APPLICATIONS OF MULTILEVEL MODELING IN PSYCHOLOGICAL SCIENCE: INTENSIVE REPEATED MEASURES DESIGNS

John B. Nezlek et Błażej Mroziński

Presses Universitaires de France | « L'Année psychologique »

2020/1 Vol. 120 | pages 39 à 72

ISSN 0003-5033
ISBN 9782130822936

Article disponible en ligne à l'adresse :

https://www.cairn.info/revue-l-annee-psychologique-2020-1-page-39.htm

Applications of multilevel modeling in psychological science: Intensive repeated measures designs

John B. Nezlek [1,2] and Błażej Mroziński[3]
[1] *Institute of Psychology, SWPS University of Social Sciences and Humanities, Poznań, Poland*
[2] *Department of Psychology, College of William & Mary, Williamsburg, VA, USA*
[3] *Institute of Psychology, SWPS University of Social Sciences and Humanities, Warsaw, Poland*

ABSTRACT

Multilevel modeling (MLM) is a statistical technique that can be used to analyze the data collected in various types of research. Although the use of and demand for MLM has increased dramatically over the past decade, instruction in MLM has not kept pace with these increases. The present paper provides an introduction to MLM that is intended to help researchers conduct MLM analyses and describe these results and to help them understand the results of MLM analyses that are presented in articles. Given the limits inherent in a single article, we do not cover all topics in depth. Nevertheless, we provide enough information so that readers should be able to conduct and understand MLM analyses. Examples of different types of analyses of diary style data (sometimes called intensive repeated measures), a design that is being used more and more often, are presented and sample data sets with worked examples are provided as on-line supplemental materials. Recommendations for best practice for conducting analyses and for reporting results are also provided.

Keywords: multilevel modeling, intensive repeated measures, nested data, within-person analyses

[1] Corresponding author: John B. Nezlek, Department of Psychology, SWPS University of Social Sciences and Humanities, Poznań, Poland. Department of Psychology, College of William & Mary, Williamsburg, VA, USA. Email: jbnezl@wm.edu

Applications du modèle multiniveau dans les sciences psychologiques : les plans à mesures répétées intensives

RESUME

Le modèle multiniveau (MLM) est une technique statistique qui peut être utilisée pour analyser les données collectées dans différents types de recherche. Bien que l'utilisation et la demande de MLM aient considérablement augmenté au cours de la dernière décennie, l'enseignement du MLM n'a pas suivi le rythme de cette évolution. Le présent article fournit une introduction au MLM qui est destinée à aider les chercheurs à effectuer des analyses MLM et à en décrire les résultats, mais aussi à les aider à comprendre les résultats des analyses MLM présentées dans les articles. Compte tenu des limites inhérentes à un seul article, nous ne couvrons pas tous les sujets en profondeur. Néanmoins, nous fournissons suffisamment d'informations pour que les lecteurs soient en mesure de conduire et de comprendre les analyses MLM. Des exemples de différents types d'analyses de données de type « journal intime » (parfois appelées mesures répétées intensives), une méthode de recueil de plus en plus utilisée, sont présentés et des séries de données avec des exemples concrets sont fournies dans le matériel supplémentaire en ligne. Des recommandations sont également données concernant les meilleures pratiques pour effectuer les analyses et pour rendre compte des résultats.

Mots-clés : modèle multiniveau, mesures répétées intensives, données emboîtées, analyses intra-sujets

1. WHY MLM AND NOT ORDINARY-LEAST-SQUARES (OLS) MULTIPLE REGRESSION?

MLM was created to analyze data structures in which observations (units of analysis) are not independent. Sometimes such data structures are described as "nested," "hierarchically nested," or clustered. Common examples of this in the social sciences are studies of students who are nested within classrooms (the subject of another paper in this special issue, **see Bressoux, 2020**), studies of workers in working groups (essentially the same data structure as students in classrooms), clients nested within clinics, people nested within countries (or counties, neighborhoods, etc.), and observations nested within persons (e.g., a diary style study). The number of possibilities is limitless. As noted by Kreft and deLeeuw (1998): "Once you know that hierarchies exist you see them everywhere."

Nested data structures cannot be analyzed using traditional OLS techniques because OLS analyses assume that errors of measurement are independent, and such independence cannot be assumed when observations are nested. Importantly, the violation of the assumption of independence

invalidates tests of significance. For example, in a study of people who work together in groups, the workers in group 1 share certain characteristics with one another, the workers in group 2 share certain characteristics with one another, and so forth. Moreover, these shared characteristics vary from group to group. This means that relationships between worker-level variables (e.g., years of experience and productivity) cannot be examined with "standard" OLS regression analyses in which workers are the units of analysis. Similarly, in an intensive repeated measures study such as a daily diary study the observations collected from a specific person are not independent, and within-person relationships such as the relationship between daily stress and daily well-being cannot be examined with OLS regression in which daily measures are the units of analysis.

In addition to the difficulties created by violations of the assumption of independence, in multilevel data structures relationships between variables can vary across levels of analysis. For example, assume a study of affect in which observations are nested within persons. At the within-person level anxiety (an active affective state) may be negatively related to sadness (a deactive affective state). People may not be anxious when they are sad and vice versa. Nevertheless, at the between-person level people who are more anxious on average may be more sad (on average) than people who are less anxious on average. As discussed in Nezlek (2001), any combination of within- and between-person relationships is possible. In fact, relationships at different levels of analysis are mathematically independent.

There is also the issue of taking into account the fact that in a multilevel data structure units of analysis at different levels of analysis are sampled from different populations. For example, in a daily diary study the people in the study constitute a sample drawn from the population of people, and the days in the study constitute a sample drawn from the population of days. Such dual sampling requires estimating two error terms in the same model, one for each sampling distribution, and this is not possible within the OLS framework because OLS can estimate only one error term at a time. Estimating multiple error terms simultaneously requires iterative algorithms, which are used in MLM. See Nezlek (2011) for an introduction to MLM that does not focus too heavily on the statistical background for MLM.

2. THE LOGIC OF MLM

Our explanation of the logic of MLM relies on the pioneering work of Bryk and Raudenbush (1991) who were the first to describe MLM using separate models for each level of analysis. Although all coefficients and all error terms at all levels are estimated simultaneously, we think presenting separate models for each level of analysis makes it easier to understand analyses, particularly for readers (and analysts) who are not familiar with MLM. This article discusses MLM in terms of two level models. Extrapolating to more levels of analysis (simply) involves adding equations for each level of analysis to the basic model.

The basic two level model is presented below. In this model there are *i* level 1 observations nested with *j* level 2 units (called groups in MLM even though they may not be groups). As can be seen from the equations below, the mean of each level 2 unit (each group) is "brought up" to level 2 where it is then analyzed. The variance of r_{ij} is the level 1 variance, which represents how much level 1 observations vary from the means of their groups. The variance of u_{0j} is the level 2 variance, which represents how much the group means vary.

Level 1: $y_{ij} = \beta_{0j} + r_{ij}$.
Level 2: $\beta_{0j} = \gamma_{00} + u_{0j}$.

This model is typically referred to as an unconditional or null model, and although it may not be used to test of hypotheses of interest, it does provide the basic descriptive statistics for a multilevel data structure. These are the mean (γ_{00}) and the two variance estimates, the variances of r_{ij} (level 1) and u_{0j} (level 2). Note that multilevel data cannot be described with a single variance (or standard deviation). There is a variance at each level of analysis.

As discussed below in the example analyses, predictors are added to these level 1 and level 2 models to test hypotheses of interest. To test hypotheses about relationships between level 2 variables and means from level 1, predictors are added to the level 2 model as below:

Level 1: $y_{ij} = \beta_{0j} + r_{ij}$.
Level 2: $\beta_{0j} = \gamma_{00} + \gamma_{01}$ (level 2 predictor) + u_{0j}.

Predictors can also be added at level 1 as illustrated below. The null hypothesis is that the mean slope across all level 2 units is 0. This hypothesis is tested at level 2 by the significance of the γ_{10} coefficient. Keep in mind that the γ_{10} coefficient is an estimate of the mean slope. Some level 2 units

are likely to have slopes that are weaker than the mean, and some are likely to have slopes that are stronger than the mean.

Level 1: $y_{ij} = \beta_{0j} + \beta_{1j}$ (level 1 predictor) + r_{ij}.
Level 2: $\beta_{0j} = \gamma_{00} + u_{0j}$.
$\beta_{1j} = \gamma_{10} + u_{1j}$.

The variability in slopes can be analyzed by including predictors at level 2, an analysis that is sometimes called a "slopes as outcomes analysis" because a slope (a coefficient representing a level 1 relationship becomes an outcome at level 2. Such a possibility can also be called a cross-level interaction or a moderating effect. Level 2 predictors can also be added to the level 2 equation for the slopes. Does the level 1 slope vary as a function of a level 2 variable? All of these possibilities are discussed in the next sections of this article.

Level 1: $y_{ij} = \beta_{0j} + \beta_{1j}$ (level 1 predictor) + r_{ij}.
Level 2: $\beta_{0j} = \gamma_{00} + \gamma_{01}$ (level 2 predictor) + u_{0j}.
$\beta_{1j} = \gamma_{10} + \gamma_{11}$ (level 2 predictor) + u_{1j}.

Note that in this model the same level 2 predictor is included in the analysis of each level 1 coefficient. The norm in MLM is to include the same level 2 predictors in all equations because the algorithms rely on covariance matrices. If a level 2 predictor is not included this assumes that the coefficient representing that coefficient is not significant and that all of the covariances involving this coefficient are 0. This is not likely to be the case, meaning that if the same predictors are not included in all equations a model may be misspecified.

3. EXAMPLE DATA SETS AND ANALYSES

To complement the presentation in the other article in this special issue we illustrate MLM using analyses of a hypothetical diary study in which observations are nested within persons. The supplemental materials contain a folder labeled Diary Data. See Nezlek and Mrozinski (2019; https:/osf.io/74m5r). The example data set is a modified version of the data presented in Nezlek and Plesko (2001). We created these data to illustrate certain aspects of MLM. In this study, days (1330) are nested within people (103). The day level measures are scc (self-concept clarity), tri (a measure of adjustment based on Beck's triadic model of depression), pa (positive

affect), na (negative affect), posevent (positive events), negevent (negative events), and negnew, a measure that was created by adding 10 to the daily negevent scores for half the participants. The person level file has a measure of trait anxiety and variables representing participant sex.

The folder has two data sets, one for each level of a two-level model. This is the file structure used by HLM. For other applications, the files can be modified as needed. The files are labeled Example-Lev1 and Example-Lev2, and the data are provided as SPSS files (.sav) and as text files (.csv). In the text files the variable names are presented in the first line of the data file. Note that there are missing data at level 1.

3.1. Standard analyses

As discussed by Nezlek (2007), the relationships examined in diary studies often consist of variants of the following four questions. More detailed coverage of these topics can be found in Nezlek (2001, 2010, 2012a).

(1) Relationships between two (or more) level 1 measures, e.g., the within-person relationship between self-concept clarity (scc) and daily negative events. A significant coefficient would indicate that daily self-concept clarity was related to daily negative events. For example, Nezlek and Plesko (2001) reported a negative coefficient between daily self-concept clarity and daily negative events. In other words, on average (across all participants) self-concept clarity was lower on days when more negative events occurred than on days when fewer negative events occurred.

(2) Relationships between the mean of a level 1 measure and a level 2 measure, e.g., the relationship between daily self-concept clarity and trait anxiety. In terms of the example data set, a significant negative coefficient would indicate that people who were more dispositionally anxious tended to have lower daily self-concept clarity than people who were less dispositionally anxious.

(3) How do level 1 relationships vary as a function of level 2 variables, e.g., does the within-person relationship between daily self-concept clarity and daily negative events vary as a function of a person's trait level of anxiety? Conceptually, such a question examines individual differences in reactions to stress (negative events). Does the daily self-concept clarity of people who are higher in dispositional anxiety fluctuate more as a function of daily stressors than the self-concept clarity of people lower in dispositional anxiety?

(4) How do these three types of relationships vary across time?

In terms of the variables provided in the example data set, these relationships would be examined using the following models.

(1) Level 1 relationships (within-person in the example data set).

Level 1: y_{ij} (scc) = $\beta_{0j} + \beta_{1j}$ (negative events) + r_{ij}.
Level 2: $\beta_{0j} = \gamma_{00} + u_{0j}$.
$\beta_{1j} = \gamma_{10} + u_{1j}$.

The null hypothesis that the mean slope is 0 (i.e., the mean level 1 coefficient representing the relationship between self-concept clarity and negative events) is tested at level 2 (the between-person level in the example data set) via the significance of the γ_{10} coefficient. Note that the presence of the u_{1j} coefficient indicates that the slope is modeled as randomly varying, a topic discussed below.

(2) Relationships between a level 1 mean and a level 2 measure.

Level 1: y_{ij} (scc) = $\beta_{0j} + r_{ij}$.
Level 2: $\beta_{0j} = \gamma_{00} + \gamma_{01}$*(trait anxiety) + u_{0j}.

The null hypothesis concerning the relationship between trait anxiety and mean daily self-concept clarity is tested at level 2 (the between-person level) via the significance of the γ_{01} coefficient.

(3) Do level 1 relationships vary as a function of a level 2 measures?

Level 1: y_{ij} (scc) = $\beta_{0j} + \beta_{1j}$ (negative events) + r_{ij}.
Level 2: $\beta_{0j} = \gamma_{00} + \gamma_{01}$ (trait anxiety) + u_{0j}.
$\beta_{1j} = \gamma_{10} + \gamma_{11}$ (trait anxiety) + u_{1j}.

Hypotheses about whether the slope between scc and negative events varies as a function of trait anxiety is tested by examining the significance of the γ_{11} coefficient. The null hypothesis is that the γ_{11} coefficient is not significantly different from 0. Note also that anxiety is included in the equation predicting the intercept from level 1. As noted previously, unless there are compelling reasons not to do so, the same predictors should be included in all level 2 equations.

The models needed to examine changes across time in the coefficients estimated in these three models will vary considerably as a function of the number of time periods and the specific hypotheses of interest. Unfortunately, explaining these possibilities is beyond the scope of this paper. Some of these possibilities are explained in more detail in Nezlek (2012b, pp. 117-118).

These equations represent the basic structural relationships that can be studied in MLM. Although these specific models are based on measures from a hypothetical diary study in which observations are nested within persons, the same models would be used if a data set contained 1330 persons nested within 103 groups, or 1330 clients nested within 103 clinics or

clinicians, or 1330 trees nested within 103 plots of land. The questions are the same (structurally), and the modeling procedures are the same. Within this context, we next consider different aspects of MLM, starting with centering and modeling error, which are critical aspects of MLM.

3.2. Centering

Centering options, which are a critical characteristic of defining predictors in MLM, will probably be unfamiliar to analysts whose primary experience is OLS regression. Centering refers to the reference value from which the deviations of a predictor are taken. A traditional OLS correlation (a slope) between x and y is a measure of how closely the deviations of x from its mean correspond to the deviations of y from its mean. In MLM, different referents can be used to calculate deviations, and choosing different referents changes the meaning of the slopes and changes the meaning of intercepts in a model. In other words, the same measured variable functions as a different variable as a function of centering.

At the top level of a model (e.g., level 2 in a two level model and level 3 in a three level model) there are two options for centering, grand-mean centering and zero-centering (also call uncentered). When a predictor is grand-mean centered the slope is based on deviations from the mean of the predictor, and the intercept represents the expected value for an observation that is at the mean of the predictor. Within rounding, this intercept has the same value as the intercept from a model in which there was no predictor. In contrast, when a predictor is entered uncentered, deviations are taken from 0, and the intercept represents the expected value for an observation that has a value of 0 on a predictor.

The differences between these two options is illustrated using the example diary data set. First, to obtain an estimate of the mean for scc, we conducted a totally unconditional (or null) model with scc as the outcome. The results of this analysis are in the file scc-null.txt in the supplemental files. The estimated mean is 4.69.

If anxiety is entered grand-mean centered at level 2, the resulting level 2 intercept is the same as it was for the totally unconditional model, although the residual level 2 variance is lower. The meaning of the change in this variance estimate is explained in the section on effect sizes. The results of this analysis are in the file scc-anx(grand).txt in the supplemental files.

When anxiety is entered uncentered, file scc-anx(unctr).txt in the supplemental files, the intercept of the level 2 coefficient is different than when anxiety was entered grand-mean centered (8.32 vs. 4.69) because now

the intercept represents the expected value for an observation (a person) who has a score of 0 on the anxiety measure. Although anx can be entered uncentered, given that 0 is not a valid value on the scale it makes little sense to estimate an intercept for an observation that cannot exist, i.e., a person who has a score of 0 on the anx measure. If anx were standardized (a topic discussed later), then entering anx as uncentered would be sensible.

At lower levels of nesting (e.g., level 1 in a two-level model and levels 1 and 2 in a three-level model) there are three options for centering: uncentered, grand-mean centered, and group-mean centered. Uncentered and grand-mean centering have the same consequences at level 1 as they had at level 2, but now the deviations are considered within each group (each level 2 unit of analysis). So, when a level 1 predictor is entered uncentered, the intercept in each group represents the expected value for an observation in each group that has a value of 0 on the predictor. When a predictor is entered grand-mean centered the intercept is the expected value for an observation that is at the grand mean of the predictor.

When predictors are entered group-mean centered the intercept represents the expected value for an observation that is at the group mean of a predictor. Within rounding, this is essentially the same as the intercept from an unconditional model. Group-mean centering is the option that is the closest to conducting a regression analysis for each group and then using the coefficients from these analyses as outcome measures at level 2.

Differences among these three options are illustrated using analyses of the example diary data set. When negnew is entered as a group-mean centered predictor of scc the intercept is the same (within rounding) as it was from the null model of scc (4.69). The variance of the intercepts from the two analyses are also the same. In other words, the same value is being "passed up" to level 2 in both analyses. See scc-negnew(group).txt in the supplemental files.

Recall that negnew was created by adding 10 to negevent for half the participants. When predictors are group-mean centered, level 2 differences in level 1 predictors do not contribute to the parameter estimates. This is illustrated by the fact that the intercept (mean and variance) from an analysis of scc in which negevent is entered as a group-mean centered predictor is exactly the same as the intercept from the analysis of negnew when entered group-mean centered. The results of this analysis are in the file scc-negevent(group).txt in the supplemental materials. Note also that the slope parameters for these two analyses are exactly the same. In other words, the fact that negevent scores were meaningfully higher for half the participants did not influence the results of the analysis.

In contrast, when negnew is entered grand-mean centered the intercept now represents the expected value for scc when negnew is at the grand-mean of negnew. As can be seen from the results of the analysis presented in scc-negnew(grand).txt in the supplemental materials, this is a meaningfully different estimate than the estimate from the group-mean centered analyses. The mean is different (4.76 vs. 4.69), and the variance of the mean is different (1.39 vs. 2.47). In other words, the fact that negnew scores were meaningfully higher for half the participants influenced the estimate of the intercept when negnew was entered a grand-mean centered predictor, and most importantly, this influence included changing the intercept that is passed up to the next level of analysis (level 2).

The last option for centering at level 1 is uncentered (zero-mean centered). The results of analyses of negevent and negnew entered as uncentered predictors are presented in the files scc-negevent(unctr).txt and scc-negnew(unctr).txt in the supplemental materials. As can be seen from these results, the estimates of the intercepts and slopes differ meaningfully between these two analyses. This is to be expected because unless level 2 mean differences in the predictors are eliminated from the model (i.e., group-mean centering) level 2 differences in predictors contribute to the parameter estimates for models with uncentered predictors. Substantively, the intercepts in these models represent the expected value of scc when the predictor is 0, i.e., for days when no negative events occurred, and the slopes represent how well deviations from 0 corresponded to within-person differences in scc.

Keeping in mind Bryk and Raudenbush's advice "That no single rule covers all cases," we recommend the following, advice which is consist with that offered by Enders and Tofighi (2007). At level 1 continuous predictors should be entered group-mean centered, which makes the intercept similar to the original intercept (null model) which helps to maintain the stability of covariance matrices. Grand-mean centering at level 1 can be used when an analyst wants to adjust the level 1 intercept for level 2 differences in a predictor. At level 2 continuous predictors should be grand-mean centered. This makes the intercept the expected value for an observation at the mean of the level 2 predictor, which corresponds to the mean of the sample for the outcome and is a value that is easy to understand. Categorical predictors (discussed below) should be entered uncentered. This makes it easier to interpret results, i.e., what coefficients represent. A more detailed discussion of centering can be found in Nezlek (2011; pp. 13-18).

NB: Not all programs require analysts to choose a centering option when entering a predictor. We strongly encourage analysts to consult program manuals and guides to ensure that predictors are centered as the analyst

intends them to be centered. Although slopes tend not to vary dramatically as a function of how a predictor is centered, they can, and the variance of intercepts can (and tend to) vary widely across centering options. MLM analyses rely on covariance matrices and so changing the variance of an intercept can change estimates of all of the parameters in a model.

3.3. Modeling error

Similar to centering, analysts whose primary experience is OLS regression may be unfamiliar with how to model error in MLM. In OLS regression there is one error term (the residual sums of squares or error variance), and it is always estimated. There are no options. In contrast, in MLM there is a residual error term at each level of analysis (at least for continuous outcomes), and each level 1 coefficient can have an error term. Moreover, although error terms and structures are typically not the focus of a study, how error is modeled can change the estimates of fixed effects, which typically are the focus of a study.

We discuss modeling error within the context of a two-level model with a continuous outcome. Level 1 coefficients can be modeled in one of three ways: fixed, which means that a fixed effect (i.e., a mean coefficient) is estimated without an accompanying random error term; randomly varying, both a mean and a random effect are estimated; and non-randomly varying, a mean is estimated with no random error term but the fixed variance of the coefficient is modeled as function of a level 2 predictor. These three possibilities are illustrated with analyses of the example diary data set.

In one analysis, pa-negevent(group)(fixed).txt, pa is regressed on negevent at level 1, and this slope is modeled as fixed—there is no random error term. Note that the mean slope is -.48 with a standard error of .105. In the next analysis, pa-negevent(group)(random).txt, the slope for negevent is modeled as randomly varying. The slope in this analysis is -.54 with a SE of .08. Although both slopes are significantly different from 0 and similar in magnitude, this is not always the case.

Regardless of whether an effect is modeled as randomly varying at level 1 the variability in this effect can be modeled at level 2. Examples of these possibilities are in the files pa-negevent(group)(random)-anxiety.txt and pa-negevent(group)(fixed)-anxiety.txt provided in the supplemental materials. Although neither of these moderating effects is significant, it is instructive to note that the significance tests of the moderating effects are quite different. This is because the two analyses are bringing different

level 1 coefficients up to level 2, i.e., the slopes that are the outcomes at level 2 are not the same because one was modeled as randomly varying and the other was not.

The differences between a model in which a predictor is model as fixed and the same model in which a predictor is modeled as randomly varying cannot be known in advance. Sometimes a fixed effect will be significant whereas the same effect modeled as randomly varying is not, and sometimes the opposite will occur. For this reason, we *strongly* recommend that analysts evaluate the error structure of their models before evaluating the fixed effects. It is inappropriate to select an error structure that provides the fixed effects someone wants.

Unless there are compelling practical or theoretical reasons to do otherwise (see below), we recommend modeling level 1 coefficients as randomly varying. As discussed previously, level 1 observations are typically randomly sampled from a population, and this error needs to be modeled. Days in a diary study are randomly sampled from the population of days, workers in work groups are randomly sampled from the population of workers, and so forth. By extension, the coefficients representing level 1 phenomena (means and slopes) are also sampled from a population of coefficients. For example, if a participant completes a diary over two different weeks, the coefficients from the first week will be similar to, but not the same as, the coefficients from the second week.

Although theoretically the level 1 coefficients in most studies should be modeled as randomly varying, this may not be possible because the data may not provide a basis for estimating a reliable error term. This is indicated by the significance test of the random error term (is the random error term significantly different from 0), a test that this distinct from the test of the fixed effect (is the fixed effect significantly different from 0). If a random error term cannot be estimated reliably, it is probably best to delete this error term from a model. Although there is no explicit guideline for the significance level that should be used to make decisions about deleting error terms, we recommend using p-values that are higher than the strict .05 used for most statistical tests, for example $p < .10$. See Nezlek (2001; 2011, pp. 19-25) for discussions of modeling random error in MLM.

It is also possible that a coefficient is not theoretically random. For example, if women provide data 30 days prior to the birth of their first child, then coefficients representing within-person relationships between measures taken across those 30 days do not need to be modeled as random. A woman has only one first birth, and the 30 days prior to the birth of this child are unique. This is a fixed effect. Note that the target of inference of such coefficients is very narrow: the 30 days prior to the birth of a first child.

Although researchers may choose to do so, it is not technically accurate to draw inferences to the 30 days prior to the birth of any child. Such a limitation highlights the relationship between the data you collect (e.g., the samples) and the population you want to describe (the target of inference).

Finally, it is important to note that the absence of a significant random error term does not mean that a coefficient does not vary. Technically speaking, the lack of a significant random error term means that the data do not provide a basis for separating true (fixed) and random variability. Fixed coefficients (coefficients without an accompanying random error term) have standard errors (they are tested against a null hypothesis), and their variability can be modeled as a function of level 2 variables. Although coefficients that are modeled as randomly varying provide a better basis for making inferences about population parameters than coefficients that are not modeled as randomly varying, analysts can draw inferences based on tests involving coefficients that are not modeled as randomly varying.

3.4. Missing data

Within the multilevel context it is important to distinguish missing cases or units of analysis from missing data. For example, assume participants are asked to provide 14 days of data in a daily diary study. In terms of MLM analyses if participants provide fewer than 14 days this would not be considered missing data. Days would be nested within participants regardless of how many days participants provide, and missing days would not be considered as missing level 1 cases. This situation is structurally similar to the possibility (the likely possibility) that classrooms will have different numbers of students. Typically, class sizes vary naturally, and students are treated as nested within classes irrespective of how many students are in each class.

In contrast, missing data occur when a measure that is meant to be collected for a case is not collected. For example, assume that participants in a daily diary study are asked to answer a question about self-esteem each day. On a particular day a participant does not answer this question although he/she answers other questions about his/her daily experience. The case is present (some data were collected for that participant for that day) but self-esteem is missing.

The standard procedure for treating missing in MLM analyses is listwise deletion. A case is deleted from an analysis if there is a missing value on a variable that is included in an analysis as either the outcome or as a predictor. At level 1, any case that has a missing value on a variable

that is included in an analysis is eliminated from the analysis. To continue the example above, if self-esteem was included in an analysis the days for which participants did not answer the question about self-esteem would be dropped from the analysis. Listwise deletion also occurs at level 2 with the additional consideration that if a level 2 case (unit of analysis) is not included in an analysis then all the level 1 cases that are nested within that level 2 case are also eliminated from the analysis. For models with more than two levels, when a case is excluded from an analysis all cases that are nested within (under) that case are also excluded.

Note: The program HLM relies upon a system file (.mdm file) to conduct analyses. When this file is created how missing data at level 1 are treated is determined. The options are listwise (as described above) or casewise – a level 1 case is not included in the system file if the case has a missing value on a variable that is included in the system file. The second option ensures that all analyses that rely on this file will use exactly the same cases. At level 2, if a case is missing value on a variable that is included in the file the case is not included in the system file and the level 1 cases associated with this level 2 case are also eliminated. See Nezlek (2011, pp. 61-63) for a more detailed discussion of missing data in MLM. The new version of HLM, HLM8, allows for missing data at level 2. See the program manual for instructions.

3.5. Standardizing variables

Unlike OLS regression, MLM analyses estimate only unstandardized coefficients. Nevertheless, measures can be standardized prior to analysis to provide estimates that closely approximate the standardized coefficients with which many analysts may be more familiar. Such a possibility is much more appropriate for level 2 variables than it is for level 1 variables. For level 2 variables standard scores can be calculated and then used in an analysis. Coefficients involving these standard scores will be standardized in the sense that a 1 unit change in a coefficient will represent the expected change in an outcome (intercept or slope) associated with a 1 *SD* change in the measure. Although the results of significance tests that involve such standard scores will be the same as the results of significance tests that involve unstandardized variables, standardizing can make it easier to interpret the results of analyses.

The situation at level 1 (or at levels 1 and 2 in a three-level data structure) is not so simple. The reason for this is that there is no single variance estimate for level 1 variables that can be used as a basis for standardization. A measure could be standardized in terms of the grand mean and the *SD*

of all observations taken together, but such standardization would be only a transformation, not a standardization. Scores could be standardized within each level 2 unit, but this would eliminate from the model the level 2 variance of the means, a variance estimate that is an important part of any analysis. At level 1, the closest an analyst can get to a standardized coefficient is to estimate scores for units of analysis that are +/- 1 *SD* in terms of the predictors in a level 1 model. This is explained in the section entitled "Interpreting and reporting results."

3.6. Categorical predictors and the analyses of groups

The algorithms that are used to conduct MLM analyses do not differentiate between categorical and continuous predictors. All predictors are treated as continuous. Moreover, the same caveats, guidelines, and principles that pertain to using continuous predictors pertain to using categorical predictors. Of particular importance is how the meaning of intercepts changes as a function of changes in how a categorical predictor is constructed and how it is entered.

Regardless, it is possible to use predictors to represent categorical variables, and categorical predictors can be used to estimate parameters for individual groups and to compare groups. When doing so, it is essential to be aware of how the meaning of coefficients and error terms can vary across combinations of coding and centering. Although the logic of using categorical predictors is the same at different levels of analysis, given the differences in implications for what coefficients represent we discuss how to use categorical predictors at different levels of analysis.

We discuss two types of categorical predictors: dummy codes in which categorical membership is represented by assigning 1 to observations that are members of a particular category and 0 for observations that are not, and contrast codes (similar to effect codes) in which memberships are represented with a series of positive and negative numbers such as assigning 1 for observations that membership of a particular category and -1 for observations that are not members. For more details about such analyses see Nezlek (2011, pp. 26-29; 2012, pp. 77-79).

We begin with a system that contains two categories. In the level 2 example data set we categorized cases as being either female or male. This categorical system is represented by two dummy-codes (one for female and one for male) and by one contrast code. These variables appear in the level 2 file as female, male, and sexcnt respectively. These variables provide the basis for three types of analyses: one that uses a dummy-code as a

predictor, a second that uses a contrast variable, and a third that uses both dummy-codes as predictors. Daily scc is the outcome for all the example analyses in this section.

Consistent with the previous recommendations about centering, in the first analysis female was entered uncentered at level 2. The results of this analysis is contained in the file scc-female(unctr).txt. The intercept is 5.64, which is the estimated score for males, i.e., when female = 0, 5.64 + 0*-1.63. The estimated score for females (i.e., when female = 1) is 5.64 + 1*-1.63 = 4.01. The same values can be obtained from an analysis in which scc is predicted by the dummy-coded variable male, the results of which are in the file scc-male(unctr).txt. In this case the intercept of 4.01 represents the expected score for female, i.e., when male = 0, and the estimated score for males is 4.01 +1*1.63 = 5.64. Note that the significance tests of these two slopes are exactly the same.

Sex differences can also be examined using a contrast code, which is the sexcnt variable in the level 2 file. The results of this analysis are in the file scc-sexcnt(unctr).txt. The t-value and p-value for the sexcnt slope are exactly the same as the values for the two dummy-coded analyses. The difference is the size of the effect and the SE, both of which are half of the values for the dummy-codes. What is distinctive about this analysis is that the intercept is "neutral." It does not represent either men or women, a characteristic that can be particularly important at level 1, as explained below.

Finally, it is possible to estimate means for men and women directly by including both dummy-coded predictors simultaneously. To avoid a singular design matrix, the intercept needs to be dropped from such analyses, which are called no-intercept or zero-intercept models. An example of this is technique is contained in the file scc-sex-zero-intercept.txt. Note that the estimated means for males and females are exactly the same as the estimates from the previous analyses. The difference in these means can be examined by testing the impact on the model fit of constraining the means to be equal. The results of this test are also in the output, and consistent with the results of the previous analyses, constraining the means to be the same significantly reduces the model fit, i.e., the means are significantly different.

Regardless of the level of analysis, when conducting analyses to estimate means for categories using the type of model described above: (1) all cases need to be classified into one and only one category, (2) predictors representing all categories need to be included in the analysis, (3) these predictors need to be entered uncentered, and (4) the intercept needs to be dropped. If *all* of these criteria are not met the analyses will either not work, the results will be difficult to interpret, or the results will be essentially meaningless.

Dummy- and contrast-codes can also be used when there are more than two groups. For dummy codes, membership in each group is represented as described for two groups. These dummy-codes can then be entered individually, in combinations, or into a no-intercept model as described above. If entering dummy-codes individually or in combinations it is important to keep in mind what the intercept represents, which will be a function of the dummy-codes that are included and how they are centered. For contrast codes, we recommend using a set of contrasts that sum to 0 (-1, -1, +2; -1, +2, -1, etc.). Coefficients are tested against a null of 0, and when the contrasts sum to 0 and the contrast is entered uncentered the null represents a direct test of the statistical significance of the contrast and the coefficient represents the estimated mean of the contrast.

The previous examples have concerned modeling intercepts at level 2, but the same techniques can be used to analyze slopes in slopes as outcomes analysis. Similar to the recommendation we made when modeling slopes as a function of continuous predictors at level 2, when modeling slopes as a function of categorical predictors we recommend using the same set of predictors for the intercept and for all slopes.

Categorical predictors can also be used at level 1 (or at levels 1 and 2 in a three-level design), and group memberships can be represented by dummy- and contrast-codes in the same ways that they can be represented at level 2 (as described above). When doing so, analysts are advised to be particularly cautious about what the coefficients represent, particularly intercepts. Unlike coefficients at level 2, coefficients at level 1 are passed up to the next level of analysis, which means that changes in coding and centering will change what is analyzed at level 2. Such changes can have dramatic implications for interpreting results.

For example, in the example level 1 data set there are three variables representing whether a case represents a weekday or a weekend. If weekday is entered as a predictor uncentered then the intercept represents the expected value for a measure on a weekend day, i.e., when weekday = 0. In contrast, if weekend is entered as an uncentered predictor then the intercept represents the expected value for a weekday. Similarly, if weekcnt is entered uncentered, then the intercept represents the expected value for an "average day," a day that is neither a weekday or a weekend day.

3.7. Non-linear outcomes

Our discussion has focused on the analysis of continuous outcomes that are normally distributed or are at least distributed sufficiently close

to normal so as not to require techniques to analyze outcomes with non-normal distributions. Nevertheless, MLM analyses can concern non-linear outcomes such as pass/fail, vaccinated or not, choice of occupation, and so forth that are explicitly non-linear. Similar to single-level analyses in which such outcomes care analyzed with logistic regression instead of OLS multiple regression, within the multilevel context, non-linear outcomes should be analyzed using multilevel logistic regression.

Similar to single level analyses, multilevel logistic regression is needed because non-linear outcomes violate a foundational assumption of statistical inference for analyzing continuous measures: the independence of means and variances. For example, the variance of a binomial outcome is the square root of *npq*, where n = number of observations, p = probability of an outcome, and $q = 1 - p$. Ensuring that means and variances are independent varies as a function of the type of outcome. For example, the algorithm that is used to analyze binomial outcomes is different from the algorithm that is used to analyze nominal three-category outcomes (e.g., picking one of three occupations), which in turn is different from the algorithm that is used to analyze an ordered three-category outcome (e.g., move to a house in the same neighborhood, move to a house in the same city, move far away).

Although a discussion of the different algorithms that are used to estimate parameters for different types of outcomes is well beyond the scope of this article, it is important to note that the logic of multilevel logistic regression is the same as the logic underlying the analysis of continuous outcomes. Within the explanatory framework we have been using for two level models, level 1 coefficients (intercepts and slopes) are estimated and then these coefficients are analyzed at level 2. An introduction to the statistical background of multilevel analyses of non-linear outcomes can be found in Raudenbush and Bryk (2002, pp. 291-335), and some practical advice for conducting and interpreting analyses of non-linear outcomes can be found in Nezlek (2011, pp. 50-52). Finally, it is important to note that for non-linear outcomes the level 1 residual variance is not estimated.

3.8. Model building

The phrase "model building" refers to finalizing the models that will provide the basis for describing the results of a study. Although norms can vary across disciplines, it is probably best to finalize the level 1 model and then add level 2 predictors of the level 1 coefficients. Finalizing in this case

refers to determining which level 1 predictors are included and how they are entered in terms of centering and random effects.

Before adding predictors analysts should conduct what is called an unconditional or null model. This is a model that has no predictors at any level of analysis. Unconditional models provide the basic summary statistics for measures in a multilevel data structure. These statistics include the mean and the variance estimates at different levels of analysis. Note that for measures collected in a multilevel design there are variance estimates for each level of analysis. A single variance estimate should not be used to describe a measure collected in a multilevel design because such estimates confound variances at different levels of analysis.

The variance estimates from an unconditional model can also be used to provide some ideas about how fruitful it might be to add predictors at different levels of a model. For example, is there is relatively little variance at level 2 compared to level 1, it may be easier to explain level 1 variability than level 2 variability, i.e., level 1 predictors may be more informative than level 2 predictors. It is important to note however, that variance distributions cannot be used to indicate that predictors at a level of analysis cannot be significant. Even small amounts of variance can be modeled.

Although there is a healthy debate about the merits of using stepping procedures in OLS regression, for MLM it is probably best to enter predictors using a "forward-stepping" procedure rather than a "backward-stepping" procedure. Forward-stepping means starting with a simple model and adding predictors one at a time to a model (or in small groups). At each step, predictors that significant are retained, and those that are not are dropped. In contrast, in backward-stepping procedures all predictors (or a large set) are included in the initial model, and predictors that are not significant are dropped until all predictors are significant.

The recommendation to use forward-stepping for MLM analysis is primarily due to the fact that MLM analyses estimate more parameters than comparable OLS regression. In MLM the number of estimated parameters increases non-linearly as a function of the number of predictors. For example, in an OLS regression in which there are two predictors 4 parameters are estimated: an intercept, a slope for each predictor, and an error term. In contrast, in MLM, in a level 1 model with two predictors 10 parameters are estimated: a fixed and random term for the intercept and the two slopes (6), a residual error, and the covariances among the error terms (3).

This means that the number of parameters can exceed the "carrying capacity of the data" more quickly in MLM than in comparable OLS analyses. The phrase "carrying capacity" refers to the number of

parameters a data set can estimate reliably. More data (more observations) provides a more stable base for estimating parameters than fewer data. Although decisions about model building need to incorporate theoretical and substantive concerns, the number of predictors in a model needs to be considered within the context of the limitations inherent in the available data. For a more detailed discussion of this topic see Nezlek (2011, pp. 29-33).

There is also the issue of what predictors are included in the level 2 equations for different level 1 coefficients. Unless there are compelling reasons to do otherwise, the same predictors should be entered into the all the level 2 equations *irrespective of whether the relationships are of theoretical or substantive interest*. As noted before, the reason for this is that MLM analyses rely on covariance matrices to estimate coefficients. If a level 2 predictor is not included in the level 2 equation for a level 1 coefficient this assumes that the coefficient for this predictor is 0 and that the covariances between this coefficient and all the other coefficients in the model are 0.

For example, in terms of the data in the example diary data set assume the following level 1 model: y_{ij} (scc) = β_{0j} + β_{1j} (posevent) + β_{1j} (negevent) + r_{ij}. If an analyst was interested in how trait anxiety moderated the relationship between scc and negevent the following level 2 model should be conducted. Note that anxiety is included in the equation for the intercept and for the posevent slope despite the fact that these relationships were not the focus of specific hypotheses.

Intercept: $\beta_{0j} = \gamma_{00} + \gamma_{01}$*(trait anxiety) + u_{0j}.
Posevent slope: $\beta_{1j} = \gamma_{10} + \gamma_{11}$*(trait anxiety) + u_{1j}.
Negevent slope: $\beta_{2j} = \gamma_{20} + \gamma_{21}$*(trait anxiety) + u_{2j}.

3.9. Model fitting

Although MLM analyses provide measures of model fit, such fit indices are usually not an important part of evaluating the results of MLM analyses. In contrast to SEM in which the goal is to evaluate how closely an observed covariance matrix corresponds to a hypothesized covariance matrix, the typical goal of MLM analyses is to test the significance of individual coefficients or sets coefficients, e.g., is a mean slope different from 0? Sometimes analysts test the significance of a coefficient by comparing the fit of a model that includes a predictor to the fit of a model without the predictor. Although this appears to be appropriate, such comparisons involve more than the predictor itself – they also include the

error structures of the models, something that is probably not of particular interest to most analysts.

Moreover, comparing models that have different fixed effects requires the use of full maximum likelihood estimation (ML) instead of restricted maximum likelihood estimation (REML). Although not a problem per se, estimates and tests of fixed effects, which are usually the focus of interest of an analysis, are more accurate using REML than ML. In sum, unless there are compelling reasons to evaluate error structures between models, we recommend that analysts evaluate individual coefficients using tests of the fixed effects: Is a coefficient significantly different from 0?

3.10. Interpreting and reporting results

Although norms vary across disciplines, we believe that it is possible to provide some general recommendations for what should be reported when describing a study that uses MLM and when describing the results of MLM analyses. Moreover, these recommendations can also provide a context for evaluating how well MLM analyses are described in papers and articles.

First, the nature of the design and analyses should be described clearly. For example, in a diary style study one might write that days were treated as nested within persons, in a study of groups one might write that persons were nested within groups, and so forth. Presumably, the number of level 1 and level 2 units in the analyses are described in the methods section. For level 1 units of analysis these descriptions should include the mean and standard deviation of the number of observations for each level 2 unit, and perhaps the minimum and maximum number of observations and the percent of level 2 units meeting a certain criterion.

How predictors were modeled should be described *explicitly*. For predictors at all levels of analysis this should include centering. Descriptions of centering are particularly important because without them it is not possible to understand what the coefficients represent and what the results mean. For predictors at lower levels of analysis (e.g., level 1 in a two-level model and levels 1 and 2 in a three-level model) the description should also include whether the coefficient was modeled as randomly varying or not. As discussed previously, most coefficients should be modeled as randomly varying. If a coefficient is not modeled as randomly varying the reasons for this should be described (Nezlek, 2001; pp. 781-782).

Under NHST (null hypothesis statistical test) tests of hypotheses of fixed effects are two-tailed tests of a null that a coefficient is 0. When describing the results of significance tests one should report the coefficient,

and the accompanying *t*-value and *p*-level. There is no need to report the standard error (se) because *t* = coefficient/se.

MLM analyses estimate unstandardized coefficients. This may present difficulties for readers (and analysts) whose experience with regression analysis is primarily OLS regression in which the norm is to report and interpret standardized coefficients. Presenting estimated values for units of analysis that are high and low on a predictor or that belong to certain groups can help clarify results. Within this context, for continuous measures, high and low are traditionally defined as +/- 1 *SD* from the mean.

To illustrate this, we calculate estimated values using the results of the worked examples for that were presented previously. The simplest example is estimating means for individuals who are high and low on a level 2 variable. This will be illustrated in terms of the results contained in the output file scc-anx(grand).txt. Because anxiety was grand-man centered, the intercept (γ_{00}, 4.69) represents the expected score for an individual who is at the mean of anxiety. The coefficient for the level 2 predictor (anxiety) is -.08. This means that for every 1-unit increase in anxiety, the average scc score decreases .08. The *SD* of anxiety is 11.94. The estimated mean scc score for a person who is 1 SD above the mean on anxiety is 4.69 + 11.94*(-.08) = 4.69 - .96 =3.73. The estimated mean scc score for a person who is 1 SD below the mean on anxiety is 4.69 - 11.94*(-.08) = 4.69 + .96 =5.65. If anxiety had been standardized, the coefficient would directly present the change associated with a 1 *SD* increase in anxiety.

Estimating values for within-person observations (e.g., days) that are high and low on a level 1 predictor is conceptually similar to the procedure described above but requires using the level 1 variance of a predictor to estimate the *SD* of the predictor. This will be illustrated using the results of the analysis contained in pa-negevent(group)(random).txt. The mean slope between pa and negevent is -.54. This means that for every 1-unit increase in negevent, daily pa decreases .54. The within-person *SD* of negevent is estimated in the model negevent-null.txt (.40). Therefore, a 1 *SD* change in negevent consists of .40 units. So, for days that are high (+1 *SD*) on negevent, the expected pa score is 3.92 + 1*(-.54*.40) = 3.92 - (.22) = 3.70, whereas for days that are low on negevent (-1 *SD*) the expected pa score is 3.92 - 1*(-.54*.40) = 3.92 + (.22) = 4.14.

3.11. Effect sizes

Similar to OLS regression and ANOVA, reductions of residual variance can be used to estimate effect sizes in MLM. At level 1, the residual variances

from two models can be compared and a percent of variance explained can be estimated. The level 1 variance of scc from the unconditional model (scc-null.txt) was .558. When negevent was added as a predictor (scc-negevent(group.txt) the level 1 variance was .506. This represents a 9% reduction in variance (.588-.506/.588). Within OLS terms, a 9% reduction in variance corresponds to a correlation of about .30.

The same procedure can be used at level 2, although effect sizes need to be estimated for each coefficient that has been "brought up" from level 1. The simplest example of this is the relationship between a level 2 predictor and a level 1 intercept. The level 2 variance of scc from the unconditional model (scc-null.txt) was 2.465. When trait anxiety was included at level 2 (scc-anx(grand).txt) the level 2 variance was 1.482. This represents a 40% reduction in variance (2.465-1.482)/2.465). Within OLS terms, a 40% reduction in variance corresponds to a correlation of about .63.

The same procedure can be done with slopes; however, note that because estimating effect sizes relies on residual variance estimates coefficients need to be modeled as randomly varying. If a coefficient is not modeled as randomly varying the residual variance is not estimated and an effect size cannot be estimated for the prediction of that coefficient. Intercepts are invariably modeled as randomly varying, whereas slopes may not be.

Finally, we urge analysts to be somewhat cautious when estimating effect sizes. For example, it is possible to add a predictor to a level 1 model that is significant without reducing the level 1 residual variance. This is because unlike in OLS analyses, in MLM significance tests are not based on reductions in variance. Along these lines, Kreft and de Leeuw (1998) noted that "In general, we suggest not setting too much store by the calculation of R^2_B [level 2 variance] or R^2_W [level 1 variance]" (p. 119). See Nezlek (2011, pp. 35-36) for a more detailed discussion of calculating effect sizes in MLM.

4. SOFTWARE OPTIONS

MLM analyses can be conducted using a wide variety of software packages. Some options, such as HLM and MLwiN, conduct only MLM analyses. MLM can also be conducted using general purpose packages such as SPSS, SAS, and Stata. Mplus, a general modeling program, can also conduct MLM analyses. There is also growing interest in conducting MLM analyses using R. It is important to note that different programs

will provide the same results if the same models are specified (and in some cases, the same estimation algorithms are used). The logic and the algorithms of MLM are reasonably well-understood and agreed upon meaning that different programs simply represent different ways of doing the same thing.

For analysts with little or no experience conducting MLM, we recommend beginning with HLM. HLM was developed to conduct MLM, and so the input and output options were designed to meet the specific needs of MLM. The interface is intuitive and relatively straightforward. For example, decisions about centering predictors are made when predictors are entered, and centering is done by the program, not as a separate process or step. Error terms are displayed on screen as part of a model before an analysis is conducted, and they can be included or deleted on an individual basis with a mouse click. Moreover, a fully functional and free shareware version of the program is available (SSI, 2019). This version of the program limits the number of cases and the number of predictors in a model, but these limits are generous. All of the analyses described in this paper could have been done using the shareware version of the program.

After using HLM, analysts may want to use a program with which they are more familiar, but HLM is perhaps the easiest program to use to learn MLM. For analysts with little or no experience conducting MLM but who do not want to use HLM, we recommend using whatever program they prefer to conduct the analyses we have presented in this paper. The results should be similar to the results of the HLM analyses we present within 4 or 5 points to the right of the decimal. Regardless of the program they use, analysts should be able to recognize where different parameters are displayed in the output of the program they have used. At a minimum, these parameters include the estimates and tests of the fixed and random effects and the estimates of the covariances between the random effects. For diagnostic purposes it can be useful to know the iteration history and convergence criteria.

5. POWER ANALYSIS

Estimating the power of a multilevel design is particularly challenging, a reality that we believe is not recognized by most editors and reviewers. In contrast to OLS based analyses in which effect sizes are well-understood and power can be estimated with pin-point accuracy (e.g., G-Power; Faul,

Erdfelder, Buchner, & Lang, 2009), in MLM a myriad of factors can affect the power of a design. These include, but are not limited to, the size of the coefficient, the distribution across the levels of the model of the variances of the outcomes and the predictors, the number of observations at each level of analysis, the error structure of the level 1 measures (and the level 2 measures in a three level data structure), the reliability of the level 1 predictors, and of course, the probability level. Moreover, making apriori assumptions about some of these parameters can be difficult.

With this in mind we present a set of techniques that we believe represent contemporary best-practice. These techniques require the use of R (R Development Core Team, 2015), and the techniques we describe can be challenging for individuals not familiar with using R modules to conduct MLM analyses. Analysts who are interested in recommendations for specific design parameters can consult Aguinis, Gottfredson, and Culpepper (2013), Richter (2006), Scherbaum and Ferreter (2009), and Snijders (2005).

Regardless, for analysts who are not experienced users of R we recommend first conducting analyses in another software package with which they are familiar and can be confident about the results and understanding them. Once the model is finalized then the analysis can be conducted in R to generate the information that is needed for the simulation. Analysts need to be certain that the models they specify in R have the appropriate and desired centering options and error structures. If the same models are specified in R and in another package (e.g., HLM), the results will not differ aside from rounding errors (i.e., in the third of fourth place decimal).

In this paper we start with what is called "observed power," the statistical power of a test that has been performed, based on the analyses of the data at hand. Although there is a debate about whether observed power should be used (Yuan & Maxwell, 2005), we thought that calculating observed power would be useful as a starting point. Readers interested in a deeper understating of power calculation in the context of MLM should consult Kain, Bolker, and McCoy (2015) and Johnson, Barry, Ferguson, and Müller (2015).

Although there are numerous R packages that can be used to estimate the power of a multilevel model, we found simR (Green & MacLeod, 2015) to be the easiest to use. SimR can be used in combination with models fit by lme4 functions: `lmer()` or `glmer()` (Bates, Machler, Bolker, & Walker, 2015). The logic of the procedure is as follows. `lmer()` estimates all the parameters of a MLM, including fixed effects, error terms, error covariances, and so forth. Such models can be based on a pilot study or a finished study. These parameters are then read by a function `powerCurve()`, which is part

of simR. `powerCurve()` then runs a series of models on simulated data, based on the parameters provided by `lmer()`. This presentation focuses on estimating fixed effects because fixed effects are typically the focus of hypotheses, and it focuses on sample sizes.

Note: In the following examples we use 100 simulations. This is meaningfully lower than the standard, which is at least 1000 and often 5000. We did this because running these simulations can take a reasonable amount of time, even on a high-powered computer (i.e., multiple cores with large memory). Running only 100 simulations can typically be done fairly quickly on a modestly powered machine. Using only 100 simulations allows analysts to determine if they are using the R modules correctly in a short period of time.

R accesses data via what is called a "dataframe", and the data we used for the R analyses are in a file named "Rdataframe.csv" in the supplemental materials. This file contains the following variables: recnumb (record number in the data set), subj (participant id), day (day number, within each participant), scc (daily self-concept clarity), tri (daily triad measure), PosEvent (daily positive events), NegEvent (daily negative events), pa (daily positive affect), na (daily negative affect), posC (daily positive events, group-mean centered), negC (daily negative events, group-mean centered), paC (daily positive affect, group-mean centered), naC (daily negative affect, group-mean centered), anx (person level, trait anxiety), and anxC (person level, trait anxiety, grand-mean centered).

To conduct the analyses we describe you will need to download the .csv file from the supplemental materials and load it to R, and you will need to install and load the lme4 and simr packages:

R syntax:
```
dataf <- read.csv("Rdataframe.csv",
stringsAsFactors=FALSE)
```

R syntax:
```
install.packages(c("lme4", "simr"),
dependencies = T)
```

R syntax:
```
library(lme4); library(simr)
```

In the first model `na` (daily negative affect) is modeled group-mean centered and as randomly varying. The equations for the comparable HLM analysis are presented below, and the results of the HLM analysis are in scc-na(group)-random.txt. The R syntax needed to estimate the same model follows the HLM model equations. The mean slope for na (γ_{10}) is -.24 in both the HLM and R analyses.

Level 1: $y_{ij}\,(scc) = \beta_{0j} + \beta_{1j}\,(na) + r_{ij}$.

Level 2: $\beta_{0j} = \gamma_{00} + u_{0j}$.

$\beta_{1j} = \gamma_{10} + u_{1j}$.

R syntax: `m1 <- lmer(scc ~ naC + (naC|subj), dataf)`

`scc` (~) is modelled as a function of `naC` with (+) a random slope of `naC` within each subject (`naC|subj`). The results of this analysis (the model parameters) are stored in an object labeled `m1`.

To estimate power, we used an estimated coefficient that was smaller than the original coefficient (-.15 vs -.24), and the following R command fixes the coefficient for na to this new value. For this model and these parameters, a smaller coefficient was needed to make the simulation for the number of days to run successfully. This will not always be the case.

R syntax: `fixef(m1)["naC"] <- -0.15`

First, we examined how the power to detect this coefficient varied as a function of the number of participants, holding the number of days constant (i.e., the number of days in the sample data set).

R syntax:
```
powerCurve(m1, along="subj", nsim=100,
        breaks = c(20, 30, 40, 50, 60, 70, 80, 90))
```

This syntax conducts 100 simulations (`nsim=100`) of a model stored in `m1`, varying the number of participants (`along="subj"`) which is specified in "`breaks=`".

R output:

```
Power for predictor 'naC', (95% confidence interval)
by value of subj:
21: 32.00% (23.02, 42.08) - 285 rows
34: 47.00% (36.94, 57.24) - 421 rows
44: 59.00% (48.71, 68.74) - 544 rows
55: 68.00% (57.92, 76.98) - 673 rows
65: 78.00% (68.61, 85.67) - 806 rows
77: 81.00% (71.93, 88.16) - 927 rows
87: 83.00% (74.18, 89.77) - 1056 rows
100: 91.00% (83.60, 95.80) - 1169 rows
```

Next, we examined how the power to detect this coefficient varied as a function of the number of days participants provide, holding constant the number of participants (i.e., the number of participants in the sample data set). Note that the same results (m1) are used, but the "`along=`" argument has changed, and the accompanying breaks have changed.

R syntax:
```
powerCurve(m1, along="day", nsim=100, breaks =
        c(7,8,9,10,14))
```

R output:

```
Power for predictor 'naC', (95% confidence interval)
by value of day:
```

```
7: 77.00% (67.51, 84.83) - 717 rows
8: 83.00% (74.18, 89.77) - 814 rows
9: 84.00% (75.32, 90.57) - 907 rows
10: 86.00% (77.63, 92.13) - 996 rows
14: 90.00% (82.38, 95.10) - 1264 rows
```

The next example has an unconditional model at level 1 (no predictors) and a single level 2 predictor. `Scc` is the outcome and anxiety is the level 2 predictor, entered grand-mean centered. The equations for the comparable HLM analysis are presented below and the results of the HLM analysis are in scc-anx(grand).txt. The R syntax following the model equations estimates the same model. The coefficient for anxiety γ_{01} is -.08 in both analyses.

Level 1: y_{ij} (scc) = $\beta_{0j} + r_{ij}$.
Level 2: $\beta_{0j} = \gamma_{00} + \gamma_{01}$ (anxiety) + u_{0j}.

R syntax: `m2 <- lmer(scc ~ anxC + (1|subj), dataf)`

First, we examined how power varied as a function of the number of participants, holding the number of days constant (i.e., the number of days in the sample data set). The value of the coefficient for anxiety was set to -.05.

R syntax: `fixef(m2)[«anxC»] <- -0.05`
R syntax:
```
powerCurve(m2, along="subj", nsim=100,
        breaks = c(20, 30, 40, 50, 60, 70, 80, 90))
```

R output:

```
Power for predictor 'anxC', (95% confidence interval)
by value of subj:
21: 67.00% (56.88, 76.08) - 285 rows
34: 85.00% (76.47, 91.35) - 421 rows
44: 89.00% (81.17, 94.38) - 544 rows
55: 96.00% (90.07, 98.90) - 673 rows
65: 99.00% (94.55, 99.97) - 806 rows
77: 99.00% (94.55, 99.97) - 927 rows
87: 98.00% (92.96, 99.76) - 1056 rows
100: 99.00% (94.55, 99.97) - 1169 rows
```

Next, we examined how power varied as a function of the number of days participants provide, holding constant the number of participants (i.e., the number of participants in the sample data set).

R syntax:
```
powerCurve(m2, along="day", nsim=100,
        breaks = c(7,8,9,10,14))
```

R output:

```
Power for predictor 'anxC', (95% confidence interval)
by largest value of day:
7: 100.0% (96.38, 100.0) - 717 rows
8: 100.0% (96.38, 100.0) - 814 rows
9: 100.0% (96.38, 100.0) - 907 rows
10: 100.0% (96.38, 100.0) - 996 rows
14: 100.0% (96.38, 100.0) - 1264 rows
```

The final example involves estimating power for a cross-level interaction: How does a level 2 variable moderate a level 1 relationship? This example focuses on the following model in equation form. Note that the outcome is now `tri`, `negevent` is a predictor entered group-mean centered and randomly varying, and anxiety is a level 2 predictor entered grand-mean centered. The R syntax follows the equations, and the parameter estimates are saved to `m3`.

Level 1: y_{ij} (tri) = β_{0j} + β_{1j} (negevent) + r_{ij}.
Level 2: $\beta_{0j} = \gamma_{00} + \gamma_{01}$ (anxiety) + u_{0j}.
$\beta_{1j} = \gamma_{10} + \gamma_{11}$ (anxiety) + u_{1j}.

R syntax:
```
m3 <- lmer(tri ~ anxC * negC + (negC|subj),
      dataf)
```

This example provides estimated power for the moderating effect for two different moderating effects (.01 and .02) and for differences in participants and for days.

The R syntax for estimating the power for different numbers of subjects is below. Note that moderating effect is represented by "`anxC:negC`".

R syntax:
```
powerCurve(m3, along="subj", nsim=100,
      breaks = c(20, 30, 40, 50, 60, 70, 80, 90),
      test=fixed("anxC:negC", method="z"))
```

The R syntax for estimating the power for different numbers of days is below. Note the addition of the parameter for method, which instructs the algorithm to use a z-test. Otherwise, the analyses will rely upon bootstrapping, which in this case could be computationally intensive.

R syntax:
```
powerCurve(m3, along="day", nsim=100,
      breaks = c(3,4,5,5,6,7, 10, 14),
      test=fixed("anxC:negC", method = "z"))
```

The interaction term is fixed using the same syntax as that used to fix the effects in the previous analyses. For an effect of .01, this is done as follows.

R syntax:
```
fixef(m3)["anxC:negC"] <- 0.01
```

```
R output power as a function of participants for
anxC:negC = .01:
```

```
21: 12.00% (6.36, 20.02) - 285 rows
34: 14.00% (7.87, 22.37) - 421 rows
44: 14.00% (7.87, 22.37) - 544 rows
55: 16.00% (9.43, 24.68) - 673 rows
65: 26.00% (17.74, 35.73) - 806 rows
77: 25.00% (16.88, 34.66) - 927 rows
87: 29.00% (20.36, 38.93) - 1056 rows
100: 27.00% (18.61, 36.80) - 1169 rows
R output power as a function of days for
anxC:negC = .01:
3: 13.00% (7.11, 21.20) - 309 rows
4: 12.00% (6.36, 20.02) - 412 rows
5: 15.00% (8.65, 23.53) - 515 rows
6: 16.00% (9.43, 24.68) - 618 rows
7: 20.00% (12.67, 29.18) - 717 rows
10: 24.00% (16.02, 33.57) - 996 rows
14: 28.00% (19.48, 37.87) - 1264 rows
```

To estimate power for an effect of .02, this is done as follows.

R syntax: `fixef(m3)["anxC:negC"] <- 0.02`

```
R output power as a function of participants for
anxC:nevgC = .02:
21: 29.00% (20.36, 38.93) - 285 rows
34: 40.00% (30.33, 50.28) - 421 rows
44: 42.00% (32.20, 52.29) - 544 rows
55: 53.00% (42.76, 63.06) - 673 rows
65: 67.00% (56.88, 76.08) - 806 rows
77: 69.00% (58.97, 77.87) - 927 rows
87: 68.00% (57.92, 76.98) - 1056 rows
100: 72.00% (62.13, 80.52) - 1169 rows
R output power as a function of days for
anxC:negC = .02:
3: 23.00% (15.17, 32.49) - 309 rows
4: 29.00% (20.36, 38.93) - 412 rows
5: 40.00% (30.33, 50.28) - 515 rows
5: 40.00% (30.33, 50.28) - 515 rows
6: 41.00% (31.26, 51.29) - 618 rows
7: 54.00% (43.74, 64.02) - 717 rows
10: 66.00% (55.85, 75.18) - 996 rows
14: 78.00% (68.61, 85.67) - 1264 rows
```

The preceding examples are intended to demonstrate how to estimate power in a MLM design. They should not be used as guidelines for how many participants or days are needed to achieve certain levels of power.

This limitation is a function of the fact that these estimates are based on numerous parameters (e.g., the error covariance matrix) that we estimated from the example data set. It is beyond the scope of this article to consider all these factors simultaneously. Nevertheless, if analysts have some way of estimating these other parameters (e.g., a pilot study or data from a previous study), these estimates can be used to provide a good-faith estimate of the power of a design. In terms of such possibilities, we underscore the importance of the word "estimate"—there are no guarantees, and the importance of the term "good faith"—using an inappropriate basis for making estimates undermines the validity of the estimates.

6. WHEN TO USE MLM

Last, and certainly not least, there is the issue of when to use MLM. Our advice about this is simple. If you have a multilevel data structure (i.e., observations at one level of analysis are nested within observations at another level of analysis), you should analyze these data with MLM. Although this article has focused on intensive repeated measures designs, we thought we would address the issue of when to use MLM in general.

First, although is it typically not relevant to intensive repeated measures designs because typically, intensive repeated measures designs have numerous level 2 units (people), we address the issue of using ICCs to determine if MLM is appropriate. Put simply, we strongly recommend ignoring advice about using intraclass correlations (ICCs) to determine if MLM is appropriate. The ICC indicates how much of the variance of a measure is at each level of analysis. Specifically, it is the ratio of the level 2 variance to the sum of the level 2 and level 1 variances. An ICC of 0 means that there is no level 2 variance in a measure, i.e., all the level 2 observations have the same mean. An ICC of 1.0 indicates that all of the variance is at level 2, i.e., there is no within-unit variance. In terms of a study of math achievement in which students were nested within schools, an ICC of 0 would occur if all schools had the same mean achievement, whereas an ICC of 1.0 would occur if all of the students in each school had the same score but schools differed.

Some argue that analysts can ignore the nesting inherent in a data structure if the ICC is low because this means that the nesting is unimportant. If there is no level 2 variance why include nesting? The answer to this is quite simple. ICCs refer to the distribution of the variance

of means at level 2. ICCs do not address the possibility that slopes (level 1 relationships) vary between level 2 units of analysis.

The shortcomings of relying on ICCs are illustrated by the sample data presented in Table I. There are six groups, and there are five observations in each group for two variables, x and y. The ICCs for both x and y are 0. The means for both measures for all groups are the same (23). Nevertheless, as can be seen from the data in Table I, the relationship between x and y is negative in groups 1, 2, and 3, whereas it is positive in groups 4, 5, and 6. If the grouping of these data is ignored, i.e., if all 30 observations are analyzed as a single group, the estimated relationship between x and y is 0. Clearly, this is inaccurate. Clearly, ICCs cannot be used to indicate when MLM is appropriate for nested data.

Nonetheless, there are situations in which observations are nested but a MLM is not appropriate. Recall that one of the advantages of MLM is that inferences can be made to two populations: the population represented by the level 2 units and the population represented by the level 1 units. Making inferences requires a sufficient number of observations to provide a basis for inference. Imagine a study in which 1,000 observations were collected in two countries. One might imagine that such data should be anlyzed using a MLM with participants nested within countires. Although technically accurate, such an analysis would not be appropriate because two countries do not provide a reasonable basis for drawing an inference about the population of countries; two is not enough.

Extending the present example, how many countries would be enough to justify MLM? In addition to questions about power (see previous section), it is difficult to provide an all-purpose estimate. Noting this, analysts can simply think how large a sample is needed to provide a basis for inference in general. MLM is not magical. If you do not have enough observations to provide a basis for inference you do not have enough.

What options are available when a data structure is nested, but there are not enough observations to provide a basis for making inferences, e.g., 1,000 people in two countries? In such cases, analysts can conduct OLS analyses that allow for the possibility that level 1 relationships vary between level 2 units. This can be done by creating interactions terms representing differences in slopes between/among the level 2 units. For example, in a study on the relationship between personal income and self-esteem conducted in two countries, differences between the two countries in this relationship can be examined using standard OLS techniques to examine moderation. Such analyses treat slopes as fixed, not random because there are not enough countries to model the random variability. As the number of level 2 units increases such techniques become unwieldy and MLM becomes

more appropriate (i.e., it is easier to estimate the random effect for level 1 coefficients). Regardless of how it is done, it is critical to take into account the possibility that level 1 slopes vary across level 2 units of analysis.

Bon voyage.

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*(6), 1490-1528.

Bates, D., Machler, M., Bolker, B. and Walker, S. (2015) Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*, 1-48.

Bressoux, P. (2020). Using multilevel models is not just a matter of statistical adjustment. Illustrations in the educational field. *L'Année Psychologique. Topics in Cognitive Psychology.*

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models*. Newbury Park, CA: Sage.

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.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*, 1149-1160.

Green, P., & MacLeod, C. J. (2016), SIMR: an R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution, 7*, 493-498.

Johnson, P. C. D., Barry, S. J. E., Ferguson, H. M., & Müller, P. (2015). Power analysis for generalized linear mixed models in ecology and evolution. *Methods in Ecology and Evolution, 6*, 133–142.

Kain, M. P., Bolker, B. M., & McCoy, M. W. (2015). A practical guide and power analysis for GLMMs: Detecting among treatment variation in random effects. *PeerJ3*, e1226

Kreft, I. G. G., & de Leeuw, J. (1998). *Introducing multilevel modeling.* Newbury Park, CA: Sage.

Nezlek, J. B., & Mrozinski, B. (2019, January 28). Annee Psychologique. Retrieved from osf.io/74m5r

Nezlek, J. B. (2001). Multilevel random coefficient analyses of event and interval contingent data in social and personality psychology research. *Personality and Social Psychology Bulletin*, *27*, 771-785. doi:10.1177/0146167201277001

Nezlek, J. B., & Plesko, R. M. (2001). Day-to-day relationships among self-concept clarity, self-esteem, daily events, and mood. *Personality and Social Psychology Bulletin*, *27*, 201-211.

Nezlek, J. B. (2003). Using multilevel random coefficient modeling to analyze social interaction diary data. *Journal of Social and Personal Relationships*, *20*, 437-469.

Nezlek, J. B. (2007). A multilevel framework for understanding relationships among

traits, states, situations, and behaviors. *European Journal of Personality, 21*, 789-810. doi: 10.1002/per.640

Nezlek, J. B. (2012a). Multilevel modeling of diary-style data. In M. R. Mehl & T. S. Conner (Eds.) *Handbook of Research Methods for Studying Daily Life*. (pp. 357-383). New York: Guilford Press.

Nezlek, J. B. (2012b). Diary methods for social and personality psychology. In J. B. Nezlek (Ed.) *The SAGE Library in Social and Personality Psychology Methods*. London: Sage Publications.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models* (2nd ed.). Newbury Park, CA: Sage.

Richter, T. (2006). What is wrong with ANOVA and multiple regression? Analyzing sentence reading times with hierarchical linear models. *Discourse Processes, 41*, 221–250.

Scherbaum, C. M., & Ferreter, J. M. (2009). Estimating statistical power and required sample sizes for organizational research using multilevel modeling. *Organizational Research Methods*, *12*, 237–367.

Snijders, T. A. (2005). Power and sample size in multilevel linear models. In B. S. Everitt and D. C. Howell (eds.), *Encyclopedia of Statistics in Behavioral Science* (V. 3), 1570-1573. Chicester: Wiley, 2005.

SSI (2019). http://ssicentral.com/index.php/products/hml/free-downloads-hlm

Yuan, K. H., & Maxwell, S. (2005). On the post hoc power in testing mean differences. *Journal of Educational and Behavioral Statistics, 30*, 141–167.

Table 1. Two level data, ICC = 0, but slopes vary.
Tableau 1. Données à deux niveaux, ICC = 0, mais les pentes varient.

Group 1		Group 2		Group 3	
x	y	x	y	x	y
21	25	21	25	21	25
22	24	22	24	22	24
23	23	23	23	23	23
24	22	24	22	24	22
25	21	25	21	25	21
Group 4		**Group 5**		**Group 6**	
x	y	x	y	x	y
21	21	21	21	21	21
22	22	22	22	22	22
23	23	23	23	23	23
24	24	24	24	24	24
25	25	25	25	25	25