Adv in Health Sci Educ (2008) 13:571–582
DOI 10.1007/s10459-007-9078-y

ORIGINAL PAPER

Conducting multilevel analyses in medical education

Michael J. Zyphur · Seth A. Kaplan · Gazi Islam · Adam P. Barsky · Michael S. Franklin

Received: 14 June 2005 / Accepted: 19 December 2005 / Published online: 12 October 2007

Abstract A significant body of education literature has begun using multilevel statistical models to examine data that reside at multiple levels of analysis. In order to provide a primer for medical education researchers, the current work gives a brief overview of some issues associated with multilevel statistical modeling. To provide an example of this technique, we then present a multilevel analysis examining the relationship between two individual-level variables and the "cross-level" interaction between this relationship and a school-level variable. In offering this discussion and example of multilevel modeling, we hope to provide medical educators with a basic introduction to multilevel statistics, including the advantages of utilizing these techniques.

Keywords Hierarchical linear modeling · Ordinary least squares · Regression · Statistics · Multilevel random coefficient modeling

Problems and issues that arise when data exist at multiple levels of analysis have long been a concern for researchers across a variety of disciplines including education, sociology, psychology and organizational studies (see Raudenbush and Bryk 2002; Klein and Kozloswski 2000). Although research endeavors across these fields often differ in terms of

M. J. Zyphur (✉)
Department of Management and Organization, National University of Singapore,
Singapore 105001, Singapore
e-mail: bizmjz@nus.edu.sg

S. A. Kaplan
Department of Psychology, George Mason University, Manassas, USA

G. Islam
Department of Management, IBMEC, Sao Paulo, USA

A. P. Barsky
Department of Management, University of Melbourne, Parkville, Australia

M. S. Franklin
Department of Psychology, University of Michigan, Ann Arbor, USA

substantive questions, theoretical perspectives, and methodological approaches, they all regularly encounter phenomena that are either hierarchically arranged and/or longitudinal in nature. In recent years, largely due to the advances in statistical computing capabilities, scholars from several of these areas have developed or refined statistical methods that allow one to appropriately analyze these complex types of data (Hox 1995). Such techniques allow for the simultaneous examination of (a) data which are grouped within individuals (i.e., multiple measures within individuals), (b) individual data nested within groups (e.g., student data nested within classes or schools), and (c) group data nested within higher-order entities (e.g., school data nested within regions or countries). Importantly, these methods allow one to assess and statistically account for the intercorrelations among data which occur within a given group, examine relationships at multiple levels of analysis, and investigate the variance of a relationship between two variables across multiple levels of analysis (Hofmann 1997).[1]

Although seldom acknowledged, most medical education data also are inherently multilevel in nature. In most cases, for instance, students are nested within course section and instructor, which further are nested within factors such as specialty and institution. This hierarchical structure presents certain conceptual and analytical considerations that, if ignored or handled without attending to potential non-independence, can yield misleading findings. In the current paper, our objectives are to introduce medical researchers to the statistical issues that arise when one encounters hierarchically nested data and to describe how one can use multilevel modeling to appropriately analyze such data structures.

In order to demonstrate the advantages that multilevel techniques offer, below, we provide a practical example highlighting the similarities and differences between multilevel techniques and the more commonly used ordinary least squares (OLS) regression. We hasten to note that, although a variety of multilevel statistical programs are available, such as Mlwin, Mplus, the "procmixed" procedure in SAS, and the mixed modeling function in SPSS, we follow the overwhelming trend in adopting the notation of Hierarchical Linear Modeling (HLM; Raudenbush and Bryk 2002), an especially popular program across various academic disciplines and literatures (for a discussion of available programs, see Kozlowski and Klein 2000). Using this notation, below, we review the concept of levels of analysis and discuss methodological and statistical considerations that multilevel data present.

Single-level modeling

A level of analysis may be defined as a place along a conceptual hierarchy which is shared by a set of data. An example of a level of analysis is that of the individual. Consider, for instance, a data set that contains 1000 entrance exam scores (EES) and first-year medical school grades (e.g., GPA) from students currently enrolled in 20 different medical schools across the U.S. (with 50 students per school). These two sets of scores both reside at the individual level of analysis, in that there is one score for each person on both variables.

[1] While this notation may be somewhat different than that employed in OLS regression, the importance of using HLM notation is threefold. First, this notation is ubiquitous in much literature dealing with multilevel modeling. Second, because this notation often is employed in statistical packages' equation-based language, our use of this notation allows one to interpret relevant output and to recognize the correspondence between results appearing in software packages and those in research articles. Third, given that multilevel models may incorporate more than two levels, our using unique symbols for each level of analysis yields less confusion in interpreting results than would using the same symbol applied at multiple levels.

Plausibly, medical school administrators, who head their school's admissions program, might wish to assess how well EES predicts students' first year grades. This relationship may be represented as the traditional OLS model:

$$y_i = \beta_0 + \beta_1(\mathrm{EES}_i) + r_i \tag{1}$$

In this model, y_i represents the first year grade of the ith student, β_0 represents the y-intercept, β_1 represents the slope or parameter estimate for the relationship between EES and GPA, EES_i represents the ith student's exam score, and r_i represents the error or residual for the ith student. Also of note here is that the variance of r_i is the residual variance for the individual-level model (Nezlek and Zyzniewski 1998).

The results from this analysis will inform the administrator of the average level of first year grades (in the form of the constant β_0) when the EES variable is "grand-mean centered" (i.e., the mean is subtracted from each score), the relationship between the variables of interest (in the form of the regression slope β_1), the amount of variance unaccounted for in GPA by these values (in the form of the variance of r_i), and will allow for the estimation of the probability that the results are due to chance alone (in the form of a statistical significance test). Relying on the results from this analysis, the administrator may be better able to predict the success of first-year medical students by knowing their scores on an entrance exam. However, this type of analysis overlooks a number of multilevel issues associated with these data and could, consequently, yield limited, and possibly erroneous, conclusions regarding the relationship of interest. Below, we discuss the two primary issues that confront researchers conducting an individual-level analysis with this or similar data.

At the conceptual level, the major problem that one encounters with such data is the fact that individual scores are nested within medical schools. This fact translates into two specific problems, one statistical and one that is more conceptual in nature. Underlying the statistical issue are a number of implicit assumptions that the researcher makes regarding the errors (i.e., residuals) when conducting OLS regression at the individual level of analysis. As Bliese (2002) notes, the researcher assumes that the errors in prediction will be (a) normally distributed with a mean of zero, (b) constant across levels of x (i.e., homoscedastic), and (c) independent of one another. Although OLS regression is fairly robust with respect to violation of the first two assumptions, the technique potentially yields erroneous results when the third assumption, that of independent errors, is violated (Kenny and Judd 1986).

In the present case, we are likely to violate this third assumption using OLS because we have not accounted for the fact that constituent scores likely will represent, in part, the schools from which they were sampled. Due either to individual factors predisposing different students to attend one school versus another (e.g., personality characteristics, desired specialty area, etc.) or to systemic characteristics that differentiate schools from each other (e.g., disparate grading procedures), errors associated with scores from a particular school are likely to be related (i.e., to co-vary) with other errors from scores for that school. In other words, the individual errors in prediction that result from OLS will not be completely independent from each other. This nonindependence is problematic because the resultant standard errors that we use in conducting tests of statistical significance will be erroneous. The larger this nonindependence, the greater the bias that exists when one computes these tests (Nezlek and Zyzniewski 1998).

Although a full discussion of the statistical nonindependence issue is beyond the scope of the current paper (for a thorough discussion see Kenny and Judd 1986), the problem essentially stems from the fact that, because errors of prediction are not independent from one another, the degrees of freedom one uses to compute the standard error must be adjusted in order to compute correct tests of statistical significance. As will be recalled from introductory statistics courses, degrees of freedom represent the number of *independent* pieces of information available to estimate a parameter (e.g., the variance). Because, in hierarchical data structures, data points generally are not independent, the number of degrees of freedom is less than they would be in a non-nested scenario (i.e., there is a correlation between nested observations). Failure to account for such nonindependence results in inappropriately small standard errors when data are treated at the lowest level of analysis. This is because the standard error calculations will be based on the number of lower-level observations *without* adjusting for the non-independence associated with the observations within each school. This will lead, in turn, to inflated Type I error rates because the standard errors will be too small. In sum, researchers are more likely to conclude that a statistically significant relationship exists when one does not.

The second problem associated with the hypothetical analysis conducted above is that the researcher ignores the fact that scores were drawn from different schools and, in doing so, is making the assumption that the relationship between EES and grades is constant across all schools. In many real-world applications, this assumption is unwarranted. In our example, one might imagine, for instance, that the relationship between EES and first year grades is stronger among schools that employ mid-term and final exams which relate to an intelligence factor and weaker at schools that utilize more effort-based exams (due to the fact that EES, often an intelligence-based measure, should predict intelligence-based tests scores). By overlooking the nesting of data within schools, the researcher forgoes the possibility of examining the variance across schools in the relationship between the variables of interest (Nezlek 1996). Below, we discuss methods of addressing the two issues mentioned above.

Multilevel modeling

In order to address the issues mentioned above, researchers have developed various statistical techniques that provide both appropriate statistical significance tests and allow one to examine and model the differences in effects (e.g., β_0 and β_1) that exist across the groups within which the data are nested (see Raudenbush and Bryk 2002; Klein and Kozlowski 2000). In the following pages, we describe how one can utilize these programs to derive more meaningful results. Here, we provide a hypothetical example involving students nested within institutions. Data for this example were generated using the "montecarlo" facility of Mplus version 3.13 (explored below).

We note that this example represents just one of the models made possible in multilevel regression designs. For example, multilevel models can be applied to three level models, such as when individuals are nested within groups and groups are nested within locations. Such models are commonplace, for instance, in management research as workers generally are nested within units (e.g., work team, branch, unit supervisor) that are further nested within higher-level units (e.g., organization, geographical location, industry; Klein and Kozlowski 2000). In models such as this, error components may be parceled at the appropriate level of analysis and cross-level effects may be examined at the level of interest.

Furthermore, in circumstances where individuals contribute scores on the same measure at different points in time, the researcher may utilize growth modeling, where intraindividual differences over time (a lower-level phenomenon) can be modeled and predicted by person-level variables (higher-level phenomena). In these types of models, it is possible to examine for different patterns of growth, predict initial status, and even model different growth patterns for individuals nested within different groups (a three-level model). Using growth modeling, one could examine, for example, the hypothesis that medical students who are very conscientiousness have greater grade improvements over time when they are in school environments which are more effort oriented than ability (i.e., intelligence) oriented. This would be a three-level model, where variance in school environment (level 3) would predict the relationship between conscientiousness (level 2) and grades over time (level 1).

Finally, more complex designs, such as cross-classified designs, multiple-membership models, and multilevel meta-analysis also are possible. In the first application, one correctly can model the fact that multiple individuals belong to one of a set of categories (such as being assigned to a given hospital rotation) and also may or may not belong to similar, other categories (such as the students' preferred specialty). These models allow addressing questions which relate to membership in a given category, while taking account of the individuals' concurrent membership in another, possibly orthogonal category. In the case of multiple-membership models, one explicitly accounts for the fact that data are linked to more than one, similar higher-level structure. For example, some students may be guided under the tutelage of more than one advisor, or may have gained experience at more than one hospital. Finally, in multilevel meta-analysis (see Hox and de Leeuw 2003), one can examine how between-study differences (e.g., differences in context) explain variance within studies. According to the approach, the variability within studies is considered to represent sampling variability and the between-study variance represents both sampling variance and systematic differences in the studies' results (Hox 1995, p. 69).

In each of these cases, multilevel models correctly model the effects of interest and provide for answers *at the same level of analysis at which they are asked* (Raudenbush and Bryk 2002). Importantly, this broadens the number, scope, and complexity of questions that health sciences education researchers can ask—questions that previously would have been either unanswerable or answerable, but with limited or inappropriate analyses and results. Below, we illustrate a multilevel model of GPA and EES data nested within universities. As noted above, we use the traditional HLM notation (see Raudenbush and Bryk 2002) in discussing multilevel approaches.

In this example, we present a general overview of a basic multilevel regression model. One of the primary advantages of multilevel analyses is that such techniques allow one to examine whether or not (and to what degree) there are differences across groups for the relationship in question. To address these issues, heuristically, multilevel approaches entail computing regression coefficients and intercepts in a regression analysis for each *j*th group. In other words, separate regression formulas are computed for each group within which individual scores are nested (e.g., a regression equation for person $i = 1$ within school $j = 1$, person $i = 2$ within school $j = 2$, etc.). This may be represented as

$$y_{ij} = \beta_{0j} + \beta_{1j}(\mathrm{EES}_{ij}) + r_{ij} \tag{2}$$

where β_{0j} represents the mean of y for the *i*th individual in the *j*th group (i.e., the group's *y*-intercept), β_{1j} represents the relationship between EES and first year medical school grades (y_{ij}) for the *j*th group, EES_{ij} is the EES value for person i in a group j, r_{ij} represents the amount of error associated with each *i*th individual's observation in the *j*th group, and the variance of r_{ij} is the residual variance for the model (Nezlek and Zyzniewski 1998; note that this formula is functionally similar to Eq. 1, except for the addition of component j). Thus, one is left with separate parameter estimates (i.e., $\beta_{0j,}$ β_{1j}) for each group. These separate estimates represent random parameters (i.e., random effects) in that they are considered to be a sample of estimates that are drawn from a probability distribution of possible parameter values for a population of groups. This approach is quite different from traditional OLS techniques, in which one assumes that the parameters do not vary, but instead are fixed parameters (i.e., fixed effects).

By examining the variance of the intercepts (β_{0j}) and regression coefficients (β_{1j}) across j groups (e.g., schools), the researcher determines the degree to which these parameters vary across groups. That is, one can examine the degree to which both average first-year medical school grades (i.e., the y-intercept, β_{0j}) and the relationship between y and EES (i.e., the slope, β_{1j}) varies across schools (a technique not often employed in medical education research, e.g., Guidon et al. 2004). For the y-intercept, this may be understood as the following regression formula (note that the normal OLS regression formula is embodied below, simply with differing terms and notation):

$$\beta_{0j} = \gamma_{00} + u_{0j} \tag{3}$$

which is termed an intercepts-as-outcome model (Pollack 1998) where γ_{00} is the average of the group means (i.e., the grand mean across all individuals and groups), u_{0j} is the error associated with the prediction of β_{0j}, and the variance of u_{0j} is the group-level residual [i.e., it is the between-groups variance in y_{ij} (e.g., GPA)]. Similarly, for the regression slope, we have:

$$\beta_{1j} = \gamma_{10} + u_{1j} \tag{4}$$

which is termed a slopes-as-outcomes model (Pollack 1998) where γ_{10} is the average regression coefficient across groups (i.e., the average relationship between EES and first-year grades, which is equivalently captured in β_1 from Eq. 1), u_{1j} is the error associated with the prediction of β_{1j}, and the variance of u_{1j} is the group-level residual (i.e., the amount of variance in β_{1j} across the j groups; Pollack 1998).

Although presented simplistically, the above formulae provide the necessary framework of multilevel regression required to understand HLM (Nezleck and Zyzniewski 1998). In order to determine whether or not a significant degree of variance exists in the average first-year grades across schools, one need only examine the variance of u_{0j} (this is accomplished in an “unconditional model”, or a model without any covariates). Interestingly, because this value represents the effect of the grouping of individuals on the variable of interest, examining the statistical significance of this error-term will provide the same information as one would discern with an ANOVA (Raudenbush and Bryk 2002). This is because the ANOVA computes between- and within-groups variance to provide information about the degree of variance at the between-groups level.

Here, however, the term u_{0j} represents the random effect, and is absent in traditional OLS. If significant variance across groups does exist, one further could seek to explain this between-group variance by introducing other variables as predictors in Eq. 3. For instance, a researcher might test the hypothesis that different grading systems result in higher or lower grades in seeking to account for the between-school differences in the intercept. Such an analysis would be a “cross-level main effect”, because mean differences in GPA would be related to a group-level variable. Further, such an examination would benefit from leaving the within- and between-groups residuals unconfounded, as would be done in a similar OLS model. This is because the school-level predictor would only account for variance in u_{0j} (the school-level residual), whereas a similar OLS model would only be able to calculate changes in r_i (the total residual).

In a parallel fashion, one need only examine the variance of u_{1j} to examine if there is a significant degree of variance in the relationship between EES and first-year grades across schools. Significant variance here would indicate that the magnitude of the relationship between EES and grades differs across schools. One could attempt to explain this variance, as noted above, by regressing the relationship between EES and first-year grades on a measure of the degree to which a university's exams are effort versus ability-oriented. This would be a "cross-level interaction", where a group-level variable explains variance in the relationship between the lower-level relationship. Again, such a model would benefit from being able to assess the changes in the school-level residual (i.e., u_{1j}).

To reiterate, while traditional OLS and multilevel regression models will provide similar parameter estimates for effects (as noted above), the primary benefits of HLM over OLS regression are that (1) multilevel models account for residual variance at the appropriate level of analysis and (2) they also automatically adjusting one's degrees of freedom based on the extent of nonindependence that is present (see above; Pollack 1998). The nonindependence of scores across groups relates to the fact that, because scores (and, therefore, residuals) from a given group are intercorrelated with one another, there may be substantial between-group, as well as within-group, variance across a sample. The degree of this nonindependence is dubbed the intraclass correlation (ICC; represented by ρ) and is computed rather simply (Pollack 1998).

To compute the intraclass correlation, a researcher must first compute the variance for the dependent variable among groups of analysis (e.g., schools; in the form of the variance of u_{0j}) and the variance within these groups (in the form of the variance of r_{ij}). Next, by dividing the amount of variance between groups by the amount of total variance (i.e., both between- and within-groups variance), the proportion of variance which is accounted for by knowing the group within which an individual's score is nested may be known. In other words, the ICC is the between-groups variance divided by total variance (similar to ANOVA's "eta-squared"). At the lowest level of analysis, the ICC is used to correct the degrees of freedom associated with statistical significance tests which assume independent observations (this correction is done automatically by most multilevel modeling software). Specifically, the ICC, by accounting for this nonindependence, corrects for the otherwise inflated Type I error rate, thereby providing accurate tests of statistical significance at the desired alpha level. In order to clarify the preceding discussion and to demonstrate the results one obtains in conducting these analyses, below, we provide a step-by-step example below using data generated to mimic the GPA and EES scenario we have constructed here.

An example using ESS and GPA

Using the montecarlo facility of Mplus 3.13, we generated multilevel data for 1,000 individuals nested within 20 schools (with 50 people per school). These data were generated with 28% of the total variance in y_{ij} (representing individual GPA) at the between-school level of analysis, a statistically significant predictor of y_{ij} at the individual-level of analysis in the form of EES (population $\beta = .80$), and a statistically significant predictor of the GPA-EES relationship, called EXAM, at the between-school level of analysis (population β = .80). Beginning with an unconditional model, we may begin to examine these multilevel data with

$$y_{ij} = \beta_{0j} + r_{ij} \tag{5}$$

where y_{ij} refers to the GPA score of the ith student within school j, β_{0j} represents the average GPA score for the students within the jth school, and r_{ij} is the error associated with the ith student within school j, and the variance of r_{ij} is the residual variance (i.e., the within-groups variance). This equation, at the group-level of analysis, may be represented as:

$$\beta_{0j} = \gamma_{00} + u_{0j} \tag{6}$$

where β_{0j} is the average GPA score for school j, γ_{00} is the intercept for the GPA scores (i.e., the average GPA score across all schools because this is an unconditional model), u_{0j} is the residual for school j (i.e., the distance between the intercept and the standing of school j), and the variance of u_{0j} is the between-groups variance. As a side, by including a predictor in Eq. 6, this would be an "intercept-as-outcome" model, where the average score for the jth school would be predicted by a school-level independent variable.

The first step in conducting HLM is to analyze such an unconditional model, where only the dependent variable is entered in the program. Output from this model allows a researcher to estimate the ICC, as without predictors, the residual terms will only occur as a function of between and within-groups variance in y_{ij}. We conducted this analysis (see Table 1, Model A) and examined the ICC, again by dividing the variance of u_{0j} (i.e., the between-individual variance) by the sum of u_{0j} and r_{ij} (i.e., total variance, being both between- and within-individual variance), which was shown to be quite large (ρ = .28). This value indicates that 28% of the variance in GPA is accounted for by knowing within which school the scores are nested (and will also allow for the automatic and proper adjustment of the degrees of freedom for subsequent statistical significance testing)—the expected value because the data were generated with a population level of between-groups variance at 28% (as noted above). Utilizing the χ^2, we conclude that a statistically significant amount of variance is accounted for by the higher-level student factor (i.e., u_{0j} has a statistically significant value; $\chi^2_{(19)}$ = 285.98, $p < .01$).

In order to examine the relationship between GPA and EES, we first ran a standard, OLS regression equation, predicting GPA with EES. Results indicate a statistically significant relationship between the variables (R^2 = .36, $F_{(1,998)}$ = 567.43, $p < .001$, β = .60, SE = .035 $t_{(1,999)}$ = 22.96, $p < .01$). However, this analysis assumes that all 1,000 individuals have been observed independently, and the standard error value has been computed with this assumption (as indicated by the very large t value and its associated degrees of freedom).

For illustrative purposes, a second OLS regression analysis was conducted in which we statistically controlled for the grouping variable (i.e., university membership). To do so, we conducted a hierarchical analysis regressing GPA on 19 dummy-coded university variables in the first step and on individual EES in the second step. Results from the first step demonstrate that, collectively, the 19 dummy-coded variables accounted for a significant amount of variance in GPA, R^2 = .27, $F_{(19,980)}$ = 15.05, $p < .01$. This result is consistent with the conclusion above that meaningful between-university variance exists and that the degree of this variance was 28% (we note here that the .27 R^2 value approximates this value of 28%). Results from the second step reveal that, after controlling for the dummy-coded variables, EES remained a significant predictor of GPA (β = .73, SE = .04). Specifically, EES accounted for 21% of the remaining variance in GPA (ΔR^2 = .21, partial $F_{(1,979)}$ = 522.00, $p < .01$; this value also roughly approximates that generated in the

Table 1 HLM of multilevel effects with intercept- and slopes-as-outcomes

Effect	Gamma	SE	t	df	p
Model A					
Model for school means					
Intercept, γ_{00}	.06	.20	.32	19	.76
Model B					
Model for school means					
Intercept, γ_{00}	–.04	.26	–.14	19	.89
Model for GPA–EES					
Intercept, γ_{10}	1.03	.27	3.82	19	<.01
Model C					
Model for school means					
Intercept, γ_{00}	–.04	.26	–.14	19	.89
Model for GPA–EES					
Intercept, γ_{10}	.82	.21	3.90	18	<.01
Slope, γ_{11}	.79	.22	4.56	18	<.01

	Parameter variance	χ^2	df	p
Model A				
Group mean, u_{0j}	.80	285.98	19	<.01
Level 1 effect, r_{ij}	2.85			
Model B				
Group mean, u_{0j}	1.43	1936.60	19	<.01
EES slope, u_{1j}	1.52	3912.16	19	<.01
Level 1 effect, r_{ij}	.37			
Model C				
Group mean, u_{0j}	1.43	1936.56	19	<.01
EES slope, u_{1j}	1.03	2333.26	18	<.01
Level 1 effect, r_{ij}	.37			

Note: HLM = hierarchical linear modeling; GPA = grade point average; EES = entrance exam score; EXAM = prevalence of intelligence based tests; Gamma = γ; SE = standard error; df = degrees of freedom

Monte Carlo facility). However, and problematically for the purposes of university-level hypothesis testing, the dummy-coded variables have accounted for the statistically significant between-university variance. This means that testing for the effects of university-level predictors of GPA is no longer possible (as, mentioned above, it is possible with an "intercept-as-outcome" multilevel model)—although we note here that interactions between a predictor and a series of dummy variables which code for group membership are possible (although, results from such tests of interaction would still assume independence of the lower-level units).

To contrast these results with those using HLM techniques, we ran another HLM by regressing GPA onto EES (see Table 1, Model B). Heuristically, this model is expressed in Eqs. 2–4. Even after accounting for the nonindependence of the observations, the relationship between GPA and EES is statistically significant, with an unstandardized regression coefficient γ_{00} = 1.03, SE = .27, $t_{(19)}$ = 3.82, $p < .01$. This estimate shows the underestimation of the standard error in the OLS results and the overvaluation of the related t-value. However, moving beyond the problem of nonindependence, the benefits of the HLM framework become clear in this model.

By examining the variance of the residual term of the GPA-EES coefficient (i.e., u_{1j}), we see that this coefficient appears to vary significantly across schools ($\chi^2_{(19)} = 3912.16$, $p < .01$). This important fact is overlooked in OLS and signifies that the relationship between GPA and EES (i.e., γ_{10}) is not constant. While this fact causes a misspecification in OLS (because the fixed effect β assumed homogeneity in the covariance), this parameter variance across schools is appropriately modeled in HLM and may even be considered serendipitous, as the variance in this relationship may be predicted by another, school-level predictor. For example, by expanding Eq. 4, we may test the notion that, as schools employ more intelligence-based examinations, there will be a stronger relationship between GPA and EES. This may be expressed as

$$\beta_{1j} = \gamma_{10} + \gamma_{11}(\mathrm{EXAM}_j) + u_{1j} \tag{7}$$

where terms are as above, except γ_{11} represents the relationship between the GPA-EES relationship and a variable EXAM, which is coded as having larger values when a school j has more intelligence-based exams (with values ranging from –1.75 to 2.1). This is a cross-level interaction, where a slopes-as-outcome model allows us to test the degree to which a school-level variable moderates the relationship between two student-level variables. Results from this analysis are presented in Table 1, Model C (as indicated in Eq. 7, see the parameter estimate γ_{11}). As can be seen, the relationship between EXAM the lower-level relational coefficient β_{1j} is positive and statistically significant ($\gamma_{11} = .79$, SE = .22, $t_{(18)} = 4.56$, $p < .01$). This means that as values along EXAM increase, so do values along β_{1j}, so that the relationship between GPA and EES increases as do scores along EXAM.

Interestingly, by contrasting Model B and Model C, we may compute the variance in β_{1j} accounted for by the predictor EXAM. This is accomplished with the same logic as in traditional OLS regression. By taking the unrestricted variance of u_{1j} from Model B (i.e., the variance when we modeled no predictors of β_{1j}), we may compare it to the same variance term in Model C (i.e., the variance when we predicted β_{1j} with the EXAM variable). Specifically, by subtracting the original variance term from the new variance term and dividing this value by the original variance term, we can conclude that 32% of the variance in the relationship between GPA and EES can be accounted for by knowing a school's score along the EXAM variable.

Conclusion

In this paper, we have presented a broad overview of multilevel statistical modeling and have demonstrated the advantages of these analyses over traditional OLS regression. In concluding, we wish to emphasize two important points regarding the use of multilevel analyses. First, we hasten to note that the utility of multilevel modeling greatly exceeds the limited number of circumstances presented here. That is, although our discussion of these techniques purposefully has been general in nature, their applicability and usefulness extends to various other situations. For instance, assessment of longitudinal data (e.g., when students' performance is tracked over time), and the analysis of data in which students have different instructors from one another or are exposed to different curricula or teaching procedures both potentially warrant multilevel approaches. In both of these latter cases, students are nested within a higher-level factor that requires both statistical and, more importantly, theoretical consideration.

In addition to using these techniques in other scenarios, one also can utilize somewhat more complex extensions of these analyses to address further questions. In fact, the information presented above only represents a broad overview of a set of related analyses (see Kozlowksi and Klein 2000). For example, as noted throughout, after detecting significant variance in the *y*-intercepts or slopes, one can attempt to explain this between-units variance as a function of many variables and may extend such a model to additional levels of analysis. Further, due to the flexibility of multilevel modeling packages, a variety of parameter estimators may be employed (e.g., full information maximum likelihood/restricted maximum likelihood), such as those which are robust to non-normal distributions. While the intricacies of these estimators is beyond the scope of the current work, researchers should be aware of their existence and should be informed about their application in the specific multilevel modeling program one chooses to use.

That said, we do wish to emphasize that some concerns exist regarding specific applications of multilevel models because of their estimation methods. For example, because many multilevel modeling programs use maximum likelihood estimators to derive parameter estimates, sample sizes are generally considered to be quite important for multilevel modeling (Raudenbush and Bryk 2002). While some techniques (e.g., the so-called “restricted maximum likelihood estimator” mentioned above) exist for reducing the importance of higher-level sample sizes, as sampling from many higher-order entities may be difficult, sampling is an important issue. While no single rule exists for sampling concerns, and having many higher- or lower-level units may compensate for sampling shortcomings in the other, at least 20 higher-level and 20 lower-level units are often recommended (however, this may vary depending upon the nature of one’s data; see Snijders and Bosker 1999).

The second point we wish to make is that, despite our general endorsement of multilevel analyses, we certainly do not advocate that investigators use these techniques without sufficient theoretical justification to do so. As an anonymous reviewer correctly indicates, “life is hierarchical.” Indeed, one always could locate phenomena at both higher and lower levels than that of the focus level. That all phenomena and relationships are, in some general sense, hierarchical, does not imply that one always needs to, or should, assess variables operating at other levels in examining the substantive research question. Decisions regarding research design, the level at which one construes a variable, and which other variables (at the same or at other levels of analysis) to assess, obviously should be driven by the question at hand, not by the availability of multilevel analyses. Similarly, having access to multilevel data does not necessarily suggest that one needs to assess hierarchical models. Too often, researchers blindly begin using new statistical techniques for any and all of their research endeavors. However, no analytical tool can replace appropriate research design, and that design must follow from proper formulation of the question or theory one seeks to address.

In our view, these tools are most valuable when the data are hierarchically nested within meaningful units, and when the neglect of such nesting potentially would yield incorrect results and, in turn, inappropriate conclusions. These situations are by no means uncommon in medical education. As noted in this manuscript, prospective students, for instance, are nested within undergraduate majors, schools, and regions, introducing nonindepdence when using GPA to predict medical school performance in traditional OLS regression. Similarly, current students are nested within instructors, years, and specialties, also potentially necessitating multilevel modeling. These are the precise situations in which multilevel analyses potentially are of the most practical use. However, to the best of our knowledge, medical educators rarely utilize such techniques.

References

Bliese, P. D. (2002). Multilevel random coefficient modeling in organizational research: Examples using SAS, S-PLUS. In F. Drasgow & N. Schmitt (Eds.), *Modeling in organizational research: Measuring and analyzing behavior in organizations* (pp. 401–445). San Francisco, CA: Jossey-Bass, Inc.

Guiton, G., Hodgson, C. S., Delandshere, G., & Wilkerson, L. (2004). Communication skills in standardized-patient assessment of final-year medical students: A psychometric study. *Advances in Health Science Education: Theory and Practice, 9*, 179–187.

Hofmann, D. A. (1997). An overview of the logic and rationale of hierarchical linear models. *Journal of Management, 23*, 723–744.

Hox, J. J. (1995). *Applied multilevel analysis*. Amsterdam: TT-Publikaties.

Hox, J. J., & de Leeuw, E. D. (2003). Multilevel models for meta-analysis. In S. P. Reise & N. Duan (Eds.), *Multilevel modeling: Methodological advances, issues, and applications* (pp. 90–111). Mahwah, NJ: Erlbaum.

Kenny, D. A., & Judd, C. M. (1986). Consequences of violating the independence assumption in analysis of variance. *Psychological Bulletin, 99*, 422–431.

Kozlowski, S. W. J., & Klein, K. J. (Eds.) (2000). *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions*. San Francisco: Jossey-Bass.

Nezlek, J. B. (1996). Multilevel random coefficient analysis of even- and interval-contingent data in social and personality psychology research. *Personality and Social Psychology Bulletin, 7*, 771–785.

Nezlek, J. B., & Zyzniewski, L. E. (1998). Using hierarchical linear modeling to analyze grouped data. *Group Dynamics, 2*, 313–320.

Pollack, B. N. (1998). Hierarchical linear modeling and the "unit of analysis" problem: A solution for analyzing responses of intact group members. *Group Dynamics, 2*, 299–312.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Newbury Park, CA: Sage.

Snijders, T. A. B., & Bosker, R. J. (1999). Standard errors and sample sizes for two-level research. *Journal of Educational Statistics, 18*, 237–259.