INTERNATIONAL REVIEW OF SOCIAL PSYCHOLOGY

Sommet, N. and 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(1), 203–218, DOI: https://doi.org/10.5334/irsp.90

RESEARCH ARTICLE

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

Nicolas Sommet and Davide Morselli

This paper aims to introduce multilevel logistic regression analysis in a simple and practical way. First, we introduce the basic principles of logistic regression analysis (conditional probability, logit transformation, odds ratio). Second, we discuss the two fundamental implications of running this kind of analysis with a nested data structure: In multilevel logistic regression, the odds that the outcome variable equals one (rather than zero) may vary from one cluster to another (i.e. the intercept may vary) and the effect of a lower-level variable may also vary from one cluster to another (i.e. the slope may vary). Third and finally, we provide a simplified three-step "turnkey" procedure for multilevel logistic regression modeling:

- Preliminary phase: Cluster- or grand-mean centering variables
- Step #1: Running an empty model and calculating the intraclass correlation coefficient (ICC)
- Step #2: Running a constrained and an augmented intermediate model and performing a likelihood ratio test to determine whether considering the cluster-based variation of the effect of the lower-level variable improves the model fit
- Step #3 Running a final model and interpreting the odds ratio and confidence intervals to determine whether data support your hypothesis

Command syntax for Stata, R, Mplus, and SPSS are included. These steps will be applied to a study on Justin Bieber, because everybody likes Justin Bieber.[1]

Keywords: Logistic regression; multilevel logistic modeling; grand-mean centering and cluster-mean centering; intraclass correlation coefficient; likelihood ratio test and random random slope variance; three-step simplified procedure; Justin Bieber

It's a bad day. You've asked your colleague whether you could run a linear regression analysis with a yes/no outcome variable. "No, you must do *logistic* regression, duh!" he replied. Then, you've asked him whether you could run this logistic regression analysis, knowing that you have surveyed various pupils from different classrooms. "No, you must do *multilevel* regression, duh!" he replied. You're infuriated. You've no idea what multilevel logistic regression is. And you don't want to ask your damned colleague, who keeps patronizing you. Well, keep calm, this article is made for you.

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

Corresponding authors: Nicolas Sommet (nicolas.sommet@unil.ch), Davide Morselli (davide.morselli@unil.ch). Stata- and SPSS-related correspondences should be addressed to Nicolas Sommet, whereas R- or Mplus-related correspondences should be addressed to Davide Morselli.

The general aim of multilevel logistic regression is to estimate the odds that an event will occur (the yes/no outcome) while taking the dependency of data into account (the fact that pupils are nested in classrooms). Practically, it will allow you to estimate such odds as a function of lower level variables (e.g. pupil's age), higher level variables (e.g. classroom size), and the way they are interrelated (cross-level interactions).

Multilevel logistic regression can be used for a variety of common situations in social psychology, such as when the outcome variable describes the presence/absence of an event or a behavior, or when the distribution of a continuous outcome is too polarized to allow linear regression. For instance, multilevel logistic regression has been used to test the influence of individuals' experience of a negative life event and the quality of their neighborhood on the odds of depression (Cutrona et al., 2005), the influence of employees' job satisfaction and the size of their department on the odds of turnover (Felps et al., 2009), or the influence of grant applicants' gender and the gender of their reviewers on the odds of funding

(Mutz, Bornmann & Daniel, 2015). Multilevel modeling can also be applied to repeated measures designs (see the first paragraph of the conclusion). For instance, if participants are primed with pictures, using such an approach will enable advanced users to take both between-stimuli and between-participant variations into account (Judd, Westfall & Kenny, 2012).

In this paper, we will first explain what logistic regression is. Second, we will explain what multilevel logistic regression is. Third, we will provide a simplified and ready-to-use three-step procedure for Stata, R, Mplus, and SPSS (n.b., SPSS is not the most suitable software for multilevel modelling and SPSS users may not be able to complete the present procedure – is it too late now to say sorry?).

"No Pressure." What Logistic Regression Is

A Very Brief Recap on Linear Regression

Assume you have conducted a study involving $N = 2{,}000$ pupils in which you wanted to test the relationship between pupil achievement and (great!) musical taste. Your predictor variable is pupil's Grade Point Average (GPA), which can range from 1 to 4. Your outcome variable is the number of hours per week pupils spent listening to Justin Bieber (see **Figure 1**). You have formulated the (pro-Justin) hypothesis that GPA should be a positive predictor of the time spent listening to Justin Bieber. In this situation, you perform a simple linear regression analysis. Make sure you are familiar with the linear regression equation below (Eq. 1).

$$Y_i = B_0 + B_1 * X_i + e_i \quad (1)$$

...in which Y_i is the observed value of the outcome variable for a pupil i (number of hours per week spent listening to Justin Bieber), whereas X_i is the observed value of the predictor variable for a pupil i (his/her GPA);

...B_0 is the predicted value of Y_i when $X_i = 0$ (i.e. the intercept), whereas B_1 is the coefficient estimate describing the relationship between X_i and Y_i (i.e. the slope);

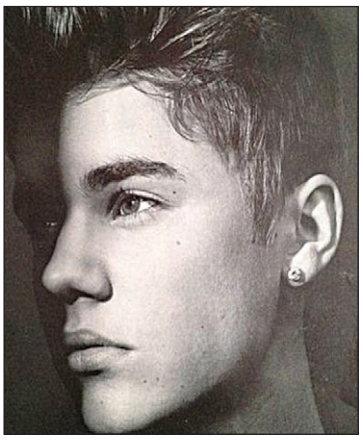

Figure 1: Justin Bieber. *Note*: This file is licensed under the Creative Commons Attribution-Share Alike 3.0 Unported license (https://commons.wikimedia.org/wiki/File:The_Bet_Justin_Bieber_y_T%C3%BA_Novela_Escrita_por_@Pretty_Jezzy_01.jpg).

...and e_i is the residual, that is, the difference between what is predicted by the regression model for a pupil i and what is *actually* observed for this pupil i.

If there were only one statistical index to remember, this would be B_1. Let's say that $B_1 = 2.00$. This indicates that an increase of one unit in GPA results in an expected increase of 2 hours per week spent listening to Justin Bieber. There are two possible scenarios:

i. B_1 is *not* significantly different from zero (formally speaking, this means that the residual e_i does *not* significantly diminish when including $B_1 * X_i$ in the model). In such a situation, you cannot reject the null hypothesis (H_0): There is no significant relationship between pupil achievement and time spent listening to Justin Bieber.
ii. B_1 is significantly different from zero (formally speaking, this means that the residual e_i does significantly diminish when including $B_1 * X_i$ in the model). In such a situation, you reject H_0 and accept the alternative hypothesis (H_1): There is a significant relationship between pupil achievement and time spent listening to Justin Bieber. Consistent with your prediction, brighter kids do seem to love Justin more and you're ready to submit your result to *Popstar! Magazine* or *Teen Vogue*. For more detailed information on linear regression analysis, see Judd, McClelland and Ryan, 2017 (French readers may see Judd, McClelland, Ryan, Muller & Yzerbyt, 2010).

From Linear to Logistic Regression

Now assume you have operationalized your outcome variable differently. Rather than self-reporting the number of hours per week spent listening to Justin Bieber, pupils have indicated whether they own *Purpose,* Justin Bieber's last album. Your outcome variable is binary, in that it can only take one of two values: 0 for "No, I don't own the album" and 1 for "Yes, I own the album."

With such a variable, a linear regression analysis is not appropriate. The main reason is that, in a linear regression analysis, the predicted value of the numeric outcome variable can take *any* value between $-\infty$ and $+\infty$ (i.e. mathematically speaking, the predicted value is not bounded). Thus, if you run a linear regression analysis using a binary outcome variable, the output might be under 0 or above 1 (i.e. it don't make no sense). To fix this, the response function should be constrained and logistic regression analysis should be used.

Whereas linear regression gives *the predicted mean value of an outcome variable* at a particular value of a predictor variable (e.g. the number of hours per week spent listening to Justin Bieber for a pupil having a GPA of 3), logistic regression gives *the conditional probability that an outcome variable equals one* at a particular value of a predictor variable (e.g. the likelihood of owning Justin's last album for a pupil having a GPA of 3). The logistic function is used to predict such a probability. It describes the relationship between a predictor variable

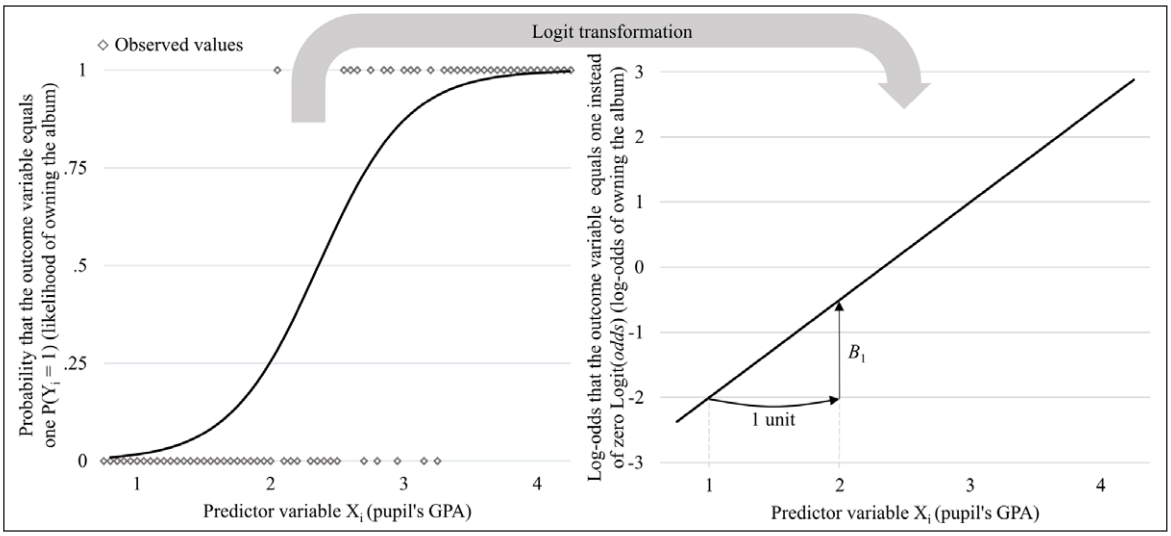

Figure 2: The logistic function describes the *s*-shaped relationship between a predictor variable X_i and the probability that an outcome variable equals one $P(Y_i = 1)$ (left panel, corresponding to Eq. 2); using the logit transformation, one can "linearize" this relationship and predict *the log-odds* that the outcome variable equals one instead of zero Logit(P(*odds*)) (right panel, corresponding to Eq. 3). *Notes*: Data are fictitious and do not correspond to the provided dataset.

X_i (or a series of predictor variables) and the conditional probability that an outcome variable Y_i equals one (owning the album). This is an *s*-shaped function: The logistic regression curve is steeper in the middle, and flatter at the beginning (when approaching 0), and at the end (when approaching 1; see **Figure 2**, left panel). The function can be represented using the equation below (Eq. 2).

$$P(Y_i = 1) = \frac{exp(B_0 + B_1 * X_i)}{1 + exp(B_0 + B_1 * X_i)} \quad (2)$$

…in which $P(Y_i = 1)$ is the conditional probability that the outcome variable equals one for a pupil *i* (that s/he owns Justin's last album);

…and *exp* is the exponent function: "$B_0 + B_1 * X_i$" are defined in the same way as in Eq. 1, although a probability is now predicted through a function.

Taking a look at the *exp* stuff, you might have the feeling that you are lost. Admittedly, the equation seems unintelligible. Fortunately, the *logit transformation* can be used to convert the *s*-shaped curve into a straight line and facilitate the reading of the results (for a graphical representation of such a transformation applied to our example, take a look at both panels of **Figure 2**). Instead of predicting the conditional probability that the outcome variable equals one, we can predict *the logit of the conditional probability* that the outcome variable equals one (owning Justin's album) over the probability that it equals zero (*not* owning Justin's album). We will refer to this as the *log-odds* (or *logit of the odds*). Odds correspond to the possibility that something will happen rather than not. For instance, the odds of being on a plane with a drunken pilot are reported to be "1 to 117" (i.e. 1:117; see Jaeger, 2008). In another example, one can calculate that the odds of an American female teenager having dated Justin Bieber are about 1 in 2,500,000.[2] However, the logit function is the natural logarithm of the odds, and the post-logit transformation logistic regression equation – which is strictly equivalent to Eq. 2 – is as follows (Eq. 3):

$$\text{Logit}(odds) = B_0 + B_1 * X_i \quad (3)$$

…in which Logit(*odds*) is the log-odds; it formally corresponds to $\text{Logit}(P(Y_i = 1)/(1 - P(Y_i = 1))$, namely the logit of the conditional probability that the outcome variable equals one (owning Justin's album) divided by the probability that it equals zero (not owning Justin's album).

Again, focus on B_1. This time, let's say that $B_1 = 1.50$. This indicates that an increase of one unit in GPA results in an expected increase of 1.5 points in the log-odds of owning Justin's last album. Hard to interpret, right?

To interpret B_1, raise it to the exponent to obtain an odds ratio, noted *OR*. Formally, the odds ratio refers to the multiplicative factor by which the predicted probability of an event occurring rather than not occurring (i.e. "$P(Y_i = 1)/1 - P(Y_i = 1)$") changes when the predictor variable X_i increases by one unit. In our example, $OR = exp(B_1) = exp(1.50) \approx 4.5$, indicates that the odds of owing Justin's album (instead of not owning it) are 4.5:1, that is, *multiplied* by 4.5/1 = 4.5 when GPA increases by one unit. Simply put, pupils are 4.5 times *more* likely to own the album when GPA increases by one unit (a 350% *increase*). Now imagine that the sign of B_1 is negative, that is, $B_1 = -1.50$. In such a case, $OR = exp(B_1) = exp(-1.50) \approx 0.22$ indicates that the odds of owning Justin's album (instead of not owning it) are 1:0.22, that is, *divided* by $1/0.22 \approx 4.5$ when GPA increases by one unit. Simply put, pupils are 4.5 times *less* likely to own the album when GPA increases by one unit (a 350% *decrease*). As earlier, there are two possible scenarios:

i. *OR* is *not* significantly different from 1 (or, equivalently, *B* is not significantly different from 0). In practice, this indicates that the odds of an event occurring is multiplied by one when the predictor variable increases by one unit (i.e. the odds remain the same). In such a situation, you cannot reject H_0.
ii. *OR* is significantly different from 1 (or, equivalently, *B* is significantly different from 0). As in the above example, if *OR* > 1, the higher the predictor variable, the *higher* the odds of the event occurring (a positive effect). Conversely, if *OR* < 1, the higher the predictor variable, the *lower* the odds of the event occurring (a negative effect). In such a situation, you reject H_0.

As you may have realized, there is another important difference between the linear and logistic regression model. This concerns residuals. With linear regression, you try to predict a *concrete* value, which may differ from what is actually observed for Y_i. As said earlier, the distance between the predicted value and the observed value is the residual e_i. Residuals can take a bunch of values (within the range of your outcome variable) and are assumed to follow a normal distribution (normality of the residual distribution is an assumption of linear regression). The residual is necessary and appears in the linear regression equation (cf. Eq. 1).

With logistic regression, you do *not* try to predict a concrete value, but a probability. Technically, the distance between this probability and the observed value can only take one of two values: "0 − $P(Y_i = 1)$" when the pupil does not own the album and "1 − $P(Y_i = 1)$" when the pupil does own the album, thereby following a binomial distribution. The residual is therefore not necessary and does not appear in the logistic regression equation (cf. Eq. 2 and, by extension, Eq. 3). For more detailed information on logistic regression analysis, see Hosmer and Lemeshow, 2000; Menard, 2002.

"What Do You Mean?" What Multilevel Logistic Regression Is

General Principles of Multilevel Logistic Regression

Now assume your study involves *N* = 2,000 pupils from *K* = 100 classrooms. That is, you have *N* participants (level-1 units) nested in *K* clusters (level-2 units; for a graphical representation of this data structure, see **Figure 3**). Classrooms pertain to a level (rather than a predictor variable), since (a) classrooms were randomly sampled from a population of units (classrooms around the world are potentially infinite and you have sampled some of them), and (b) classrooms have no intrinsic meaning *per se* (classrooms are interchangeable units without theoretical content). On the contrary, socioeconomic status would for instance pertain to a predictor variable (rather than a level) since its categories are both non-random and theoretically meaningful (e.g. lower, middle, and upper class are *not* "atheoretical" random units). Other examples of nested data are employers nested in firms, inhabitants nested in provinces and even observations nested in participants in repeated measure designs.

With such a data structure, you cannot run a standard logistic regression analysis. The reason is that this violates one of the most important assumptions in the linear model, namely the assumption of independence (or lack of correlation) of the residuals (Bressoux, 2010). Observations are interdependent: Participants nested in the same cluster are more likely to function in the same way than participants nested in different clusters. In our example, there might be some classrooms in which Justin Bieber is worshipped (with pupils having more chances to own Justin's last album) and other classrooms in which Justin Bieber is abhorred (with pupils having less chances to own Justin's last album). Multilevel (logistic) modeling notably aims to disentangle the within-cluster effects (the extent to which some participant characteristics are associated with the odds of owning Justin's last album) from the between-cluster effects (the extent to which some classroom characteristics are associated with the odds of owning Justin's last album).

What about sample size? Sufficient sample size is one of the first indications of research quality (Świątkowski & Dompnier, 2017). In multilevel modelling, the number of clusters is more important than the number of observations per cluster (Swaminathan, Rogers & Sen, 2011). In multilevel linear modeling, simulation studies show that 50 or more level-2 units are necessary to accurately estimate standard errors (Maas & Hox, 2005; see also Paccagnella, 2011). More to the point, in multilevel logistic modeling, Schoeneberger (2016) showed that a minimum of 50 level-1 units and 40 level-2 units are needed to accurately estimate small fixed effects (set at *OR* = 1.70) when intercept variance is small (set at $\text{var}(u_{0j}) \approx 0.1$), whereas 100 level-1 units and 80 level-2 units are needed when estimating cross-level interaction effects and/or when intercept variance is large (set at $\text{var}(u_{0j})$

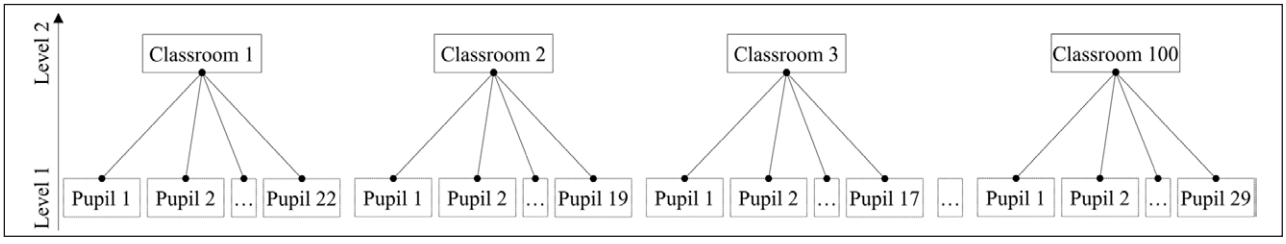

Figure 3: Example of a hierarchical data structure, in which *N* participants (pupils, lower-level units) are nested in *K* clusters (classrooms, higher-level units). *Notes*: Multilevel modeling is flexible enough to deal with this kind of unbalanced data, that is, having unequal numbers of participants within clusters.

Table 1: Summary of main notation and definition (level-1 and level-2 sample size and variables, as well as fixed and random intercept and slope).

Sample size	N Level-1 sample size (number of observations)	K Level-2 sample size (number of clusters)
Variables	$x1_{ij}$, $x2_{ij}$, ..., xN_{ij} Level-1 variables (observation-related characteristics)	$\mathbf{X1}_j$, $\mathbf{X2}_j$, ..., $\mathbf{XK}_j$ Level-2 variables (cluster-related characteristics)
Intercept	B_{00} Fixed intercept (average log-odds that the outcome variable equals one instead of zero Logit(P(*odds*)), when all predictor variables are set to zero)	u_{0j} Level-2 residual (deviation of the cluster-specific log-odds that the outcome variable equals one instead of zero from the fixed intercept; the variance component var(u_{0j}) is the random intercept variance)
Level-1 effect	B_{10}, B_{20}, ..., B_{N0}, Fixed slope (average effect of a level-1 variable in the overall sample; it becomes the odds ratio when raised to the exponent $exp(B) = OR$)	u_{1j}, u_{2j}, ..., u_{Nj} Residual term associated with the level-1 predictor $x1_{ij}$, $x2_{ij}$, ..., xN_{ij} (deviation of the cluster-specific the effect of the level-1 variable from the fixed intercept; the variance component var(u_{1j}) is the random slope variance)
Level-2 effect	B_{01}, B_{02}, ..., B_{0K}, Necessarily fixed slope (average effect of a level-2 variable in the overall sample; it becomes the odds ratio when raised to the exponent $exp(B) = OR$)	

Notes: For the sake of simplicity, no distinction is made between sample and population parameters and only Latin letters are used.

≈ 0.5). Insufficient sample size obviously reduces statistical power (the probability of "detecting" a true effect); moreover, insufficient sample size at level 2 increases Type I error rates pertaining to level-2 fixed effect (the risk of "detecting" a false effect; for another simulation study, see Moineddin, Matheson & Glazier, 2007). To more accurately detect the bias in the regression coefficients and standard errors due to sample size at both levels, advanced users should consider doing a Monte Carlo study (e.g. Muthèn & Muthèn, 2002).

Having two levels has two implications. First, the (log-) odds that the outcome variable equals one instead of zero will be allowed to vary between clusters (in our example the chances of owning Justin's last album may be allowed to vary from one classroom to another). Specifically, we will differentiate between the average log-odds that the outcome variable equals one in the overall sample (later referred to as the fixed intercept) and the variation of this log-odds from one specific cluster to another (later referred to as forming the random[3] intercept variance). Second, the *effect* of a lower-level variable on the (log-) odds that the outcome variable equals one instead of zero will also be allowed to vary between clusters (in our example the effect of GPA may also be allowed to vary from one classroom to another). Specifically, we will differentiate between the average effect of the lower-level variable in the overall sample (later referred to as the fixed slope) and the variation of this effect from one specific cluster to another (later referred to as forming the random slope variance; see **Table 1** for a summary and a definition of the key notions and notations).

A First Implication: The Log-Odds May Vary Between Clusters

The first difference between simple and multilevel logistic regression is that the log-odds that the outcome variable equals one instead of zero is allowed to vary from one cluster to another. To illustrate this, go back to your study and imagine building an empty multilevel logistic model. This model still aims to estimate the log-odds of owning Justin's album, while including no predictors. This empty multilevel logistic regression equation is shown below (Eq. 4):

$$\text{Logit}(odds) = B_{00} + u_{0j} \quad (4)$$

...in which Logit(*odds*) is the log-odds that the outcome variable equals one instead of zero (i.e. the chance that a pupil *i* from a classroom *j* owns Justin's last album).

...and B_{00} is the fixed intercept, whereas u_{0j} is the deviation of the cluster-specific intercept from the fixed intercept (i.e. the level-2 residual).

First, remember that we are not trying to predict the log-odds of owning Justin's album for a simple participant *i*; we are trying to predict such log-odds for a participant *i* in a classroom *j*.

Second, we are now estimating two types of parameters pertaining to the intercept: The *fixed intercept* and the *random intercept variance*. Let's take things one step at a time. The fixed intercept B_{00} is a general constant term. Since there are no predictors here, the fixed intercept B_{00} corresponds to the overall log-odds of owning Justin's album (instead of not owning it) for a typical pupil belonging to a typical classroom. Keep in mind that we are still estimating the log-odds (or *the logit of* the odds). If you want to calculate the average probability of owning the album, you must convert the fixed intercept using the logit transformation (see Eq. 2 and Eq. 3). In your study, $B_{00} = 0.10$, thus $P(Y_{ij} = 1) = exp(B_{00})/(1 + exp(B_{00}) = 1.10/(1 + 1.10) \approx 0.52$, that is, pupils have on average 52% chances of owning Justin's album *across all classrooms*.

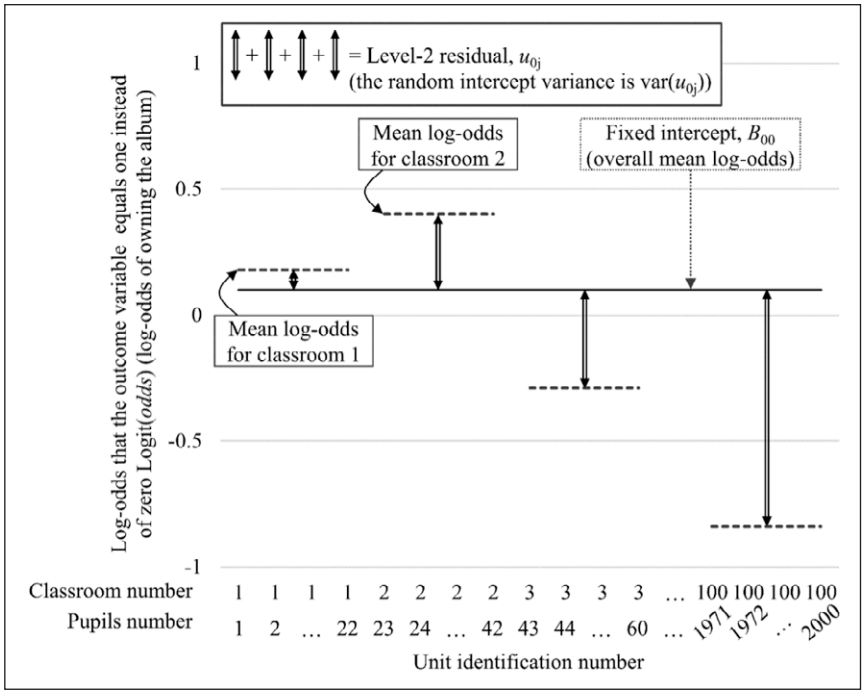

Figure 4: Graphical representation of the fixed intercept B_{00} and the level-2 residual u_{0j} (cf. Eq. 4); the fixed intercept B_{00} corresponds to the overall mean log-odds of owning Justin's album across classrooms; the random intercept variance var(u_{0j}) corresponds to the variance of the deviation of the *classroom-specific* mean log-odds from the *overall* mean log-odds (here represented by the double-headed arrow for the 1st, 2nd, 3rd, and 200th classrooms only). *Notes*: Data are fictitious and do not correspond to the provided dataset.

But as noted earlier, the log-odds may vary from one cluster to another. In other words, the intercept is not the same in every cluster. The level-2 residual u_{0j} will provide information regarding the extent of the intercept variation. Since there are no predictors here, the level-2 residual u_{0j} corresponds to the deviation of the *specific* log-odds of owning Justin's album in a given classroom from the *overall* log-odds of owning Justin's album across all classrooms (the mean of these deviations is assumed to be zero). The variance component of such a deviation is the random intercept variance var(u_{0j}). This is the key element here: The higher the random intercept variance, the larger the variation of the log-odds of owning Justin's album from a cluster to another; this indicates that pupils have more chances of owning Justin's album in some classrooms than in others (for a graphical representation of the fixed intercept and random intercept variance, see **Figure 4**).

A Second Implication: The Effect of a Lower-level Variable May Vary Between Clusters

The second change with multilevel logistic modeling is that the effect of a lower-level variable is allowed to vary from one cluster to another. Before going into details, we should distinguish between level-1 variables, noted x_{ij} (in lowercase), and level-2 variables noted $\mathbf{X}_j$ (in uppercase and bold). On the one hand, level-1 variables are lower-level observation characteristics (e.g. pupil's age). The value of a level-1 variable may change within a given cluster (there might be pupils of different ages within the same classroom). On the other hand, level-2 variables are higher-level cluster characteristics (e.g. class size). The value of a level-2 variable *cannot* change within clusters (class size is obviously the same for all pupils within a classroom).

We can understand that the *effect* of a level-1 variable – but not that of a level-2 variable(!) – may vary from one cluster to another. For instance, the effect of pupil's age on some outcome variable may be positive in some classrooms and negative in others. This also means that the average effect could be statistically non-significant, because it is positive in half of the classrooms and negative in the other half. Hence, considering only the fixed effect in the presence of between-classroom differences may wrongly lead one to conclude that the effect is negligible, when in fact the effect is positive (or stronger) in some clusters and negative (or weaker) in others.

To illustrate this, go back to your study and imagine building a simple multilevel logistic regression model. This model aims to estimate the log-odds of owning Justin's album using GPA as the sole predictor. This simple multilevel logistic regression equation is shown below (Eq. 5):

$$\text{Logit}(odds) = B_{00} + (B_{10} + u_{1j}) * x_{ij} + u_{0j} \qquad (5)$$

…in which x_{ij} is the observed value of the predictor variable for a pupil *i* in a classroom *j* (his/her GPA);

…and B_{10} is the fixed slope, whereas u_{1j} is the deviation of the cluster-specific slope from the fixed slope (i.e. the residual term associated with the level-1 variable).

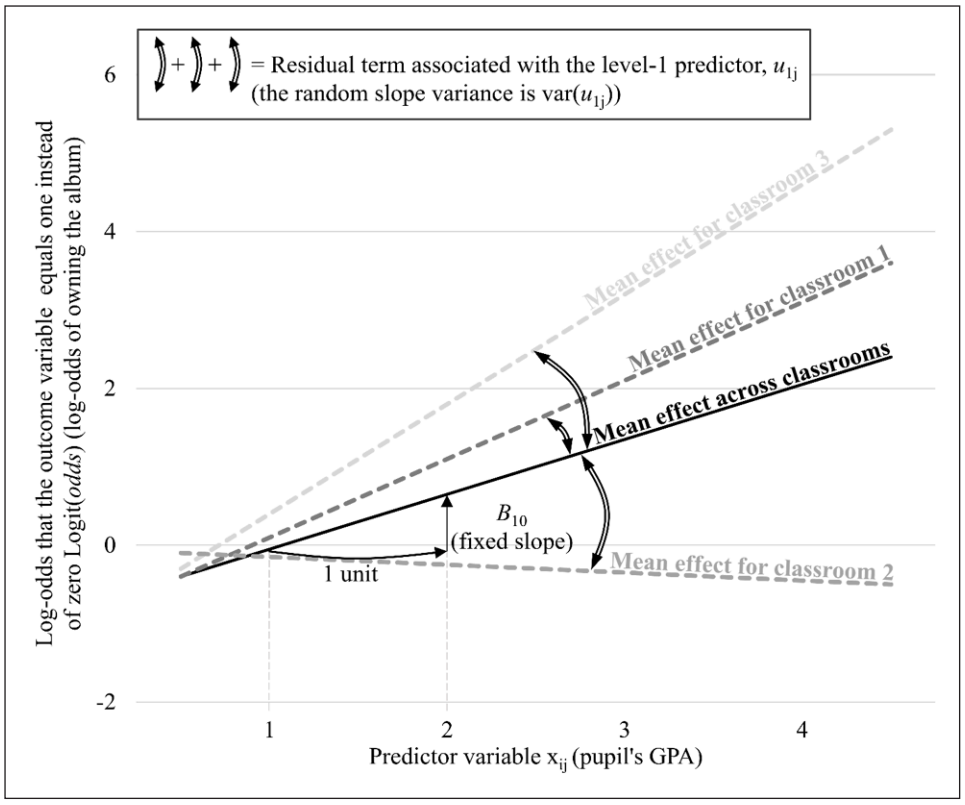

Figure 5: Graphical representation of the fixed slope B_{10} and the residual term associated with the level-1 predictor u_{1j} (cf. Eq. 5); the fixed slope B_{10} corresponds to the overall effect of pupil's GPA on the log-odds of owning Justin's album across classrooms; the random intercept variance var(u_{0j}) corresponds to the variance of the deviation of the *classroom-specific* effects of pupil's GPA from the *overall* effect of pupil's GPA (here represented by the double-headed arrow for the 1st, 2nd, and 3rd classroom). *Notes*: Data are fictitious and do not correspond to the provided dataset.

In addition to (still) having two types of parameters pertaining to the intercept (the fixed intercept B_{00} and the random intercept variance var(u_{0j}), we now have two types of parameters pertaining to the level-1 effect: The *fixed slope* and the *random slope variance*. Again, let's take things one step at a time.

The fixed slope B_{10} is the general effect of the level-1 variable x_{ij}. The interpretation is similar to the case of a single-level logistic regression analysis: An increase of one unit in GPA results in a change of B_{10} in the overall log-odds of owning Justin's album for a typical pupil belonging to a typical classroom. Once again, in order to interpret B_{10}, raise it to the exponent to obtain the odds ratio. In your study, $B_{10} = 0.70$, $OR = exp(B_{10}) = exp(0.70) \approx 2$, that is, when GPA increases by one unit, pupils are twice as likely to own Justin's album instead of not owning it *across all classrooms* (i.e. a 100% increase).

Just as for the intercept, this effect may vary from one cluster to another. The residual term associated with the level-1 predictor u_{1j} will provide information regarding the extent of the effect variation. Specifically, this residual u_{1j} corresponds to the deviation of the *specific* effects of the level-1 variable x_{ij} in a given classroom from the *overall* effect of the level-1 variable x_{ij} across all classrooms (the mean of these deviations is assumed to be zero). The variance component of such a deviation is the random slope variance var(u_{1j}). This is the key element here: The higher the random slope variance, the larger the variation of the effect of GPA from a cluster to another (for a graphical representation of the fixed intercept and the random slope variance, see **Figure 5**). Note that a non-significant random slope variance would mean that the variation of the effect of GPA is very close to zero and that B_{10} is virtually the same in all the classrooms. For more detailed information on multilevel logistic regression, see Heck, Thomas & Tabata, 2013; Rabe-Hesketh & Skrondal, 2012b; Snijders & Bosker, 2004.

"I'll Show You." A Three-Step Simplified Procedure for Multilevel Logistic Regression

You should have understood that: (a) multilevel logistic regression enables one to predict the log-odds that an outcome variable equals one instead of zero (mark my words: some software packages, e.g. SPSS, do the opposite and estimate the probability of the outcome being zero instead of one),[4] (b) the average log-odds is allowed to vary from one cluster to another (forming the random intercept variance), and (c) a lower-level effect may also be allowed to vary from one cluster to another (forming the random slope variance).

But where should we begin when running the analysis? We propose a three-step "turnkey" procedure for multilevel logistic regression modeling (summarized in **Figure 6**), including the command syntax for Stata

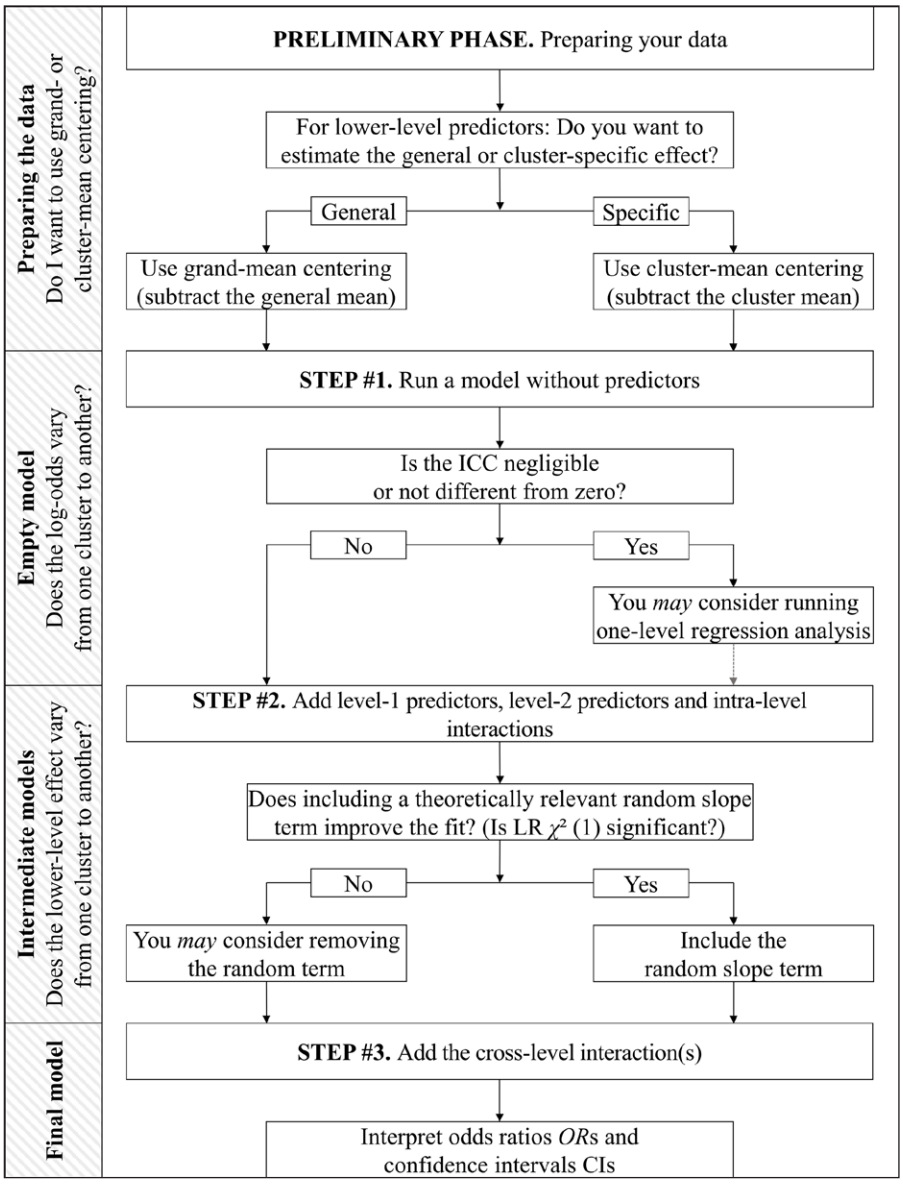

Figure 6: Summary of the three-step simplified procedure for multilevel logistic regression.

(Stata/SE version 13.1), R (using the lme4 library; Bates, Maechler, Bolker & Walker, 2015; version 1.1–12), Mplus (version 8), and SPSS (version 24, although having several limitations). The steps of the procedure are as follows:

- Preliminary phase: Preparing the data (centering variables)
- Step #1: Building an empty model, so as to assess the variation of the log-odds from one cluster to another
- Step #2: Building an intermediate model, so as to assess the variation of the lower-level effect(s) from one cluster to another
- Step #3: Building a final model, so as to test the hypothesis(/-es)

It should be stressed that this is a simplified version of the procedure usually found in the literature (e.g. Aguinis, Gottfredson & Culpepper, 2013; Hox, 2010). The few limitations due to such a simplification are footnoted.

Examples, Dataset, and Syntax Files

Let's go back to our example, that is, your N = 2,000 pupils nested from K = 100 classrooms. Now imagine you have two predictor variables. The first predictor variable is still GPA (ranging from 1 to 4). Again, this is a level-1 variable, since it may vary *within* clusters (i.e. pupils in any one classroom may have different GPA). The second predictor is called "teacher's fondness for Bieber": This is whether the classroom teacher is a "belieber" (i.e. a fan of Justin Bieber), coded 0 for "the teacher is not a belieber" and 1 for "the teacher is a belieber." This is a level-2 variable, since it *cannot* vary within clusters (i.e. pupils in any one classroom necessarily have the same teacher). For the sake of simplicity, we will assume that there is only one teacher per classroom and that teachers are all from the same schools. Note that the independence assumption should be met for level-2 residuals (e.g. classroom teachers need not to be nested in different higher-level units, such as schools, neighborhoods, or countries).

You're still trying to predict the odds of owning Justin's last album and you formulate two hypotheses. First, you predict that pupils' music taste will be influenced by their teacher's music taste (a teacher-to-pupils socialization hypothesis):

> *The main effect hypothesis*: Pupils have more chance to own Justin's album when their teacher is a belieber than when s/he is not.

Second, you posit that high-achievers tend to self-identify more with their teachers and, as such, that they should be particularly influenced by their teacher's music taste; conversely, low-achievers should be less influenced by their teacher:

> *The cross-level interaction hypothesis*: Pupils with a high GPA have more chance to own Justin's album when their teacher is a belieber than when s/he is not; this effect will be weaker for pupils with a low GPA.

The (fictitious) dataset is provided as supplementary material, in .csv format, .dta (for Stata), .rdata (for R), .dat (for Mplus), and .sav (for SPSS). You will find the syntax files in .do format (for Stata), .R (for R), .inp (for Mplus), and .sps (for SPSS), all of which provide the commands to be used at each stage of the procedure. In running the syntax file you will obtain the same estimates as those reported in the main text. In addition, for each software, a series of sub-appendices also provided in supplementary material describes the way to handle each stage of the procedure, namely:

- The preliminary phase: Sub-Appendix A
- Step #1: Sub-Appendix B
- Step #2: Sub-Appendix C
- Step #3: Sub-Appendix D

Thus, you can read the three-step simplified procedure while working on your favorite software and/or going back and forth between the main text and the Sub-Appendixes A–D. As mentioned in the opening paragraph, SPSS users may not be able complete the procedure as the software often, if not always, fails to estimate the random slope variance (in Step #2). Supplementary material (i.e. datasets, syntax files, and appendices) can be found at https://figshare.com/s/78dd44e7a56dc19d6eaa (DOI: https://doi.org/10.6084/m9.figshare.5350786).

Preliminary Phase. Preparing the Data: Centering

First and foremost, you might decide to center the predictor variable(s). Although not strictly speaking necessary, such a decision may facilitate the interpretation of some estimates. In particular, the fixed intercept will become the log-odds that your outcome variable equals one when predictor variables are all set to their mean (B_{00} is the value of Logit(*odds*) when $x1_{ij}$, $x2_{ij}$, ..., xN_{ij}, $\mathbf{X1}_j$, $\mathbf{X2}_j$, ..., and $\mathbf{XK}_j = 0$).

Centering a predictor variable depends on the level to which it is located. A level-2 predictor variable $\mathbf{X}_j$ can only be grand-mean centered (i.e. one should subtract the *general* mean across level-2 units from the predictor variable), whereas a level-1 variable could either be (a) grand-mean centered or (b) cluster-mean centered (for Stata, R, Mplus, and SPSS commands, see the relevant Sub-Appendix A).

When grand-mean centering a level-1 variable $x1_{ij}$, that is, when subtracting the *general* mean of the predictor variable ($x1gc_{ij} = x_{ij} - \bar{x}_{00}$), the fixed slope B_{10} corresponds to the average general effect. A one-unit increase in the grand-mean centered level-1 variable $x1gc_{ij}$ results in an average change of B_{10} in the log-odds that the outcome variable equals one *for the overall sample*. In our example, the fixed slope of the grand-mean centered GPA would pertain to the estimation of *the general between-pupil effect* of GPA, regardless of the classroom.

However, when cluster-mean centering a level-1 variable $x1_{ij}$, that is, when subtracting the *cluster-specific* mean of the predictor variable ($x1cc_{ij} = x_{ij} - \bar{x}_{0j}$), the fixed slope B_{10} corresponds to the cluster-specific effect. A one-unit increase in the cluster-mean centered level-1 variable $x1cc_{ij}$ results in an average change of B_{10} in the log-odds that the outcome variable equals one *for a typical cluster*. In our example, the fixed slope of the cluster-mean centered GPA would pertain to the estimation of the *within-classroom effect* of GPA, comparing the pupils nested in the same classroom (the difference between the higher and lower achievers from one class).

Beware that the type of centering (cluster- vs. grand-mean) may affect your model and results. The choice of one over the other should be done depending on your specific research question. For instance, grand-mean centering is recommended if you are interested in the effect of a level-2 predictor variable or the absolute (between-observation) effect of a level-1 predictor variable, whereas cluster-mean centering is recommended when the focus is on the relative (within-cluster) effect of a level-1 variable. In our example, you decide to cluster-mean center pupils' GPA (i.e. subtracting the classroom-specific mean, to estimate the within-classroom effect) and to center teacher's fondness for Bieber using –0.5 for "not a belieber" and +0.5 for "belieber" because you aim to test the idea that the highest-achieving students of a given classroom are more likely to be influenced by their teacher.

Note that grand- and cluster-mean centering are applicable to level-1 dichotomous predictors. In such a case, grand-mean centering entails removing the general mean of the dichotomous predictor (i.e. the proportion of cases across cluster; e.g. that female = 1), whereas cluster-mean centering entails removing the cluster-specific mean (i.e. the proportion of cases within cluster; e.g. testing the effect of the deviation of one's gender from the proportion of females in a given cluster). For more detailed information on centering decision, see Enders and Tofighi (2007), as well as some cautionary recommendations on group-mean centering by Kelley, Evans, Lowman and Lykes (2017).

Step #1. Building an Empty Model: To What Extent Do the Log-Odds Vary Between Clusters?

Now that you have centered your variables, you want to know the extent to which the odds that the outcome

variable equals one instead of zero varies from one cluster to another. In our example, you want to estimate the proportion of variability in the chance of owning an album rather than not owning it that lies between classrooms.

To do so, you need to run an empty model, that is, a model containing no predictors (also referred to as an "unconditional mean model"; cf. Eq. 3), and calculate the intraclass correlation coefficient (for Stata, R, Mplus, and SPSS commands, see the relevant Sub-Appendix B). Below is the formula of the Intraclass Correlation Coefficient (ICC; Eq. 6):

$$\mathrm{ICC} = \frac{\mathrm{var}(u_{0j})}{\mathrm{var}(u_{0j}) + \left(\pi^2 / 3\right)} \quad (6)$$

...in which $\mathrm{var}(u_{0j})$ is the random intercept variance, that is, the level-2 variance component: The higher $\mathrm{var}(u_{0j})$, the larger the variation of the average log-odds between clusters;

...and $(\pi^2/3) \approx 3.29$ refers to the standard logistic distribution, that is, the assumed level-1 variance component: We take this assumed value, as the logistic regression model does not include level-1 residual (cf. Eq. 3).

The ICC quantifies the degree of homogeneity of the outcome within clusters. The ICC represents the proportion of the between-cluster variation $\mathrm{var}(u_{0j})$ (in your case: the between-classroom variation of the chances of owning the album) in the total variation (in your case: the between- *plus* the within-classroom variation of the chances of owning an album).

The ICC may range from 0 to 1. ICC = 0 indicates perfect independence of residuals: The observations do not depend on cluster membership. The chance of owning Justin's album does not differ from one classroom to another (there is no between-classroom variation). When the ICC is not different from zero or negligible, one could consider running traditional one-level regression analysis.[5] However, ICC = 1 indicates perfect interdependence of residuals: The observations only vary between clusters. In a given classroom, either everyone or nobody owns Justin's album (there is no within-classroom variation).

In your study $\mathrm{var}(u_{0j}) = 1.27$. Thus, $\mathrm{ICC} = 1.27/(1.27 + 3.29) \approx 0.28$. This indicates that 28% of the chances of owning an album is explained by between-classroom differences (and – conversely – that 72% is explained by within-classroom differences). For more detailed information on intraclass correlation coefficient in multilevel logistic regression, see Wu, Crespi, and Wong (2012).

Step #2. Building an Intermediate Model: To What Extent Does the Effect of a Relevant Lower-Level Variable Vary Between Clusters?

Now that you know the extent to which the odds vary from one cluster to another, you want to know the extent to which the *effect of the relevant lower-level variable(s)* varies from one cluster to another.

There is a debate in the literature, with some authors advocating the use of maximal model estimating all random slope variance parameters (Barr et al., 2013) and others underlining the risk of overparametrization, failure of convergence, and uninterpretable findings (Bates, Kliegl, Vasishth & Baayen, 2015). Our position is that random variations should primarily be tested when having *theoretical reasons* to do so. In our example, you surely want to estimate the variation of the effect of GPA on the odds of owning the album from one classroom to another, since you expect the effect of GPA to depend on some teacher characteristics.

To do so, you need to (a) run a constrained intermediate model (CIM), (b) run an augmented intermediate model (AIM), and (c) compare both by performing a likelihood-ratio test (for Stata, R, Mplus, and SPSS commands, see the relevant Sub-Appendix C).

The constrained intermediate model contains *all* level-1 variables (in our case: GPA), *all* level-2 variables (in our case: teacher's fondness for Bieber), as well as *all* intra-level interactions (level-1 * level-1 or level-2 * level-2 interactions; in our case: none). Note that the constrained intermediate model does *not* contain cross-level interactions, since the model precisely aims to estimate the *unexplained* variation of lower-level effects. For instance, if your study included pupils' sex (a second level-1 variable) and classroom size (a second level-2 variable), the constrained intermediate model would *only* contain the *intra*-level interactions "GPA * sex" and "teacher's fondness for Bieber * classroom size," not the cross-level interaction like "GPA * classroom size" or "sex * teacher's fondness for Bieber").

This constrained intermediate model equation is shown below (Eq. 7):

$$\mathrm{Logit}(odds) = B_{00} + B_{10} * \mathrm{x}_{ij} + B_{01} * \mathbf{X}_j + u_{0j} \quad (7)$$

...in which x_{ij} refers to GPA (the level-1 variable), whereas $\mathbf{X}_j$ refers to teacher's fondness for Bieber (the level-2 variable);

...B_{10} is the fixed slope of x_{ij} (the overall effect of GPA), and B_{01} is the (necessarily fixed) slope of $\mathbf{X}_j$ (the overall effect of teacher's fondness for Bieber).

The augmented intermediate model is similar to the constrained intermediate model, with the exception that it includes the residual term associated with the relevant level-1 variable, thereby estimating the random slope variance (if you have several relevant lower-level variables, test them one at a time; for the procedure see the Notes of the relevant Sub-Appendix C).[6] Remember that the random slope variance corresponds to the extent of the variation of the effect of the lower-level variable from one cluster to another (in our case: the extent of the variation of the effect of GPA from one classroom to another). Note that only main level-1 terms are thought to vary, not interaction terms.

The augmented intermediate model equation is shown below (Eq. 8):

$$\mathrm{Logit}(odds) = B_{00} + \left(B_{10} + u_{1j}\right) * \mathrm{x}_{ij} + B_{01} * \mathbf{X}_j + u_{0j} \quad (8)$$

...in which u_{1j} is the deviation of the cluster-specific slope (i.e. the specific effect of GPA within a given classroom)

from the fixed slope (i.e. the average effect of GPA regardless of classrooms).

No need to not look at the coefficient estimates or variance components of the intermediate models. Your goal is to determine whether the augmented intermediate model achieves a better fit to the data than the constrained intermediate model. In other words, your goal is to determine whether considering the cluster-based variation of the effect of the lower-level variable improves the model. To do so, after gathering or storing the deviance of the CIM and AIM, you will have to perform a likelihood-ratio test, noted LR χ^2 (1). Below is the formula of the likelihood-ratio test (Eq. 9):

$$\text{LR } \chi^2\ (1) = \textit{deviance}(\text{CIM}) - \textit{deviance}(\text{AIM}) \quad (9)$$

...in which *deviance*(CIM) is the deviance of the constrained intermediate model, whereas *deviance*(AIM) is the deviance of the augmented intermediate model;

...and "(1)" corresponds to the number of degrees of freedom (this equals one because there is only one additional parameter to be estimated in the AIM compared to the CIM, namely the random slope variance).

The deviance is a quality-of-(mis)fit index: The smaller the deviance, the better the fit. There are two possible scenarios:

i. The deviance of the augmented intermediated model is significantly lower than the deviance of the constrained model. That is, including the residual term u_{1j} significantly improves the fit. In this case, the residual term u_{1j} should be kept in the final model. In your study *deviance*(CIM) = 2,341 and *deviance*(AIM) = 2,312.[7] Thus, LR χ^2 (1) = 2,341 – 2,312 = 29, $p < .001$ (you can find the *p*-value using a common chi-square distribution table). This indicates that allowing the effect of GPA to vary between classrooms improves the fit and that it is better to take such variation into account.
ii. The deviance of the augmented intermediated model is not significantly lower than the deviance of the constrained model. That is, including the residual term u_{1j} does not significantly improve the fit. In this case, the term could perhaps be discarded to avoid overparametrization (Bates et al., 2015). However, this does not necessarily mean that the effect of the lower-level variable does not vary from one cluster to another (absence of evidence of variation is not evidence of absence of variation; see Nezlek, 2008). Importantly, a non-significant LR χ^2 (1) should not prevent you from testing cross-level interactions.[8]

Step #3. Building the Final Model: Do the Data Provide Support for Your Hypotheses?

Now that you know the extent to which the effect of the relevant lower-level variable varies from one cluster to another (and have decided whether to consider the variation of the level-1 effect and keep estimating the random slope variance in the final model or not), you can finally test your hypotheses.

To do so, you need to run the final model, adding the cross-level interaction(s) (for the Stata, R, Mplus, and SPSS commands, see the relevant Sub-Appendix D). The predictor variables are the same as that in the intermediated models (level-1 variable, level-2 variable, and *intra*-level interactions), except that the level-1 * level-2 interactions are now included (in our case: the GPA * teacher fondness for Bieber interaction). The final model equation is shown below (Eq. 10):

$$\begin{aligned}\text{Logit}(\textit{odds}) = {} & B_{00} + \left(B_{10} + u_{1j}\right) * \text{x}_{ij} + B_{01} \\ & * \mathbf{X}_j + B_{11} * \text{x}_{ij} * \mathbf{X}_j + u_{0j}\end{aligned} \quad (10)$$

...in which B_{11} is the coefficient estimate associated with the cross-level interaction, that is, the GPA * teacher fondness for Bieber interaction.

What about the children? It is now time to take a look at the odds ratios and to discover how pupils behave and whether the data support your hypotheses. With the provided commands, each of the odds ratios comes with its 95% Confidence Interval (CI). Let's first interpret the odds ratio of the hypothesized main effect and then the odds ratio of the hypothesized interaction effect.

Interpretation of the main effect. Regarding your main effect hypothesis, $\exp(B_{01}) = OR = 7.50$, 95% CI [5.00, 11.25]. Congruent with your teacher-to-pupil socialization hypothesis, this indicates that pupils whose classroom teacher is a belieber have 7.50 times more chance of owning Justin's album than pupils whose teacher is not a fan a Justin Bieber (i.e. the interpretation of the *OR*).

Since the 95% confidence interval ranges from 5.00 to 11.25, we decide that the effect lies about between 5 and 12 more chances of owning the album. When a 95% confidence interval does not contain 1, the effect is significant (i.e. $p < 0.05$) and we reject H_0, whereas when a 95% confidence interval does contain 1, the effect is non-significant (i.e. $p > 0.05$) and we cannot reject H_0 (with a 95% CI, the alpha level = 0.05). For more detailed information on confidence intervals, see Cumming, 2014.[9]

Interpretation of the interaction effect. Regarding your interaction hypothesis, this is a bit more complicated. In logistic regression, there is a debate in the literature regarding the procedure to be followed for calculating the interaction effect (see Kolasinski & Siegel, 2010).

In (multilevel) logistic regression, the coefficient estimate of the product term does not correspond mathematically to the interaction effect. Technically, your software calculates the coefficient estimate of the product term as for any main effect (i.e. your software calculates the *marginal* effect), despite the fact that this calculation does not apply to interaction effect in logistic regression (i.e. the *marginal* effect does not equal the interaction effect in logistic regression; Ai & Norton, 2003; see also Karaca-Mandic, Norton & Dowd, 2012).

In logistic regression, the sign, the value, and the significance of the product term is likely to be biased, which has made some authors advocate calculating the correct interaction effect using special statistical package (e.g. the inteff command in Stata or the intEff function in R; see Norton, Wang & Ai, 2004). However,

the calculation of the correct interaction effect (or the correct *cross-partial derivative*) is quite complex and there is no statistical package available for multilevel modeling. Moreover, other authors have shown that interpreting the coefficient estimate of product term was appropriate most (but not all) of the time (Kolasinski & Siegelm, 2010; see also, Greene, 2010). Pending better approach, scholars might rely on the simple significance-of-the-product-term approach. This is what we do here.

In your study, $\exp(B_{11}) = OR = 3.01$, 95% CI [1.86, 4.86]. Since the 95% confidence intervals does not contain 1, the effect is statistically significant. This means that the effect of teacher's fondness for Bieber significantly differs as a function of pupils' GPA. To be interpreted, the interaction needs to be decomposed: We want to know the effect of the level-2 predictor variable for each category of the level-1 variable (this could have been vice versa). Decomposing the interaction may be done using two dummy-coding models (e.g. see Preacher, Curran & Bauer, 2004):

i. The first dummy coding model aims to estimate the effect of teacher's fondness for Bieber for pupils having a low GPA (by convention: 1 SD below the cluster-mean). To do that, you have to *add* one standard deviation from cluster-mean centered GPA (with a dichotomized variable, you may fix the condition of interest at 0 and the other at 1). Both the main term x_{ij} (GPA) and the product term $x_{ij} * \mathbf{X}_j$ (GPA * teacher's fondness for Bieber) need to be changed accordingly. In such a model, the coefficient estimate B_{10} of teacher's fondness for Bieber will become *the simple slope* of teacher's fondness for Bieber when GPA = 0, i.e. in this case, when GPA = −1 SD. In this first dummy coding model, $\exp(B_{10}) = OR = 3.47$, 95% CI [2.13, 5.64]. This indicates that for the lowest achievers of their classroom (−1 SD), having a teacher who is a belieber (versus not) results in a 3.50 times higher chance of owning Justin's last album.[10]
ii. A second dummy coding model aims to estimate the effect of teacher's fondness for Bieber for pupils having a high GPA (by convention: 1 SD above the cluster-mean). To do that, you have to *remove* one standard deviation from cluster-mean centered GPA (with a dichotomized variable you may fix the condition of interest at 0 and the other at −1). Again, both the main term and the product term need to be changed accordingly. In such a model, the coefficient estimate B_{01} will become the *simple slope* of teacher's fondness for Bieber when GPA = 0, i.e. this time, when GPA = +1 SD. In this second dummy coding model, $\exp(B_{10}) = OR = 16.20$, 95% CI [9.21, 28.49]. This indicates that for the highest achievers of their classroom (+1 SD), having a teacher who is a belieber (versus not) results in a 16.20 times higher chance of owning Justin's last album the significant interaction effect suggests that this second simple slope effect is stronger than the first one.

In both models, the random slope component will have to remain the same. In other words, the residual term associated with the level-1 predictor u_{1j} will have to remain centered. It is worth noting that adding the GPA * teacher fondness for Bieber interaction term may result in the reduction of the random slope variance (from the intermediate to the final model). This is due to that fact that there now are fewer *unexplained* variations of the effect of GPA from one classroom to another, since teacher fondness for Bieber accounts for part of these variations. Likewise, specifically adding the teacher fondness for Bieber term may result in the reduction of the random intercept variance, because there now are less unexplained variations of the odds of owning Justin's album from one classroom to another. However, importantly, adding a significant fixed term sometimes does not result in the decrease of residual variance (sometimes, it may even result in an increase) because of the way fixed and random effects are estimated (Snijders & Bosker, 1994; see also, LaHuis, Hartman, Hakoyama & Clark, 2014). If you observe such a phenomenon, it is not necessarily an issue.

Finally, using one of the syntax files provided with the article, you can compare the coefficient estimates obtained in the final model, with or without the use of multilevel modelling. You will realize that standard errors are deflated when using the traditional one-level logistic regression, thereby increasing the risk of Type I error.

"Where Are Ü Now." (Your) Future Challenges and Conclusion

For the brave readers who wish to go further, let's keep each other company for another couple of paragraphs. Know that multilevel (logistic) regression may be applied to other types of research designs, data structures, or outcome variables. First, multilevel logistic regression may be applied to *repeated measure designs and/or longitudinal data* (Quené & Van den Bergh, 2004). In such a situation, observations are nested in participants (e.g. right or wrong test answers nested in examinees; but a more rigorous approach would be using a model in which observations are cross-classified by stimuli and participants; see Baayen, Davidson & Bates, 2008). Thus, participants are treated as higher-level units and the analysis aims to disentangle the within-participant effects from the between-participant effects (in such a case, one may cluster-mean center the level-1 predictors so as to estimate the pooled within-participant fixed effects; Enders & Tofighi, 2007). Our three-step procedure may well be used in this case (with participant number as the level-2 identifier), although the database will have to be rearranged in the preliminary phase in order to have one line per lower-level units (for the Stata, R, and SPSS commands, see the relevant Sub-Appendix E; the current version of Mplus does not perform data reshaping).

Second, multilevel logistic regression may be applied to *three- (or more) level hierarchical or cross-classified data structure* (see Rabe-Hesketh & Skrondal, 2012a). In a two-level cross-classified data structure, pupils (level 1) could for example be nested in two non-hierarchical clusters: the school they attend (level 2a) and the neighborhood they live in (level 2b; see Goldstein, 2003). This a cross-classified data structure in the sense

that pupils in a given cluster (school or neighborhood) are not "sub-classified" by the other type of cluster (i.e. pupils do not necessarily attend to the school of their neighborhood). Our three-step procedure is incomplete in this case, as two ICCs would have to be calculated in Step #1 (there is level-2a and a level-2b random intercept variance) and various random slope variance could be estimated in Step #2 (for a given level-1 variable, there are level-2a and level-2b random slopes variance; for the Stata, R, and Mplus commands, see the relevant Sub-Appendix F; SPSS commands are not given due to software limitation).

Third, multilevel non-linear regression may be applied to a wide range of (non-normally distributed) discrete outcome variables, such as multinomial outcomes (three or more response categories), ordinal outcomes (three or more ordered response categories), or count outcomes (three or more counts of events; see Rabe-Hesketh & Skrondal, 2012b). As it would take too long to cover all the different cases, let's focus on the last example. Count outcome variables typically correspond to a number of occurrences (e.g. the number of murders per year and per neighborhood) and are often right-skewed, that is, follow a Poisson distribution (e.g. the number of murders per year is zero for most neighborhoods, one for some rare neighborhoods, two for even rarer neighborhoods, and so on; see King, 1988). Our three-step procedure is to be modified in this case, as multilevel Poisson regression or negative binomial multilevel regression have to be carried out. Note that these regression models give incidence rate ratio rather odds ratio (for the Stata, R, and Mplus commands, see the relevant Sub-Appendix G; SPSS commands are not given due to software limitation).

In conclusion, notwithstanding the aforementioned future challenges, it's a good day. Reading this article, you have understood that logistic regression enables the estimation of odds ratio and confidence interval, describing the strength and the significance of the relationship between a variable and the odds that an outcome variable equals one instead of zero. Moreover, now you know that multilevel logistic regression enables to estimate the fixed intercept and random intercept variance (i.e. the average general log-odds and its variation from one cluster to another), as well as the estimation of fixed slope and random slope variance (i.e. the average general effect of a lower-level variable and its variation from one cluster to another). And while your condescending colleague struggles with complex multilevel procedures, you calmly use the three-step simplified procedure for multilevel logistic regression analysis presented in this article: In a preliminary phase, you may choose to grand- or cluster-mean center your variables; in Step #1, you run an empty model estimating the random intercept variance and calculating the ICC; in Step #2, you run a series of intermediate models determining whether including the residual term associated with the level-1 predictor (and estimating the random slope variance) improves the model fit with a LR χ^2 (1); in Step #3, you run a final model testing the hypotheses. Yes, finally, it's a very good day. Life is worth living, so live another day.

Notes

[1] …and if you think otherwise, oh baby you should go and love yourself. By the way, we should add that we have hidden the names of all the songs of Justin's last album in the manuscript, often along with some lyrics. Just for fun.

[2] The odds are about 1 in 2,500,000 since there are roughly 10,000,000 female teenagers in the US (US Census Bureau, 2016) and that Justin Bieber has so far had four relationships (allegedly), namely with Jasmine Villegas, Selena Gomez, Hailey Baldwin, and Sofia Richie.

[3] In multilevel modeling, the term "random" indicates that a coefficient (intercept or slope) varies across clusters, as opposed to the fixed effect which is the average coefficient across clusters. It should not be understood in terms of mathematical randomness.

[4] By default, some software packages, as SPSS or Statistica, will estimate the log-odds that an outcome variable equals zero instead of one (rather than the other way around). For SPSS, we encourage you to make a small change to the syntax command so as to avoid any confusion (see Sub-Appendix A). Alternatively, you can recode the variable so that "0" corresponds to the event occurring and "1" to the event not occurring.

[5] The assumption of independence of residuals can be understood as the assumption that ICC = 0, which is why some authors have argued that a non-significant and negligible ICC may lead to the decision of treating the individual as the sole unit of analysis (e.g. Kenny, Kashy & Cook, 2006; SPSS users and R users working with the lme4 library will not be able to estimate the level of significance of the ICC). However, the independence of residuals does not rule out the presence of variation in the effect of a lower-level variable and – by extension – the need of multilevel modeling (see Barr, Levy, Scheepers & Tily, 2013). An alternative to the ICC would be to calculate the design effect with the formula Design effect = 1 + (average group size – 1) * ICC, as suggested by Muthén and Satorra (1995). A design effect > 2 is considered as suggesting that clustering should not be ignored and that multilevel analysis is required.

[6] The covariance between the random intercept variance and the random slope variance is assumed to be zero in this procedure. Although the covariance structure is usually tested in multilevel modeling procedures, the results are rarely interpreted (Hox, 1995). In the same way as for the random slope variance, we argue that this covariation should be primarily tested when having theoretical reasons to do so. To determine whether including the covariance parameter improves the model, one should include it in the augmented intermediate model. In doing so, Stata users have to add, cov(uns) at the end of the command; R users have to replace (1 + **lvl1_predict** || **id_cluster**) by (1 + **lvl1_predict** | **id_cluster**) in the glmer function; Mplus users have to add s1 WITH **outcome** (s1 is the name given to the random slope) to the %BETWEEN% part of the model; and SPSS users have to specify COVARIANCE_TYPE = UNSTRUCTURED in the com-

mand. In this situation, the augmented intermediate model estimates two more terms than the constrained intermediate model (i.e. the random slope variance and the covariance parameter). Hence, the likelihood-ratio test will have two degrees of freedom instead of one.

[7] Deviance estimates may marginally vary from one software to another, as the implementation of the algorithms is not exactly identical. For instance, in R deviance(CIM) = 2342 and deviance(AIM) = 2315.

[8] A similar approach could be used to test whether the inclusion of a predictor variable improves the model fit (i.e. to test whether a fixed effect is significant): A likelihood ratio test can be used to compare a constrained model not including the fixed effect with an augmented intermediated model including the fixed effect (Gelman & Hill, 2007).

[9] Formally speaking, the interpretation of the confidence interval is as such: If we repeated the study an infinite number of time, and computed a 95% confidence interval each time, then 95% of these confidence intervals would contain the true population odds ratio and 5% of them would miss it (see Morey, Hoekstra, Rouder, Lee & Wagenmakers, 2016).

[10] The standard errors of the simple slopes should normally be calculated using the Delta method (Greene, 2010); this method is not covered herein for the sake of simplicity.

Acknowledgements

This publication is based on research conducted at the Swiss National Centre of Competence in Research LIVES – Overcoming vulnerability: Life course perspectives (NCCR LIVES), which is financed by the Swiss National Science Foundation. We wish to thank Anatolia ('niita) Batruch and Wojciech Świątkowski for reading the first version of this article. None of the authors is actually a fan of Justin Bieber.

Competing Interests

The authors have no competing interests to declare.

References

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

Ai, C., & **Norton, E. C.** (2003). Interaction terms in logit and probit models. *Economics Letters*, *80*, 123–129. Get rights and content. DOI: https://doi.org/10.1016/S0165-1765(03)00032-6

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.

Bates, D., Maechler, M., Bolker, B., & **Walker, S.** (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*, 1–48. DOI: https://doi.org/10.18637/jss.v067.i01

Bressoux, P. (2010). *Modélisation Statistique Appliquée aux Sciences Sociales* [Statistical modelling applied to social sciences]. Bruxelles, Belgium: De Boeck. DOI: https://doi.org/10.3917/dbu.bress.2010.01

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*, 7–29. DOI: https://doi.org/10.1177/0956797613504966

Cutrona, C. E., Russell, D. W., Brown, P. A., Clark, L. A., Hessling, R. M., & **Gardner, K. A.** (2005). Neighborhood context, personality, and stressful life events as predictors of depression among African American women. *Journal of Abnormal Psychology*, *114*, 3–15. DOI: https://doi.org/10.1037/0021-843X.114.1.3

Enders, C. K., & **Tofighi, D.** (2007). Centering predictor variables in cross-sectional multilevel models: a new look at an old issue. *Psychological Methods*, *12*, 121–138. DOI: https://doi.org/10.1037/1082-989X.12.2.121

Felps, W., Mitchell, T. R., Hekman, D. R., Lee, T. W., Holtom, B. C., & **Harman, W. S.** (2009). Turnover contagion: How coworkers' job embeddedness and job search behaviors influence quitting. *Academy of Management Journal*, *52*, 545–561. DOI: https://doi.org/10.5465/AMJ.2009.41331075

Gelman, A., & **Hill, J.** (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. New York, NY: Cambridge University Press.

Goldstein, H. (2003). *Multilevel Statistical Models* (3rd ed.). London, UK: Arnold.

Greene, W. (2010). Testing hypotheses about interaction terms in non-linear models. *Economics Letters*, *107*, 291–296. DOI: https://doi.org/10.1016/j.econlet.2010.02.014

Heck, R. H., Thomas, S. L., & **Tabata, L. N.** (2013). *Multilevel and Longitudinal Modeling with IBM SPSS: Quantitative Methodology Series* (Second Edition). New York, NY: Routledge.

Hosmer, D. W., & **Lemeshow, S.** (2000). *Applied Logistic Regression*. New York, NY: John Wiley & Sons, Inc. DOI: https://doi.org/10.1002/0471722146

Hox, J. (2010). *Multilevel Analysis: Techniques and Applications* (2nd ed.). Hove, UK: Routledge.

Hox, J. J. (1995). *Applied Multilevel Analysis*. Amsterdam, Netherland: TT-publikaties.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*, 434–446. DOI: https://doi.org/10.1016/j.jml.2007.11.007

Judd, C. M., McClelland, G. H., & **Ryan, C. S.** (2017). *Data Analysis: A Model Comparison Approach to Regression, ANOVA, and Beyond*. Routledge (3rd ed.). Abingdon, UK: Routledge.

Judd, C. M., McClelland, G. H., Ryan, C. S., Muller, D., & **Yzerbyt, V.** (2010). *Analyse des Données: Une Approche par Comparaison de Modèles* [Data analysis: A model comparison approach]. Bruxelles, Belgium: De Boeck.

Judd, C. M., Westfall, J., & **Kenny, D. A.** (2012). Treating stimuli as a random factor in social psychology: a new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, *103,* 54–69. DOI: https://doi.org/10.1037/a0028347

Karaca-Mandic, P., Norton, E. C., & **Dowd, B.** (2012). Interaction terms in nonlinear models. *Health Services Research*, *47*, 255–274. DOI: https://doi.org/10.1111/j.1475-6773.2011.01314.x

Kelley, J., Evans, M. D. R., Lowman, J., & **Lykes, V.** (2017). Group-mean-centering independent variables in multi-level models is dangerous. *Quality & Quantity*, *51,* 261–283. DOI: https://doi.org/10.1007/s11135-015-0304-z

Kenny, D. A., Kashy, D. A., & **Cook, W. L.** (2006). *Dyadic Data Analysis*. New York, NY: Guilford Press.

King, G. (1988). Statistical models for political science event counts: Bias in conventional procedures and evidence for the exponential Poisson regression model. *American Journal of Political Science*, *32*, 838–863. DOI: https://doi.org/10.2307/2111248

Kolasinski, A., & **Siegelm, A.** (2010). *On the Economic Meaning of Interaction Term Coefficients in Nonlinear Binary Response Regression Models (Working paper)*. Seattle, WA: University of Washington.

LaHuis, D. M., Hartman, M. J., Hakoyama, S., & **Clark, P. C.** (2014). Explained variance measures for multilevel models. *Organizational Research Methods*, *17*, 433–451. DOI: https://doi.org/10.1177/1094428114541701

Maas, C. J., & **Hox, J. J.** (2005). Sufficient sample sizes for multilevel modeling. *Methodology*, *1*, 86–92. DOI: https://doi.org/10.1027/1614-2241.1.3.86

Menard, S. (2002). *Applied Logistic Regression Analysis* (2nd ed.) (Sage University Series on Quantitative Applications in the Social Sciences, series no. 07–106). Thousand Oaks, CA: Sage.

Moineddin, R., Matheson, F. I., & **Glazier, R. H.** (2007). A simulation study of sample size for multilevel logistic regression models. *BMC Medical Research Methodology*, *7*, 34. DOI: https://doi.org/10.1186/1471-2288-7-34

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

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

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

Mutz, R., Bornmann, L., & **Daniel, H. D.** (2015). Does gender matter in grant peer review? *Zeitschrift für Psychologie*, *220*, 121–129. DOI: https://doi.org/10.1207/2151-2604/a000103

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

Norton, E. C., Wang, H., & **Ai, C.** (2004). Computing interaction effects and standard errors in logit and probit models. *Stata Journal*, *4*, 154–167.

Paccagnella, O. (2011). Sample size and accuracy of estimates in multilevel models: New simulation results. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *7*, 111–120. DOI: https://doi.org/10.1027/1614-2241/a000029

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.

Quené, H., & **Van den Bergh, H.** (2004). On multi-level modeling of data from repeated measures designs: A tutorial. *Speech Communication*, *43*, 103–121. DOI: https://doi.org/10.1016/j.specom.2004.02.004

Rabe-Hesketh, S., & **Skrondal, A.** (2012a). *Multilevel and Longitudinal Modeling Using Stata, Volume II: Categorical Responses, Counts, and Survival* (3rd ed.). College Station, TX: Stata Press.

Rabe-Hesketh, S., & **Skrondal, A.** (2012b). *Multilevel and Longitudinal Modeling Using Stata, Volume II: Categorical Responses, Counts, and Survival* (3rd ed.). College Station, TX: Stata Press.

Schoeneberger, J. A. (2016). The impact of sample size and other factors when estimating multilevel logistic models. *The Journal of Experimental Education*, *84*, 373–397. DOI: https://doi.org/10.1080/00220973.2015.1027805

Snijders, T. A. B., & **Bosker, R. J.** (1994). Modeled variance in two-level models. *Sociological Methods & Research*, *22*, 342–363. DOI: https://doi.org/10.1177/0049124194022003004

Snijders, T. A. B., & **Bosker, R. J.** (2004). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. London, UK: Sage. DOI: https://doi.org/10.1080/00220973.2015.1027805

Swaminathan, H., Rogers, H. J., & **Sen, R.** (2011). Research Methodology for Decision-Making in School Psychology. In: Bray, M. A., & Kehle, T. J. (Eds.), *The Oxford Handbook of School Psychology*, 103–139. New York, NY: Oxford University Press. DOI: https://doi.org/10.1093/oxfordhb/9780195369809.013.0038

Świątkowski, W., & **Dompnier, B.** (2017). Replicability Crisis in Social Psychology: Looking at the Past to Find New Pathways for the Future. *International*

Review of Social Psychology, 30(1), 111–124, DOI: https://doi.org/10.5334/irsp.66

U.S. Census Bureau. (2016). Annual Estimates of the Resident Population for Selected Age Groups by Sex for the United States, States, Counties, and Puerto Rico Commonwealth and Municipios: April 1, 2010 to July 1, 2015. Retrieved June 27, 2017, from: factfinder.census.gov/bkmk/table/1.0/en/PEP/2015/PEPAGESEX?slice=GEO~0400000US36.

Wu, S., Crespi, C. M., & **Wong, W. K.** (2012). Comparison of methods for estimating the intraclass correlation coefficient for binary responses in cancer prevention cluster randomized trials. *Contemporary Clinical Trials*, *33*, 869–880. DOI: https://doi.org/10.1016/j.cct.2012.05.004

How to cite this article: Sommet, N. and 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(1), 203–218, DOI: https://doi.org/10.5334/irsp.90

Published: 08 September 2017

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

OPEN ACCESS