

Psychological Methods

Centering Categorical Predictors in Multilevel Models: Best Practices and Interpretation

Haley E. Yaremych, Kristopher J. Preacher, and Donald Hedeker

Online First Publication, December 16, 2021. <http://dx.doi.org/10.1037/met0000434>

CITATION

Yaremych, H. E., Preacher, K. J., & Hedeker, D. (2021, December 16). Centering Categorical Predictors in Multilevel Models: Best Practices and Interpretation. *Psychological Methods*. Advance online publication. <http://dx.doi.org/10.1037/met0000434>

Centering Categorical Predictors in Multilevel Models: Best Practices and Interpretation

Haley E. Yaremych¹, Kristopher J. Preacher¹, and Donald Hedeker²

¹ Department of Psychology and Human Development, Vanderbilt University

² Department of Public Health Sciences, University of Chicago

Abstract




The topic of centering in multilevel modeling (MLM) has received substantial attention from methodologists, as different centering choices for lower-level predictors present important ramifications for the estimation and interpretation of model parameters. However, the centering literature has focused almost exclusively on continuous predictors, with little attention paid to whether and how categorical predictors should be centered, despite their ubiquity across applied fields. Alongside this gap in the methodological literature, a review of applied articles showed that researchers center categorical predictors infrequently and inconsistently. Algebraically and statistically, continuous and categorical predictors behave the same, but researchers using them do not, and for many, interpreting the effects of categorical predictors is not intuitive. Thus, the goals of this tutorial article are twofold: to clarify why and how categorical predictors should be centered in MLM, and to explain how multilevel regression coefficients resulting from centered categorical predictors should be interpreted. We first provide algebraic support showing that uncentered coding variables result in a conflated blend of the within- and between-cluster effects of a multicategorical predictor, whereas appropriate centering techniques yield level-specific effects. Next, we provide algebraic derivations to illuminate precisely how the within- and between-cluster effects of a multicategorical predictor should be interpreted under dummy, contrast, and effect coding schemes. Finally, we provide a detailed demonstration of our conclusions with an empirical example. Implications for practice, including relevance of our findings to categorical control variables (i.e., covariates), interaction terms with categorical focal predictors, and multilevel latent variable models, are discussed.

Translational Abstract

Multilevel modeling (MLM) is frequently used in the social sciences when data are nested or clustered (e.g., students nested within classrooms; clients nested within therapists). Centering is an important topic in MLM because it can be conducted in different ways, each of which yields slightly different parameter estimates that also must be interpreted differently. However, work regarding centering has focused almost exclusively on continuous predictors. Little attention has been paid to categorical predictors, whether and how they should be centered, and how their resulting coefficients should be interpreted. This is problematic, because categorical predictors and covariates are ubiquitous across all fields wherein MLM is used. Thus, the goals of this report are to clarify why and how categorical predictors should be centered in MLM, and to explain how multilevel regression coefficients resulting from centered categorical predictors should be interpreted. We present an overview of popular centering options and provide best-practice recommendations for centering and interpretation of binary and multicategorical predictors. We provide a detailed demonstration of our conclusions with an empirical example from the education literature. In addition, we discuss the practical implications of our work at length; topics include multicategorical covariates, interaction terms with categorical focal predictors, and multilevel latent variable models.

Keywords: multilevel modeling, hierarchical linear modeling, centering, categorical predictors, binary predictors

Supplemental materials: <https://doi.org/10.1037/met0000434.supp>

Haley E. Yaremych  <https://orcid.org/0000-0002-5963-0758>
Kristopher J. Preacher  <https://orcid.org/0000-0003-4099-3636>
Donald Hedeker  <https://orcid.org/0000-0001-8134-6094>
Components of this project were previously presented at the Association

for Psychological Science Annual Convention, 2020.

Correspondence concerning this article should be addressed to Haley E. Yaremych, Department of Psychology and Human Development, Vanderbilt University, PMB 552, 230 Appleton Place Nashville, TN 37203-5721, United States. Email: haley.e.yaremych@vanderbilt.edu

In the multilevel modeling (MLM) literature, the topic of centering has been discussed and debated at great length, as different centering choices for lower-level predictors present important ramifications for the estimation and interpretation of parameter estimates (Hofmann & Gavin, 1998; Kreft et al., 1995). However, such work has focused almost exclusively on continuous predictors, with little attention paid to whether and how categorical predictors should be centered. This is problematic given the ubiquity of categorical predictors in multilevel data across the applied fields of psychology, education, organizational research, and more. Thus, the goals of this tutorial article are twofold: to clarify why and how categorical predictors should be centered in MLM, and to explain how multilevel regression coefficients resulting from centered categorical predictors should be interpreted.

Notation and Equivalence

We use the following notation common in MLM literature and textbooks (e.g., Raudenbush & Bryk, 2002; Snijders & Bosker, 2012): The subscript i refers to a Level-1 unit within a cluster, whereas j is a cluster indicator. Thus, we observe individual i nested within cluster j . In MLM, Level-1 predictors are often centered in one of three ways. First, the predictor may be uncentered (UN; denoted x_{ij}). Second, centering at the grand mean (CGM; denoted $x_{ij} - \bar{x}_{..}$) involves subtracting the overall sample's "grand" mean ($\bar{x}_{..}$) from each observation. Third, centering at the cluster mean,¹ also known as centering within context (CWC; denoted $x_{ij} - \bar{x}_{.j}$), involves subtracting the cluster-specific mean ($\bar{x}_{.j}$) from each observation. The centering technique chosen for the Level-1 predictor(s) influences the interpretation of parameter estimates, and often, the estimated values themselves, yielded by the multilevel model.

As a preliminary note, models containing UN and CGM predictors are equivalent, meaning that all the parameter estimates of one model are either the same as, or a simple linear transformation of, those from the other model. Both models will fit the data identically. The primary difference between UN and CGM predictors pertains to interpretation of the intercept. For UN predictors, the intercept is the predicted value of the outcome when all predictors are zero. For CGM predictors, the intercept is the predicted value of the outcome when all predictors are equal to the grand mean of the sample; this interpretation is often more useful, as continuous psychological scales typically do not contain meaningful zero-points. Nevertheless, UN predictors are used most frequently in practice, and therefore will be a focus of this report for practical relevance. We will not address CGM predictors in detail, though conclusions drawn for UN predictors will apply to CGM predictors as well. Finally, models with CWC predictors are *not* equivalent to those with either UN or CGM predictors, for reasons described in detail below.

Running Empirical Example

Throughout this tutorial, we will draw from a data set originally reported on by Paterson (1991) to investigate relationships between students' socioeconomic status (SES) and academic achievement in Scotland. The data set was obtained from the Centre for Multilevel Modeling at the University of Bristol (<http://www.bristol.ac.uk/cmm/learning/mmssoftware/data-rev.html#lev-xc>). The sample consists of 3,435 children nested within 148 primary schools; cluster size ranged from 1 to 72, with a mean cluster size of 23.21 ($SD = 16.78$). The Level-1 outcome variable, $ATTAIN_{ij}$, is students' academic

achievement score at age 16, measured on a scale from 1 to 10 ($M = 5.69$, $SD = 3.06$). We constructed a four-category predictor of academic achievement, PED_{ij} , based on two indicators of parental education: Group 1 = neither parent has a high level of education; Group 2 = mother is educated, father is not; Group 3 = father is educated, mother is not; Group 4 = both parents are educated. These categories allow us to distinguish between mothers' and fathers' education, two unique markers of SES, and how they may differentially and jointly influence children's academic outcomes. Throughout the tutorial we use "educated" to describe parents who attended school until they were at least 16 years of age, and "not educated" to describe parents who left school at age 15 or younger. In total, 58.54% of students were in Group 1; 13.92% were in Group 2; 7.25% were in Group 3; and 20.29% were in Group 4.

Coding Schemes

An important precursor to the examination of nominal categorical predictors is the discussion of coding schemes. Categorical independent variables may be represented by a variety of coding systems (for a thorough overview of the most common systems, including those used here, see Cohen et al., 2003; Chapter 8). In all cases, a k -category predictor is represented by $k - 1$ coding variables. Regardless of the coding system chosen, equivalent results will be obtained for the omnibus effect of the predictor; the researcher's choice of coding system does not fundamentally alter the model or the information carried within the coding variables. However, each coding system will produce different sets of regression coefficients, each of which must be interpreted differently and answers different central questions. In this tutorial we will focus on dummy codes, contrast codes, and effect codes, as these are most frequently used in practice. We will refer to "coding variables" as the general case, as the conclusions and recommendations presented here will apply equivalently to any coding system.

All coding schemes require the researcher to choose a *reference group*. This choice is statistically arbitrary, but substantively important—the reference group should allow for useful comparisons. Each coding variable makes a specific comparison between the reference group and what we refer to as the *focal group*, or Group f . How this comparison is interpreted will vary according to the coding scheme. All coding schemes used in our example are shown in Table 1.

In our example data set, we first created three dummy codes, denoted d_{1ij} , d_{2ij} , d_{3ij} . In typical multiple regression (i.e., not a multilevel setting), each dummy code's slope is interpreted as the mean difference on y between the focal group and the reference group. We chose children with no educated parents as the reference group because it would yield logical interpretations, allowing us to identify academic achievement gains for children with one or two educated parents over those with none. The first dummy code, d_{1ij} , was coded 1 for children with only an educated mother and 0 for all other children. The others, d_{2ij} and d_{3ij} , were coded 1 for

¹ Throughout this report we use *cluster* to refer to the Level-2 unit of nesting, and *group* to refer to a given category of a categorical variable. For clarity and simplicity, we restrict our focus to two-level models, although we expect that our conclusions will generalize intuitively to three-level models, as has been shown for continuous predictors (Brincks et al., 2017).

Table 1
Coding Schemes Used in Empirical Example

Dummy coding			
Educated?	d_{1ij}	d_{2ij}	d_{3ij}
Mom no, dad no	0	0	0
Mom yes, dad no	1	0	0
Mom no, dad yes	0	1	0
Mom yes, dad yes	0	0	1
Contrast coding			
Educated?	c_{1ij}	c_{2ij}	c_{3ij}
Mom no, dad no	-3/4	0	0
Mom yes, dad no	1/4	-1/3	-1/2
Mom no, dad yes	1/4	-1/3	1/2
Mom yes, dad yes	1/4	2/3	0
Effect coding			
Educated?	e_{1ij}	e_{2ij}	e_{3ij}
Mom no, dad no	-1	-1	-1
Mom yes, dad no	1	0	0
Mom no, dad yes	0	1	0
Mom yes, dad yes	0	0	1

children with only an educated father and 1 for children with two educated parents, respectively.

Next, we created three contrast codes, denoted c_{1ij} , c_{2ij} , c_{3ij} . Each contrast code can be constructed in a variety of ways in order to compare a particular focal group (or set of groups) against a particular reference group (or set of groups). As long as there are $k - 1$ contrast codes in total, a nearly infinite set of codes can be created to test various hypotheses. For each contrast code, any groups not involved in the contrast are coded 0, whereas the focal and reference groups receive codes that are weighted according to how many groups are involved in the contrast (see Table 1). First, c_{1ij} probes the effect of having *any* educated parents by comparing the mean of Group 1 to the overall mean of Groups 2–4. Second, c_{2ij} probes the effect of having one versus two educated parents by comparing the mean of Group 4 with the overall mean of Groups 2 and 3 (Group 1 is not involved in this contrast). Finally, c_{3ij} probes the effect of having an educated mother versus an educated father by comparing the mean of Group 2 to that of Group 3 (Groups 1 and 4 are not involved in this contrast). These contrasts were chosen for their substantive utility and to demonstrate interpretations for codes involving varying subsets of groups.

Finally, we created unweighted effect codes² denoted e_{1ij} , e_{2ij} , e_{3ij} . The slope of an unweighted effect code reflects the difference between the mean of the focal group and the unweighted *mean of all group means* in the sample. The reference group is coded -1, the focal group is coded 1, and all other groups are coded 0. Mean differences are no longer estimated with respect to the reference group, but the model will not yield a specific estimate for this group. We again chose children with no educated parents as the reference group so we could identify the effects of either/both parents' education relative to the overall sample. Thus, children with no educated parents received -1 on all effect codes. Children with only an educated mother were coded 1 on e_{1ij} , children with only an educated father were coded 1 on e_{2ij} , and children with two educated parents were coded 1 on e_{3ij} . We will return to each

of these coding schemes, and their associated interpretations in a multilevel setting, in the Empirical Example section.

Prior Work on Centering Categorical Predictors

In general, methodologists have advocated the use of centering techniques that partition the effects of a predictor in a manner that is consistent with the research question at hand, emphasizing that parameter estimates must be interpreted differently depending upon the centering method chosen (Enders & Tofighi, 2007; Grilli & Rampichini, 2018; Hofmann & Gavin, 1998; Kreft et al., 1995; Van Landeghem et al., 1999). Methodologists also encourage researchers to exercise transparency in describing their centering choices and the motivation preceding them (Enders & Tofighi, 2007).

However, previous work regarding centering has focused almost entirely on continuous predictors. Little work has been dedicated to categorical coding variables, whether they, too, should be centered, and if centered, how their resulting coefficients may be interpreted. Raudenbush and Bryk (2002) and Enders and Tofighi (2007) have briefly addressed centering binary predictors, showing algebraically that regardless of whether binary dummy coding or binary effect coding is used, intercepts can be interpreted similarly to how they are interpreted for continuous predictors. Under CGM, a random intercept β_{0j} can be interpreted as the predicted value of the outcome for cluster j when the predictor equals the grand mean, which in the context of binary predictors may be best understood as an "adjusted" cluster mean (adjusted for the proportion of comparison-group cases across the entire sample; in other words, β_{0j} is the cluster mean that would result if the proportion of comparison-group cases were equal across all clusters). Under CWC, the intercept should be interpreted as the unadjusted cluster mean.

This supporting algebra by Enders and Tofighi (2007) has been cited elsewhere in the methodological literature, typically informing most discussion of centering categorical predictors (Enders, 2013; Nezlek, 2012b; Peugh, 2010), but has not led to a consensus in the recommendations provided by methodologists (Nezlek, 2012a). Additionally, existing treatments have three important limitations: first, they are restricted to the discussion of binary predictors; second, they are restricted to the interpretation of intercepts; and third, they do not address UN binary or categorical predictors, which are frequently used in practice. To our knowledge, no methodological work has addressed centering multicategorical predictors (i.e., those reflecting more than two groups) or addressed the interpretation of resulting slope coefficients at both the within- and between-cluster levels for such predictors.

Literature Review

The lack of attention to centering categorical predictors is concerning, given that categorical predictors are used ubiquitously in MLM across many fields. As a precursor to the current tutorial, we

² We could have instead constructed *weighted* effect codes, which are often recommended over unweighted effect codes in nonexperimental contexts where group proportions are assumed to be representative of the population from which the sample was drawn. However, because unweighted effect codes are used far more frequently in practice, we will focus on them throughout the tutorial. Weighted effect codes will be addressed briefly. For more detail on weighted versus unweighted effect codes, see Cohen et al. (2003, Chapter 8).

briefly surveyed the applied literature to explore how categorical predictors are most often treated in practice. We assessed whether and how authors typically conducted centering and, if conducted, whether logical rationales and interpretations were provided. To identify relevant articles, we used two approaches. First, we forward-searched Enders and Tofighi (2007), a highly-cited resource on centering in MLM. Second, we simply searched “multilevel model” and “hierarchical linear model” in applied databases. In all cases we selected the highest-cited articles yielded by the search and that included at least one categorical predictor, in order to assess whether important substantive conclusions being drawn across applied fields may be helped or hindered by current centering practices. Identified articles came from important outlets including the *Journal of Educational Psychology*, *Journal of Applied Psychology*, and *Journal of Personality & Social Psychology*.

In line with expectation, we found that centering is applied to categorical predictors infrequently and inconsistently. Some authors appropriately centered continuous predictors, but explicitly left categorical predictors uncentered, arguing that this would facilitate interpretation of results (Murayama & Elliot, 2009; Reyes et al., 2012). Many articles included categorical predictors for precisely the same purpose (e.g., as a covariate), but conducted centering differently (Galindo & Sheldon, 2012; Kärnä et al., 2010; Littell & Tajima, 2000; Lüdtke et al., 2009; Merritt et al., 2012), and many were vague about how centering was conducted, making it unclear as to whether the appropriate conclusions were drawn from their results (Charbonnier-Voirin et al., 2010; Hofmann et al., 2012; Morrison et al., 2011; Trautwein & Lüdtke, 2009). However, the majority simply did not discuss whether categorical predictors were centered, suggesting that centering was not conducted at all (Aryee et al., 2012; Bowers & Urlick, 2011; Dettmers et al., 2011; Gong et al., 2013; Kuo et al., 2000; Liu et al., 2010; Major et al., 2008; McCoach et al., 2006; Powell et al., 2010; Sacco & Schmitt, 2005). When centering was conducted, corresponding rationales and interpretations were consistently lacking. Only two articles were found wherein the centering of categorical predictors was accompanied by thorough and coherent reasoning and interpretation of results (Kärnä et al., 2013; Montague et al., 2011). Taken together, the literature suggests that applied researchers are unfamiliar with whether and how categorical predictors should be centered and interpreted, and are left to rely on intuition or avoid centering altogether. A comprehensive treatment of the subject is clearly warranted.

Current Aims

Identified gaps in the methodological literature, alongside problematic practices by applied researchers, illuminate the need for the current tutorial paper. First, we aim to clarify why and how categorical predictors should be centered in multilevel models. Second, we aim to explicate how multilevel regression coefficients resulting from centered categorical predictors should be interpreted. We focus on slope coefficients, especially for multicategorical predictors, as to our knowledge these have never been addressed. To maximize applicability for applied researchers, we demonstrate our conclusions with an empirical example. We focus on the interpretation of coefficients from three models that are commonly used in practice: the UN Model, the CWC(M) Model, and the UN(M) Model (defined in the next section).

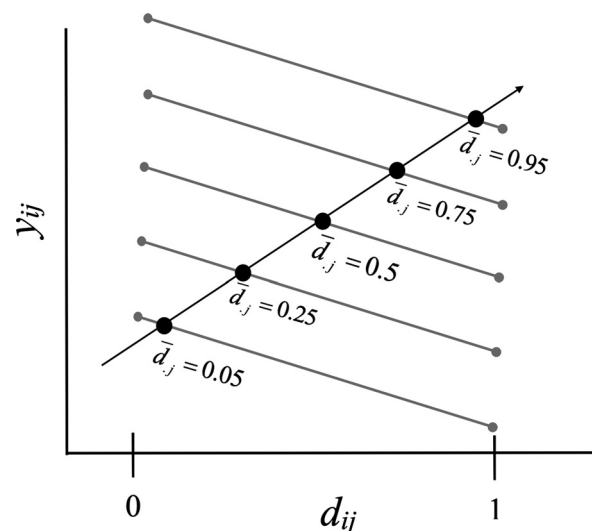
Logic and Algebra of Centering

We begin with logic and algebra to demonstrate the importance of centering, and discuss how this applies analogously to both continuous and categorical predictors. Generally, a Level-1 predictor, x_{ij} , will contain two parts: a between-cluster component, which is the cluster mean, \bar{x}_j , and a within-cluster component, $x_{ij} - \bar{x}_j$. Between-cluster variance arises when cluster means fluctuate around the grand mean, and within-cluster variance arises when Level-1 units fluctuate around their respective cluster means.

Because Level-1 predictors contain level-specific components that can be separated, the same is true for their effects on the outcome variable y . These level-specific effects can differ drastically (Curran & Bauer, 2011; Robinson, 1950). As an example, consider Figure 1, which provides a visual aid from a random-intercept, fixed-slope scenario concerning the relation between a binary dummy-coded predictor d_{ij} and a continuous outcome. Here, the within-cluster effect is negative and denoted with a separate regression line for each cluster. In contrast, the between-cluster effect is positive and denoted with a black arrow that cuts through all clusters. Such divergent within- and between-cluster effects are not uncommon. Failure to separately estimate these effects can often result in erroneous conclusions (e.g., the *ecological fallacy*; see Diez Roux, 2002, for more detail). A well-established body of work has shown that centering is needed to effectively separate and estimate these level-specific effects.

Importantly, the algebra underlying centering does not require us to distinguish between continuous versus categorical predictors, and in univariate MLM, there are no distributional assumptions placed on the predictors. All this suggests that the guidelines in place for continuous predictors will carry over to categorical predictors; however, this has never been demonstrated. To do so, we will use a k -group categorical predictor, expressed with $k - 1$ coding variables, as the basis for our models. The algebra underlying

Figure 1
Illustration of Within- and Between-Cluster Effects of a Binary Predictor



Note. Line with positive slope represents the between-cluster effect. Lines with negative slopes represent within-cluster effects.

centering also does not require us to define the coding scheme (e.g., dummy vs. contrast codes). Different coding schemes, however, introduce different interpretational and conceptual considerations, which we address in later sections.

As with continuous predictors, each coding variable for a categorical predictor can be partitioned into its within-cluster part and its between-cluster part by the typical method of subtracting, and then reintroducing, each cluster mean (Raudenbush & Bryk, 2002). In this section we use d_{ij} to denote each coding variable; we reiterate that the following logic and algebra applies to any coding variable regardless of coding scheme.

$$\begin{aligned} d_{1ij} &= (d_{1ij} - \bar{d}_{1,j}) + \bar{d}_{1,j} \\ d_{2ij} &= (d_{2ij} - \bar{d}_{2,j}) + \bar{d}_{2,j} \\ &\dots \\ d_{(k-1)ij} &= (d_{(k-1)ij} - \bar{d}_{(k-1),j}) + \bar{d}_{(k-1),j} \end{aligned}$$

Partitioning categorical predictors in this way introduces a few notable departures from the continuous predictor setting. First, cluster means are now related to the *proportions* of people in cluster j that belong to each group of the multigroup predictor. Consider the first dummy code in our empirical example, keeping in mind that $d_{1ij} = 0$ for children with no educated parents and $d_{1ij} = 1$ for children with only an educated mother. In School 1, $\bar{d}_{1,j} = .0926$, indicating that 9.26% of the children in this school have an educated mother. Similarly, in School 5, $\bar{d}_{1,j} = .0755$, indicating that 7.55% of children in this school have an educated mother. For each dummy code, the cluster mean equals the proportion of cases in cluster j that belong to the focal group. This arises because units in the focal group are coded 1 whereas all other units are coded 0. However, relationships between cluster means and proportions become a bit more complicated under other coding schemes. Consider the cluster means of the first effect code, where $e_{1ij} = -1$ for children with no educated parents, $e_{1ij} = 1$ for children with an educated mother, and $e_{1ij} = 0$ for all other children. In School 1, $\bar{e}_{1,j} = -.611$. This arises from the fact that 9.26% of children in this school are coded 1 on e_{1ij} , whereas 70.37% of children in this school are coded -1 (i.e., have no educated parents and thus belong to the reference group). The average is therefore: $.0926(1) + .7037(-1) = -.6111$ in School 1. Similarly, in School 5, $\bar{e}_{1,j} = -.6604$, which follows from the fact that 7.55% of children in this school have an educated mother and 73.58% have no educated parents. To summarize, cluster means of coding variables are related to the proportion of units that belong to each group, but the exact relationships between these proportions and the cluster means will vary depending on the coding scheme.

Second, both the original coding variable and the CWC coding variable can take on just two possible values in a given cluster. For example, suppose $\bar{d}_{1,j} = 0.3$. Because the uncentered dummy code takes on the values (0, 1), it follows that the CWC dummy code, $d_{1ij} - \bar{d}_{1,j}$, will take on the value of either $-.3$ or $.7$. Similarly, when the uncentered effect code takes on the values $(-1, 1)$ the CWC effect code will take on the value of either -1.3 or $.7$. This can also be seen in our empirical data example. In School 1, $\bar{d}_{1,j} = .0926$, and therefore $d_{1ij} - \bar{d}_{1,j}$ is equal to either $-.0926$ or $.9074$ in this cluster. The same pattern follows for all the CWC coding variables, which take on unique pairs of values in each

cluster. Although it may seem unintuitive that each CWC coding variable is still dichotomous, Raudenbush and Bryk (2002) note that this centering is indeed appropriate and still functions to partition the variable into level-specific parts.

Between-cluster variability of the categorical predictor is reflected in the variation of cluster means, and variation of CWC coding variables around their respective cluster means represents within-cluster variability. In our empirical data set, the value of $\bar{d}_{1,j}$ varies because in each school, a different proportion of students have only an educated mother. Similarly, the value of $\bar{d}_{2,j}$ varies because in each school, a different proportion of students have only an educated father. A similar pattern follows for all the coding variables; for example, $\bar{e}_{1,j}$ fluctuates as a function of the proportion of students with an educated mother *and* the proportion of students with no educated parents in each school. These fluctuations reflect between-school variability in student SES. Within each school, the values of the CWC coding variable fluctuate around their respective cluster mean; this reflects within-school variability in student SES.

To summarize, variables representing a Level-1 predictor will contain a within-cluster and a between-cluster component regardless of whether the predictor is continuous or categorical. It follows that centering decisions for categorical coding variables will yield effects that are similar to those found for continuous predictors. To demonstrate, we present the UN Model, the CWC(M) Model, and the UN(M) Model, altered to contain our categorical predictor represented by $k - 1$ coding variables. Again, d_{ij} is used to denote each coding variable, but the conclusions drawn here will apply to any coding scheme.

The UN Model

The UN Model:

$$\begin{aligned} y_{ij} &= \beta_{0j}^* + \beta_{1j}^* d_{1ij} + \beta_{2j}^* d_{2ij} + \dots + \beta_{(k-1)j}^* d_{(k-1)ij} + e_{ij} \\ \beta_{0j}^* &= \gamma_{00}^* + u_{0j} \\ \beta_{1j}^* &= \gamma_{10}^* \\ \beta_{2j}^* &= \gamma_{20}^* \\ &\dots \\ \beta_{(k-1)j}^* &= \gamma_{(k-1)0}^* \end{aligned} \quad (1)$$

Reduced form:

$$y_{ij} = \gamma_{00}^* + \gamma_{10}^* d_{1ij} + \gamma_{20}^* d_{2ij} + \dots + \gamma_{(k-1)0}^* d_{(k-1)ij} + u_{0j} + e_{ij} \quad (2)$$

Here, β_{0j}^* denotes the intercept for cluster j , β_{1j}^* denotes the conflated slope of d_{1ij} in cluster j , β_{2j}^* denotes the conflated slope of d_{2ij} in cluster j , and so on. In the reduced-form equation, γ_{00}^* is the intercept, γ_{10}^* is the conflated slope of d_{1ij} , γ_{20}^* is the conflated slope of d_{2ij} , and so on. We also estimate the variance of u_{0j} (denoted τ_{00}^*) which quantifies between-cluster variance in intercepts, and the Level-1 residual variance (σ_e^{2*}). The asterisk (*) is used to differentiate conflated estimates in this model from the unconflated estimates described in later models. Here and throughout the tutorial, we restrict our focus to random-intercept models, so none of the slope parameters has a corresponding random error term.

The use of UN Level-1 predictors in isolation is perhaps the most common approach in practice, though this model possesses important drawbacks. Crucially, in the UN Model, effects are not partitioned into Level-1 and Level-2 components, and only one slope is estimated for each predictor. Thus, a slope in the UN Model is an uninterpretable, conflated blend of the within- and between-cluster effects of the predictor (Cronbach, 1976; Cronbach & Webb, 1975; Grice, 1966; Härmqvist, 1978; Kenny & La Voie, 1985; Sirotnik, 1980). Failing to partition the level-specific effects of a predictor can often have serious consequences and lead to erroneous conclusions.

The same logic applies to coding variables: although each coding variable can be decomposed into level-specific parts, the UN Model is constrained to estimate only one parameter for each, rather than separate estimates for each within-cluster and between-cluster effect. Equation 2 can be re-expressed as:

$$\begin{aligned} y_{ij} = & \gamma_{00}^* \\ & + \gamma_{10}^*(d_{1ij} - \bar{d}_{1,j}) + \gamma_{10}^*\bar{d}_{1,j} \\ & + \gamma_{20}^*(d_{2ij} - \bar{d}_{2,j}) + \gamma_{20}^*\bar{d}_{2,j} \\ & + \dots \\ & + \gamma_{(k-1)0}^*(d_{(k-1)ij} - \bar{d}_{(k-1),j}) + \gamma_{(k-1)0}^*\bar{d}_{(k-1),j} \\ & + u_{0j} + e_{ij} \end{aligned}$$

This reveals the implicit equality constraint placed on the coefficients associated with each coding variable.

As a result of this constraint, Raudenbush and Bryk (2002) have shown in the context of a single predictor that the conflated slope estimate will be an uninterpretable mix of within-cluster and between-cluster effects. Assuming a balanced design, they show that:

$$\hat{\gamma}_{10}^* = \frac{W_1\hat{\beta}_b + W_2\hat{\beta}_w}{W_1 + W_2} \quad (3)$$

where $\hat{\beta}_b$ is the between-cluster effect and $\hat{\beta}_w$ is the within-cluster effect of the predictor. W_1 and W_2 are weights that reflect the precision of the estimates of $\hat{\beta}_b$ and $\hat{\beta}_w$, respectively:

$$W_1 = \frac{1}{\text{var}(\hat{\beta}_b)}; \quad W_2 = \frac{1}{\text{var}(\hat{\beta}_w)}. \quad (4)$$

Applying this logic to coding variables, past work suggests that $\hat{\gamma}_{10}^*, \hat{\gamma}_{20}^*, \dots, \hat{\gamma}_{(k-1)0}^*$ will be precision-weighted averages of the within-cluster and between-cluster effects of each of their respective codes (Raudenbush & Bryk, 2002; Raudenbush & Willms, 1995). Stated differently, holding the predictor's total variance constant, a conflated slope estimate should be pulled closer to β_b as the intraclass correlation of the predictor (ICC_X) increases, and closer to β_w as ICC_X decreases, all else being equal.³

In unbalanced designs the derivations become more complex, but the same principle applies in that an uncentered predictor yields a conflated blend of within-cluster and between-cluster effects. Even more problematically, other extraneous characteristics of the data, including average cluster size and ICC_X , also influence the nature of conflation observed (Raudenbush & Bryk,

2002; Raudenbush & Willms, 1995; Van de Pol & Wright, 2009).

Methodological investigation into conflation has focused on single variables in isolation. However, when employing multicategorical predictors, at least two coding variables must be used simultaneously. These UN Level-1 coding variables will necessarily be correlated, and the covariance of the within-cluster components, for example, $\text{cov}((d_{1ij} - \bar{d}_{1,j}), (d_{2ij} - \bar{d}_{2,j}))$, and the between-cluster components, $\text{cov}(\bar{d}_{1,j}, \bar{d}_{2,j})$, may differ. In extending Equations 3 and 4 to two or more predictors, it may be necessary to consider covariances at both the within- and between-cluster levels, rather than only the variances of $\hat{\beta}_b$ and $\hat{\beta}_w$. Therefore, we may expect more complex patterns of conflation to arise for multicategorical predictors. Though some researchers have investigated the effects of multicollinearity in multilevel models (Clark, 2013; Shieh & Fouladi, 2003; Yu et al., 2015), none have done so in the context of categorical predictors or conflated estimates. In summary, the inclusion of multiple coding variables may influence conflation in yet-unknown ways, supplying yet another reason to avoid the UN Model.

Finally, we note that there are some special cases wherein the UN Model may be acceptable; however, these special cases are rarely realized in practice. First, if $ICC_X = 0$, then the predictor has no Level-2 variance and has no ability to exert a Level-2 effect. Therefore, conflation cannot occur and UN Level-1 predictors yield accurate estimates of Level-1 effects (Asparouhov & Muthén, 2019). For a categorical predictor, it is possible for ICC_X to be zero only if each cluster has an identical composition of groups (e.g., all clusters have 10% Group *a*, 30% Group *b*, and 60% Group *c*). Second, if the within- and between-cluster effects of the predictor are exactly the same, then the “conflated” estimate is equal to both the within- and between-cluster effect (Rights et al., 2019) and is therefore interpretable. Outside of simulated data, it is impossible to know the true level-specific effects of a predictor, and identical level-specific effects are virtually nonexistent. Therefore, in the vast majority of cases, the UN Model is not recommended.

The CWC(M) Model

The CWC(M) Model:

$$\begin{aligned} y_{ij} = & \beta_{0j} + \beta_{1j}(d_{1ij} - \bar{d}_{1,j}) + \beta_{2j}(d_{2ij} - \bar{d}_{2,j}) + \dots + \beta_{(k-1)j}(d_{(k-1)ij} - \bar{d}_{(k-1),j}) + e_{ij} \\ \beta_{0j} = & \gamma_{00} + \gamma_{01}\bar{d}_{1,j} + \gamma_{02}\bar{d}_{2,j} + \dots + \gamma_{0(k-1)}\bar{d}_{(k-1),j} + u_{0j} \\ \beta_{1j} = & \gamma_{10} \\ \beta_{2j} = & \gamma_{20} \\ \dots & \\ \beta_{(k-1)j} = & \gamma_{(k-1)0} \end{aligned} \quad (5)$$

Reduced form:

³ In two-level models, ICC_X is defined as the proportion of a predictor's total variance that is attributable to between-cluster variance. The definition of ICC_Y is analogous for the outcome variable. In the null model with no predictors, $ICC_Y = \tau_{00}/(\tau_{00} + \sigma_e^2)$, where τ_{00} is the Level-2 variance of y_{ij} and σ_e^2 is the Level-1 variance of y_{ij} .

$$y_{ij} = \gamma_{00} + \gamma_{01}\bar{d}_{1j} + \gamma_{10}(d_{1ij} - \bar{d}_{1j}) + \gamma_{02}\bar{d}_{2j} + \gamma_{20}(d_{2ij} - \bar{d}_{2j}) + \dots \\ + \gamma_{0(k-1)}\bar{d}_{(k-1)j} + \gamma_{(k-1)0}(d_{(k-1)ij} - \bar{d}_{(k-1)j}) + u_{0j} + e_{ij} \quad (6)$$

Here, β_{0j} is the intercept for cluster j , β_{1j} is the slope of $(d_{1ij} - \bar{d}_{1j})$ in cluster j , β_{2j} is the slope of $(d_{2ij} - \bar{d}_{2j})$ in cluster j , and so on. In the reduced-form equation, γ_{00} is the intercept, γ_{10} is the slope of $(d_{1ij} - \bar{d}_{1j})$, γ_{20} is the slope of $(d_{2ij} - \bar{d}_{2j})$, and so on. Next, γ_{01} is the slope of \bar{d}_{1j} , γ_{02} is the slope of \bar{d}_{2j} , and so on. We again estimate the variance of u_{0j} (τ_{00}) and the Level-1 residual variance (σ_e^2). In this model, we obtain separate within- and between-cluster slope estimates and these estimates are no longer denoted with asterisks.

The CWC approach, $x_{ij} - \bar{x}_{.j}$, is often recommended by methodologists. This approach removes all Level-2 variance from the variable because the cluster mean of a CWC predictor is always zero. Thus, CWC predictors (e.g., $x_{ij} - \bar{x}_{.j}$) are “pure” Level-1 variables, and are necessarily orthogonal to all Level-2 predictors, including their own cluster means. As a result, the Level-1 slope of a CWC predictor is an estimate of its within-cluster effect, independent of any Level-2 influence the uncentered predictor may exert. Methodologists have thus encouraged the use of CWC when a Level-1 effect is of primary interest (Enders & Tofighi, 2007; Raudenbush, 2009).

It may also be of interest to include cluster means of Level-1 predictors, $\bar{x}_{.j}$, as Level-2 predictors themselves. Here, methodologists argue for including CWC Level-1 predictors alongside their corresponding cluster means, yielding the CWC(M) Model (Kreft & de Leeuw, 1998). Because $x_{ij} - \bar{x}_{.j}$ and $\bar{x}_{.j}$ are necessarily uncorrelated, the CWC(M) Model yields within- and between-cluster effects that are mutually independent, and is therefore frequently advocated (Asparouhov & Muthen, 2006; Hedeker & Gibbons, 2006; Kreft et al., 1995; Lüdtke et al., 2008; Neuhaus & Kalbfleisch, 1998; Neuhaus & McCulloch, 2006; Preacher et al., 2010, 2016). In summary, the CWC(M) Model allows the researcher to separately estimate the within- and between-cluster effects of the predictor, and each estimate is unaffected by the other. The independence of these estimates means the problems encountered in the UN Model (i.e., bias, conflation, lack of interpretability) can largely be avoided with the CWC(M) Model.

The algebra also implies that the CWC(M) Model will yield estimates of the independent within- and between-cluster effects of a categorical predictor. CWC coding variables contain no between-cluster variance because their cluster means will be zero, and therefore they will be orthogonal to all Level-2 cluster means. Additionally, their values will still reflect individual differences *relative* to other cases in the same cluster, as is true for CWC continuous predictors (Enders & Tofighi, 2007). The CWC(M) Model partitions the categorical predictor into its uncorrelated within- and between-cluster parts. Thus, the CWC(M) Model will effectively separate a categorical predictor’s level-specific effects, just as it does for continuous predictors.

The UN(M) Model

The UN(M) Model:

$$y_{ij} = \beta_{0j} + \beta_{1j}d_{1ij} + \beta_{2j}d_{2ij} + \dots + \beta_{(k-1)j}d_{(k-1)ij} + e_{ij} \\ \beta_{0j} = \gamma_{00} + \gamma_{01}^c\bar{d}_{1j} + \gamma_{02}^c\bar{d}_{2j} + \dots + \gamma_{0(k-1)}^c\bar{d}_{(k-1)j} + u_{0j} \\ \beta_{1j} = \gamma_{10} \\ \beta_{2j} = \gamma_{20} \\ \dots \\ \beta_{(k-1)j} = \gamma_{(k-1)0} \quad (7)$$

Reduced form:

$$y_{ij} = \gamma_{00} + \gamma_{01}^c\bar{d}_{1j} + \gamma_{10}d_{1ij} + \gamma_{02}^c\bar{d}_{2j} + \gamma_{20}d_{2ij} + \dots \\ + \gamma_{0(k-1)}^c\bar{d}_{(k-1)j} + \gamma_{(k-1)0}d_{(k-1)ij} + u_{0j} + e_{ij} \quad (8)$$

Here, β_{0j} is the intercept for cluster j , β_{1j} is the slope of $(d_{1ij} - \bar{d}_{1j})$ in cluster j , β_{2j} is the slope of $(d_{2ij} - \bar{d}_{2j})$ in cluster j , and so on. In the reduced-form equation, γ_{00} is the intercept, γ_{10} is the slope of $(d_{1ij} - \bar{d}_{1j})$, γ_{20} is the slope of $(d_{2ij} - \bar{d}_{2j})$, and so on. Next, γ_{01}^c is the slope of \bar{d}_{1j} , γ_{02}^c is the slope of \bar{d}_{2j} , and so on. We again estimate the variance of u_{0j} (τ_{00}) and the Level-1 residual variance (σ_e^2). In this model, the Level-2 slope parameters include a c superscript to denote the contextual effect (to be defined shortly).

The UN(M) Model contains UN Level-1 predictors alongside their corresponding cluster means. UN predictors still contain both Level-1 and Level-2 variance, and thus are not orthogonal to their Level-2 cluster means. Thus, regression coefficients associated with x_{ij} and $\bar{x}_{.j}$ become *partial* regression coefficients, representing the effect of each predictor after partialing out the effect of the other. The Level-1 slope still represents the within-cluster effect of the predictor (for an algebraic proof, see Kreft et al., 1995, p. 12). However, the slope of $\bar{x}_{.j}$ reflects the between-cluster effect of the predictor that is above and beyond its within-cluster effect. This is often referred to as the *contextual effect*. The contextual effect has been defined in a variety of ways, but most simply, it is equal to the difference between a predictor’s between- and within-cluster effects (Raudenbush & Bryk, 2002; Snijders & Bosker, 2012). The contextual effect can be conceptualized as the between-cluster effect of a Level-2 variable that remains after controlling for its Level-1 counterpart. For some research questions, this interpretation at Level 2 is better-suited and more useful than the “pure” Level-2 effect supplied by the CWC(M) Model (Begg & Parides, 2003). More detail on interpretational considerations is provided in later sections.

The UN(M) Model is equivalent to the CWC(M) Model, with the change in centering of Level-1 predictors resulting in a slight reparameterization and new interpretation of Level-2 coefficients (Enders & Tofighi, 2007; Kreft et al., 1995; Raudenbush & Bryk, 2002). Again, this equivalence holds only in random-intercept models that do not include random slopes.

The above logic will follow for the UN(M) Model with categorical predictors. UN coding variables contain both Level-1 and Level-2 variance, and therefore will not be orthogonal to their Level-2 cluster means. Thus, the UN(M) Model yields contextual effects at level 2. $\hat{\gamma}_{01}^c, \hat{\gamma}_{02}^c, \dots, \hat{\gamma}_{0(k-1)}^c$ are estimates of the contextual effect for each group, denoted $\beta_{c1}, \beta_{c2}, \dots, \beta_{c(k-1)}$. Due to the equivalency of the CWC(M) and UN(M) Models, each

contextual effect will equal the between-cluster effect minus the within-cluster effect (Kreft et al., 1995).

Summary

In this section we have demonstrated that centering guidelines for continuous predictors should be applied analogously to categorical predictors. Importantly, a conflated slope estimate, resulting from an uncentered coding variable used in isolation, will equal neither the within-cluster effect nor the between-cluster effect. The estimate carries little interpretational value and is heavily influenced by extraneous characteristics of the data and study design, including ICC_X and cluster size. This logic extends analogously to multigroup categorical predictors, regardless of coding scheme. Algebraically and statistically, multilevel models and centering principles behave the same whether predictors are continuous or categorical. However, for many researchers, interpretation of a categorical predictor's effects in MLM will not be intuitive. Thus, we provide an in-depth discussion of parameter interpretations in the following sections.

Parameter Interpretations

Derivations

The second goal of this report is to clarify how parameter estimates should be interpreted in multilevel models with categorical predictors, with a particular focus on slope coefficients, as prior work has focused on intercept interpretation (Enders, 2013; Enders & Tofighi, 2007; Nezlek, 2012b). To accomplish this, we conducted expected-value derivations. When categorical predictors are used in the single-level regression setting, group mean differences in terms of the outcome variable y are most often of interest (e.g., the mean difference on y between the focal group and the reference group). Therefore, our goal was to derive the expected value (i.e., mean) of y separately for each group of interest, in order to clarify relationships between multilevel slope coefficients and group mean differences. Derivations were conducted for dummy codes, contrast codes, and both unweighted and weighted effect codes. All derivations were conducted such that they apply to any number of groups (i.e., any k), and do not require the assumption of equal cluster sizes or group balance (i.e., equal group proportions).

We focused our derivations on the CWC(M) Model. First, this model effectively separates within- and between-cluster effects, and interpretations of these effects will be of particular utility in practice. Second, interpretations of the UN(M) Model are a straightforward extension of those from the CWC(M) Model. Under each coding scheme, we derived expected values of y as a function of expected values of the coding variables and their associated cluster means. See Appendices A–D in online supplemental materials for full derivations. For a reminder of what each coding scheme looks like in real data, see Table 1.

Dummy Codes

Our derivations clarify precisely how within- and between-cluster effects should be interpreted for dummy-coded multicategorical predictors. For details, see Appendix A in online supplemental

materials. We first show that when there is *no* between-cluster variability with respect to the categorical predictor (i.e., each cluster has identical composition, yielding $ICC = 0$ for all dummy codes), the within-cluster slope γ_{f0} is equal to the mean difference on y between the reference group (Group k) and the focal group (Group f). The $g = f$ notation indicates that we are conditioning on the focal group, whereas the $g = k$ notation indicates that we are conditioning on the reference group. We show that when ICC of all dummy codes is zero, $E(y_{ij})|_{g=f} - E(y_{ij})|_{g=k} = \gamma_{f0}$. Based on this correspondence between the coefficient and within-cluster mean difference, we conclude that in the CWC(M) Model, for any Group f , the within-cluster slope γ_{f0} is interpreted as the mean difference on y between Group f and the reference group, within clusters, on average.

Second, we show that when the categorical predictor has *no* within-cluster variability (i.e., each cluster is composed entirely of a single group, yielding $ICC = 1$ for all dummy codes), the between-cluster slope γ_{0f} is equal to the mean difference on y between the reference group and Group f . When $ICC = 1$ for all dummy codes, $E(y_{ij})|_{g=f} - E(y_{ij})|_{g=k} = \gamma_{0f}$. Thus, for any Group f , the between-cluster effect γ_{0f} is interpreted as the mean difference on y when moving from a cluster composed entirely of the reference group to a cluster composed entirely of Group f .

Contrast Codes

Derivations for contrast codes were conducted for the general case, such that they apply to any total number of groups, and to any number of groups in the reference group versus the focal group for the contrast. Here, we denote $f1$ as the focal group(s) and $f2$ as the reference group(s). The $g \in f1$ notation indicates that we are conditioning on a group that is either the sole focal group or part of the set of focal groups. Similarly, the $g \in f2$ notation indicates that we are conditioning on a group that is either the sole reference group or part of the set of reference groups. We show that when $ICC = 0$ for all contrast codes, the within-cluster slope γ_{f0} is equal to the mean difference on y between the focal group(s) and the reference group(s). When $ICC = 0$ for all contrast codes, $E(y_{ij})|_{g \in f1} - E(y_{ij})|_{g \in f2} = \gamma_{f0}$. Thus, for any contrast code, the within-cluster slope γ_{f0} is interpreted as the mean difference on y between the focal group(s) and the reference group(s), within clusters, on average. If more than one group is involved in the reference and/or focal group, this interpretation involves unweighted means of those groups (Cohen et al., 2003). Examples are provided in the following section.

Between-cluster effects extend similarly. We show that when $ICC = 1$ for all contrast codes, the between-cluster slope γ_{0f} is equal to the mean difference on y between the focal group(s) and the reference groups(s). When $ICC = 1$ for all codes, $E(y_{ij})|_{g \in f1} - E(y_{ij})|_{g \in f2} = \gamma_{0f}$. Thus, for any contrast code, the between-cluster slope γ_{0f} is interpreted as the mean difference on y upon moving from a cluster composed entirely of reference group(s) to a cluster composed entirely of focal group(s). See Appendix B in online supplemental materials for details.

Effect Codes

Appendix C in online supplemental materials contains derivations for unweighted effect codes. Assuming $ICC = 0$ for all effect

codes, we derive the expected value of y for a generic focal Group f , then we derive the unweighted mean of all group means in the sample. We show that the within-cluster slope γ_{f0} is equal to the difference between these two quantities. Thus, the within-cluster slope γ_{f0} is interpreted as the mean difference on y between Group f and the unweighted mean of all group means in the sample.

Next, assuming $ICC = 1$ for all effect codes, we derive the expected value of y for a generic focal Group f and the unweighted mean of all group means in the sample. We then show that the between-cluster slope γ_{0f} is equal to the difference between these two quantities. Thus, the between-cluster slope γ_{0f} is interpreted as the difference on y as we go from the unweighted mean of all group means to the mean in a cluster composed entirely of Group f . Our derivations did not require us to commit to balance with respect to group proportions or cluster sizes.

Appendix D in online supplemental materials contains similar derivations for weighted effect codes. Using a similar approach, we show that when $ICC = 0$ for all codes, the within-cluster slope γ_{f0} is equal to the difference between the expected value of y for Group f and the weighted mean of all group means (which is the grand mean of the sample). Similarly, when $ICC = 1$ for all codes, the between-cluster slope γ_{0f} is equal to the difference on y as we go from the weighted mean of all group means, to the mean in a cluster composed entirely of Group f .

Implications for Practice

Our derivations indicate direct correspondence of within-cluster slopes and mean differences when $ICC = 0$ for all coding variables, and direct correspondence of between-cluster slopes and mean differences when $ICC = 1$ for all coding variables. Though these equivalencies are useful for illuminating how each coefficient should be interpreted, it is important to note that ICC will rarely be equal to 0 or 1 in practice. $ICC = 0$ will occur only if each cluster has an identical composition of the categorical predictor (e.g., all clusters have 10% Group a , 30% Group b , and 60% Group c). Such a pattern is unlikely to arise in real data, especially with naturally occurring categories, though could arise in an experimental context. In contrast, $ICC = 1$ if each cluster contains entirely one category. Here, the categorical predictor is by definition a Level-2 predictor. In all other situations, ICC will lie between 0 and 1, and parameter estimates will no longer correspond to raw group mean differences (i.e., those that could be identified from descriptive statistics of the raw data).

For a more technical demonstration of the relationship between ICC of the coding variables and group mean differences, see Appendix E in online supplemental materials.

Summary

In this section, we have derived how slopes of categorical predictors should be interpreted in multilevel models, and demonstrated how those slopes relate to group mean differences on y . It is important to understand that coefficients reflecting within- and between-cluster effects are equal to *actual* group mean differences on y (i.e., those that can be calculated from group means in the raw data) only under very particular data conditions, when the ICC of the coding variables is either 0 or 1. However, the direct

correspondence between the coefficients and group mean differences under these conditions illuminates what the coefficients represent, and therefore how these level-specific effects should be correctly interpreted. These level-specific effects can be obtained directly through use of the CWC(M) Model.

Interpretation of coefficients in the UN(M) Model is a straightforward extension of those from the CWC(M) Model, so their derivation is not necessary. Interpretational considerations for the CWC(M) Model versus the UN(M) Model will be discussed at length in the following section.

Importantly, the interpretations that we have shown to be correct stand in contrast to the interpretations that are often employed in practice. Our literature review revealed that researchers often employ UN coding variables, which yield uninterpretable coefficients, but subsequently interpret those coefficients as though they represent within-cluster effects. In reality, the categorical predictor and its corresponding effects must be appropriately partitioned into level-specific parts before such interpretation is warranted. Next, we return to our empirical example to anchor these derived interpretations with real-world context.

Illustration With Empirical Data

We return to our running empirical example of 3,435 children nested within 148 primary schools. As a reminder, the goal of these analyses was to assess the relationship between student SES, as indexed by parental education (PED_{ij}), and student academic achievement ($ATTAIN_{ij}$). PED_{ij} was a four-group categorical predictor: Group 1 = neither parent has a high level of education; Group 2 = mother is educated, father is not; Group 3 = father is educated, mother is not; Group 4 = both parents are educated. As described above, PED_{ij} was coded in multiple ways (dummy codes, contrast codes, effect codes) to demonstrate that our conclusions about centering will apply to any categorical predictor regardless of coding scheme. Coding schemes are displayed in Table 1. All analyses were conducted in *R*, Version 3.6.2 (R Core Team, 2019) using the *lme4* package (Bates et al., 2015). *R* code is provided in Appendix F in online supplemental materials.

Analyses and Interpretations

We fit the UN Model, the CWC(M) Model, and the UN(M) Model under each coding scheme; all parameter estimates are reported in Table 2. Next, we detail how these estimates should be interpreted. Due to space constraints, we present interpretations for dummy codes and contrast codes, as these are particularly common for multicategorical predictors (i.e., those with more than two categories). Interpretations associated with unweighted effect codes are in Appendix G in online supplemental materials; interpretations for weighted effect codes are a very similar extension of those for unweighted effect codes, and thus we do not address them in detail here.

The UN Model

The UN Model yields parameter estimates that are uninterpretable. Notice that for all three coding variables and all three coding schemes, the single slope estimate in the UN Model lies between the within-cluster slope and the between-cluster slope that are obtained in the CWC(M) Model. For example, consider the first dummy code.

Table 2
Parameter Estimates

UN Model										
Coding scheme	$\hat{\gamma}_{00}^*$	$\hat{\gamma}_{10}^*$	$\hat{\gamma}_{20}^*$	$\hat{\gamma}_{30}^*$			τ_{00}^*	σ_e^{2*}	logL	
Dummy	5.284 (0.12)	0.638 (0.15)	0.880 (0.19)	0.865 (0.13)			1.118	8.080	−8,554.4	
Contrast	5.879 (0.11)	0.598 (0.08)	0.070 (0.10)	0.121 (0.11)			1.118	8.080	−8,554.4	
Effect	5.879 (0.11)	0.042 (0.11)	0.285 (0.14)	0.269 (0.10)			1.118	8.080	−8,554.4	
CWC(M) Model										
			Level 1			Level 2				
Coding scheme	$\hat{\gamma}_{00}$	$\hat{\gamma}_{10}$	$\hat{\gamma}_{20}$	$\hat{\gamma}_{30}$	$\hat{\gamma}_{01}$	$\hat{\gamma}_{02}$	$\hat{\gamma}_{03}$	τ_{00}	σ_e^2	logL
Dummy	3.991 (0.30)	0.605 (0.15)	0.801 (0.20)	0.799 (0.13)	2.419 (0.99)	6.240 (1.51)	4.071 (0.83)	0.864	8.092	−8,543.1
Contrast	7.174 (0.33)	0.551 (0.08)	0.064 (0.11)	0.098 (0.11)	3.183 (0.56)	−0.172 (0.73)	1.911 (0.86)	0.864	8.092	−8,543.1
Effect	7.174 (0.33)	0.053 (0.11)	0.249 (0.15)	0.248 (0.10)	−0.764 (0.80)	3.057 (1.13)	0.889 (0.69)	0.864	8.092	−8,543.1
UN(M) Model										
			Level 1			Level 2				
Coding scheme	$\hat{\gamma}_{00}$	$\hat{\gamma}_{10}$	$\hat{\gamma}_{20}$	$\hat{\gamma}_{30}$	$\hat{\gamma}_{01}^c$	$\hat{\gamma}_{02}^c$	$\hat{\gamma}_{03}^c$	τ_{00}	σ_e^2	logL
Dummy	3.991 (0.30)	0.605 (0.15)	0.801 (0.20)	0.799 (0.13)	1.814 (1.01)	5.439 (1.52)	3.272 (0.84)	0.864	8.092	−8,543.1
Contrast	7.174 (0.33)	0.551 (0.08)	0.064 (0.11)	0.098 (0.11)	2.631 (0.57)	−0.237 (0.74)	1.812 (0.87)	0.864	8.092	−8,543.1
Effect	7.174 (0.33)	0.053 (0.11)	0.249 (0.15)	0.248 (0.10)	−0.817 (0.81)	2.808 (1.14)	0.641 (0.69)	0.864	8.092	−8,543.1

Note. See Table 1 for details on each coding scheme. Numbers in parentheses are standard errors.

* = estimates from the UN Model; c = contextual effect.

In the UN Model, we obtain a conflated slope estimate, $\hat{\gamma}_{10}^* = .638$. In the CWC(M) Model, we obtain its within-cluster slope estimate, $\hat{\gamma}_{10} = .605$, and its between-cluster slope estimate, $\hat{\gamma}_{01} = 2.419$. The conflated slope in the UN Model lies between its within- and between-cluster effect, but is equal to neither. The conflated estimate is therefore meaningless and cannot be interpreted. In this case, the UN Model results in an *overestimation* of the within-cluster effect, but the reverse may also occur, depending on the magnitude of within- and between-cluster effects (e.g., notice the estimates associated with the first effect code). Finally, estimates from the UN Model are influenced by arbitrary, irrelevant design factors such as cluster size and ICC_X. Upon fitting the UN Model to a slightly different data set, we would obtain different estimates.

In some situations, a conflated effect as estimated in the UN Model may be nearly identical to either the level-specific “within” or “between” effect. For example, notice the estimates associated with the second contrast code. In the UN Model, the conflated estimate is $\hat{\gamma}_{20}^* = .070$, and in the CWC(M) Model, the within-cluster estimate is $\hat{\gamma}_{20} = .064$. The similarity of these coefficients may suggest that estimating a conflated slope is not problematic because it is such a close approximation to the within-cluster effect. However, even in these situations, estimates associated with UN coding variables are still flawed. The model is misspecified at both levels, and even if conflated and unconflated estimates are nearly identical, their interpretations differ.

In our literature review, we found that UN coding variables were used most frequently by applied researchers. As this example makes clear, UN coding variables used in isolation are not appropriate for use under any coding scheme, and their parameter estimates are not substantively meaningful.

The CWC(M) Model

Dummy Codes. The within-cluster effect of a dummy code is interpreted as the mean difference on y between Group f and the reference group, within clusters, on average. As a reminder, our

reference group is children with no educated parents. d_{1ij} is coded 1 for children whose mother is educated, and its within-cluster slope is $\hat{\gamma}_{10} = .605$. Thus, within a given school, on average, children with an educated mother score .605 points higher on academic achievement than children with no educated parents. The within-cluster slope of the second dummy code, $\hat{\gamma}_{20} = .801$, shows that within schools, on average, children with an educated father score .801 points higher than children with no educated parents. Finally, the third within-cluster slope, $\hat{\gamma}_{30} = .799$, indicates that children with two educated parents score .799 points higher than children with no educated parents, within schools, on average.

These coefficients are particularly useful because they are “pure” Level-1 slopes; they reflect mean differences between students in the same school. Thus, any school-level factors that may also influence achievement scores, such as funding, resources, or teacher quality, are not confounding factors in these slopes, and we can isolate key effects of interest.

Level-2 slopes correspond to the mean difference on y when moving from a cluster composed entirely of the reference group to a cluster composed entirely of Group f . Thus, the first between-cluster slope, $\hat{\gamma}_{01} = 2.419$, indicates that a school where all children have an educated mother will have a mean academic achievement score that is 2.419 points higher than a school where all children have no educated parents. Similarly, $\hat{\gamma}_{02}$ shows that a school where all children have an educated father will have a mean achievement score that is 6.240 points higher than a school where all children have no educated parents. Finally, $\hat{\gamma}_{03}$ indicates that a school where all children have two educated parents will have a mean achievement score that is 4.071 points higher than a school where all children have no educated parents.

These Level-2 slopes are school-level effects. Rather than addressing the effect of one’s own parents’ education, they shed light on the effect of attending a school where many children have educated parents. At both levels, we observe the greatest difference between children with no educated parents and children with only an educated

father. It appears that, at the individual level, the unique influence of an educated father results in the greatest academic achievement gains over children who have no educated parents. Additionally, at the school level, children surrounded by peers whose fathers are educated have greater achievement scores on average.

Initial interpretations of Level-2 effects may seem strange. We are required to compare homogeneous clusters (e.g., clusters containing *only* the reference group, or *only* Group *f*). However, such clusters are usually hypothetical, as heterogeneity within clusters is typically expected. The fact that Level-2 interpretations revolve around hypothetical clusters is not ideal. Instead, we may wish to divide these coefficients by 10 to facilitate more useful interpretation.⁴ Consider the between-cluster slope of the third dummy code, which is $\hat{\gamma}_{03} = 4.071$. Instead of using the raw coefficient to compare two hypothetical schools, we could instead say that as the percentage of students with two educated parents increases by 10% (and the percentage of students with no educated parents decreases by 10%), we expect the mean school achievement score to increase by .4071 points.

Contrast Codes. Interpretations of each contrast code will change depending on how the code was constructed. The contrasts presented here are not an exhaustive set of all that could have been created, but instead provide examples that readers may carry forward to their own work. First, c_{1ij} involves all four groups by comparing the mean of Group 1 (no parents educated) with Groups 2–4 (one or two parents educated). Its within-cluster slope is $\hat{\gamma}_{10} = .551$. We interpret this as the difference between the unweighted mean of *y* across Groups 2, 3, and 4, and the mean of *y* in Group 1. Thus, children with one or two educated parents are expected to score .551 points higher than children with no educated parents, within schools, on average.

Next, c_{2ij} involves three groups in the contrast. Its within-cluster slope, $\hat{\gamma}_{20}$, is interpreted as the difference between the mean of *y* in Group 4 and the unweighted mean of *y* across Groups 2 and 3, within clusters, on average. Within a given school, the mean achievement score among children with two educated parents will be .064 points higher than the mean achievement score across students with one educated parent. This slope suggests there is little influence of having two educated parents versus just one. Taken together with $\hat{\gamma}_{10}$, it appears that within schools, having *any* educated parent is the most influential for students' academic achievement, and it matters less whether one or both of the student's parents are educated.

Finally, c_{3ij} involves just two groups in the contrast. Its within-cluster slope, $\hat{\gamma}_{30}$, is simply interpreted as the within-cluster difference between the mean of *y* in Group 3 and the mean of *y* in Group 2. Within schools, students with only an educated father will score about .098 points higher than students with only an educated mother, on average. This suggests that among children with one educated parent, *which* parent is educated is minimally influential over academic achievement.

The between-cluster slope of the first contrast code is $\hat{\gamma}_{01} = 3.183$. This reflects the difference on *y* as we go from clusters composed entirely of Group 1 to the unweighted mean across clusters composed entirely of Groups 2, 3, or 4. Thus, the mean achievement score across schools where all children have one or two educated parents is 3.183 points higher than the mean achievement score in schools where all children have no educated parents.

Attending a school where many students have *any* educated parents yields notable gains in academic achievement.

The between-cluster slope of the second contrast code is $\hat{\gamma}_{02} = -.172$. This is the difference on *y* as we go from the unweighted mean across clusters composed entirely of Groups 2 or 3 to clusters composed entirely of Group 4. Thus, the mean achievement score in schools where all children have two educated parents is .172 points *lower* than the mean achievement score across schools where all children have one educated parent. Attending a school where many children have two educated parents does *not* yield academic achievement gains over attending a school where many children have just one educated parent. Taken together with $\hat{\gamma}_{01}$, it appears that attending a school where children have *any* educated parents leads to academic achievement gains, and whether fellow students have one or two educated parents is not influential.

The between-cluster slope of the third contrast code is $\hat{\gamma}_{03} = 1.911$. This reflects the mean difference on *y* as we go from a cluster composed entirely of Group 2 to a cluster composed entirely of Group 3. Thus, as we go from a school where all children have an educated mother to a school where all children have an educated father, the mean academic achievement score increases by 1.911 points. This contrast points to a unique benefit of attending a school where many children have educated fathers. The within-school slope of this contrast code was near zero (.098). It appears that, although having only an educated father does not yield academic gains over having only an educated mother for individual students *within* schools, students *across* schools perform better when more of their classmates have educated fathers rather than educated mothers. Many educated fathers may be linked to increased SES and financial resources of a school, and/or to greater value placed on education in the broader community.

Interpreting Level-2 slopes involves comparing means of hypothetical, homogeneous schools. Again, we divide these slopes by 10. The between-cluster slope of the first contrast code, $\hat{\gamma}_{01} = 3.183$, shows that as the percentage of students with one or two educated parents increases by 10% (and the percentage of students with no educated parents decreases by 10%), mean school achievement increases by .3183 points. The between-cluster slope of the second contrast code, $\hat{\gamma}_{02} = -.237$, indicates that as the percentage of students with two educated parents increases by 10% (and the percentage of students with one educated parent decreases by 10%), mean school achievement decreases by .0237 points. Finally, the between-cluster slope of the third contrast code, $\hat{\gamma}_{03} = 1.911$, suggests that as the percentage of students with educated fathers increases by 10% (and the percentage of students with educated mothers decreases by 10%), the mean school achievement increases by .1911 points.

The UN(M) Model vs. the CWC(M) Model. As shown in Table 2, the UN(M) Model and the CWC(M) Model yield precisely the same estimates at Level 1. In both cases, these slopes reflect “pure” within-cluster effects and are interpreted identically. Additionally, the two models are equivalent; each yields identical estimates of variance components, as well as identical log-likelihood values, thus yielding the same fit statistics. The sole difference between these models pertains to Level-2 slopes, which are different in both magnitude and interpretation. The UN(M) Model is a

⁴ We thank an anonymous reviewer for this useful suggestion.

simple reparameterization of the CWC(M) Model, but importantly, this is only true when random slopes are not included. When including random slopes and/or interaction terms, the CWC(M) Model is recommended over the UN(M) Model (Kreft et al., 1995).

At Level 2, the UN(M) Model yields *contextual effects* whereas the CWC(M) Model yields between-cluster effects. For any predictor, the contextual effect is equal to the between effect minus the within effect: $\beta_c = \beta_b - \beta_w$. In Table 2, this is evident for all three coding schemes. For example, consider the first contrast code. In the CWC(M) Model, its within slope is $\hat{\gamma}_{10} = .551$ and its between slope is $\hat{\gamma}_{01} = 3.183$. In the UN(M) Model, its within slope is again $\hat{\gamma}_{10} = .551$, whereas its contextual effect is $\hat{\gamma}_{01}^c = 2.631$. Indeed, $3.183 - .551 = 2.631$, within rounding error. The key difference between these coefficients is their interpretation. The contextual effect is the Level-2 effect that is *above and beyond* the within-cluster effect. This can also be thought of as the Level-2 effect that remains after partialing out the overall effect of the Level-1 predictor. Indeed, the statistical significance of the contextual effect reveals whether the between-cluster effect is significant over and above the within-cluster effect.

Why might the UN(M) Model be chosen over the CWC(M) Model, and vice versa? This depends largely on the researcher's questions and goals. In our example, the between effect may be most useful to the researcher who aims to inform theory about macro-level influences of the school environment on academic achievement. Between-cluster effects answer questions concerning Level-2 effects *in isolation*, irrespective of any Level-1 effects that may also be present.

In contrast, the contextual effect isolates the *unique effect* of macro-level factors, ruling out any potential confounding or contradictory effects that may be present at the micro-level. In our example, the contextual effect of the second dummy code describes the effect of attending a school where many children have educated fathers, *above and beyond* the individual effect of having an educated father (vs. no educated parents). This coefficient can be interpreted as the expected difference in achievement scores between two hypothetical students: both have an educated father, but one attends a school where no children have educated fathers and the other attends a school where all children have educated fathers. The contextual effect answers: Beyond the individual-level effect, what is the additional effect of attending a school where all students have an educated father, compared with a school where no students do? Even though the CWC(M) and UN(M) Models are equivalent, some methodologists (e.g., Begg & Parides, 2003) argue that the parameterization offered by the UN(M) Model is the desirable choice in most situations because it allows the researcher to isolate unique Level-2 effects. However, extra care must be taken in the interpretation stage.

The UN(M) Model

Dummy Codes. The first dummy code has a contextual effect of $\hat{\gamma}_{01}^c = 1.814$. Beyond the individual effect of having an educated mother, there is an additional positive effect of attending a school where many children have educated mothers. If we chose two hypothetical students who both had an educated mother, but one attended a school where all children had no educated parents and the other attended a school where all children had educated mothers, the latter child would score about 1.814 points higher on

achievement. An even stronger contextual effect is present for educated fathers, $\hat{\gamma}_{02}^c = 5.439$ points. The contextual effect of two educated parents is $\hat{\gamma}_{03}^c = 3.272$ points.

Again, it is undesirable that interpreting contextual effects involves comparing students from hypothetical schools. Let us again divide these coefficients by 10. Now, using the contextual effect of the second dummy code, $\hat{\gamma}_{02}^c$, as an example, we can say that holding an individual's parental education constant, as the percentage of students with educated fathers increases by 10% in a given school (and the percentage of students with no educated parents decreases by 10%), we expect a student's achievement score to increase by about .54 points.

Contrast Codes. The contextual effect of the first contrast code is $\hat{\gamma}_{01}^c = 2.631$. Beyond the individual-level effect of having any educated parents, there is an additional effect of attending a school where many children have any educated parents. Holding a student's parental education constant, we expect that attending a school where all students have one or two educated parents will lead to an academic achievement score that is 2.631 points higher, compared with attending a school where all students have no educated parents. Next, the contextual effect of the second contrast code is $\hat{\gamma}_{02}^c = -.237$. Holding a student's parental education constant, attending a school where all students have two educated parents will result in an academic achievement score that is .237 points *lower*, compared with attending a school where all students have just one educated parent. Finally, the contextual effect of the third contrast code is $\hat{\gamma}_{03}^c = 1.812$. Holding a student's parental education constant, attending a school where all students have only an educated father predicts an academic achievement score that is 1.812 points higher, compared with attending a school where all students have only an educated mother.

We can again divide these coefficients by 10. Beginning with the first contextual effect $\hat{\gamma}_{01}^c$, holding a student's parental education constant, as the percentage of students with one or two educated parents increases by 10% (and the percentage of students with no educated parents decreases by 10%), academic achievement scores increase by .2631 points. Second, $\hat{\gamma}_{02}^c$ indicates that holding a student's parental education constant, as the percentage of students with two educated parents increases by 10% (and the percentage of students with just one educated parent decreases by 10%), academic achievement scores decrease by .237 points. Finally, $\hat{\gamma}_{03}^c$ shows that holding a student's parental education constant, as the percentage of students with only an educated father increases by 10% (and the percentage of students with only an educated mother decreases by 10%), academic achievement scores increase by .1812 points. Note that these interpretations are nearly identical to those obtained from the CWC(M) Model, with the added caveat that we are holding the original predictor constant. Because the UN(M) Model isolates the *unique* macro-level effects of parental education on academic achievement, above and beyond any individual effects, one may argue that the parameterization of the UN(M) Model is preferable to that of the CWC(M) Model in this example.

Special Considerations for Interpretation

The use of categorical predictors in MLM presents many interpretational nuances that must be approached with care. To our

knowledge, these nuances have not been previously addressed in methodological work. First, we discuss cluster homogeneity. As described above, a major interpretational oddity of categorical predictors involves Level-2 slopes; we are required to compare two (often hypothetical) clusters, each of which is homogeneous. To improve upon this interpretation, we suggest dividing the coefficient by 10, which permits one to interpret the coefficient in terms of linear increase. However, the reader may wonder whether this approach is sufficient for gaining a full understanding of macro-level effects. In our example, suppose we are first interested in the effects of a linear increase in the percentage of students with two educated parents. However, suppose we *also* suspect that there is a distinct qualitative effect of attending a school where *all* students have two educated parents (or, vice versa, a school where *all* students have no educated parents). In certain cases, we can explicitly model this in MLM. In addition to interpreting our Level-2 slope as described above, we may also create a Level-2 indicator of cluster homogeneity with respect to parent education (e.g., we can code this indicator 1 if the cluster is homogeneous, 0 otherwise, assuming there is a sufficient number of homogeneous clusters). Then, we can obtain the Level-2 effect of the cluster proportion as usual, and also introduce this indicator as an additional Level-2 predictor of achievement, thus obtaining two separate effects. We can then answer: Beyond the effect of increasing proportions of educated parents in a school, is there a distinct qualitative effect of attending a homogeneous school wherein *all* children have two educated parents? MLM permits the investigation of such nuanced questions with categorical predictors.

We next discuss the interpretation of intercepts, which may be notably different from intercept interpretation with continuous predictors. In a model with Level-2 predictors (for example, the CWC(M) Model), the intercept term $\hat{\gamma}_{00}$ must be interpreted as the expected value of y in a cluster where all the Level-2 predictors are zero. Note that this interpretation rests on the Level-2 (not Level-1) predictors equaling zero, and depending on the coding scheme employed, this will take on various meanings. For dummy codes, the intercept is the expected value of y for a cluster that is composed entirely of the reference group. In some situations, this intercept interpretation may be substantively meaningful (e.g., in our example, it may be useful to describe the predicted achievement score in a school where all children have no educated parents). For contrast and effect codes, interpretation of the intercept is less straightforward. According to our derivations, the intercept maps neatly onto a substantively useful quantity only under particular data conditions (e.g., when $ICC_X = 1$). In all other conditions, the unweighted mean of all group means is equal to the intercept term plus an “adjustment” term that may vary in complexity from sample to sample. Thus, if the unweighted (or weighted) grand mean is desired, we recommend relying on descriptive statistics rather than attempting to interpret the intercept estimate.

General Discussion

Our goals in this tutorial article were to clarify why and how categorical predictors should be centered in multilevel models, to explain how their corresponding coefficients should be interpreted, and to demonstrate these conclusions with a real-world example. First, we have shown that centering guidelines for continuous predictors do indeed apply to categorical predictors. Perhaps most importantly, we demonstrate that slope coefficients associated with uncentered coding

variables yield conflated, uninterpretable blends of within- and between-cluster effects. Second, we have demonstrated via algebraic derivations precisely how multilevel slope coefficients associated with dummy codes, contrast codes, and effect codes should be interpreted at both levels, and demonstrated these interpretations with an empirical example. Next, we address special considerations and issues that applied researchers may face when including categorical predictors in multilevel models. Throughout, we draw on core conclusions made in previous sections and use the empirical example to demonstrate our recommendations.

Working With Categorical Covariates

Our literature review showed that categorical predictors are often included in multilevel models as covariates (e.g., to explore the effects of focal predictors after controlling for gender, race, etc.). Here, we explain how our findings apply to categorical predictors whose sole function is that of a covariate. Regardless of whether the focal predictor(s) are at Level 1, Level 2, or involve cross-level interactions, the Level-1 categorical covariate should be centered in such a way that isolates its relevant level-specific effects. Further, the researcher should consider what is most substantively relevant to control for: within-cluster differences in the categorical predictor (e.g., individual differences in race within classrooms), across-cluster differences in the categorical predictor (e.g., differences in racial makeup across classrooms), or both.

First, we consider the scenario where a Level-1 predictor is of primary interest; the goal is to estimate the within-cluster effect of the focal predictor, after controlling for a categorical covariate. Here it is necessary to control for the within-cluster component of the covariate, whether it be continuous or categorical. However, UN (and by extension, CGM) coding variables capture a conflated mix of within- and between-cluster effects. Thus, inclusion of a UN or CGM coding variable as a covariate will not effectively isolate and control for its within-cluster component. Therefore, we argue that CWC is most appropriate for a categorical covariate in this scenario.

Second, we consider the scenario where a Level-2 predictor is of focal interest; the goal is to estimate the between-cluster effect of the focal predictor while controlling for individual differences in the Level-1 covariate. In the past, methodologists have recommended using CGM—and, by extension, UN—covariates in this scenario (Enders & Tofghi, 2007). However, Rights et al. (2019) show that including a UN or CGM covariate is *not* sufficient to control for it in this situation, and that it often yields a biased estimate of the Level-2 effect of interest. Sole inclusion of the UN or CGM covariate is a misspecification which lets unwanted bias propagate throughout the model (Rights et al., 2019). The authors then show that inclusion of the *cluster mean* of the covariate will yield an unbiased estimate of the Level-2 effect of focal interest; inclusion of the covariate at Level 1, whether CWC, CGM, or UN, is not necessary. The algebraic support for this conclusion does not require distinguishing between continuous and categorical predictors, so it immediately follows that the same guidelines apply for categorical covariates.

In some applications, it may be necessary to control for a covariate at multiple levels. Rights et al. (2019) show that inclusion of a UN or CGM covariate does not successfully control for a covariate across multiple levels. Ultimately, inclusion of the covariate

without effectively partitioning its level-specific effects is a model misspecification and is therefore inappropriate. If it is necessary to control for a categorical covariate at multiple levels, we recommend inclusion of CWC alongside cluster means of the covariate, as in the CWC(M) Model.

Estimating and Interpreting Interactions

In practice, it may also be of interest to include a categorical predictor in a cross-level or same-level interaction. We consider scenarios where the within-cluster effect of a categorical Level-1 predictor is moderated by a Level-2 variable (i.e., cross-level interaction) or another Level-1 variable (i.e., same-level interaction), as this has not been previously addressed. Inclusion of a Level-2 categorical moderator is a straightforward extension of single-level regression and has been addressed elsewhere (e.g., Preacher et al., 2006), so we do not address it here.

When including a Level-1 categorical predictor in a cross-level interaction, centering the categorical predictor will again be appropriate. With interactions, it is important to obtain an unbiased estimate of the within-cluster effect of the predictor, because *this* is the effect that is hypothesized to be moderated. Just as is the case for slopes, interaction effects involving UN predictors will be a conflated mix of separate interactions occurring at the within- and between-cluster levels (Cronbach & Webb, 1975; Preacher et al., 2016). Thus, it is necessary to isolate the within-cluster effect of a coding variable via CWC before introducing a cross-level interaction, and this is also true for same-level interactions. For details, see Enders and Tofghi (2007, pp. 132–134), keeping in mind that their algebraic demonstrations do not require distinguishing between continuous and categorical focal predictors.

Testing the Overall Significance of a Categorical Predictor's Effects

It is also possible to test the significance of the between-cluster, within-cluster, and overall effects of a multicategorical predictor. Because CWC coding variables are uncorrelated with their cluster means, nested model comparisons using deviance tests can be undertaken (for details, see Snijders & Bosker, 2012, Chapter 6). For instance, the deviance test can assess the omnibus within-cluster effect of the categorical predictor by comparing the null model to the model with all CWC coding variables. Similarly, the omnibus between-cluster effect of the predictor can be assessed by introducing cluster means of all coding variables and conducting a deviance test. Finally, by the same procedure, we can test the omnibus significance of the predictor at both levels by adding within- and between-cluster components simultaneously.

We conducted each of these deviance tests on our empirical data example; *R* code is provided in Appendix F in online supplemental materials. The omnibus test of within-cluster effects yields a χ^2 statistic with 3 degrees of freedom (*df*), because compared to the null model, we introduced three CWC coding variables. This test showed that inclusion of the CWC coding variables significantly improved model fit, $\chi^2(3) = 53.23$, $p < .001$, indicating that, overall, parental education status is significantly predictive of students' academic achievement within schools. Next, the omnibus test of between-cluster effects also yielded a χ^2 statistic with 3 *df*; compared to the null model, we introduced three cluster means. This test indicated that,

overall, parental education status is also significantly predictive of mean academic achievement outcomes across schools, $\chi^2(3) = 32.61$, $p < .001$. Last, we conducted an omnibus test of the categorical predictor at both levels; this test yielded a χ^2 statistic with 6 *df*, because the null model was compared to a model with three CWC coding variables and three cluster means. Parental education was significantly predictive of student academic achievement outcomes, $\chi^2(6) = 85.76$, $p < .001$, though with this test alone, we cannot conclude which effects were significant. Separate level-specific tests of the coding variables are therefore more useful. Of note, these tests yielded the same results regardless of coding scheme, reiterating that precisely the same information is carried within any set of coding variables.

However, deviance tests for fixed effects are valid only when full information maximum likelihood (FIML) estimation is used. FIML estimation is sometimes appropriate, such as when sample size is large, but restricted maximum likelihood (REML) is the preferred estimation method under many conditions, especially when sample size is smaller. Indeed, REML is the default estimation method in many popular software programs, including the *lme4* package in *R*. Deviance tests for fixed effects are no longer valid under REML estimation. Instead, we can use multivariate Wald-style *F*-tests. These are available through the *contest* function in the *lmerTest* package (Kuznetsova et al., 2017); example *R* code is provided in Appendix F in online supplemental materials. Here, we test the same hypotheses as above (omnibus significance of all coding variable effects at Level 1, Level 2, and both levels), but rather than obtaining a χ^2 statistic, we obtain an *F* statistic and corresponding *p*-value. The same multivariate tests are available in most software packages (e.g., CONTRAST in SAS MIXED; TEST in STATA MIXED, and TEST in SPSS MIXED).

Finally, to obtain effect sizes associated with these omnibus level-specific effects, various R^2 measures exist for quantifying variance explained at Level 1, Level 2, or both levels (Rights & Sterba, 2019)—these measures can be applied analogously to categorical predictors.

Latent Variables and MSEM

When working with categorical predictors in practice, one must consider whether multilevel structural equation modeling (MSEM) is appropriate based on the variable(s) under study. It is important to acknowledge whether latent constructs should be considered with respect to the Level-1 categorical variable itself. In this tutorial article, for simplicity, we assumed that our categorical predictor represents a variable that is *truly* categorical in nature. However, categorical predictors may not always take this form in practice; in other situations, a categorical predictor may be better conceptualized as an inaccurate, error-prone representation of a latent continuous construct. In fact, our empirical example may be better suited for this conceptualization. Although we used a four-category indicator of parental education, underlying this indicator is likely a continuous range of both mothers' and fathers' education level. For example, a given student may have a mother who was educated until age 14 and a father who was educated until age 16, whereas another student's mother was educated until age 17 and father was educated until age 13. Many of these fine-grained differences in parental education were likely washed out by our

categorical indicator. Conceptualizing the categorical predictor as a rough indicator of a truly continuous construct may lead to different results with respect to estimation bias, and different methodological steps that should be taken in practice.

Indeed, Asparouhov and Muthén (2019) simulated multilevel data with a binary predictor that was created to be a crude representation of a latent continuous construct, and subsequently observed biased Level-2 estimates. The authors concluded that in such scenarios, it is necessary to use MSEM and group-mean-center the binary predictor around its *latent* cluster mean rather than its observed cluster mean. In summary, categorical predictors that are error-prone representations of latent continuous constructs (e.g., illness status, SES), and largely error-free categorical predictors that represent a truly categorical construct (e.g., experimental group, school grade year, blood type) may need to be treated differently in practice.

Randomly Assigned Versus Naturally-Occurring Categories

In addition to distinguishing whether the categorical variable should be treated as manifest or latent, it is also useful to consider whether the categorical predictor arises from random assignment or naturally-occurring groups. This distinction may have implications for how the effects of the categorical predictor will be interpreted and generalized. When naturally-occurring categories are under study (e.g., in our empirical example), the researcher will likely not be able to manipulate within-cluster proportions of the predictor; rather, the natural variation of group proportions across clusters will bolster the representativeness of the sample, and in turn, the generalizability of conclusions. In these situations, as demonstrated above, the researcher is in a position to study the effects of linear change of group proportions on the outcome variable, as well as distinct qualitative effects of cluster homogeneity versus heterogeneity.

In contrast, random assignment to experimental groups at Level 1 allows the researcher to intentionally manipulate group proportions within and across clusters. Indeed, MLM approaches are increasingly used to study social and psychological phenomena in experimental contexts (e.g., Judd et al., 2012). In these situations, we argue that there are two opposing approaches that present different benefits. First, consider a situation in which the Level-2 effect of group proportion is of substantive interest. For example, in assigning students within classrooms to a reading intervention, the researcher is interested in the individual-level effect of the intervention *as well as* classroom-level effects: Does the proportion of students who receive the intervention influence mean reading outcomes? In this case, it is advantageous to manipulate cluster proportions such that they are highly variable across clusters (i.e., the proportion of students who receive the intervention is very different from classroom to classroom). Thus, estimates of Level-2 effects can be obtained with reasonable precision.

Second, when the Level-2 effect of group proportions is not of substantive interest, the researcher may instead allocate subjects to groups in order to achieve an *optimal design*. One class of optimal design minimizes the standard error (*SE*) of the effect(s) of interest (e.g., a within-cluster treatment effect), thereby maximizing the precision of the estimate. In single-level settings with two experimental groups, the *SE* of the treatment effect is minimized when

the proportion of subjects in each group is .5 (McClelland, 1997), and this principle extends to multilevel designs (Moerbeek et al., 2008). Thus, in this case, it is advantageous to manipulate cluster proportions such that they are identical (i.e., minimally variable) across clusters. When multiple experimental groups are involved and there is no single contrast that is of greatest substantive interest, more complex approaches are available to determine optimal allocation into groups (Aufenanger, 2017). However, to our knowledge, these approaches have been developed only for single-level settings. In summary, in experimental contexts, the substantive relevance of Level-2 cluster proportions should be considered in the study design.

Categorical Predictors and Multilevel Model Assumptions

In past work, some methodologists have posited that the inclusion of binary predictors results in violations of many of MLM's fundamental assumptions. It is important to understand these model assumptions, and how the properties of categorical variables are in line with them. Indeed, we argue that the inclusion of centered categorical (including binary) predictors is warranted. First, Asparouhov and Muthén (2019) note that the variance of a CWC binary predictor is not constant across clusters, as it is directly determined by the value of the cluster mean; namely, its variance will be $p_j(1 - p_j)$ where p_j is the proportion of Group 1 members in cluster j . However, multilevel models do not invoke assumptions about the variances of predictors, as assumptions pertain only to error variances (Davidian & Giltinan, 1995; Dedrick et al., 2009; Pinheiro & Bates, 2000; Raudenbush & Bryk, 2002; Snijders & Bosker, 2012). Moreover, a predictor is considered fixed for a given observation at Level-1 or Level-2, not random. Fixed predictors (considered case-wise) do not vary. All this suggests that nonconstant variance of the within parts of coding variables is inconsequential.

Second, Asparouhov and Muthén (2019) argue that the within- and between-cluster components of a binary predictor are not independent, in that the value of the cluster mean directly determines the variance, range, and values that the CWC binary predictor can take on. They further argue that this lack of independence undermines the idea of a between-cluster or contextual effect, and that separate level-specific effects can never be estimated because of this violation. However, MLM assumptions do not pertain to the independence of the within- and between-cluster parts of the predictor, but rather to the independence of errors across levels (Dedrick et al., 2009; Pinheiro & Bates, 2000; Raudenbush & Bryk, 2002). This assumption is still supported when the predictor is categorical. Additionally, whereas \bar{d}_j does determine the values of $(d_{ij} - \bar{d}_j)$ that can exist for cluster j , crucially, \bar{d}_j does not determine whether a *particular observation* is equal to $(0 - \bar{d}_j)$ or $(1 - \bar{d}_j)$; that is determined stochastically. Thus, we argue that this potential lack of independence across levels is also inconsequential.

Third, Asparouhov and Muthén (2019) note that the assumption of normality of the within- and between-cluster components of a predictor is violated, because the distribution of a binary Level-1 variable may be skewed. Here, we note that distributional assumptions need not be applied to predictors in MLM, provided the predictors are on an interval scale (Hoffman, 2015; Snijders & Bosker, 2012). This suggests that no assumption violation is present.

Finally, Asparouhov and Muthén (2019) argue that the most significant problem posed by binary predictors is that they are, necessarily, imperfect and error-filled representations of latent underlying constructs that are continuous. However, as discussed previously, this is not always true of a binary predictor, and especially of a multicategorical predictor. Indeed, such variables often do represent constructs that are truly categorical in nature, either because they are manipulated (e.g., experimental condition) or observed natural categories (e.g., blood type). As noted above, we agree that additional methodological steps should be taken when there is substantive reason to believe that the categorical predictor is a crude representation of a latent continuous construct. However, such steps are not always necessary.

Limitations and Future Directions

In terms of limitations of this report, first, we restricted our focus to random intercept, fixed slope models, though in practice it is often of interest to include categorical predictors with random slopes. A deeper exploration of the behavior of random slope models with categorical predictors is warranted. Second, in our empirical example we do not address the treatment of cluster means associated with categorical predictors as latent variables (i.e., MSEM). Third, our focus was restricted to continuous outcomes. Future work could examine the issues discussed here in the context of outcome variables that are binary, count, ordinal, and so on. Finally, the use of categorical predictors in scenarios with partial nesting, three-level structures, or cross-classification remains unexplored. All of these topics will be important directions for future work.

Central Takeaways

In this tutorial article, we have shown that the algebra and principles underlying centering remain the same whether a predictor is continuous or categorical. However, there are important differences between continuous and categorical predictors when it comes to the interpretation of their effects. When including multicategorical predictors in multilevel models, researchers must be intentional throughout each stage of model specification and interpretation.

First, the coding scheme that best fits the researcher's theory and hypotheses should be used. In our empirical example, we demonstrate the coefficients that can be obtained through use of dummy, contrast, and unweighted effect codes, and how the interpretations of those coefficients map onto various substantive questions, though there are even more potential coding schemes not addressed here. Second, the model that isolates the most useful level-specific effects should be chosen, such as the CWC(M) Model for within- and between-cluster effects, or the UN(M) Model for within-cluster and contextual effects. We provide considerations and examples from which researchers may draw while deciding on a model specification. Third, and perhaps most importantly, the resulting slope estimates must be interpreted such that they match the chosen coding scheme and model specification. To accomplish this, the algebraic support and empirical examples provided here may be carried forward to new contexts. Finally, other important conceptual considerations may require attention (e.g., interpreting intercepts; whether the use of latent variable modeling [MSEM] is warranted). Guidelines for approaching these considerations are outlined throughout the tutorial.

Centering a coding variable may initially seem counterintuitive because its new values are unfamiliar (e.g., no longer 0 and 1), and could be seen as further *compromising* the interpretability of results rather than enhancing it. Such criticisms have been raised in regard to continuous predictors (e.g., Kelley et al., 2017; Plewis, 1989) but have been widely refuted (e.g., Kenny & La Voie, 1985; Neuhaus & McCulloch, 2006). Here, we have demonstrated that these conclusions carry over to categorical predictors. In practice, the flawed interpretation of conflated estimates has likely led to many spurious conclusions, and will continue to do so unless appropriate treatment of categorical predictors is employed. We seek to aid future researchers in more accurately estimating and interpreting important effects in multilevel models across the fields of psychology, education, and more.

References

- Aryee, S., Walumbwa, F. O., Seidu, E. Y. M., & Otaye, L. E. (2012). Impact of high-performance work systems on individual- and branch-level performance: Test of a multilevel model of intermediate linkages. *Journal of Applied Psychology, 97*(2), 287–300. <https://doi.org/10.1037/a0025739>
- Asparouhov, T., & Muthén, B. (2006). Constructing covariates in multilevel regression. *Mplus Web Notes, 11*, 1–8.
- Asparouhov, T., & Muthén, B. (2019). Latent variable centering of predictors and mediators in multilevel and time-series models. *Structural Equation Modeling, 26*(1), 119–142. <https://doi.org/10.1080/10705511.2018.1511375>
- Aufenganger, T. (2017). *Treatment allocation for linear models with covariate information*. FAU Discussion Papers in Economics, No. 14/2017. <https://www.iwf.rw.fau.de/files/2015/12/14-2017.pdf>
- Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1–48. <https://doi.org/10.18637/jss.v067.i01>
- Begg, M. D., & Parides, M. K. (2003). Separation of individual-level and cluster-level covariate effects in regression analysis of correlated data. *Statistics in Medicine, 22*(16), 2591–2602. <https://doi.org/10.1002/sim.1524>
- Bowers, A. J., & Urlick, A. (2011). Does high school facility quality affect student achievement? A two-level hierarchical linear model. *Journal of Education Finance, 37*(1), 72–94. <https://www.jstor.org/stable/23018141>
- Brincks, A. M., Enders, C. K., Llabre, M. M., Bulotsky-Shearer, R. J., Prado, G., & Feaster, D. J. (2017). Centering predictor variables in three-level contextual models. *Multivariate Behavioral Research, 52*(2), 149–163. <https://doi.org/10.1080/00273171.2016.1256753>
- Charbonnier-Voirin, A., El Akremi, A., & Vandenberghe, C. (2010). A multilevel model of transformational leadership and adaptive performance and the moderating role of climate for innovation. *Group & Organization Management, 35*(6), 699–726. <https://doi.org/10.1177/1059601110390833>
- Clark, P. C. (2013). *The effects of multicollinearity in multilevel models* (Doctoral dissertation). Wright State University.
- Cronbach, L. J. (1976). *Research on classrooms and schools: Formulation of questions, design and analysis*. Stanford University Evaluation Consortium.
- Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Lawrence Erlbaum Associates.
- Chronbach, L. J., & Webb, N. (1975). Between-class and within-class effects in a reported aptitude* treatment interaction: Reanalysis of a study by GL Anderson. *Journal of Educational Psychology, 67*, 717–724.
- Curran, P. J., & Bauer, D. J. (2011). The disaggregation of within-person and between-person effects in longitudinal models of change. *Annual*

- Review of Psychology*, 62(1), 583–619. <https://doi.org/10.1146/annurev.psych.093008.100356>
- Davidian, M., & Giltinan, D. M. (1995). *Nonlinear models for repeated measurement data*. CRC Press. <https://doi.org/10.1201/9780203745502>
- Dedrick, R. F., Ferron, J. M., Hess, M. R., Hogarty, K. Y., Kromrey, J. D., Lang, T. R., Niles, J. D., & Lee, R. S. (2009). Multilevel modeling: A review of methodological issues and applications. *Review of Educational Research*, 79(1), 69–102. <https://doi.org/10.3102/0034654308325581>
- Dettmers, S., Trautwein, U., Lüdtke, O., Goetz, T., Frenzel, A. C., & Pekrun, R. (2011). Students' emotions during homework in mathematics: Testing a theoretical model of antecedents and achievement outcomes. *Contemporary Educational Psychology*, 36(1), 25–35. <https://doi.org/10.1016/j.cedpsych.2010.10.001>
- Diez Roux, A. V. (2002). A glossary for multilevel analysis. *Journal of Epidemiology and Community Health*, 56(8), 588–594. <https://doi.org/10.1136/jech.56.8.588>
- Enders, C. K. (2013). Centering predictors and contextual effects. In M. A. Scott, J. S. Simonoff, & B. D. Marx (Eds.), *The SAGE handbook of multilevel modeling* (pp. 89–108). Sage. <https://doi.org/10.4135/9781446247600.n6>
- Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods*, 12(2), 121–138. <https://doi.org/10.1037/1082-989X.12.2.121>
- Galindo, C., & Sheldon, S. B. (2012). School and home connections and children's kindergarten achievement gains: The mediating role of family involvement. *Early Childhood Research Quarterly*, 27(1), 90–103. <https://doi.org/10.1016/j.ecresq.2011.05.004>
- Gong, Y., Kim, T.-Y., Lee, D.-R., & Zhu, J. (2013). A multilevel model of team goal orientation, information exchange, and creativity. *Academy of Management Journal*, 56(3), 827–851. <https://doi.org/10.5465/amj.2011.0177>
- Grice, G. R. (1966). Dependence of empirical laws upon the source of experimental variation. *Psychological Bulletin*, 66(6), 488–498. <https://doi.org/10.1037/h0023914>
- Grilli, L., & Rampichini, C. (2018). A handful of critical choices in multilevel modelling. *BEIO: Boletín de Estadística e Investigación Operativa*, 34(1), 7–24.
- Härnqvist, K. (1978). Primary mental abilities at collective and individual levels. *Journal of Educational Psychology*, 70(5), 706. <https://doi.org/10.1037/0022-0663.70.5.706>
- Hedeker, D., & Gibbons, R. D. (2006). *Longitudinal data analysis*. Wiley. <https://doi.org/10.1002/0470036486>
- Hoffman, L. (2015). *Longitudinal analysis: Modeling within-person fluctuation and change*. Routledge. <https://doi.org/10.4324/9781315744094>
- Hofmann, W., Baumeister, R. F., Förster, G., & Vohs, K. D. (2012). Everyday temptations: An experience sampling study of desire, conflict, and self-control. *Journal of Personality and Social Psychology*, 102(6), 1318–1335. <https://doi.org/10.1037/a0026545>
- Hofmann, D. A., & Gavin, M. B. (1998). Centering decisions in hierarchical linear models: Implications for research in organizations. *Journal of Management*, 24(5), 623–641. <https://doi.org/10.1177/014920639802400504>
- 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(1), 54–69. <https://doi.org/10.1037/a0028347>
- Kärnä, A., Voeten, M., Little, T. D., Alanen, E., Poskiparta, E., & Salmivalli, C. (2013). Effectiveness of the KiVa antibullying program: Grades 1–3 and 7–9. *Journal of Educational Psychology*, 105(2), 535–551. <https://doi.org/10.1037/a0030417>
- Kärnä, A., Voeten, M., Poskiparta, E., & Salmivalli, C. (2010). Vulnerable children in varying classroom contexts: Bystanders' behaviors moderate the effects of risk factors on victimization. *Merrill-Palmer Quarterly*, 56(3), 261–282. <https://www.jstor.org/stable/23098070>
- Kelley, J., Evans, M. D. R., Lowman, J., & Lykes, V. (2017). Group-mean-centering independent variables in multi-level models is dangerous. *Quality & Quantity: International Journal of Methodology*, 51(1), 261–283. <https://doi.org/10.1007/s11135-015-0304-z>
- Kenny, D. A., & La Voie, L. (1985). Separating individual and group effects. *Journal of Personality and Social Psychology*, 48(2), 339–348. <https://doi.org/10.1037/0022-3514.48.2.339>
- Kreft, I. G. G., & De Leeuw, J. (1998). *Introducing multilevel modeling*. Sage.
- Kreft, I. G. G., de Leeuw, J., & Aiken, L. S. (1995). The effect of different forms of centering in hierarchical linear models. *Multivariate Behavioral Research*, 30(1), 1–21. https://doi.org/10.1207/s15327906mbr3001_1
- Kuo, M., Mohler, B., Raudenbush, S. L., & Earls, F. J. (2000). Assessing exposure to violence using multiple informants: Application of hierarchical linear model. *Journal of Child Psychology and Psychiatry*, 41(8), 1049–1056. <https://doi.org/10.1111/1469-7610.00692>
- Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82(13), 1–26. <https://doi.org/10.18637/jss.v082.i13>
- Littell, J. H., & Tajima, E. A. (2000). A multilevel model of client participation in intensive family preservation services. *The Social Service Review*, 74(3), 405–435. <https://doi.org/10.1086/516411>
- Liu, O. L., Lee, H., & Linn, M. C. (2010). An investigation of teacher impact on student inquiry science performance using a hierarchical linear model. *Journal of Research in Science Teaching*, 47(7), 807–819. <https://doi.org/10.1002/tea.20372>
- Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods*, 13(3), 203–229. <https://doi.org/10.1037/a0012869>
- Lüdtke, O., Robitzsch, A., Trautwein, U., & Kunter, M. (2009). Assessing the impact of learning environments: How to use student ratings of classroom or school characteristics in multilevel modeling. *Contemporary Educational Psychology*, 34(2), 120–131. <https://doi.org/10.1016/j.cedpsych.2008.12.001>
- Major, D. A., Fletcher, T. D., Davis, D. D., & Germano, L. M. (2008). The influence of work-family culture and workplace relationships on work interference with family: A multilevel model. *Journal of Organizational Behavior*, 29(7), 881–897. <https://doi.org/10.1002/job.502>
- McClelland, G. H. (1997). Optimal design in psychological research. *Psychological Methods*, 2(1), 3–19. <https://doi.org/10.1037/1082-989X.2.1.3>
- McCoach, D. B., O'Connell, A. A., Reis, S. M., & Levitt, H. A. (2006). Growing readers: A hierarchical linear model of children's reading growth during the first 2 years of school. *Journal of Educational Psychology*, 98(1), 14–28. <https://doi.org/10.1037/0022-0663.98.1.14>
- Merritt, E. G., Wanless, S. B., Rimm-Kaufman, S. E., Cameron, C., & Peugh, J. L. (2012). The contribution of teachers' emotional support to children's social behaviors and self-regulatory skills in first grade. *School Psychology Review*, 41(2), 141–159. <https://doi.org/10.1080/02796015.2012.12087517>
- Moerbeek, M., Van Breukelen, G. J. P., & Berger, M. P. F. (2008). Optimal designs for multilevel studies. In J. DeLeeuw & E. Meijer (Eds.), *Handbook of multilevel analysis* (pp. 177–205). Springer. https://doi.org/10.1007/978-0-387-73186-5_4
- Montague, M., Enders, C., & Dietz, S. (2011). Effects of cognitive strategy instruction on math problem solving of middle school students with learning disabilities. *Learning Disability Quarterly*, 34(4), 262–272. <https://doi.org/10.1177/0731948711421762>
- Morrison, E. W., Wheeler-Smith, S. L., & Kamdar, D. (2011). Speaking up in groups: A cross-level study of group voice climate and voice. *Journal of Applied Psychology*, 96(1), 183–191. <https://doi.org/10.1037/a0020744>
- Murayama, K., & Elliot, A. J. (2009). The joint influence of personal achievement goals and classroom goal structures on achievement-

- relevant outcomes. *Journal of Educational Psychology*, 101(2), 432–447. <https://doi.org/10.1037/a0014221>
- Neuhaus, J. M., & Kalbfleisch, J. D. (1998). Between- and within-cluster covariate effects in the analysis of clustered data. *Biometrics*, 54(2), 638–645. <https://doi.org/10.2307/3109770>
- Neuhaus, J. M., & McCulloch, C. E. (2006). Separating between- and within-cluster covariate effects by using conditional and partitioning methods. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 68(5), 859–872. <https://doi.org/10.1111/j.1467-9868.2006.00570.x>
- Nezlek, J. B. (2012a). Multilevel modeling analyses of diary-style data. In M. R. Mehl & T. S. Conner (Eds.), *Handbook of Research Methods for Studying Daily Life* (pp. 357–383). Guilford.
- Nezlek, J. B. (2012b). Multilevel modeling for psychologists. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), *APA handbook of research methods in psychology*, Vol. 3: *Data analysis and research publication* (pp. 219–241). American Psychological Association. <https://doi.org/10.1037/13621-011>
- Paterson, L. (1991). Socio-economic status and educational attainment: a multi-dimensional and multi-level study. *Evaluation & Research in Education*, 5(3), 97–121.
- Peugh, J. L. (2010). A practical guide to multilevel modeling. *Journal of School Psychology*, 48(1), 85–112. <https://doi.org/10.1016/j.jsp.2009.09.002>
- Pinheiro, J. C., & Bates, D. M. (2000). Linear mixed effects models: Basic concepts and examples. *Mixed-effects models in S and S-plus* (pp. 3–56). Springer. <https://doi.org/10.1007/b98882>
- Plewis, I. (1989). Comment on “centering” predictors in multilevel analysis. *Multilevel Modelling Newsletter*, 1(3), 6. <https://doi.org/10.1177/0193841X05275649>
- Powell, D. R., Son, S.-H., File, N., & San Juan, R. R. (2010). Parent-school relationships and children’s academic and social outcomes in public school pre-kindergarten. *Journal of School Psychology*, 48(4), 269–292. <https://doi.org/10.1016/j.jsp.2010.03.002>
- Preacher, K. J., Curran, P. J., & Bauer, D. J. (2006). Computational tools for probing interaction effects in multiple linear regression, multilevel modeling, and latent curve analysis. *Journal of Educational and Behavioral Statistics*, 31(4), 437–448. <https://doi.org/10.3102/10769986031004437>
- Preacher, K. J., Zhang, Z., & Zyphur, M. J. (2016). Multilevel structural equation models for assessing moderation within and across levels of analysis. *Psychological Methods*, 21(2), 189–205. <https://doi.org/10.1037/met0000052>
- Preacher, K. J., Zyphur, M. J., & Zhang, Z. (2010). A general multilevel SEM framework for assessing multilevel mediation. *Psychological Methods*, 15(3), 209–233. <https://doi.org/10.1037/a0020141>
- R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Raudenbush, S. W. (2009). Adaptive centering with random effects: An alternative to the fixed effects model for studying time-varying treatments in school settings. *Education Finance and Policy*, 4(4), 468–491. <https://doi.org/10.1162/edfp.2009.4.4.468>
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Sage.
- Raudenbush, S. W., & Willms, J. (1995). The estimation of school effects. *Journal of Educational and Behavioral Statistics*, 20(4), 307–335. <https://doi.org/10.3102/10769986020004307>
- Reyes, M. R., Brackett, M. A., Rivers, S. E., White, M., & Salovey, P. (2012). Classroom emotional climate, student engagement, and academic achievement. *Journal of Educational Psychology*, 104(3), 700–712. <https://doi.org/10.1037/a0027268>
- Rights, J. D., & Sterba, S. K. (2019). Quantifying explained variance in multilevel models: An integrative framework for defining R-squared measures. *Psychological Methods*, 24(3), 309–338. <https://doi.org/10.1037/met0000184>
- Rights, J. D., Preacher, K. J., & Cole, D. A. (2019). The danger of conflating level-specific effects of control variables when primary interest lies in Level-2 effects. *British Journal of Mathematical & Statistical Psychology*, 73(1), 194–211. <https://doi.org/10.1111/bmsp.12194>
- Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, 15(3), 351–357. <https://doi.org/10.2307/2087176>
- Sacco, J. M., & Schmitt, N. (2005). A dynamic multilevel model of demographic diversity and misfit effects. *Journal of Applied Psychology*, 90(2), 203–231. <https://doi.org/10.1037/0021-9010.90.2.203>
- Shieh, Y.-Y., & Fouladi, R. T. (2003). The effect of multicollinearity on multilevel modeling parameter estimates and standard errors. *Educational and Psychological Measurement*, 63(6), 951–985. <https://doi.org/10.1177/0013164403258402>
- Sirotnik, K. A. (1980). Psychometric implications of the unit-of-analysis problem (with examples from the measurement of organizational climate). *Journal of Educational Measurement*, 17(4), 245–282. <https://doi.org/10.1111/j.1745-3984.1980.tb00831.x>
- Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Sage.
- Trautwein, U., & Lüdtke, O. (2009). Predicting homework motivation and homework effort in six school subjects: The role of person and family characteristics, classroom factors, and school track. *Learning and Instruction*, 19(3), 243–258. <https://doi.org/10.1016/j.learninstruc.2008.05.001>
- Van de Pol, M., & Wright, J. (2009). A simple method for distinguishing within-versus between-subject effects using mixed models. *Animal Behaviour*, 77(3), 753–758. <https://doi.org/10.1016/j.anbehav.2008.11.006>
- Van Landeghem, G., Onghena, P., Van Damme, J., & Opdenakker, M.-C. (1999). The effect of different centering methods in multilevel analysis. *UCS Colloquium on Current Issues in Statistics*. <https://lirias.kuleuven.be/retrieve/8762>
- Yu, H., Jiang, S., & Land, K. C. (2015). Multicollinearity in hierarchical linear models. *Social Science Research*, 53, 118–136. <https://doi.org/10.1016/j.ssresearch.2015.04.008>

Received October 16, 2020

Revision received July 14, 2021

Accepted August 10, 2021 ■