Using Multilevel Regression Mixture Models to Identify Level-1 Heterogeneity in Level-2 Effects

M. Lee Van Horn – University of South Carolina

Yuling Feng – University of South Carolina

Minjung Kim – University of South Carolina

Andrea Lamont – University of South Carolina

Daniel Feaster – University of Miami

Thomas Jaki – Lancaster University

This research was supported by grant number R01HD054736, M. Lee Van Horn (PI), funded by the National Institute of Child Health and Human Development. Dr. Van Horn is the senior and
corresponding author for this paper, questions or comments should be addressed to vanhorn@sc.edu.

Abstract

This paper proposes a novel exploratory approach for assessing how the effects of level-2 predictors differ across level-1 units. Multilevel regression mixture models are used to identify latent classes at level-1 that differ in the effect of one or more level-2 predictors. Monte Carlo simulations are used to demonstrate the approach with different sample sizes and to demonstrate the consequences of constraining 1 of the random effects to zero. An application of the method to evaluate heterogeneity in the effects of classroom practices on students is used to show the types of research questions which can be answered with this method and the issues faced when estimating multilevel regression mixtures.

A common research objective is to assess heterogeneity in the effects of a predictor on an outcome. Take, for example, a study looking at the effects of teaching style on student achievement that finds no average effects on student outcomes. A logical next question is to examine whether the effects of teaching differs across students (Van Horn & Ramey, 2003). The standard approach would be to test cross-level interactions between student-level predictors and the classroom-level variable teaching style. This yields an understanding of the impact of specified variables on specific students. However, this is not the same thing as a global assessment of heterogeneity in the effects of teaching style. An alternative approach would be to use a regression mixture (also known as mixture regression or latent class regression) model to explore for latent classes of students who respond differently to teaching style. Latent classes which are different in the effect of a predictor can be identified without a prori identification of moderator variables, which is a much broader question than the typical moderation analyses that assesses whether the effects of a predictor vary as a function of a specific moderator. However, currently available regression mixture models are only able to assess heterogeneity in the effects of a level-1 predictor, thus they cannot be used to assess level-1 variability (between students) in the effects of a level-2 predictor (teaching style).

Regression mixture models are an established method in the area of marketing research and an increasingly popular approach in the social sciences for examining heterogeneous effects (Desarbo, Jedidi, & Sinha, 2001; Van Horn et al., 2009; Wedel & DeSarbo, 1995). Multilevel extensions of regression mixtures allow for the identification of latent classes at level-1, which differ in the effects of a level-1predictor on a level-1 outcome (B. O. Muthén & Asparouhov, 2009; Vermunt, 2010; Vermunt & Van Dijk, 2001), for example, the effects of student level poverty on student performance. This paper extends the multilevel regression mixture model to allow for level-1 latent classes that differ in the effects of a level-2 predictor such as teaching style on level-1 outcomes. This allows us to answer a new type of research question which cannot be assessed with other mixture or multilevel approaches: how do the effects of level-2 predictors differ across level-1 units?

Consider a continuous outcome, y, and let y_{ij} be the observation for individual i in cluster j. Within each cluster (which defines level-2 in the model), the regression mixture contains K latent classes. The latent class variable is denoted as C with K categories labeled $c = 1,2,\ldots,K$. Each latent class is defined by its unique effects of the cluster-level (level-2) covariate on the outcome. The level-1 model can be written:

$$y_{ikj|C_{ij}=k_j} = \beta_{0kj} + r_{ikj}, \tag{1}$$

where the residual $r_{ikj} \sim N(0, \sigma_k^2)$. Note that unlike previous multilevel regression mixtures (B. O. Muthén & Asparouhov, 2009) this equation contains only a class-specific intercept and random error; there need be no individual-level (level-1) covariates in (1).

Differences amongst individuals in level-2 predictors are modeled as class specific regression weights:

$$\beta_{0kj} = \gamma_{0k0} + \gamma_{0k1} w_j + u_{0kj}, \tag{2}$$

where the intercept of each mixture class within each level-2 cluster is modeled as the function of the class-specific intercept (γ_{0k0}) and the class-specific effects of a cluster-level covariate (γ_{0k1}). We use the parametric parameterization of the model in which the between-level residual variance $u_{0kj} \sim N(0, \tau_k)$, note that it is possible to use a non-parametric model to represent any of the random variances (Vermunt, 2003). There are K 'average' effects of each cluster-level covariate (one for each latent class); this is what allows for heterogeneity in level-2 effects and what distinguishes this approach from previous models. Differences across classes in the effects of a cluster-level variable on individuals within the cluster (i.e., differences represented by the *K* regression weights; γ_{0k1}) are indicative of level-1 heterogeneity in the effects of a level-2 variable. Additionally, there are *K* random error terms u_{0kj} which allow for differences in class specific intercepts between clusters. These errors are assumed to be normal with mean zero and variance covariance matrix τ_{0k}.

The probability that an individual is in a particular latent class is modeled by a two-level multinomial logistic regression function:

$$P\left(C_{ij} = k_j c\right) = \frac{\exp(\alpha_{kj})}{\sum_{s=1}^{K} \exp(\alpha_{sj})} \tag{3}$$

where for the last class *K*, $\alpha_{Kj} = 0$, for identification . The model presented is an intercept only model, which we recommend in practice for latent class enumeration because misspecification of the predictors of latent class membership may result in bias in latent class enumeration and parameter estimates. Additional predictors will typically be added in later analysis steps with particular attention paid to changes in other model parameters. In this case, the intercept represents the log-odds that an individual in cluster *j* is in class *c* versus the reference class (typically defined as class *K*). Across level-2 clusters, the intercept is a function of the overall intercept and the cluster-level random variation (cluster-level predictors of latent class membership would be included here):

$$\alpha_{kj} = \gamma_{1k0} + u_{1kj} \tag{4}$$

The residuals, u_{1kj}, represent differences between clusters in the probability of being in class k versus the reference class, they allow clusters to differ in the percentage of respondents in each class. In this application cluster level residuals are assumed to follow a multivariate normal distribution and their variances and covariances are included in τ matrix. Because this matrix is quite difficult to estimate, restricted forms are often considered, such as a diagonal matrix, constraining certain variances or covariances to zero, or placing equality constraints on particular parameters. The unconstrained variance-covariance τ matrix for a 2-class model can be written as:

$$\text{var}\begin{bmatrix} u_{01j} \\ u_{02j} \\ u_{11j} \end{bmatrix} \sim N(0, \begin{bmatrix} \tau_{00} & \tau_{01} & \tau_{02} \\ \tau_{01} & \tau_{11} & \tau_{12} \\ \tau_{02} & \tau_{12} & \tau_{22} \end{bmatrix}, \quad (5)$$

where τ_{00} and τ_{11} refers to the intercept variance of class-1 and class-2, respectively, τ_{22} refers to the variances between clusters in the probability of being in class-1 versus the class-2 (the reference class), τ_{01} refers the covariance between the intercept variance of two classes, and τ_{02} and τ_{12} represents the covariance between the variance of the intercept and the class proportion for each class. The logic for class specific variance estimates is that if the effect size for a predictor is larger in one class then it is reasonable to expect the residual variance to be lower in that class.

An interesting feature of this model is that although the latent class variable operates primarily at level-2, it works by differentiating individuals at level-1 and can be used to obtain predictions of latent class membership for each individual. Latent classes are defined by differences between classes in the effects of a level-2 variable (W) on the outcome (Y) as well as differences between classes in the conditional mean of the outcome. Substantively, these are the important parts of the model. They allow for different level-2 effects across classes as well as different means for the outcome. The model also includes several random effects: σ^2_k is the class specific variance of r which allows for differences between classes in the residual variance of the outcome; τ_{00} is the variance of u_0 which allows for class specific differences across clusters in level-1 intercepts. The intraclass correlation coefficient (ICC) is a common assessment of the extent to which an outcome differs between clusters. In this case the ICC for each intercept can be estimated separately for each class as: $\tau_{kk}/(\sigma^2_k + \tau_{kk})$, thus this model allows the extent of clustering to vary across latent classes. Additionally, τ_{22} is the variance of u_1 which allows each cluster to differ in the proportion of respondents in each class; omitting this term would result in the class probabilities (the distribution of respondents across the different classes) being identical across all clusters. ICCs for the latent class equation predicting the probability of class membership can also be calculated. The level-1 variance of a logistic outcome is the variance for the logistic distribution ($\pi^2/3$). Because it is a constant which does not depend on the data, it is not estimated. The formula is then: $\rho = \frac{\tau_{22}}{\tau_{22} + \pi^2/3}$ where π is the constant 3.142 (Snijders & Bosker, 1999).

Because the proposed model has not been previously tested, the current paper uses Monte Carlo simulations and applied analyses to demonstrate the use of these models and examine model performance. Our first aim uses simulations to demonstrate that multilevel regression mixture models can successfully find level-1 heterogeneity in level-2 effects at sample sizes that are realistic for many multilevel studies. We examine latent class enumeration, the ability to determine that there are multiple classes of individuals using penalized information criteria, as well as bias in parameter estimates. We hypothesize that model results will be less stable with smaller samples, with extreme parameter estimates for a larger number of simulated datasets than expected given the theoretical sampling distribution of the parameters. We expect that multilevel regression mixtures will require large samples in terms of both numbers of clusters and number of observations per cluster to achieve stable results. Our second simulation aim is to evaluate the effects of simplifying the random components of the multilevel regression mixture model, specifically focusing on model performance when random effects for the latent class means are included or excluded. Based on previous work with multilevel mixtures, we hypothesize that constraining the level-2 variance of the latent class intercepts to zero will not seriously impact model results, given that these variances are not large (Van Horn et al., 2008). This is important because, if confirmed, it provides guidance for the model building process.

The final aim of this paper is to demonstrate the use of multilevel regression mixtures for finding heterogeneity between students in the effects of classroom practices on achievement.

Simulation Study: Methods

Data Generation. The first aims of this study are addressed using Monte Carlo simulations (Mooney, 1997). Data were generated from two populations (latent classes) within each cluster. Slopes and intercepts in (3) are chosen as

$$\gamma_{0k0} = \begin{cases} 0, & k = 1 \\ 0.5, & k = 2 \end{cases}$$

$$\gamma_{0k1} = \begin{cases} 0.2, & k = 1 \\ 0.7, & k = 2 \end{cases}$$

Then,

$$\beta_{01j} = 0.2 * w_j + u_{01j}$$

$$\beta_{02j} = 0.5 + 0.7 * w_j + u_{02j}$$

where, $w_j \sim N(0,1)$, $u_{01j} \sim N(0, \sqrt{0.096})$, $u_{02j} \sim N(0, \sqrt{0.051})$, the variance was chosen to maintain an ICC for the intercept of .10 in each class. The covariance between u_{01j} and u_{02j} is set to be zero, and the residual errors are assumed independent of u_{1kj} in (4). Thus the variance covariance matrix for random error terms, τ, is diagonal.

Therefore,

$$y_{ij|C_{ij}=1_j} = 0.2 * w_j + u_{01j} + r_{i1j}$$

$$y_{ij|C_{ij}=2_j} = 0.5 + 0.7 * w_j + u_{02j} + r_{i2j}$$

where, $r_{i1j} \sim N(0, \sqrt{0.864})$, $r_{i2j} \sim N(0, \sqrt{0.459})$. Values for the residual variances were chosen so that the total variance of y in each of the two populations (latent classes) would be equal to 1, thus the regression weights are interpreted as correlations and difference in intercepts between classes is scaled to be Cohen's D. The probability of being in class 1 and class 2 both are equal to .50 in the population resulting in the true value for γ_{110} from equation 4 being zero. Analyses were run with the value of α_{1j} for each cluster *j* drawn from a normal distribution with mean zero and variance of 0.3656, resulting in an ICC of 0.1.

The outcome variable *Y* was generated for either 50 or 100 observations per cluster and for 50, 100, or 200 clusters. Therefore, there are 3(number of clusters)*2(number of people per cluster) = 6 simulation conditions. 500 data sets were generated for each simulation condition using R (R Development Core Team, 2010).

Model estimation. The two level mixture model is estimated in Mplus (Version 6.1, L. K. Muthén & Muthén, 2010) using the maximum likelihood estimator with robust standard errors (MLR). For each simulation results were estimated with 48 different starting values with 24 starting values completed till convergence. Sample code for estimating this model is included in the Appendix. An identifiability constraint (the larger regression weight was always in class 2) was used to sort results into class 1 and class 2 so that they can be compared across simulations. Penalized information criteria, in this case the Bayesian information criterion (BIC; Schwarz, 1978) and sample-size adjusted BIC (Sclove, 1987) were used to decide the optimal number of classes. Sample size is included in the calculation of both criteria, for multilevel models an issue is whether the level-1 or level-2 sample sizes are most appropriate. (Lukociene, Varriale, & Vermunt, 2010) found that level-2 sample size is more appropriate when the latent classes are at level-2 with results being more ambiguous when the latent classes are at level-1. In this case the classes are at level-1 and so we used the level-1 sample size; however, we checked the results of several simulations using the level-2 sample size and found no substantive changes.

Simulation Study: Results

Latent Class Enumeration. Initial simulations examined class enumeration when the probability of class membership was allowed to vary randomly across clusters. The convergence rate for the 3-class model was about 50%. We interpret convergence problems when the number of classes being estimated is too large as an indication that the 3-class model is not supported by the data. Results in Table 1 are reported for the 1-class and 2-class models. The 2-class model is selected over the 1-class model in nearly all of the simulations unless there are 50 clusters with 50 respondents per cluster where it is still selected in 90% of the simulations. The estimated class probabilities across simulations is fairly wide for the smallest sample size although no very small classes (which may indicate selecting the 1-class model) were found.

Next class enumeration was assessed for the analysis model which was misspecified by fixing the class probabilities to be equal across clusters. Both BIC and adjusted BIC choose the 2-class over the 1-class and 3-class models for almost all replications of data simulated. This constraint resulted in no problems in estimating the 3-class models and now the worst case scenario resulted in the 2-class model being chosen over the 1-class and 3-class models in over 95% of the simulations. When the models are misspecified by fixing the probability of latent class membership across clusters these models do a good job of finding the correct number of differential effects across all sample sizes examined.

Identification of Differential Effects. Given that two classes were found, analyses turned to whether those two classes represent the true differential effects. Analyses were run for each sample size with both random and fixed probabilities of class membership. Results for simulations with a random variance for class membership (Table 2) show that across all conditions there is minimal bias in parameter estimates. While average parameter estimates look good, sampling distributions become quite large at the smaller sample sizes (note the three-fold increase in average standard errors). Of more concern is that the empirical standard errors appear to be underestimating the true sampling variation and that this effect appears to increase with small sample sizes. This is seen in Table 2 as the difference between the average of the empirical standard errors and the standard deviation of the parameters across all simulations and by the degree to which coverage estimates (the proportion of simulations for which the 95% confidence interval contained the true value) are below .95. The parameters with the most problems are the level-2 residuals for the two classes, E1var and E2var, and the probability of class membership. The variance of the probability of class member ship across clusters is especially hard to estimate with coverage under 0.6 for all sample sizes. We believe that there are two causes for the problems seen with the empirical standard errors. First, with small sample sizes the regression mixture results appear to be less stable leading to more extreme solutions than would be expected given the sampling distribution. This can be seen by the fact that coverage rates decrease with smaller samples and by the increasingly large outliers seen with smaller sample sizes. Second, Mplus confidence intervals for variances are estimated from a symmetric t-distribution which only approximates the true sampling distribution of a variance. To test this, we ran one simulation condition in which the variances were constrained to be equal to their true values and used a likelihood ratio test to compare models with the variances freely estimated to those in which they were constrained to their population values. This test found significant differences just over 5% of the time indicating that the Wald confidence intervals for variance components of these models should be seen as only rough approximations. Finally, results for the models in which the random effect for the class probabilities was constrained to zero were quite similar to the results reported here. There was no bias seen in any of the model parameters that were estimated, and there was less variability across simulations in model parameters and outliers were less extreme although coverage rates were still less than .95.

Simulation Study: Discussion

The most important objective of these simulations was to demonstrate that multilevel regression mixtures are capable of finding level-1 heterogeneity in level-2 effects with realistic sample sizes. Although previous work has shown that the regression mixture can be applied to clustered data, these models only assessed heterogenetiy in level-1 predictors. This is the first study to test whether these models can assess level-1 heterogeneity of level-2 effects. Results of these simulations were very encouraging across a range of sample sizes the BIC and aBIC were reliably able to find the true number of latent classes and the level-2 effects in those classes were well estimated. Additionally, the simulations in which the between cluster variance of the latent class mean to was fixed to zero provided some useful guidance for the model building process. Results showed that this constraint did not lead to bias in other model parameters and resulted in somewhat more stable estimates. This suggests that a reasonable first step in estimating multilevel regression mixtures is to simplify the model by excluding the random variability in class probabilities. It is prudent to ultimately verify that this restriction is reasonable in the final model, but this simplification can facilitate the model building process as parameter estimates are more stable the models run up to 10 times faster without this parameter included.

These methods work with sample sizes which we found to be surprisingly low. Across simulations there are signs of problems starting to arise with a sample of 50 clusters and 50 individuals per cluster for a total sample size of 2500. This was especially evident in the number of extreme outlying estimates found. However, on average the models still appear viable with this sample size. Given some evidence that single level regression mixture models require large samples (Park, Lord, & Hart, 2010) and that level-2 effects in multilevel models are typically limited by the number of clusters available (Raudenbush & Bryk, 2002), we found it encouraging that it appears to be possible to estimate these models with as few as 50 clusters.

While these results are encouraging, they also suggest areas of further investigation. First, empirical confidence intervals are underestimated and there is evidence for extreme parameter estimates. While rare, this shows that even under ideal conditions confidence intervals should be taken with some caution. Second, the simple model tested here included 5 random effects and 6 fixed effects with only one misspecification tested (the effect of constraining the random effect for the class mean to zero). We do not know how the models respond to other misspecifications, particularly important would seem to be the assumption that all error terms follow a multivariate normal distribution. While these initial results show promise, further experience using these models in applied analyses and additional simulations are needed to help better understand the conditions under which multilevel regression mixtures work.

Applied Study: Heterogeneity in the Effects of Developmentally Appropriate Practices

In the 1980's the National Association for the Education of Young Children, published a set of guidelines promoting the use of Developmentally Appropriate Practices (DAP) (Bredekamp, 1987; Bredekamp & Copple, 1997; National Association for the Education of Young Children, 1986). DAP guidelines emphasized the use of open classrooms where children are actively engaged in learning; move between different learning centers; have choice in what activities they engage in; learn in the context of social groups; and where curriculum is integrated across multiple areas. However, decades of research in the area have produced ambiguous results with some studies finding positive effects of DAP, others finding negative effects, and many others finding no effects (for a review see Van Horn, Karlin, Ramey, Aldridge,

& Snyder, 2005). The two largest studies found no average effects of DAP on achievement (Van Horn & Ramey, 2003) or psycho-social outcomes (Van Horn, Karlin, Ramey, & Wetter, 2012) in 1st through 3rd grades.

Existing research has also found no consistent evidence for interactions between level-1 predictors such as child sex, ethnicity, and poverty and the level-2 DAP measures, however, this may be because heterogeneity in DAP is due to more complex, possibly latent, processes which cannot be easily modeled using traditional interactions (for a review see Van Horn et al., 2005). Regression mixtures which can assess heterogeneity beyond interactions with observed variables are a natural choice. However, because there are multiple students in a classroom, regression mixtures have not previously been a viable method for assessing heterogeneity in the effects of DAP. In this study, we illustrate the use of cross-level regression mixtures to explore for previously undetected heterogeneity in the effects of DAP in one cohort of students (just finishing first or second grades) on reading achievement. Based on the ambiguity of previous research we hypothesize that there will be no total effect of DAP across all students but that groups of students for which there is a positive effect of DAP as well as groups for which there is a negative (or a large group with no effect) will be identified.

Applied Study: Methods

Data for this illustration come from 879 classrooms across the US which were part of the National Head Start Public School Early Childhood Transition Demonstration Project in the 1995 year (for a full description of the study see C. T. Ramey, Ramey, & Phillips, 1996; S. L. Ramey et al., 2001). DAP was measured using A Developmentally Appropriate Practices Template (ADAPT; Gottlieb, 1995) rated by trained observers in the 1994-1995 school year. ADAPT has three factors including: integrated curriculum, social/emotional emphasis, and child-centered approach. Reading achievement was assessed for students from the same classrooms, 3247 of whom were available for testing in both spring of 1994 and spring of 1995. Single imputation was used for any data missing within a given year. Achievement was assessed using Woodcock Johnson broad reading scores, administered to students individually by trained evaluators at the end of 1994-1995 school year, at which point students in the first cohort were completing second grade and those in the second cohort were completing first grade.

To investigate the differential effects of the three domains of DAP on students' achievement in reading, we used a series of multilevel regression mixture models which differed in the number of classes (i.e., one through three) and equality constraints for model parameters (i.e., variance of class means and regression coefficients). The multilevel regression mixture model used in this analysis is:

Level-1 (within-cluster):

$$\text{Reading}_{ij|C_{ij}=k_j} = \beta_{0kj} + \beta_{1j}\text{Baseline}_{ij} + r_{ikj} \qquad (5),$$

where $Baseline_{ij}$ represents student's prior reading achievement the effect of which is assumed to be class invariant, r_{ikj} indicates the residual which is assumed to be normally distributed with class specific variance τ_{0k}. The probability of an individual being in a particular latent class is modeled using equation (3).

Level-2 (between-cluster):

$$\beta_{0kj} = \gamma_{0k0} + \gamma_{0k1}\text{Integrated curriculum}_j + \gamma_{0k2}\text{Social/Emotional emphasis}_j$$
$$+ \gamma_{0k3}\text{Child-centered approach}_j + u_{0kj}, \qquad \text{and}$$
$$\beta_{1j} = \gamma_{10} \qquad (6),$$

where γ_{0k0} represents the average reading achievement score for each cluster at the mean of DAP (given that the other predictors are centered). The regression coefficients of the three DAP measures are γ_{0k1} to γ_{0k3}, represent the effects of each DAP component within latent class k, holding other predictors constant, and u_{0kj} corresponds to the class-specific between-level residual variances for each intercept; the correlations of the residual variances between latent classes were freely estimated given no prior assumptions on those parameters in this application. γ_{10} represents the average score of the baseline reading achievement for all classes. Latent class membership is modeled with equation 4 where γ_{1k0} denotes the average log-odds that an individual in cluster *j* is in class *k* versus the reference class and u_{1kj} represents differences between clusters in the probability of being in class k versus the reference class. As recommended in the above simulations, we started the estimation process by fixing u_{1kj} so that all clusters have equal class proportions throughout the data. In this example, the 3-class model with no random effect for class probabilities took 30 minutes to estimate, while the model with this random effect took two days. Subsequently, we included the random effects of the latent class means for the 2-class and 3-class models and, again, compared those with the traditional single-class regression model to see whether the inclusion of random effects of class means affected the results of class enumeration and other parameter estimates.

After selecting the best fitting model we tested the statistical significance of individual predictors by individually constraining each parameter to be the same across classes. Given three predictors of DAP, we compared a total of four different models: (1) freely estimating all three predictors differed by classes, (2) constraining a path of integrated curriculum, (3) constraining a path of social/emotional emphasis, and (4) constraining a path of child-centered approach. Given that the simpler model was nested within the more complex model, we conducted a likelihood ratio test (LRT) employing Satorra-Bentler (SB) scaled difference test. We used SB LRT because a difference between the two scaled goodness-of-fit statistics values does not follow a chi-square distribution (Satorra, 2000; Satorra & Bentler, 2001).

Applied Study: Results

Analyses begin by finding the optimal number of latent classes defined by the relationship of the three DAP subscales with reading and the conditional means (intercepts) and residual variances for reading. We compared the traditional (1-class) multilevel regression model to the 2-class and 3-class with fixed probabilities of class membership. Penalized information criteria (both the BIC and aBIC) selected the 2-class (see Table 6) over the 1-class and 3-class models, in the 2-class solution the classes were split 48% to 52% and the entropy was .14. We next added the random probability of class membership to the selected 2-class model and found improved fit. An examination of the variance for the class means shows very large differences between classrooms in the probability that students are in each of the two classes (ICC = .68). This may be a function of the relatively small number of students per classroom (3.7 on average). This suggests also estimating the 3-class model with random class means to verify the selection of the two classes. The results showed that the 2-class model was again selected over the 1-class and 3-class models using penalized information criteria. In addition, one class of the 3-class solution one contained 0.47% of the students, indicating that third class added little.

The next step is to examine what distinguishes the different classes. Table 3 presents the parameter estimates for each model. The 1-class solution replicates previous research looking for effects of DAP in this and other datasets using traditional multilevel models, there is no evidence for the effects of any of the three DAP subscales on reading achievement. In contrast, the 2-class mixture model shows large differences between classes in the effects of two of the three DAP subscales (integrated curriculum and child-centered approach). Thus, for model simplicity and efficiency, we constrained the regression coefficient of social/emotional emphasis to be the same between the two classes and tested whether the restricted model still represents the data appropriately. Because the constrained model (with *p* free parameters) was nested within the relaxed model (with *p+1* free parameters), we were able to use the likelihood ratio test. The results showed that the relationship between social/emotional emphasis and student's reading achievement did not differ across classes (SB test statistic of 1.89, *df*=1). We also assessed heterogeneity in the other two predictors between the classes. The results showed a different in the relationship of child-centered approach and reading achievement (SB chi-square = 3.87, *df*=1) but no difference across classes in the impact of integrated curriculum (SB chi-square = 1.19, *df*=1). This result is interesting given that the class specific parameter estimates and standard errors in Table 3 show large differences between classes in the child-centered approach and integrated curriculum. The above simulations suggest that the standard errors for regression weights in the multilevel regression mixtures are at least close to the nominal values; it may be that the SB test is overly conservative. We report results of the 2-class model with social/emotional emphasis constrained to have no differences between classes. Of note are the fairly large changes in model parameters and especially in standard errors as a result of including random class means. While the substantive results don't change, standard errors are substantially reduced by estimating the random class means and entropy is a bit higher.

The last step in this demonstration is to interpret the results of the best fitting model, shown in Table 3 and Figure 1. Overall, social/emotional emphasis had no impact on student's reading achievement. The effects of having an integrated curriculum and a child-centered approach to learning tell an interesting story. Holding the other DAP constructs constant, for class-2 the effects of integrated curriculum are negative and moderately strong and in class-1 the effects are not different from zero. When combined, the two effects are reduced in the traditional multilevel model such that there is a small and not significant negative effect of integrated curriculum. The effects of using a child-centered approach were approximately equally strong and in the opposite direction across the two classes with children in class-1 benefiting from these practices and children in class-2 showing negative effects. The effect size for the child-centered approach is quite large if considered across the two classes where a 1 unit increase in child-centered approach is expected to move the two classes 7.5 units apart in reading achievement. The different direction of these two effects cancel each other out when averaged in the traditional (1-class) model. Additionally, we found a strong negative correlation of the intercepts between the two latent classes (r = -.84). This makes sense given the regression weights differing in sign between the classes, when students in class-1 do better, those in class-2 do worse and vice versa.

Applied Study: Discussion

This study proved to be an interesting application of multilevel regression mixture models for finding cross-level differential effects. Unlike previous research using traditional multilevel models where the effects of DAP have typically been zero, we found evidence for two groups of students who respond differently to different aspects of DAP. The findings were more complicated than hypothesized with no evidence for any group of students who universally benefited from DAP, and one class that showed negative effects. We suspect that results like this (which raise more questions than they provide answers) will likely to common in the applied use of these models. These are exploratory methods which are used to find level-1 heterogeneity which was not previously assessed and about which there is little theory. We see this as the start of a research process which focuses on assessing individual differences in level-2 effects and ultimately explaining these differences. Additionally, in this applied example there were strong differences between classrooms in the proportion of students in each latent class, differences which were much greater than those used in the simulations for this parameter. This emphasizes the importance of testing simplifying assumptions. There are also important implications if there are truly large differences between classrooms in the effects of teaching style on students.

Conclusions

This study proposed a new exploratory method for finding level-1 heterogeneity in level-2 effects. For those familiar with multilevel models looking at heterogeneity at level-2 in level-1 effects, this approach turns the traditional approach on its head. For those familiar with regression mixtures which examine heterogeneity in effects in a single level analyses, this is an important extension of the methods proposed by Vermunt and colleagues (Vermunt & Van Dijk, 2001). This study demonstrated that this method works under ideal conditions with sample sizes as low as 2500, suggested an approach for implementing the method involving constraining one of the random effects, and showed the use of the method to a dataset where differential effects were expected but not previously found. In both the simulations and the applied data we found that this method worked better than we had initially expected, requiring smaller samples and being less prone to misspecification than anticipated. However, multilevel regression mixtures remain complicated models which typically often involve estimating many more parameters than variables. Of particular concern in multilevel regression mixtures is the number of variance parameters being estimated. While much is now known about the effects of model assumptions in single level regression mixtures, the effect of model assumptions on parameter estimation in multilevel regression mixtures is still an open question. Answers to this and other questions will determine the ultimate utility of the method.

References

Bredekamp, S. (1987). *Developmentally appropriate practice in early childhood programs serving children from birth through age 8: Expanded edition*. Washington, DC: NAEYC.

Bredekamp, S., & Copple, C. (Eds.). (1997). *Developmentally appropriate practice in early childhood programs* (Revised ed.). Washington, D. C.: National Association for the Education of Young Children.

Desarbo, W. S., Jedidi, K., & Sinha, I. (2001). Customer value analysis in a heterogeneous market. *Strategic Management Journal, 22*(9), 846. doi: 10.1002/smj.191

Gottlieb, M. (1995). A developmentally appropriate practice template. Des Plaines, IL: Illinois Resource Center.

Lukociene, O., Varriale, R., & Vermunt, J. (2010). The simultaneous decision(s) about the number of lower- and higher- level classes in multilevel latent class analysis. *Sociological Methodology, 40*, 247-283.

Mooney, C. Z. (1997). *Monte Carlo simulation*. Thousand Oaks, Sage: Sage.

Muthén, B. O., & Asparouhov, T. (2009). Multilevel regression mixture analysis. *Journal of the Royal Statistical Society, Series A, 172*, 639-657.

Muthén, L. K., & Muthén, B. O. (2010). Mplus (Version 6). Los Angeles: Muthén & Muthén.

National Association for the Education of Young Children. (1986). NAEYC position statement on developmentally appropriate practice in early childhood programs serving children from birth to age 8. *Young Children, 41*(6), 3-19.

Park, B. J., Lord, D., & Hart, J. (2010). Bias Properties of Bayesian Statistics in Finite Mixture of Negative Regression Models for Crash Data Analysis. *Accident Analysis & Prevention, 42*, 741-749.

R Development Core Team. (2010). R: A language and environment for statistical computing (Version 2.10). Vienna, Austria: R Foundation for Statistical Computing.

Ramey, C. T., Ramey, S. L., & Phillips, M. M. (1996). Head Start children's entry into public school: An interim report on the National Head Start-Public School Early Childhood Transition Demonstration Study. Washington, DC: Report prepared for the U.S. Department of Health and Human Services, Head Start Bureau.

Ramey, S. L., Ramey, C. T., Phillips, M. M., Lanzi, R. G., Brezausek, C., Katholi, C. R., & Snyder, S. W. (2001). Head Start children's entry into public school: A report on the National Head Start/Public School Early Childhood Transition Demonstration Study. Washington, DC: Department of Health and Human Services, Administration on Children, Youth, and Families.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: applications and data analysis methods* (Second ed.). Thousand Oaks, CA: Sage Publications.

Satorra, A. (2000). Scaled and Adjusted Restricted Tests in Multi-Sample Analysis of Moment Structures. In R. D. H. Heijmans, D. S. G. Pollock, & A. Satorra (Eds.), *Innovations in Multivariate Statistical Analysis* (Vol. 36, pp. 233-247): Springer US.

Satorra, A., & Bentler, P. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika, 66*(4), 507-514. doi: 10.1007/BF02296192

Schwarz, G. (1978). Estimating the Dimension of a Model. *Ann. Statist., 6*(2), 461-464.

Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika, 52*, 333-343.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Thousand Oaks, CA: Sage Publications.

Van Horn, M. L., Fagan, A. A., Jaki, T., Brown, E. C., Hawkins, J. D., Arthur, M. W., . . . Catalano, R. F. (2008). Using multilevel mixtures to evaluate intervention effects in group randomized trials. *Multivariate Behavioral Research, 43*(2), 289-326. doi: 10.1080/00273170802034893

Van Horn, M. L., Jaki, T., Masyn, K., Ramey, S. L., Smith, J., A., & Antaramian, S. (2009). Assessing differential effects: Applying regression mixture models to identify variations in the influence of family resources on academic achievement. *Developmental Psychology, 45*(5), 1298-1313.

Van Horn, M. L., Karlin, E. O., Ramey, S. L., Aldridge, J., & Snyder, S. W. (2005). Effects of Developmentally Appropriate Practices on Children's Development: A Review of Research and Discussion of Methodological and Analytic Issues. *The Elementary School Journal, 105*, 325-352.

Van Horn, M. L., Karlin, E. O., Ramey, S. L., & Wetter, E. (2012). Effects of Developmentally Appropriate Practices on social skills and problem behaviors in first through third grades. *Journal of Research in Childhood Education, 26*, 18-39.

Van Horn, M. L., & Ramey, S. L. (2003). The effects of Developmentally Appropriate Practices on academic outcomes among former Head Start students and classmates from first through third grades. *American Educational Research Journal, 40*, 961-990.

Vermunt, J. K. (2003). Multilevel latent class models. *Sociological Methodology, 33*, 213-239.

Vermunt, J. K. (2010). Mixture models for multilevel data. In J. Hox & J. K. Roberts (Eds.), *Handbook of Advanced Multilevel Analysis* (pp. 59-81). New York: Routledge.

Vermunt, J. K., & Van Dijk, L. (2001). A nonparameteric random-coefficients approach: the latent class regression model. *Multilevel Modeling Newsletter, 13*, 6-13.

Wedel, M., & DeSarbo, W. S. (1995). A mixture likelihood approach for generalized linear models. *Journal of Classification, 12*(1), 21-55. doi: 10.1007/bf01202266

Table 1: Deciding the optimal classes using BIC and adjusted BIC for simulated data with random probabilities of class membership across clusters.

# of clusters	# of people per cluster	%BIC 2 v.s. 1	% aBIC 2 v.s. 1	lower class probability 10th percentile	50th percentile	90th percentile
50	50	90.60%	99.20%	33.98%	50.53%	65.08%
	100	99.80%	100.00%	38.58%	49.51%	60.74%
100	50	99.80%	100.00%	40.26%	50.87%	58.76%
	100	100.00%	100.00%	42.42%	49.62%	56.70%
200	50	100.00%	100.00%	44.38%	50.53%	57.13%
	100	100.00%	100.00%	45.23%	49.74%	54.34%

%BIC : the proportion out of 500 replications in which two-class model has a smaller BIC value. %aBIC : the proportion out of 500 replications in which two-class model has a smaller adjusted BIC value. Lower class probability: probability that a randomly selected individual belongs to the first latent class when data was modeled by a two-level model with two latent classes.

Table 2: Model parameter estimates over 500 replications for simulated data with random probabilities of class membership across clusters.

# of clusters	Parameter	True value	# of people per cluster=50 *M*	*SE*	*SD*	Coverg	Max	Min	# of people per cluster=100 *M*	*SE*	*SD*	Coverg	Max	Min
200	$Resid_1$	0.864	0.859	0.029	0.031	0.934	0.953	0.731	0.862	0.021	0.023	0.926	0.927	0.790
	$Resid_2$	0.459	0.455	0.028	0.032	0.918	0.533	0.318	0.456	0.019	0.023	0.896	0.532	0.382
	$Interpt_1$	0	0.000	0.041	0.045	0.924	0.115	-0.195	0.001	0.033	0.035	0.936	0.113	-0.114
	$Slope_1$	0.2	0.202	0.037	0.040	0.920	0.328	0.079	0.198	0.030	0.034	0.924	0.296	0.082
	C1var	0.366	0.279	0.065	0.164	0.486	0.976	0.000	0.316	0.072	0.117	0.674	0.711	0.050
	E1var	0.096	0.096	0.018	0.022	0.874	0.157	0.029	0.093	0.015	0.016	0.890	0.154	0.053
	E2var	0.051	0.048	0.010	0.011	0.868	0.082	0.012	0.049	0.008	0.009	0.902	0.081	0.027
	$Interpt_2$	0.5	0.505	0.029	0.033	0.918	0.617	0.409	0.500	0.023	0.026	0.928	0.574	0.411
	$Slope_2$	0.7	0.699	0.027	0.029	0.928	0.778	0.604	0.700	0.022	0.023	0.938	0.774	0.630
	C1mean	0	0.028	0.171	0.202	0.890	0.749	-0.593	-0.005	0.126	0.150	0.884	0.417	-0.486
100	$Resid_1$	0.864	0.856	0.043	0.048	0.940	1.003	0.689	0.861	0.030	0.033	0.936	0.954	0.689
	$Resid_2$	0.459	0.458	0.040	0.050	0.866	0.583	0.298	0.457	0.027	0.034	0.894	0.572	0.357
	$Interpt_1$	0	-0.005	0.059	0.067	0.922	0.170	-0.284	-0.003	0.047	0.053	0.920	0.131	-0.215
	$Slope_1$	0.2	0.193	0.053	0.058	0.908	0.362	-0.010	0.199	0.043	0.049	0.908	0.335	0.048
	C1var	0.366	0.288	0.093	0.243	0.510	1.790	0.000	0.332	0.106	0.185	0.708	1.433	0.000
	E1var	0.096	0.094	0.024	0.033	0.842	0.204	0.000	0.093	0.020	0.026	0.846	0.195	0.025
	E2var	0.051	0.047	0.013	0.018	0.810	0.113	0.000	0.048	0.011	0.013	0.858	0.092	0.016

	Interpt$_2$	0.5	0.502	0.042	0.045	0.934	0.700	0.383	0.502	0.033	0.034	0.942	0.640	0.410
	Slope$_2$	0.7	0.701	0.039	0.044	0.900	0.832	0.577	0.700	0.031	0.034	0.918	0.811	0.573
	C1mean	0	0.012	0.241	0.308	0.874	1.225	-1.052	-0.024	0.180	0.221	0.888	0.660	-0.996
	Resid$_1$	0.864	0.849	0.067	0.087	0.908	1.188	0.426	0.860	0.045	0.059	0.916	1.007	0.325
	Resid$_2$	0.459	0.456	0.059	0.075	0.832	0.672	0.127	0.459	0.039	0.048	0.862	0.647	0.313
	Interpt$_1$	0	-0.034	0.093	0.161	0.880	0.261	-0.916	-0.019	0.068	0.116	0.888	0.192	-1.488
	Slope$_1$	0.2	0.198	0.076	0.093	0.864	0.433	-0.112	0.196	0.061	0.070	0.884	0.422	-0.057
50	C1var	0.366	0.360	0.156	0.404	0.482	2.629	0.000	0.343	0.151	0.281	0.656	1.839	0.000
	E1var	0.096	0.087	0.032	0.051	0.756	0.424	0.000	0.087	0.029	0.036	0.814	0.227	0.000
	E2var	0.051	0.044	0.018	0.026	0.728	0.138	0.000	0.047	0.016	0.017	0.856	0.109	0.009
	Interpt$_2$	0.5	0.501	0.059	0.069	0.872	0.723	0.321	0.500	0.047	0.052	0.918	0.667	0.318
	Slope$_2$	0.7	0.696	0.057	0.071	0.846	0.916	0.425	0.699	0.045	0.055	0.870	0.921	0.484
	C1mean	0	0.008	0.370	0.569	0.798	2.346	-2.735	-0.027	0.257	0.397	0.830	1.021	-3.526

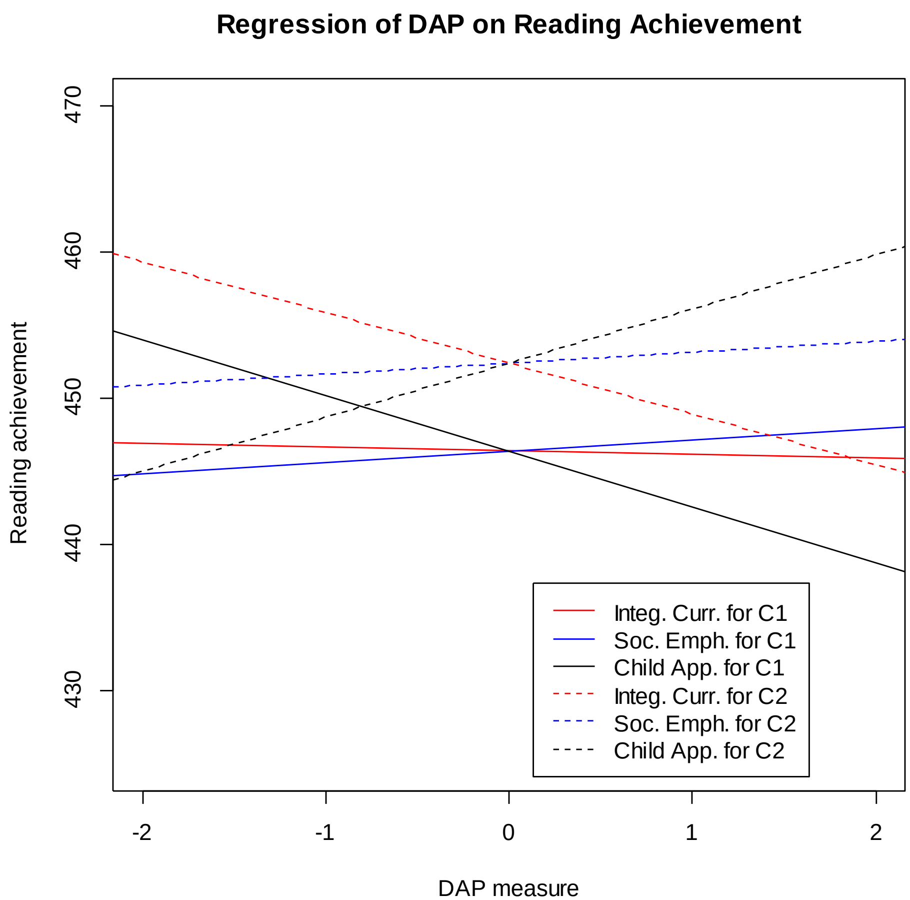
Note: The “True Value” column lists the values of model parameters used to generate the simulated data. The “mean” column is the average of parameter estimates over 500 replications. The “S.E.” column is the mean standard errors over 500 replications. The “S.D.” column lists the sample standard deviations of model parameter estimates over 500 replications. The “RMSE” column is calculated as the square root of the squares of the difference between the true parameter value and parameter estimates mean over 500 replications. The “Coverg” column is the proportion out of 500 replications in which the true parameter values fall in the 95% C.I.s for the model parameters.

Table 3. Parameter estimates and standard errors for the two-class model solution

Model	1 class model		2 class with fixed u_{1cj}				2 class with random u_{1cj}				2 class with random u_{1cj} and fixed γ_{02c}			
			class 1		class 2		class 1		class 2		class 1		class 2	
Parameters	*B*	*SE*	*B*	*SE*	*B*	*SE*	*B*	*SE*	*B*	*SE*	*B*	*SE*	*B*	*SE*
Between-level														
Intercept	448.96**	0.44	444.90**	1.05	453.01**	3.39	446.49**	1.00	452.27**	1.21	446.39**	1.05	452.40**	1.49
Integrated curriculum	-1.60	1.13	2.13	2.64	-5.10**	1.82	0.79	2.42	-4.17*	2.02	-0.24	2.29	-3.48†	2.06
Social/emotional emphasis	0.10	0.97	-0.94	2.14	1.93	1.41	-1.75	1.87	2.82*	1.40	0.76	0.95	-	-
Child-centered approach	-0.36	1.00	-5.77*	2.31	4.45†	2.38	-3.01	2.08	3.00	1.73	-3.81†	2.11	3.70*	1.83
Residual variance	81.58**	9.22	245.93**	51.63	85.71	87.83	161.92**	24.19	75.71**	13.37	156.74**	28.16	78.74**	18.25
Residual covariance[a]	-	-	-14.15	22.31	-	-	-103.77**	9.35	-	-	-92.84**	7.50	-	-
Within-level														
Baseline	0.72**	0.01	0.68**	0.04	-	-	0.68**	0.02	-	-	0.68**	0.02	-	-
Residual variance	246.02**	10.10	183.84**	65.72	99.01**	16.08	283.70**	23.01	80.80**	10.52	286.59**	26.91	79.77**	9.66

Note. [a]Residual covariance between latent classes; **Significant at $p<.01$, *significant at $p<.05$, †significant at $p<.10$.

Figure 1. Regression of three DAP measures on reading achievement for two latent classes



Appendix

Mplus code for estimating a multilevel regression mixtures with two latent classes with fixed probabilities of class membership across clusters.

```
title:  a two-level mixture regression for a continuous dependent variable;
data: file is C:\example.txt;

variable:
  names are cluscov y class clus;
  cluster=clus;
  usevariables are
  cluscov y;
  between = cluscov;
  classes=c(2);

  analysis: type=twolevel mixture;
        starts=48 24;  ! This should be made larger if there is any evidence that most solutions do not arrive at a common ;
                       ! LL value   ;
        processors=24 (starts);
        integration = standard (5);
        stscale=1;
        stiterations=20;

  model:
  %within%
  %overall%
  y;  !  Estimatimates the residual variance of y;
  %c#2%
  y;  ! Frees the residual variance of y to be independently estimated in each class;

  %between%
  %overall%
  y on cluscov;
  c#1@0;
  e1 by y*1; ! e1 and e2 are used to allow the between level variances of y to differ across classes  ;
  y@0;       ! the variance of y is fixed to zero, all error variance is in e1 and e2;
  [e1@0];    !  e1 and e2 have means of zero;
  e2 by y*1;
```

```
y@0;
[e2@0];
e1*0.096;
e2*0.051;
e1 with e2@0;  ! between level residual variances have no residual correlation in the data and so this parameter can!
                 ! not be estimated;

%c#1%
y on cluscov*0.2;  ! Class specific effect of the cluster level covariate;
[y*0];
e1 by y@1;  ! only e1 has variability across clusters for class 1;
e2 by y@0;

%c#2%
y on cluscov*0.7;
[y*0.5];
e1 by y@0;
e2 by y@1;
```