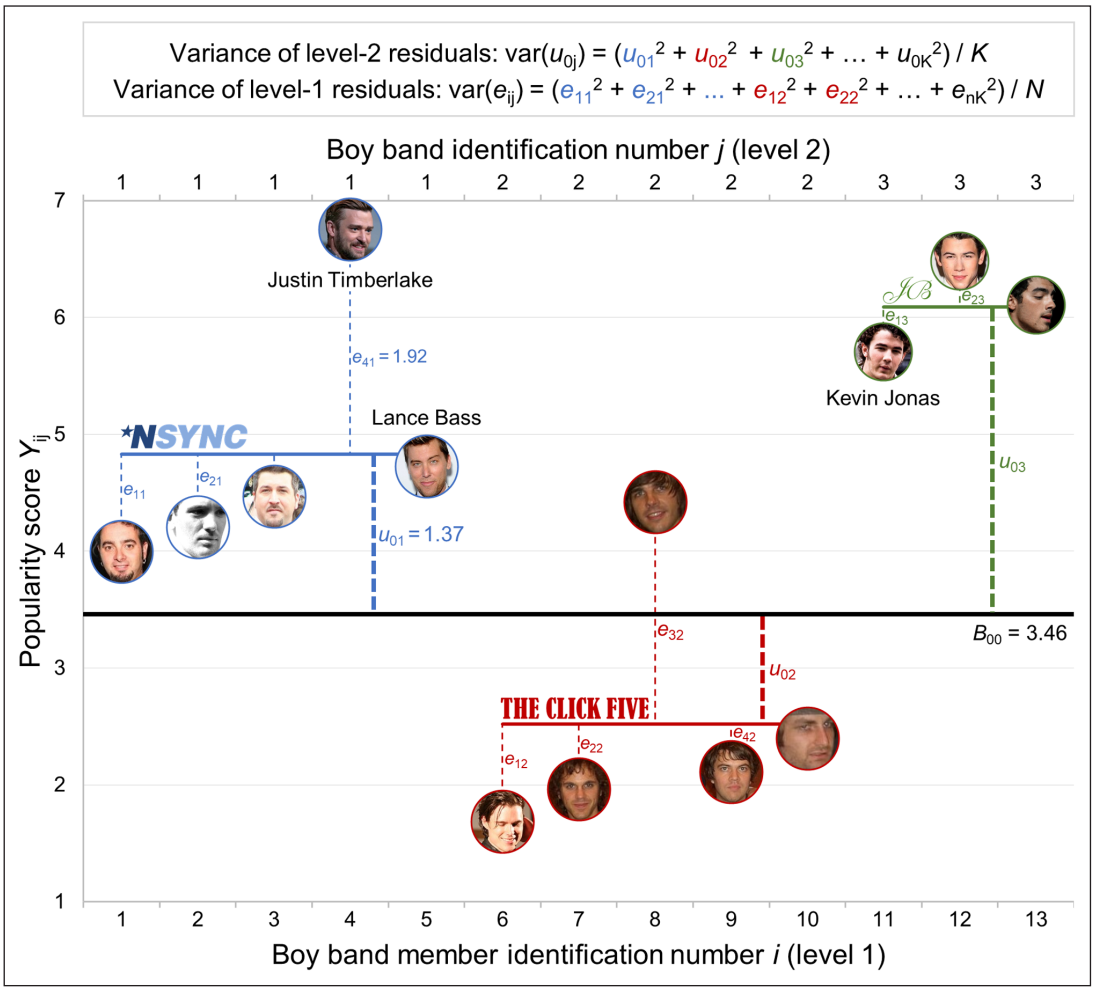


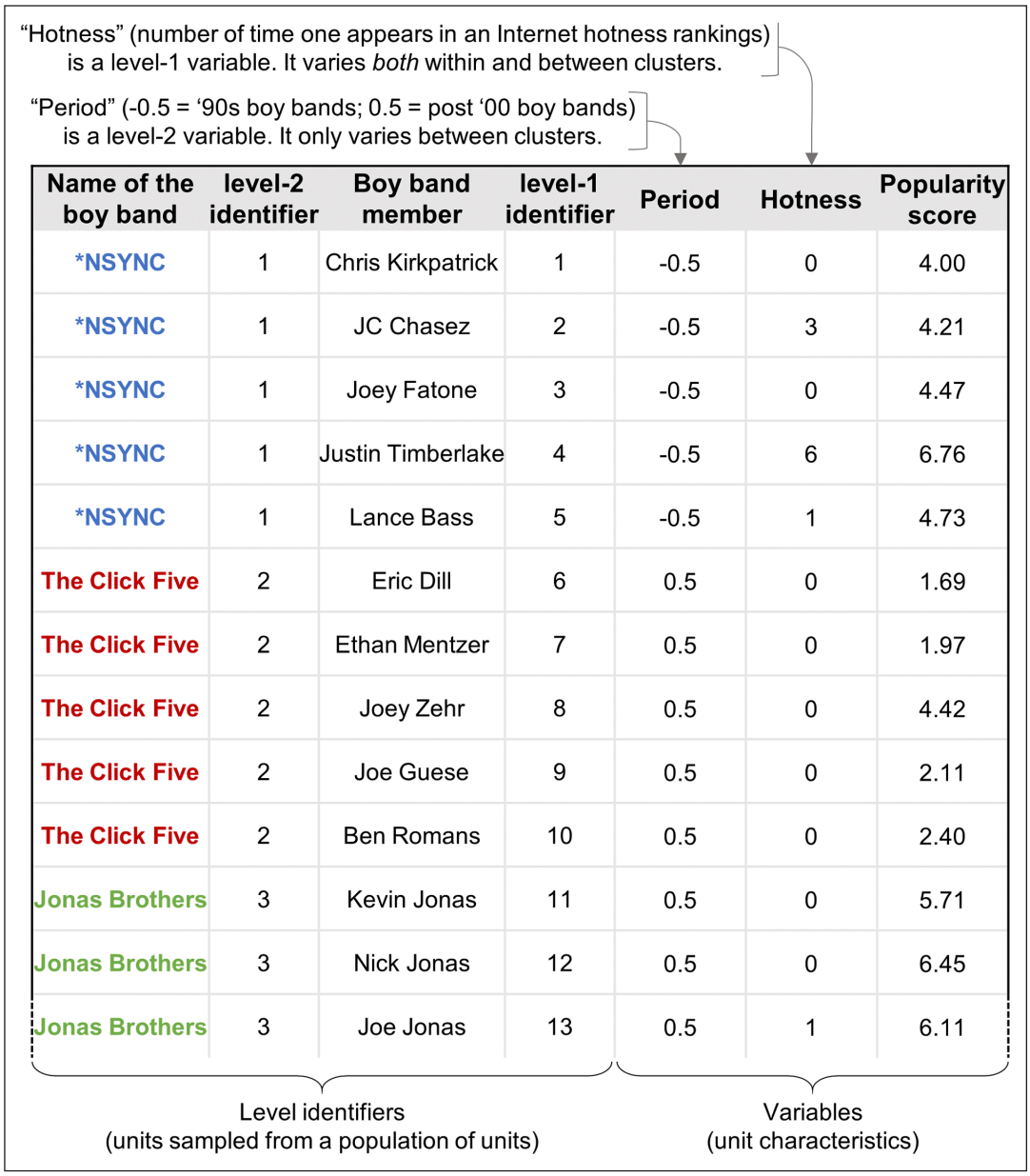
Figure 3: Graphical representation of a two-level linear regression with no predictor (Equation 4) in which the observed popularity score Y_{ij} (*y*-axis) of a particular boy band member *i* (bottom *x*-axis) from a particular boy band *j* (top *x*-axis) corresponds to the overall mean popularity score B_{00} (horizontal thick line) plus the level-2 residuals u_{0j} (vertical thick dotted lines) and level-1 residuals e_{ij} (vertical thin dotted lines). *Notes*: Only the first 13 observations are represented; normally, B_{00} is only equivalent to the arithmetic grand-mean when the design is completely balanced (i.e., when the number of participants is the same for each cluster); the first author is somewhat prosopagnosic and apologizes in advance if he has erroneously interchanged two faces.

average squared differences between the predicted values and the data), whereas multilevel modeling typically uses the maximum likelihood (ML) estimator (the coefficients and variance terms are jointly estimated by maximizing the likelihood of the predicted values given the data).[4] However, the general goal of multilevel modeling is the same as traditional regression, that is, estimating whether a predictor for which you formulated a prediction contributes to explaining between- and/or within-cluster changes in the value of your outcome.

The Second Principle of Two-Level Modeling: Two Levels Mean Two Types of Predictors

Now imagine you focus on two particular predictors: boy band period of success, and boy band member hotness. As we are about to see, these predictors represent the *two types of predictors* in two-level modeling: level-2 predictors and level-1 predictors.

First, let's focus on *period of success*. You operationalized this variable by distinguishing '90s boy bands (whose greatest year of success fell between 1990 and 2000; coded '–0.5') from post-'00s boy bands (whose greatest year of success came after 2000; coded '+0.5'). This is a higher-level unit characteristic or a level-2 variable. There is a straightforward rule for recognizing such a variable: The value of a level-2 variable CANNOT change within clusters, but CAN ONLY change between clusters (see **Figure 4**, column 5). Level-2 variables are noted X_j (uppercase) and—as you can see—the letter X only comes with a *j* subscript (no *i* subscript) because: (i) X_j CANNOT vary from one level-1 unit *i* to another within a given cluster (e.g., because *NSYNC's period of success was the '90s, Justin Timberlake and his buddy Lance Bass's automatically have the value) and (ii) X_j CAN ONLY vary from one level-2 unit *j* to another (e.g., from *NSYNC [a '90s boy band] to The Click Five [a post-'00s boy band]). The boy band dataset uploaded on the OSF includes other examples of level-2 variables: boy band number of weeks in the U.S. chart, boy band biggest hit, and number of YouTube views of biggest hits.



Name of the boy band	level-2 identifier	Boy band member	level-1 identifier	Period	Hotness	Popularity score
*NSYNC	1	Chris Kirkpatrick	1	-0.5	0	4.00
*NSYNC	1	JC Chasez	2	-0.5	3	4.21
*NSYNC	1	Joey Fatone	3	-0.5	0	4.47
*NSYNC	1	Justin Timberlake	4	-0.5	6	6.76
*NSYNC	1	Lance Bass	5	-0.5	1	4.73
The Click Five	2	Eric Dill	6	0.5	0	1.69
The Click Five	2	Ethan Mentzer	7	0.5	0	1.97
The Click Five	2	Joey Zehr	8	0.5	0	4.42
The Click Five	2	Joe Guese	9	0.5	0	2.11
The Click Five	2	Ben Romans	10	0.5	0	2.40
Jonas Brothers	3	Kevin Jonas	11	0.5	0	5.71
Jonas Brothers	3	Nick Jonas	12	0.5	0	6.45
Jonas Brothers	3	Joe Jonas	13	0.5	1	6.11

Figure 4: The 13 first lines of your boy band datasets. *Notes*: Fans often argue that 'The Click Five' is a pop rock band, not a boy band; we respectfully disagree with them, and we invite readers to watch The Click Five's video clip 'Kidnap My Heart' on YouTube and make up their own mind (this song does not count as a hidden song).

Second, let's focus on *hotness*. You operationalized this variable by counting the number of time(s) a given boy band member appears in Internet hotness rankings, such as *The Hollywood Gossip's Hottest Boy Band Members of All-Time*. This is a lower-level unit characteristic or a level-1 variable. As before, there is a straightforward rule for recognizing such a variable: The value of a level-1 variable CAN change within clusters, and CAN ALSO vary between clusters (see **Figure 4**, column 6). Level-1 variables are noted x_{ij} (lowercase) and—as you can see—the letter x comes with an i and j subscript because (i) x_{ij} CAN vary from one level-1 unit i to another within a given cluster (e.g., from Justin Timberlake [6 rankings] to his buddy Lance Bass [1 ranking]) and (ii) x_{ij} CAN ALSO vary from one level-2 unit j to another (from *NSYNC [10 rankings] to The Jonas Brothers [1 ranking]). The boy band dataset uploaded on the OSF includes other examples of level-1 variables: boy band member height, boy band member skin color, or boy band member hair style (e.g., spiked, swept, or shaved). The popularity score (as any outcome in two-level modeling) is another example of a level-1 variable.

The Third Principle of Two-Level Modeling: Two Levels Mean Two Types of Level-1 Effect Parameters

Now imagine you want to estimate the effect of your level-1 predictor x_{ij} (hotness) on the popularity score Y_{ij}, so you build the following one-predictor two-level model:

$$Y_{ij} = B_{00} + \left(B_{10} + u_{1j}\right) \times x_{ij} + u_{0j} + e_{ij} \qquad \text{(Eq. 5)}$$

Don't freak out just yet, youngblood! It's a lot of information, so let's unpack the terms of the equation together:

Y_{ij}, B_{00}, u_{0j}, and e_{ij} → First, you should know from Equation 4 and **Figure 3** that Y_{ij} is the outcome (the popularity score of member *i* within boy band *j*), B_{00} is the fixed intercept (the overall value of Y_{ij} when predictor x_{ij} is set at zero), u_{0j} is the level-2 residuals (the distance between the observations and model predictions at the boy band level), and e_{ij} is the level-1 residuals (the distance between the observations and model predictions at the boy band member level).

$(B_{10} + u_{1j}) \times x_{ij}$ → Now, focus on your level-1 predictor x_{ij} (boy band member hotness). Things are a tad more complicated because there are now *two level-1 effect parameters*: the coefficient estimate B_{10} and the slope residuals u_{1j}. To make sense of this, take a look at **Figure 5**:

① $B_{10} \times x_{ij}$. The thick slope in **Figure 5** represents the coefficient estimate B_{10} (also called the 'fixed slope'). It is the *overall mean effect* of your level-1 predictor x_{ij} *across all clusters*. It has the same meaning as in any regular linear regression: An increase of one unit in x_{ij} is associated with a change of B_{10} in the value of the outcome Y_{ij}, *regardless of clustering*. In your case, you can see that $B_{10} = 0.23$, which means the following: When hotness increases by one unit, popularity score increases by 0.23 points on average, *while leaving aside boy band membership*.

② $u_{1j} \times x_{ij}$. The vertical thick dotted curves in **Figure 5** represent the slope residuals u_{1j} (sometimes called the 'random slope'). Each curve corresponds to the difference between (i) the specific effect of hotness for boy band *j* and (ii) the *overall mean effect* of hotness B_{10}. In the same way that the mean of the outcome can vary from one cluster to another (forming the intercept residuals), the effect of a level-1 variable can vary from one cluster to another (forming the slope residuals). As an illustration, **Figure 5** shows that the effect of hotness is positive for *NSYNC, negative for The Click Five, and null

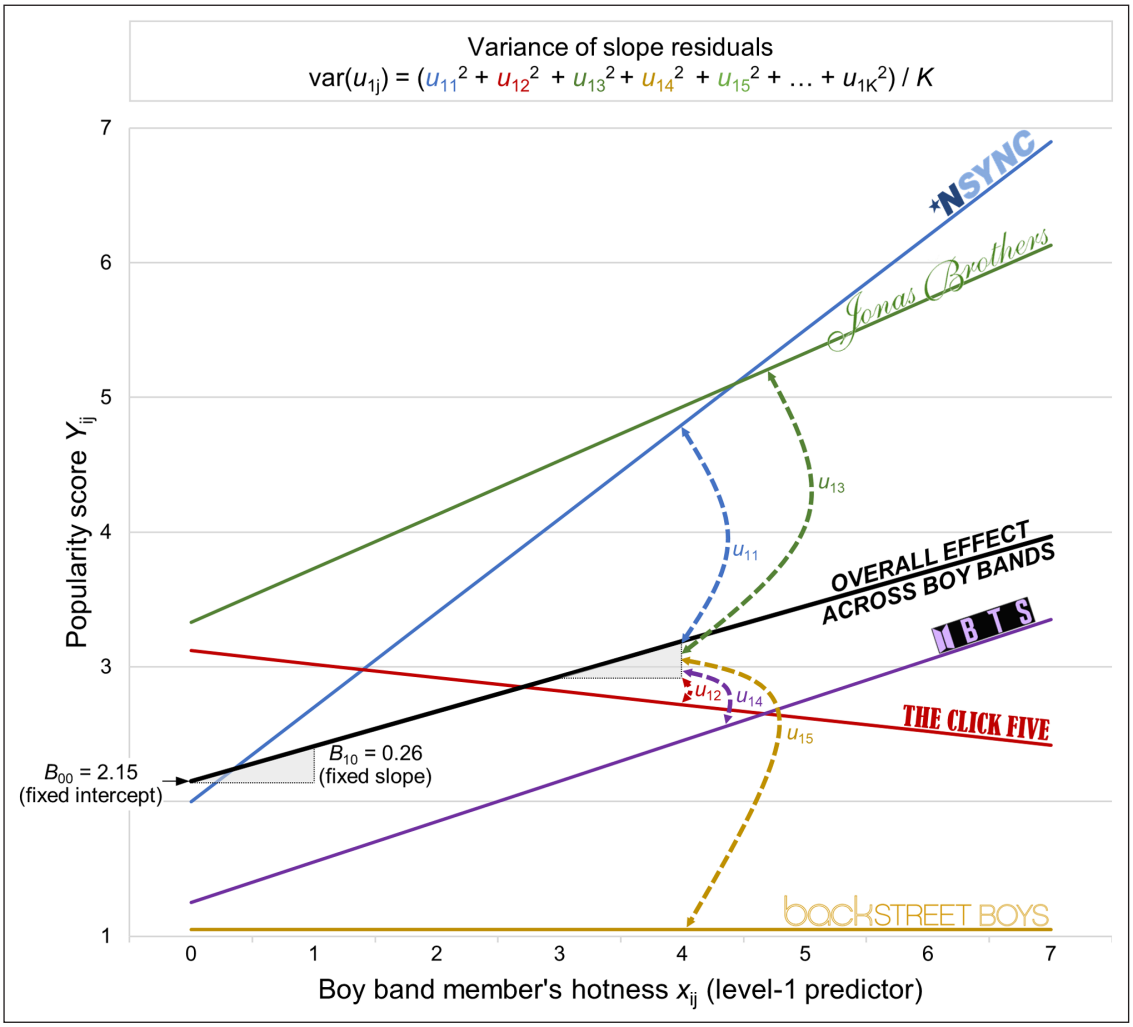


Figure 5: Graphical representation of the coefficient estimate or fixed slope B_{10} (the thick slope – the overall mean effect of hotness across boy bands) and the slope residuals u_{1j} (vertical thick dotted curves – the differences between the specific effects of hotness in a given boy band compared to the overall mean effect). *Notes*: Only the first five boy bands are represented; obviously, the data for this figure are fictitious (the Backstreet Boys are still way more popular than that!).

for The Backstreet Boys. Simply put, although the overall mean effect of hotness on popularity is globally positive, the effect appears to be stronger for some bands, weaker for others, and even reversed for others. The aggregate of the slope residuals is *the variance of the slope residuals*, written as $\text{var}(u_{1j})$. It is calculated by taking the mean of the squared slope residuals: $\text{var}(u_{1j}) = (u_{11}^2 + u_{12}^2 + \ldots + u_{1K}^2)/K$ (the mean of the squared differences between the slopes of *NSYNC, The Click Fives, The Jonas Brothers, etc., and the over mean effect). This captures the unexplained between-cluster slope variations. When $\text{var}(u_{1j})$ is larger than zero, this indicates that the effect of hotness on popularity varies between bands.

There is one last thing that adds to the complexity (the hardest thing, to be honest): The intercept residuals u_{0j} (or level-2 residuals) and the slope residuals u_{1j} can covary (Robson & Pevalin, 2016). The degree to which these two parameters covary is *the covariance term*, written as $\text{cov}(u_{0j}, u_{1j})$. Taking this covariance term into account is important because $\text{cov}(u_{0j}, u_{1j})$ cannot be assumed to be zero, and assuming otherwise can inflate the false-positive rate (Wang et al., 2019). There are three possible situations. First, if $\text{cov}(u_{0j}, u_{1j})$ is approximately *zero*, this means that the popularity score at $x_{ij} = 0$ for a given boy band *j* is not systematically related to the strength of the within-boy band effect of hotness; there is simply no pattern to be found here (**Figure 6**, left panel). Second, if $\text{cov}(u_{0j}, u_{1j})$ is *positive*, this means that a larger popularity score at $x_{ij} = 0$ (a larger intercept) tends to be associated with a stronger effect (a larger slope). In other words, there is a pattern of 'fanning out' (there are greater between-boy band differences when focusing on hotter members, e.g., a floor effect; **Figure 6**, middle panel). Third, if $\text{cov}(u_{0j}, u_{1j})$ is *negative*, this means that a larger popularity score at $x_{ij} = 0$ (a larger intercept) tends to be associated with a weaker effect (a smaller slope). In other words, there is a pattern of 'fanning in' (there are slighter between-boy band differences for hotter members, e.g., a ceiling effect; **Figure 6**, right panel).

Finally, we want to draw your attention to the fact that estimating the effect of a level-2 predictor X_j (e.g., period of success) is much more straightforward than estimating the effect of a level-1 predictor x_{ij} (e.g., hotness). In that case, the interpretation of the level-2 coefficient estimate $B_{01} \times X_j$ is the same as in any traditional linear regression: An increase of one unit in X_j is associated with a change of B_{01} in the value of the outcome Y_{ij} (in our example, compared to members from '90s boy bands, the popularity score of members from post-'00s boy bands is higher by B_{01} points on average). Importantly, there are no slope residuals here because it is impossible for the effect of a level-2 predictor to vary within clusters (in our example, because members of a given boy band are either all from a '90s boy band or all from a post-'00s boy band, the effect of period of success cannot vary from one boy band to the next).

Brief Summary of PART 1

After reading Part 1, you should have a good grasp on the three key principles of two-level linear modeling. The first principle is that 'two levels mean two types of residuals.' This is illustrated by **Figure 3**: Observations can vary both between clusters (forming the variance of level-2 residuals or the variance of the intercept residuals) and within clusters (forming the variance of level-1 residuals). The second principle is that 'two levels mean two types of predictors.' This is illustrated by **Figure 4**: Predictors can be level-2 variables (higher-level characteristics that CANNOT vary

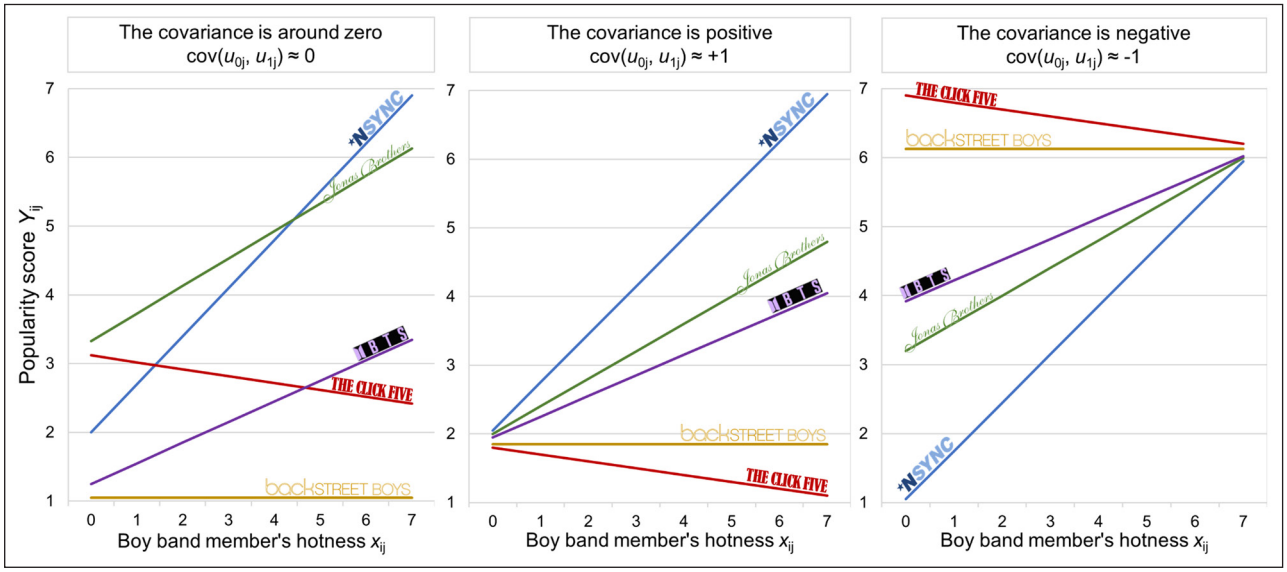


Figure 6: Graphical representations of the covariance between the intercept residuals (or level-2 residuals) u_{0j} and the slope residuals u_{1j}. In the left panel, the covariance is equal to zero (no pattern), in the middle panel, the covariance is positive (higher boy band-specific intercepts come with larger slopes), and in the right panel, the covariance is negative (higher boy band-specific intercepts come with smaller slopes). *Note*: Again, the data for this figure are fictitious.

within clusters) or level-1 variables (lower-level characteristics that CAN vary within clusters). The third principle is that 'two levels mean two types of level-1 effect parameters.' This is illustrated by **Figure 5**: The effect of a level-1 variable is described by a coefficient estimate or a fixed slope (the overall mean effect across clusters) and the variance of slope residuals (the variations of the effect from one cluster to another). Moreover—and as illustrated by **Figure 6**—the intercept residuals and slope residuals can covary (e.g., larger cluster-specific intercepts may imply larger cluster-specific effects). We know it's a lot to digest, but no diggity: Once you understand these three key principles, what you have yet to learn about multilevel modeling is truly a matter of details! **Table 1** provides a summary of the main notations and definitions of two-level modeling concepts that you will encounter in the present paper.

PART 2. A Three-Step Procedure for Conducting Two-Level Linear Modeling

The Aim of This Part Is for You to Learn How to Perform Two-Level Linear Modeling

You're pretty happy. You've read Part 1, and you understand the three key principles of two-level linear modeling. But now that you're in front of your computer… you don't know what to do!

Stay with us a little bit longer: Part 2 is a ready-to-use three-step procedure for conducting two-level linear modeling using SPSS, Stata, R, or Mplus. Before reading it, download the contents of the folder named after your favorite statistical software from the OSF (https://osf.io/4yhbm/ DOI: 10.17605/OSF.IO/4YHBM). In this folder, you will find: (i) the complete software-specific instructions, (ii) the boy band dataset and (iii) the script to perform our three-step procedure.

While reading Part 2, we *strongly* recommend you try to reproduce the procedure using the relevant script. When doing this, remember that your dataset has $N = 175$ members (level-1 units) nested in $K = 50$ boy bands (level-2 units), and imagine you formulated the following three hypotheses:

The level-2 main effect hypothesis. Compared to '90s boy band members, post-'00s boy band members have a higher popularity score.

The level-1 main effect hypothesis. The higher the boy band member hotness, the higher the boy band member popularity score.

The cross-level interaction hypothesis. For '90s boy bands, the higher the member hotness, the higher the member popularity score; for post-'00s boy bands, this link is attenuated.

The three-step procedure used to test these hypotheses is organized as follows:

STEP #0. Centering variables
STEP #1. Building an empty model to determine if multilevel modeling is needed.
STEP #2. Building intermediate models to estimate the (co)variance terms
STEP #3. Building the final model and interpreting the 95% CIs

Figure 7 presents a decision tree summarizing the three-step procedure.

Table 1: Summary of the main notations and definitions of two-level modeling concepts.

	Level 2 ***K* level-2 units (clusters)** **with *n* observations per cluster (mean cluster size)**	**Level 1** ***N* level-1 units (observations)**
The first principle: Two types of residuals	u_{0j} Level-2 residuals or intercept residuals ("random intercept") *Distance of the cluster-specific means from the overall mean* **Tip:** The aggregated index of level-2 residuals is var(u_{0j})	e_{ij} Level-1 residuals *Distance of the observations from the cluster-specific means* **Tip:** The aggregated index of level-1 residuals is var(e_{ij})
The second principle: Two types of variable	$X1_j$, $X2_j$, $X3_j$, etc. Level-2 predictors *Cluster characteristics* **Tip:** They CANNOT vary within clusters	$x1_{ij}$, $x2_{ij}$, $x3_{ij}$, etc. Level-1 predictors *Observation characteristics* **Tip:** They CAN vary within clusters
The third principle: Two types of level-1 effects parameters	B_{00}, B_{01}, B_{02}, B_{03}, etc. Fixed intercept (B_{00}) and level-2 coefficient estimates (B_{01}…) *Overall mean/intercept and effects of $X1_{ij}$, $X2_{ij}$, $X3_{ij}$, etc.*	B_{10}, B_{20}, B_{30}, etc. Level-1 coefficient estimates or fixed slopes *Overall mean effect of $x1_{ij}$, $x2_{ij}$, $x3_{ij}$, etc., across all clusters*
	N/a Slope residuals are not possible for level-2 predictors	u_{1j}, u_{2j}, u_{3j}, etc. Variation of the effect of the level-1 predictors or slope residuals ("random slopes") *Differences between the cluster-specific slopes and the fixed slope* **Tip 1:** The variance term is var(u_{1j}), var(u_{2j}), etc. **Tip 2**: The covariance term is cov(u_{0j}, u_{1j}), etc.

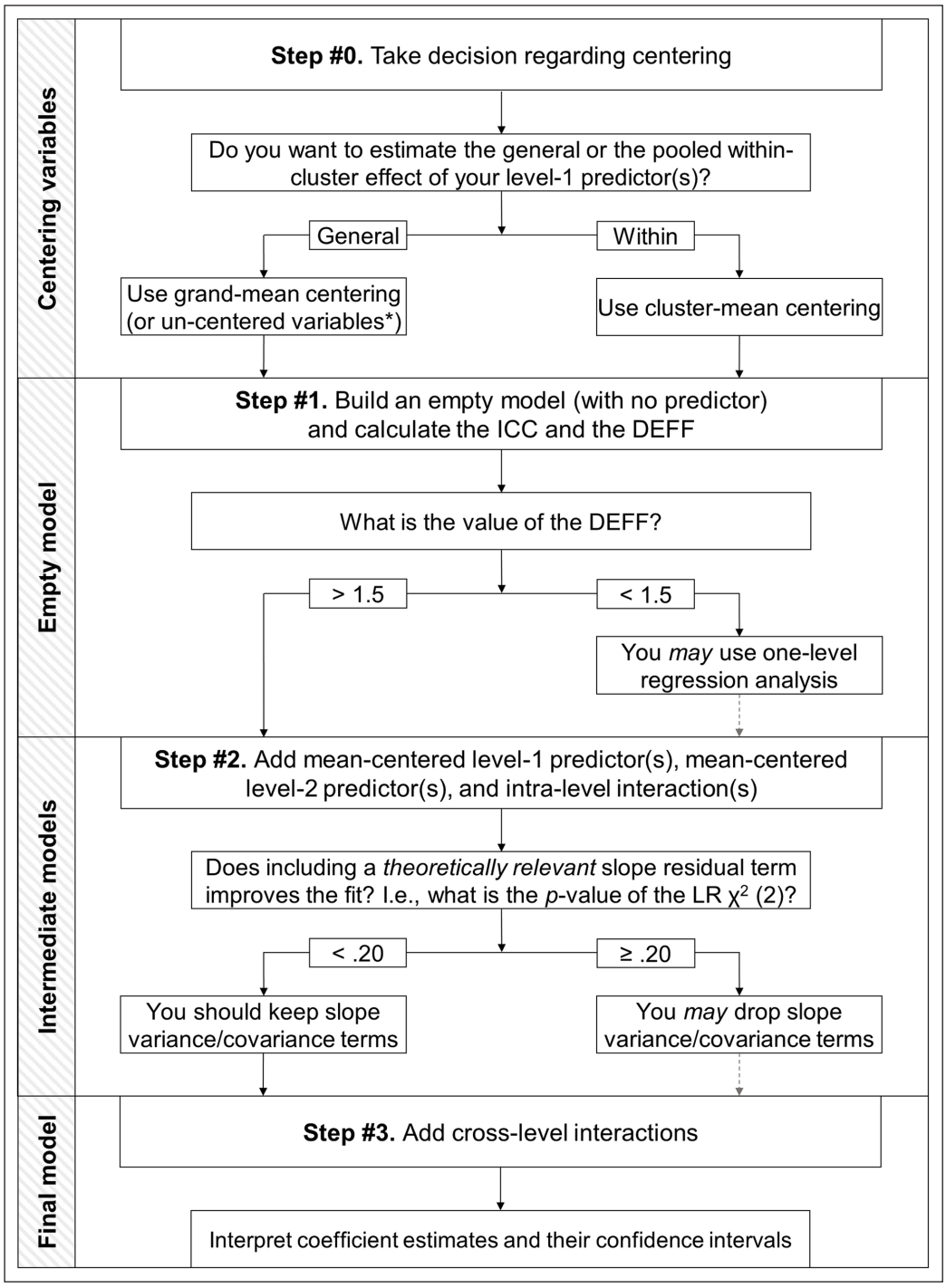


Figure 7: Decision tree illustrating the three-step procedure for two-level linear modeling. *Note*: *We recommend always centering your predictor when your model includes an interaction term.

STEP #0. Centering Variables

Main question to be answered: Do you want to estimate the general effect or the pooled within-cluster effect of your level-1 predictor?

First, you need to reflect on the way you will center your predictor(s). Different types of centering will lead you to estimate different effects, particularly concerning your level-1 predictor(s). There are basically two centering approaches when it comes to level-1 predictors: grand-mean centering and cluster-mean centering (Myers et al., 2010; beware, some software programs other than the software programs discussed in this primer may automatically center the variables for you).

Grand-mean centering a level-1 predictor means subtracting the *overall* mean of the level-1 predictor M_{00} from each individual observation x_{ij}, namely:

$$x_{ij}^{gmc} = x_{ij} - M_{00} \quad \text{(Eq. 6)}$$

In your dataset, grand-mean centering hotness means subtracting the overall hotness mean from each boy band member's hotness value (subtracting the same overall sample mean from Justin Timberlake's, Lance Bass', or Kevin Jonas' hotness value). Thus, a positive value indicates that the boy band member is hotter than the typical boy band member *in the overall sample* (and reciprocally

for a negative value). For instance, Justin Timberlake's grand-mean centered hotness value is $x_{41}^{gmc} = +7.53$, signaling that he is, on average, hotter than the other singers.

With interaction terms, grand-mean centering is convenient for estimating main effects, although it is not a strict requirement. Without interaction terms, grand-mean centering will neither change the value nor the interpretation of the coefficient estimate B_{10}. Uncentered and grand-mean centered level-1 predictors will both lead you to estimate *the general between-observation effect*: A deviation of one unit in hotness from the overall sample mean will be associated with a change of B_{10} in the popularity score. Because the coefficient estimate corresponds to the general between-observation effect, it is a mixture of the within- and between-boy band effects. However, grand-mean centering will also change the value of the fixed intercept $B_{00,}$ which—as in a traditional regression—will become the overall value of Y_{ij} when predictor x_{ij}^{gmc} is set at zero, that is, the predicted popularity score of a boy band member with an average level of hotness *across boy bands*.

Cluster-mean centering a level-1 predictor means subtracting the *cluster-specific* mean of the level-1 predictor $M(x_{0j})$ from each individual observation x_{ij}, namely:

$$x_{ij}^{cmc} = x_{ij} - M\left(x_{0j}\right) \quad \text{(Eq. 7)}$$

In your dataset, cluster-mean centering hotness means subtracting the boy band-specific hotness mean from each boy band member's hotness value (subtracting the specific *NSYNC-mean from Justin Timberlake's hotness value, subtracting the specific Jonas Brother-mean from Kevin Jonas's hotness value, etc.). Thus, a positive value indicates that the boy band member is hotter than the average hotness *of his band* (and reciprocally for a negative value). For instance, Justin Timberlake's cluster-mean centered hotness value is $x_{41}^{cmc} = +5.60$, signaling that he is, on average, hotter than the rest of *NSYNC members.

Cluster-mean centering will change both the value and the interpretation of the coefficient estimate B_{10}. A cluster-mean centered level-1 predictor will lead you to estimate *the pooled within-cluster effect*: A deviation of one unit in hotness from the boy band-specific mean will be associated with a change of B_{10} in the popularity score. The coefficient estimate here corresponds to the aggregated within-boy band slopes (the typical within-boy band effect). Note that dichotomous level-1 predictors can also be cluster-mean centered (Enders & Tofighi, 2007) and that level-2 predictors can *only* be grand-mean centered (because they are constant within each cluster). However, cluster-mean centering will also change the value of the fixed intercept B_{00}, which will now become the overall value of Y_{ij} when predictor x_{ij}^{cmc} is set at zero, that is, the predicted popularity score of a boy band member with the average level of hotness *within his boy band*.[5]

Summary and recommendations. The centering decision pertaining to a level-1 predictor depends on the kind of effect you want to test: Grand-mean center the predictor (or keep it uncentered) if you are interested in the absolute, between-observation effect and cluster-mean center the predictor if you are interested in the relative, within-cluster effect.

Here's an extract of our mock paper pertaining to STEP #0:

> 'All of the variables were centered. We cluster-mean centered our level-1 variable, namely, hotness (subtracting the boy band-specific hotness mean from each observation), to obtain the estimation of the pooled within-boy band effect.'

STEP #1. Building an Empty Model to Calculate the ICC/DEFF

Main questions to be answered: How much of the variation in your outcome is related to between-cluster differences and do you really need multilevel modeling?

Now that you have made a decision regarding centering, the next thing you want to do is to build an empty model (i.e., a model with no predictor [see Equation 4], also known as an 'unconditional mean model' or 'random-intercept model') and calculate the Intraclass Correlation Coefficient (ICC) (Hox, 2017; Snijders & Bosker, 2011).

$$\text{ICC} = \frac{\text{Between-cluster variance}}{\text{Total variance}} = \frac{var\left(u_{0j}\right)}{\text{var}\left(u_{0j}\right) + \text{var}\left(e_{ij}\right)} \quad \text{(Eq. 8)}$$

As you can see in the above equation, the ICC corresponds to the proportion of the between-cluster variance $\text{var}(u_{0j})$ (in your case, the between-boy band variations) in the total variance $\text{var}(u_{0j}) + \text{var}(e_{ij})$ (in your case, the between-boy band variations plus the within-boy band variations; if the meaning of these variance terms is not clear, go back to **Figure 3**).

The ICC quantifies *the degree of resemblance of the observations belonging to the same cluster*, and can range from 0 to 1. An ICC of 0 indicates perfect independence of the residuals. In this case, the observations are completely independent of cluster membership: Each and every boy band has the same mean popularity score (there is no between-boy band variation). However, an ICC of 1 indicates perfect interdependence of the residuals. In that case, the observations are completely dependent on cluster membership: Each and every member of any boy band has the same popularity score (there is no within-boy band variation).

ICCs of 0.01, 0.05, and 0.20 can be considered as small, medium, and large levels of within-cluster homogeneity, respectively (Kreft & de Leeuw, 1998). In your dataset, ICC = 0.82 (it is *very* large), meaning that 82% of the variance in popularity score can be attributed to between-boy band differences; conversely, this means that 18% of the variance in popularity score can be attributed to within-boy band differences.

Authors sometimes argue that when the ICC falls below a certain threshold (e.g., ICC < 0.05), one can ignore the hierarchical structure of their data and use traditional regression (for a relevant discussion, see Hayes, 2006). However, simulation studies show that an ICC as low as 0.01 can multiply the false-positive rate by four when

using traditional regression (Musca et al., 2011), revealing that a non-zero, small ICC *cannot* be taken as an indication that multilevel modeling is unwarranted (Huang, 2018).

To determine whether or not multilevel modeling is needed, the Design EFFect is more informative (Kish, 1965; Muthén & Satorra, 1995):

$$\text{DEFF} = 1 + (n - 1) \times \text{ICC} \quad \text{(Eq. 9)}$$

The DEFF takes both the mean cluster size (n) and within-cluster homogeneity (ICC) into account in order to quantify *the degree to which a multilevel sample differs from a simple random sample* (with perfectly independent residuals). The DEFF can range from 1 (no difference) to n (a maximal difference). In your dataset, DEFF = 3.04, meaning that the sampling variance of the popularity score (the population sampling error) is about three times larger than if your 175 members belonged to 175 different boy bands (Dattalo, 2008).

Authors usually argue that when the DEFF falls below 2, one can simply ignore the hierarchical structure of their data and use traditional regression (Peugh, 2010). However, such a threshold may be too liberal, as a more recent simulation study showed that when the DEFF is as small as 1.5, the estimation of standard errors from traditional regressions is sometimes biased (Lai & Kwok, 2015).

Summary and recommendations. In STEP #1, you need to build an empty model to calculate (i) the ICC (to estimate the proportion of the variance accounted for by clustering) and (ii) the DEFF (to determine whether or not multilevel modeling is needed). We recommend that if DEFF < 1.5, clustering *may* be ignored and traditional regression *may* be used.

Here is the extract of our mock paper pertaining to STEP #1.

> 'As a first step, we built an empty model and calculated the ICC and the DEFF. The ICC was 0.82, meaning that 82% of the variance in the popularity score was explained by between boy band differences (a large within-cluster homogeneity). The DEFF was above 1.5, meaning that multilevel modeling was warranted.'

STEP #2. Building Intermediate Models to Estimate the (Co)Variance Terms

Main question to be answered: Does the effect of your level-1 predictor vary between clusters, and should you estimate the residual slope variance or covariance terms?

Now that you have made a decision regarding the need to use multilevel modeling, ask yourself whether you have *theoretical* reasons to expect the effect of your level-1 predictor to vary between clusters? (Maxwell & Delaney, 2004). If the answer is 'YEP,' you have to figure out whether you need to estimate this kind of variation. If the answer is 'NOPE,' you can directly go to STEP #3 (there is no need to test for this kind of variation).

In your case, the answer is an unequivocal 'YEP.' Because you formulated a cross-level interaction hypothesis, you have *theoretical* reasons to expect the effect of hotness to differ between boy bands (at least, between '90s and post-'00s boy bands). Now to determine the need to estimate this expected variation, you have to build two intermediate models: (i) a constrained intermediate model (*not* taking between-cluster variation of the level-1 effect into account) and (ii) an augmented intermediate model (taking this variation into account). Then you will have to compare the two intermediate models (Aguinis et al., 2013).

First, let's focus on the *constrained* intermediate model. This model includes *all* of your predictors *except* the cross-level interactions (because your goal is to estimate the crude slope residuals and the cross-level interactions are likely to explain a part of the residual variance).

$$Y_{ij} = B_{00} + B_{10} \times x_{ij}^{cmc} + B_{01} \times X_j + u_{0j} + e_{ij} \quad \text{(Eq. 10)}$$

In the above constrained intermediate model equation the coefficient estimate B_{10} (the fixed slope) corresponds to the overall effect of your level-1 predictor x_{ij}^{cmc} (cluster-mean centered hotness), whereas the coefficient estimate B_{01} corresponds to the effect of your level-2 predictor X_j (period of success;).

Second, let's focus on the *augmented* intermediate model:

$$Y_{ij} = B_{00} + (B_{10} + u_{1j}) \times x_{ij}^{cmc} + B_{01} \times X_j + u_{0j} + e_{ij} \quad \text{(Eq. 11)}$$

The only new thing in Equation 11 is the slope residuals u_{1j}, which corresponds to the differences between the cluster-specific effects of your level-1 predictor x_{ij}^{cmc} (the boy band-specific effect of hotness) and the overall effect of x_{ij}^{cmc}, that is, the fixed slope B_{10} (the overall effect of hotness). This implies that two more terms will be estimated: (i) the variance of the slope residuals var(u_{1j}) (the amount of variation between the boy band-specific slopes) and (ii) the covariance term between the intercept and slope residuals cov(u_{0j}, u_{1j}) (the association between the boy band-specific intercepts and slopes; if this is not clear, go back to **Figures 5** and **6**, respectively).

The next thing you want to do is compare the two models and test whether the augmented intermediate model (including u_{1j}) achieves a better fit than the constrained intermediate model (excluding u_{1j}). In your case, this means you want to know whether including the between-boy band variation of the effect of hotness improves the accuracy of estimation. To do so, you have to gather a (mis)fit index for each model named the 'deviance' (the smaller deviance, the better the fit) and perform the following likelihood-ratio test:[6]

$$\text{LR}\,\chi^2(2) = \text{deviance}_{\text{constrained}} - \text{deviance}_{\text{augmented}} \quad \text{(Eq. 12)}$$

In Equation 12, the likelihood-ratio test LR χ^2 has two degrees of freedom. This is because the augmented

intermediate model estimates *two* more terms than the constrained intermediate model (var(u_{1j}) and cov(u_{0j}, u_{1j})). There are essentially two possible scenarios here:

SCENARIO A. Deviance$_{augmented}$ is substantially smaller than Deviance$_{constrained}$ (LR χ^2 (2) is positive). This means that estimating var(u_{1j}) and cov(u_{0j}, u_{1j}) matters (it improves the fit!). Thus, u_{1j} needs to be kept in the final model.

SCENARIO B. Deviance$_{augmented}$ is not substantially different than deviance$_{constrained}$ (LR χ^2 (2) is null). This means that estimating var(u_{1j}) and cov(u_{0j}, u_{1j}) does not necessarily matter (it does not improve the fit), and u_{1j} could be discarded.

In your dataset, the result of the LR χ^2 is somewhat ambiguous, namely, LR χ^2 (2) = deviance$_{constrained}$ – deviance$_{augmented}$ = 415.96–412.08 = 3.88, p = 0.144 (you can find the p-value using an online chi-square online calculator).

So, where do you go with this p = 0.144? Well, not everybody agrees... Some authors argue that models should *always* be maximal (Barr et al., 2013). If you follow their guidelines, regardless of the p-value of your LR χ^2 (2), the slope residuals u_{1j} need to be kept in the final model. Other authors argue that models should be as parsimonious as possible in order to avoid overparametrization and convergence issues (Bates et al., 2015). If you follow their guidelines, given that the p-value of the LR χ^2 (2) is above the alpha level of 0.05, the slope residuals u_{1j} may be discarded. We believe that these guidelines may fall at one of two extremes. A more nuanced criterion for accepting the significance of the LR χ^2 (2) may be setting the alpha level at 0.20 instead of at 0.05 (as suggested by Matuschek et al., 2017).[7]

However, our criterion is certainly not a miracle solution. You should know that discarding the slope residuals of your focal variable(s) may sometimes substantially inflate the false-positive rate (for relevant simulations, see Schielzeth & Forstmeier, 2009). Thus, in the context of a small sample size or theoretical uncertainty (e.g., when testing novel effects or when running exploratory analyses), it may be more reasonable to embrace a maximalist approach and include all the random slopes that are justified by the study design.

Summary and recommendations. In STEP #2, you need to build two intermediate models: (i) a constrained model (*not* including the slope residuals u_{1j}) and (ii) an augmented model (including the slope residuals u_{1j}). Then you need to compare the deviance of the two models using a two-degree-of-freedom likelihood-ratio test, noted as LR χ^2. We recommend that if the p-value of the LR χ^2 (2) is less than .20, then the variance and covariance terms var(u_{1j}) and cov(u_{0j}, u_{1j}) should be kept in the model. If you have several slope residuals to test (u_{1j}, u_{2j}, u_{3j}, etc.), we advise you to calculate an LR χ^2 (2) for each of them.

Here is the extract of our mock paper pertaining to STEP #2.

> 'As a second step, we built an intermediate model using hotness and period of success as predictors, and we performed a likelihood-ratio test to see whether estimating the slope residuals improved the fit. The p-value of the LR χ^2 (2) was below 0.20, meaning that estimating the slope residual variance and the covariance terms was warranted.'

STEP #3. Building the Final Model and Interpreting the Confidence Intervals

Main question to be answered: Are your hypotheses supported?

Now that you have made a decision regarding the need to include slope residuals, you can finally include your cross-level interaction(s) (if you have one) and build your final model:

$$Y_{ij} = B_{00} + \left(B_{10} + u_{1j}\right) \times x_{ij}^{cmc} + B_{01} \times X_j + B_{11} \times x_{ij}^{cmc} \times X_j + u_{0j} + e_{ij} \quad \text{(Eq. 13)}$$

Interpretation of the main effects

The coefficient estimate of your level-1 main effect (hotness) is B_{10} = 0.03, 95% CI [–0.14, 0.21]. Because period of success is coded –0.5 = ''90s boy bands' and +0.5 = 'post-'00s boy bands,' this coefficient estimate pertains to the pooled within-boy band effect of hotness between '90s and post-'00s boy bands (i.e., when X_j = 0). Moreover, the coefficient estimate of your level-2 main effect (period of success) is B_{01} = 1.59, 95% CI [0.94, 2.24]. Because hotness is cluster-mean centered, this coefficient estimate pertains to the average effect of period of success for the typical member of a given boy band in terms of hotness (when x_{ij}^{cmc} = 0). Now that we are clear about the interpretation of the coefficient estimates, let's focus on the interpretation of the 95% confidence intervals (Cumming, 2014).

First, the 95% CI of your level-1 effect can be interpreted as follows: If we repeated the boy band study an infinite number of times, 95% of all CIs will contain the true population parameter (Morey et al., 2016). An easier (but less precise) interpretation is as follows: We can be 95% confident that the pooled within-boy band effect of hotness lies between B_{10} = –0.14 (the lower bound) and B_{10} = 0.21 (the upper bound). Here the fact that the 95% CI includes zero means that *the effect is not statistically significant* at the traditional alpha level ($p > 0.05$); we cannot be confident that the effect is negative ($-0.14 \leq B_{10} < 0$) or positive ($0 < B_{10} \leq 0.21$). Thus, we fail to reject the null hypothesis according to which there is no relationship between hotness and popularity.

Second, the 95% CI of your level-2 effect can be roughly interpreted as follows: We can be 95% confident that the effect of period of success lies between B_{01} = 0.94 (the lower bound) and B_{01} = 2.24 (the upper bound). The fact that the 95% CI does not include zero means that *the effect is statistically significant* at the traditional alpha level ($p < 0.05$); we decide that the popularity score of members from post-'00s boy bands (coded +0.5) is between 0.94 and 2.24 higher than the popularity score of members from '90s boy bands (coded –0.5). Thus, we have evidence to support the alternative hypothesis, according to which there is a positive effect of period of success on popularity.

Interpretation of the interaction effects

The coefficient estimate of your cross-level interaction (hotness × period of success) is B_{11} = –0.39, 95% CI [–0.73, –0.04]. The fact that the 95% CI does not include zero means that the pooled within-boy band effect of hotness is statistically significantly different between '90s and post-'00s boy bands. This means that we are not at the end of the road yet. The cross-level interaction now needs to be decomposed, which can be done using two dummy-coding models (e.g., see Preacher et al., 2004):

1. A first dummy-coding model aims to estimate the effect of hotness within '90s boy bands. To build this model, you have to recode period of success using '0 = '90s boy band' and '1 = post-'00s boy band,' re-compute the product terms, and then re-run the final model. In doing so, the coefficient estimate B_{10} will become the simple fixed slope of hotness when 'period of success = 0,' that is, the pooled within-boy band effect of hotness *for '90s boy bands*. Note that if period of success were treated as a continuous variable (ranging from 1990 and 2020), you would have to *add* one standard deviation to the variable to obtain the simple fixed slope of hotness when 'period of success = –1 SD' (older boy bands).
2. A second dummy-coding model aims to estimate the effect of hotness within post-'00s boy bands. To build this model, you have to recode period of success using '–1 = '90s boy band' and '0 = post-'00s boy band,' re-compute the product terms, and then re-run the final model. In doing so, the coefficient estimate B_{10} will become the simple fixed slope of hotness when 'period of success = 0,' that is, the pooled within-boy band effect of hotness *for post-'00s boy bands*. Note that if period of success were treated as a continuous variable (ranging from 1990 and 2020), you would have to *remove* one standard deviation from the variable to obtain the simple fixed slope of hotness when 'period of success = +1 SD' (newer boy bands).

Summary and recommendations. In STEP #3, you need to include the cross-level interaction(s) to build the final model. Interpret the 95% CIs: (i) if including zero, then $p > 0.05$ (H_0 is maintained); (ii) if excluding zero, then $p < 0.05$ (H_0 is rejected). When having a significant interaction, build a series of dummy-coding models to test simple slopes.

Here is the extract of our mock paper pertaining to STEP #3:

> 'As a third step, we built the final model using hotness (cluster-mean centered), period of success (–0.5 = ''90s boy bands' vs. 0.5 = 'post-'00s boy bands'), and the cross-level interaction as predictors. [...]
>
> Consistent with our third hypothesis, we observed a significant cross-level interaction between hotness and period of success, B = –0.39, 95% CI [–0.73, –0.04]. A simple slope analysis revealed that the pooled within-boy band effect of hotness was positive for '90s boy bands, B = 0.23, 95% CI [0.05, 0.41], whereas the effect was null for post-'00s boy bands, B = –0.16, 95% [–0.45, 0.14]. We called this phenomenon "The Justin Timberlake Effect."'

PART 3. A Simulation Comparing the Performances of SPSS, Stata, R, and Mplus when Running Two-Level Linear Models

Let us be honest with you: We thought that a tutorial was not a sufficient contribution for this paper to be published, so we felt compelled to run a bunch of simulations comparing the performances of SPSS, Stata, R, and Mplus. Running these simulations was nevertheless important as researchers use various statistical software programs for a variety of reasons. However, when it comes to multilevel modeling, different software programs rely on different computational and optimization techniques, which can exert an influence on statistical outcomes and—by extension—on research conclusions (Dedrick et al., 2009).

To our knowledge, there is only one simulation study that compared the performance of various statistical software programs (McCoach et al., 2018), but this study focused on differences in the variance term estimation and left aside the issue of coefficient estimation. Here we ran a simulation study that compared the performances of SPSS, Stata, R, and Mplus when using our three-step procedure to estimate a cross-level interaction coefficient. Does the choice of statistical software impact the outcome of the likelihood-ratio test estimating slope residual variance, the type II error rate (false negative), and the type I error rate (false positive) when testing a cross-level interaction?

Simulation Conditions

To answer these questions, we simulated a series of two-level datasets. The following factors were fixed across datasets:

- *Sample sizes.* Each dataset was comprised of N = 12,500 level-1 units nested in K = 50 level-2 units (resembling a small secondary survey dataset). This sample size was sufficient to detect a small cross-level interaction of $\beta_{11} = 0.10$ with $\text{var}(u_{1j}) = 0.01$, with a power of 0.80 and an alpha of 0.05 (we performed the power analysis using the R package simglm; LeBeau, 2020).
- *Variables and ICC.* Each dataset was comprised of an outcome variable, a level-1 predictor, and a level-2 predictor. All variables were drawn from a normal distribution with a mean of 0 and a standard deviation of 1. In the population, the links between the outcome and predictor variables were zero, and the ICC was 0.05 (a medium-sized ICC).

The following factors varied across the datasets:

- *Size of the cross-level interaction.* There were three conditions: The size of the cross-level interaction in the population could be small ($\beta_{11} = 0.10$), very small ($\beta_{11} = 0.05$), or zero ($\beta_{11} = 0.00$). We decided to focus on small- or less-than-small-sized interactions because

real-data interactive effects are usually much smaller than main effects (especially for attenuated interactions; see Blake & Gangestad, 2020).

- *Magnitude of the slope residual variance*. There were again three conditions: The magnitude of the slope residual variance in the population could be small ($\text{var}(u_{1j}) = 0.01$), very small ($\text{var}(u_{1j}) = 0.005$), or near zero ($\text{var}(u_{1j}) = 0.001$). We decided to focus on small- or less-than-small-sized slope residual variance because statistical software typically struggles to estimate variance parameters that are close to zero (McCoach et al., 2018). For practical reasons, the covariance parameter was set to zero.

We simulated 3 (size of the cross-level interaction: small vs. very small vs. zero) × 3 (magnitude of the slope residual variance: small vs. very small vs. near zero) × 1,000 (datasets per condition) = 9,000 datasets. The R script used to simulate the data, the complete simulated datasets, and the SPSS, Stata, R, and Mplus scripts used to perform the analysis can be found on the OSF.

Results

For each software and each simulated dataset, we built a two-level model and regressed the outcome on the level-1 predictor, the level-2 predictor, and the cross-level interaction. For each software and each condition, we calculated (i) the convergence rate (the proportion of models converging in 100 iterations, since a nonconvergence issue is a recurrent problem in multilevel modeling), (ii) the slope residual detection rate (the proportion of the significant likelihood-ratio tests with $\alpha = 0.20$), and (iii) the type I and type II error rates for the cross-level interaction (the proportion of [non]significant cross-level interactions with $\alpha = 0.05$). **Table 2** present the full set of results.

Convergence rates

Mplus, Stata, and R showed perfect or near-perfect convergence rates, regardless of the conditions (Mplus: 100%; Stata: 99%–100%; R: 97%–98%). SPSS showed perfect convergence rates when the magnitude of residual slope variance was small or even very small, but the convergence rates dropped to 88%–90% when the residual slope variance was near zero (for similar conclusions, see McCoach et al., 2018).

Slope residual variance detection rates

SPSS, Stata, R, and Mplus showed the same slope residual variance detection rates. Overall, the likelihood-ratio test correctly detected small and very small residual slope variances ≤99% of the time. However, the likelihood-ratio test was not always reliable for near-zero residual slope variance: The detection rates were satisfying when the cross-level interaction was small (≤99%) or very small (81%), but *not* when it was zero (≈53%). This means that the likelihood-ratio test may be limited when trying to detect tiny between-cluster variations of a level-1 effect. Thus, cautious analysts may favor a maximalist approach (i.e., *always* estimating random slope components when testing a cross-level interaction; Heisig & Schaeffer, 2019).

Type II and type I error rates

Conditions #1–3: When the cross-level interaction was small ($\beta_{11} = 0.10$), Stata, R, and Mplus showed similar type II error rates. Unsurprisingly, when the residual slope variance was small (Condition #1), the type II error rate was 20% (because in this condition, the statistical power was 80%). Unsurprisingly, when the residual slope variance was very small or near zero, the type II error rates dropped to 6% and 0%, respectively (generally speaking, the power to detect a cross-level interaction increases as the residual slope variance decreases; see Arend & Schäfer, 2019). Descriptively speaking, SPSS performed slightly worse than Stata, R, and Mplus (approximately +1% in terms of the type II error rate).

Conditions #4–6: When the cross-level interaction was very small ($\beta_{11} = 0.05$), Stata, R, and Mplus showed similar type II error rates. When the residual slope variance was small (Condition #4), very small (Condition #5), and near zero (Condition #6), the type II error rates were 70%, 55%, and 23%, respectively. SPSS again performed slightly worse than Stata, R, and Mplus (approximately +2–3% in terms of the type II error rate).

Conditions #7–9: When the cross-level interaction was zero ($\beta_{11} = 0.00$), all software showed similar type I error rates. When the residual slope variance was small (Condition #7), very small (Condition #8), and near zero (Condition #9), the type I error rates were 3%, 3%, and 2% (which for some reason were slightly above the alpha level), respectively.

Summary and software recommendations. The simulations revealed two critical software differences: (i) SPSS is more likely to encounter convergence issues than Stata, R, or Mplus when the slope residual variance is near zero (however, convergence issues rarely affect coefficient estimations) and (ii) SPSS was more likely than Stata, R, or Mplus to miss very small cross-level interactions (though the differences were not significant with n = 1,000 datasets per condition). In summary, SPSS is *marginally* worse than Stata, R, and Mplus in estimating two-level models. However, let's quit playing games:[8] SPSS still performs reasonably well, and the results from this simulation cannot be used to recommend one statistical software over another.

PART 4. A Q&A Addressing Multilevel Modeling Issues

Oooops... It seems we are over the IRPS word limit! Before saying bye bye bye, we would like to bring your attention to our Supplementary Materials (available on the OSF), in which you will find answers to the following four questions:

Q1. What is a sufficient sample size in multilevel modeling? (for further reading on multilevel power analysis, see Arend and Schafer, 2019)
Q2. How can I calculate the effect size in multilevel modeling? (for further reading on effect size measures for multilevel models, see LaHuis et al., 2014)
Q3. How do I handle three-level modeling and other complex multilevel designs? (for further reading

Table 2: Convergence rates, slope residual variance detection rates, type II (when $\beta_{11} \geq 0.05$) and type I (when $\beta_{11} = 0.00$) error rates as a function of the statistical software (SPSS vs. Stata vs. R vs. Mplus), the size of the cross-level interaction (small vs. very small vs. zero), and the magnitude of the slope residuals variance (small vs. very small vs. near zero).

#	Condition 3 × 3		Convergence rates Proportion of models, in 100 iterations				Slope residual variance detection rates Proportion of sig. LR χ^2 with $\alpha = 0.20$				Type II and type I error rates Proportion of [non]sig. B_{11} with $\alpha = 0.05$				
	Interaction (β_{11})	**Residuals (var(u_{1j}))**	**SPSS**	**Stata**	**R**	**Mplus**	**SPSS**	**Stata**	**R**	**Mplus**	**SPSS**	**Stata**	**R**	**Mplus**	
1.	Small (0.10)	Small (0.01)	1.00 [1.0, 1.0]	1.00 [1.0, 1.0]	.97 [0.96, 0.98]	1.00 [1.0, 1.0]	1.00 [1.0, 1.0]	1.00 [1.0, 1.0]	1.00 [1.0, 1.0]	1.00 [1.0, 1.0]	0.21 [0.19, 0.24]	0.20 [0.17, 0.22]	0.20 [0.17, 0.22]	0.20 [0.17, 0.23]	**Type II error**
2.	Small (0.10)	Very small (0.005)	1.00 [1.0, 1.0]	1.00 [1.0, 1.0]	0.98 [0.97, 0.99]	1.00 [1.0, 1.0]	1.00 [1.0, 1.0]	1.00 [1.0, 1.0]	1.00 [1.0, 1.0]	1.00 [1.0, 1.0]	0.07 [0.05, 0.08]	0.06 [0.05, 0.08]	0.06 [0.05, 0.08]	0.06 [0.05, 0.08]	
3.	Small (0.10)	Near zero (0.001)	0.89 [0.87, 0.91]	0.99 [0.98, 0.99]	0.98 [0.97, 0.99]	1.00 [1.0, 1.0]	0.99 [1.0, 1.0]	0.99 [1.0, 1.0]	0.99 [1.0, 1.0]	0.99 [1.0, 1.0]	0.00 [0.00, 0.01]	0.00 [0.00, 0.01]	0.00 [0.00, 0.01]	0.00 [0.00, 0.01]	
4.	Very small (0.05)	Small (0.01)	1.00 [1.0, 1.0]	1.00 [1.0, 1.0]	0.98 [0.97, 0.99]	1.00 [1.0, 1.0]	1.00 [1.0, 1.0]	1.00 [1.0, 1.0]	1.00 [1.0, 1.0]	1.00 [1.0, 1.0]	0.73 [0.70, 0.75]	0.70 [0.67, 0.73]	0.70 [0.67, 0.73]	0.70 [0.67, 0.73]	
5.	Very small (0.05)	Very small (0.005)	1.00 [1.0, 1.0]	1.00 [1.0, 1.0]	0.99 [0.98, 0.99]	1.00 [1.0, 1.0]	1.00 [1.0, 1.0]	1.00 [1.0, 1.0]	1.00 [1.0, 1.0]	1.00 [1.0, 1.0]	0.57 [0.54, 0.60]	0.55 [0.52, 0.58]	0.55 [0.52, 0.58]	0.55 [0.52, 0.58]	
6.	Very small (0.05)	Near zero (0.001)	0.90 [0.87, 0.91]	0.99 [0.98, 1.0]	0.98 [0.97, 0.99]	1.00 [1.0, 1.0]	0.81 [0.85, 0.83]	0.81 [0.85, 0.83]	0.81 [0.86, 0.83]	0.81 [0.85, 0.83]	0.25 [0.22, 0.28]	0.23 [0.20, 0.26]	0.23 [0.20, 0.26]	0.23 [0.21, 0.26]	
7.	Zero (0.00)	Small (0.01)	1.00 [1.0, 1.0]	1.00 [1.0, 1.0]	0.98 [0.97, 0.99]	1.00 [1.0, 1.0]	1.00 [1.0, 1.0]	1.00 [1.0, 1.0]	1.00 [1.0, 1.0]	1.00 [1.0, 1.0]	0.03 [0.02, 0.04]	0.03 [0.02, 0.04]	0.03 [0.02, 0.04]	0.03 [0.02, 0.04]	**Type I error**
8.	Zero (0.00)	Very small (0.005)	1.00 [1.0, 1.0]	1.00 [1.0, 1.0]	0.98 [0.97, 0.99]	1.00 [1.0, 1.0]	0.99 [1.0, 1.0]	0.99 [1.0, 1.0]	0.99 [1.0, 1.0]	0.99 [1.0, 1.0]	0.03 [0.02, 0.04]	0.03 [0.02, 0.05]	0.03 [0.02, 0.05]	0.03 [0.02, 0.05]	
9.	Zero (.00)	Near zero (.001)	0.88 [0.86, 0.90]	0.99 [0.99, 1.0]	0.97 [0.96, 0.98]	1.00 [1.0, 1.0]	0.53 [0.60, 0.56]	0.53 [0.60, 0.56]	0.54 [0.60, 0.57]	0.53 [0.60, 0.56]	0.02 [0.01, 0.03]	0.02 [0.01, 0.03]	0.02 [0.01, 0.03]	0.02 [0.01, 0.03]	

Notes: 95% CIs are given in brackets.

on three-level modeling, cross-classified modeling for repeated measures design, and multiple membership structures, see Peugh, 2014; Baayen et al., 2008; and Browne et al. 2001, respectively).

Q4. How do I run nonlinear two-level regression (logistic, ordered logistic, Poisson)? (for further reading on multilevel logistic modeling [binary outcome], multilevel ordered logistic modeling [ordinal outcome], and multilevel Poisson modeling [count outcome], see Sommet & Morselli, 2017; Stawski, 2013, chapter 17; and Aiken et al., 2015, respectively).

Notes

[1] For the curious minds who want to find out more about cat overweightness, this finding is explained by the fact that 'dogs might intimidate cats when they are eating and drive them away from their food' (p. 194).

[2] For boy band enthusiasts who want to know more about our sources, the data are based on the Internet Boy Band Database, which contains information about boy bands with at least one song in the U.S. Billboard Hot 100 between 1980 and 2018 (Goldenberg et al., 2018).

[3] For the smarty-pants who think that linear regression cannot be used here because the outcome is not continuous, the popularity score corresponds to the common logarithm of the number of Instagram followers (in hundreds) plus one (thus, a continuous outcome): 1 ≤ 100 Instagram followers, 2 = 1,000 followers, 3 = 10,000 followers, 4 = 100,000 followers, 5 = 1,000,000 followers, 6 = 10,000,000 followers, and 7 = 100,000,000 followers.

[4] For the attentive users who realized that their software could also use the restricted maximum likelihood (REML) estimator, know that this estimator has the particularity of generating regression coefficients and variance terms *separately* rather than jointly (i.e., in two stages rather than one). Given the way it works, REML cannot compare models with different fixed components (Peugh, 2010). Generally speaking, REML-produced estimates are less biased than ML-produced estimates when the sample is small (McNeish, 2017), but the difference between the two methods should be negligible when having $K > 25$–30 clusters (Elff et al., 2021).

[5] For the stat nerds who want to deepen their knowledge, cluster-mean centering will also change the value of the variance components (i.e., the variance of level-2 residuals, the variance of slope residuals, and the covariance term) because cluster-mean-centered variance cannot explain variance at the cluster-level (by definition), which logically results in a change in the variance partitioning (see Bell et al., 2018).

[6] For the speedsters who are tempted to cut corners and use the *p*-values given by your statistical software rather than performing the LR χ^2 (2): don't do it. These *p*-values are often biased because they are derived from tests assuming a normal distribution, whereas the distribution of variance is left-skewed (Hox, 2017).

[7] One way or another, note that a nonsignificant LR χ^2 (2) should *not* prevent you from examining a cross-level interaction and proceeding to STEP #3: The fact that the between-cluster variations of a level-1 effect are nonsignificant does not necessarily mean that these variations are absent (absence of evidence is not evidence of absence; Nezlek, 2008; for a relevant simulation study, see LaHuis & Ferguson, 2009).

[8] (with my heart).

Additional File

The additional file for this article can be found as follows:

- **Supplementary Materials.** Questions and answers Q1 to Q4. DOI: https://doi.org/10.5334/irsp.555.s1

Acknowledgements

This publication is based on research conducted at the Swiss National Center of Competence in Research LIVES—Overcoming vulnerability: Life course perspectives (NCCR LIVES), which is financed by the Swiss National Science Foundation. This work was also funded by a SFNS Ambizione fellowship awarded to the first author (subside #PZ00P1_185979). We wish to thank Anatolia ('niita) Batruch, Wojciech Świątkowski, Mengling Chen, David Weissman, and Nele Claes for their helpful comments on an earlier version of this manuscript.

Competing Interests

The authors have no competing interests to declare.

References

Abelkermit, J. R., Hazesc, J. S., Prikkitrack, C. A., Etafon, J. A., & **Ssab, J. B.** (2021a). The Justin Timberlake effect. *Journal of Humanities and Cultural Studies R&D*, *6*, 6.

Abelkermit, J. R., Hazesc, J. S., Prikkitrack, C. A., Etafon, J. A., & **Ssab, J. B.** (2021b). The Justin Timberlake effect. *International Journal of Business and Social Science Research*, *2*, 8.

Abelkermit, J. R., Hazesc, J. S., Prikkitrack, C. A., Etafon, J. A., & **Ssab, J. B.** (2021c). The Justin Timberlake effect. *Education, Society and Human Studies*, *2*, 3.

Aguinis, H., Gottfredson, R. K., & **Culpepper, S. A.** (2013). Best-practice recommendations for estimating cross-level interaction effects using multilevel modeling. *Journal of Management*, *39*, 1490–1528. DOI: https://doi.org/10.1177/0149206313478188

Aiken, L. S., Mistler, S. A., Coxe, S., & **West, S. G.** (2015). Analyzing count variables in individuals and groups: Single level and multilevel models. *Group Processes & Intergroup Relations*, *18*, 290–314. DOI: https://doi.org/10.1177/1368430214556702

Allan, F. J., Pfeiffer, D. U., Jones, B. R., Esslemont, D. H. B., & **Wiseman, M. S.** (2000). A cross-sectional study of risk factors for obesity in cats in New Zealand. *Preventive Veterinary Medicine*, *46*, 183–196. DOI: https://doi.org/10.1016/S0167-5877(00)00147-1

Arend, M. G., & **Schäfer, T.** (2019). Statistical power in two-level models: A tutorial based on Monte Carlo simulation. *Psychological Methods*, *24*, 1–19. DOI: https://doi.org/10.1037/met0000195

Baayen, R. H., Davidson, D. J., & **Bates, D. M.** (2008). Mixed-effects modeling with crossed random

effects for subjects and items. *Journal of Memory and Language*, *59*, 390–412. DOI: https://doi.org/10.1016/j.jml.2007.12.005

Barr, D. J., Levy, R., Scheepers, C., & **Tily, H. J.** (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*, 255–278. DOI: https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Kliegl, R., Vasishth, S., & **Baayen, H.** (2015). *Parsimonious mixed models*. arXiv preprint, arXiv:1506.04967. DOI: https://doi.org/10.1177/0146167220913363

Blake, K. R., & **Gangestad, S.** (2020). On attenuated interactions, measurement error, and statistical power: Guidelines for social and personality psychologists. *Personality and Social Psychology Bulletin*, *46*, 1702–1711. DOI: https://doi.org/10.1007/s11135-017-0593-5

Bell, A., Jones, K., & **Fairbrother, M.** (2018). Understanding and misunderstanding group mean centering: A commentary on Kelley et al.'s dangerous practice. *Quality & Quantity*, *52*, 2031–2036. DOI: https://doi.org/10.1177/1471082X0100100202

Browne, W. J., Goldstein, H., & **Rasbash, J.** (2001). Multiple membership multiple classification (MMMC) models. *Statistical Modelling*, *1*, 103–124. DOI: https://doi.org/10.1177/0956797613504966

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*, 7–29. DOI: https://doi.org/10.1093/acprof:oso/9780195315493.001.0001

Dattalo, P. (2008). *Determining Sample Size: Balancing Power, Precision, and Practicality*. New York, NY: Oxford University Press. DOI: https://doi.org/10.3102/0034654308325581

Dedrick, R. F., Ferron, J. M., Hess, M. R., Hogarty, K. Y., Kromrey, J. D., Lang, T. R., ..., & **Lee, R. S.** (2009). Multilevel modeling: A review of methodological issues and applications. *Review of Educational Research*, *79*, 69–102. DOI: https://doi.org/10.1017/S0007123419000097

Elff, M., Heisig, J. P., Schaeffer, M., & **Shikano, S.** (2021). Multilevel analysis with few clusters: Improving likelihood-based methods to provide unbiased estimates and accurate inference. *British Journal of Political Science*, *51*(1), 412–426. DOI: https://doi.org/10.1037/1082-989X.12.2.121

Enders, C. K., & **Tofighi, D.** (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods*, *12*, 121–128.

Goldenberg, R., Amber, T., & **Malik, Y.** (2018). *Internet boy band database: An audio-visual history of every bod band to chart on the Billboard Hot 100 since 1980*. Retrieved from https://data.world/the-pudding/internet-boy-band-database

Goldstein, H. (2013). Likelihood Estimation in Multilevel Models. In M. A. Scott, J. S. Simonoff & B. D. Marx (Eds.). *The SAGE Handbook of Multilevel Modeling* (pp. 39–52). London, UK: Sage Publications.

Hayes, A. F. (2006). A primer on multilevel modeling. *Human Communication Research*, *32*, 385–410. DOI: https://doi.org/10.1111/j.1468-2958.2006.00281.x

Heisig, J. P., & **Schaeffer, M.** (2019). Why you should always include a random slope for the lower-level variable involved in a cross-level interaction. *European Sociological Review*, *35*, 258–279. DOI: https://doi.org/10.1093/esr/jcy053

Hox, J. (2017). *Multilevel analysis: Techniques and applications* (3rd edition). New York, NY: Routledge. DOI: https://doi.org/10.4324/9781315650982

Huang, F. L. (2018). Multilevel modeling myths. *School Psychology Quarterly*, *33*, 492–499. DOI: https://doi.org/10.1037/spq0000272

Judd, C. M., McClelland, G. H., & **Ryan, C. S.** (2017). *Data analysis: A model comparison approach to regression, ANOVA, and beyond* (3rd ed.). Abingdon, UK: Routledge. DOI: https://doi.org/10.4324/9781315744131

Kish, L. (1965). Survey Sampling. New York, NY: John Wiley & Sons, Inc.

Kreft, I. G. G., & **De Leeuw, J.** (1998). *Introducing multilevel modeling*. Thousand Oaks, CA: Sage. DOI: https://doi.org/10.4135/9781849209366

Lai, M. H., & **Kwok, O. M.** (2015). Examining the rule of thumb of not using multilevel modeling: The "design effect smaller than two" rule. *The Journal of Experimental Education*, *83*, 423–438. DOI: https://doi.org/10.1080/00220973.2014.907229

LaHuis, D. M., & **Ferguson, M. W.** (2009). The accuracy of significance tests for slope variance components in multilevel random coefficient models. *Organizational Research Methods*, *12*, 418–435. DOI: https://doi.org/10.1177/1094428107308984

LeBeau, B. (2020). *simglm: Simulate models based on the generalized linear model* (R package version 0.8.0). https://CRAN.R-project.org/package=simglm

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & **Bates, D.** (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315. DOI: https://doi.org/10.1016/j.jml.2017.01.001

Maxwell, S. E., & **Delaney, H. D.** (2004). *Designing experiments and analyzing data: A model comparison perspective*. New York, NY: Psychology Press. DOI: https://doi.org/10.4324/9781410609243

McCoach, D. B., Rifenbark, G. G., Newton, S. D., Li, X., Kooken, J., Yomtov, D., ..., & **Bellara, A.** (2018). Does the package matter? A comparison of five common multilevel modeling software packages. *Journal of Educational and Behavioral Statistics*, *43*, 594–627. DOI: https://doi.org/10.3102/1076998618776348

McNeish, D. (2017). Small sample methods for multilevel modeling: A colloquial elucidation of REML and the Kenward-Roger correction. *Multivariate Behavioral Research*, *52*, 661–670. DOI: https://doi.org/10.1080/00273171.2017.1344538

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & **Wagenmakers, E. J.** (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic*

Bulletin & Review, *23*, 103–123. DOI: https://doi.org/10.3758/s13423-015-0947-8

Musca, S. C., Kamiejski, R., Nugier, A., Méot, A., Er-Rafiy, A., & **Brauer, M.** (2011). Data with hierarchical structure: impact of intraclass correlation and sample size on type-I error. *Frontiers in Psychology*, *2*, 74. DOI: https://doi.org/10.3389/fpsyg.2011.00074

Muthén, B., & **Satorra, A.** (1995). Complex sample data in structural equation modeling. *Sociological Methodology*, *25*, 267–316. DOI: https://doi.org/10.2307/271070

Myers, N. D., Brincks, A. M., & **Beauchamp, M. R.** (2010). A tutorial on centering in cross-sectional two-level models. *Measurement in Physical Education and Exercise Science*, *14*, 275–294. DOI: https://doi.org/10.1080/1091367X.2010.520247

Nezlek, J. B. (2008). An introduction to multilevel modeling for social and personality psychology. *Social and Personality Psychology Compass*, *2*, 842–860. DOI: https://doi.org/10.1111/j.1751-9004.2007.00059.x

Peugh, J. L. (2010). A practical guide to multilevel modeling. *Journal of School Psychology*, *48*, 85–112. DOI: https://doi.org/10.1016/j.jsp.2009.09.002

Peugh, J. L. (2014). Conducting three-level cross-sectional analyses. *The Journal of Early Adolescence*, *34*, 7–37. DOI: https://doi.org/10.1177/0272431613498646

Preacher, K. J., Curran, P. J., & **Bauer, D. J.** (2004). *Simple intercepts, simple slopes, and regions of significance in MLR 2-way interactions* [Computer software]. Retrieved from: http://quantpsy.org/interact/hlm2_instructions.pdf

Robson, K., & **Pevalin, D.** (2016). *Multilevel modelling in plain language*. London, UK: Sage Publications. DOI: https://doi.org/10.4135/9781473920712

Scariano, S. M., & **Davenport, J. M.** (1987). The effects of violations of independence assumptions in the one-way ANOVA. *The American Statistician*, *41*, 123–129. DOI: https://doi.org/10.1080/00031305.1987.10475459

Snijders, T., & **Bosker, R.** (2011). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd edition). Thousand Oaks, CA: Sage.

Sommet, N., & **Morselli, D.** (2017). Keep calm and learn multilevel logistic modeling: A simplified three-step procedure using Stata, R, Mplus, and SPSS. *International Review of Social Psychology*, *30*, 203–218. DOI: https://doi.org/10.5334/irsp.90

Stawski, R. S. (2013). *Multilevel analysis: An introduction to basic and advanced multilevel modeling, 2nd ed.* London, UK: Sage. DOI: https://doi.org/10.1080/10705511.2013.797841

Wang, L., Yang, M., & **Liu, X.** (2019). The impact of oversimplifying the between-subject covariance structure on inferences of fixed effects in modeling nested data. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*, 1–11. DOI: https://doi.org/10.1080/10705511.2018.1489725

How to cite this article: Sommet, N., & Morselli, D. (2021). Keep Calm and Learn Multilevel Linear Modeling: A Three-Step Procedure Using SPSS, Stata, R, and Mplus. *International Review of Social Psychology*, 34(1): 24, 1–19. DOI: https://doi.org/10.5334/irsp.555

Submitted: 23 December 2020 **Accepted:** 08 July 2021 **Published:** 09 September 2021

]u[*International Review of Social Psychology* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS