An Introduction to Latent Variable Mixture Modeling (Part 1): Overview and Cross-Sectional Latent Class and Latent Profile Analyses

Kristoffer S. Berlin,[1] PhD, Natalie A. Williams,[2] PhD, and Gilbert R. Parra,[3] PhD

[1]Department of Psychology, The University of Memphis, [2]Department of Child, Youth and Family Studies, The University of Nebraska-Lincoln, and [3]Department of Psychology, The University of Southern Mississippi

All correspondence concerning this article should be addressed to Kristoffer S. Berlin, PhD, Department of Psychology, The University of Memphis, 202 Psychology Building, Memphis, TN 38152, USA. E-mail: kristoffer.berlin@gmail.com.

Objective Pediatric psychologists are often interested in finding patterns in heterogeneous cross-sectional data. Latent variable mixture modeling is an emerging person-centered statistical approach that models heterogeneity by classifying individuals into unobserved groupings (latent classes) with similar (more homogenous) patterns. The purpose of this article is to offer a nontechnical introduction to cross-sectional mixture modeling. **Method** An overview of latent variable mixture modeling is provided and 2 cross-sectional examples are reviewed and distinguished. **Results** Step-by-step pediatric psychology examples of latent class and latent profile analyses are provided using the Early Childhood Longitudinal Study–Kindergarten Class of 1998–1999 data file. **Conclusions** Latent variable mixture modeling is a technique that is useful to pediatric psychologists who wish to find groupings of individuals who share similar data patterns to determine the extent to which these patterns may relate to variables of interest.

Key words cross-sectional data analysis; latent class; latent profile; person-centered; statistical analysis; structural equation modeling.

Latent variable mixture modeling (LVMM) is a flexible analytic tool that allows researchers to investigate questions about patterns of data and to determine the extent to which identified patterns relate to important variables. For example, do patterns of co-occurring developmental and medical diagnoses influence the severity of pediatric feeding problems (Berlin, Lobato, Pinkos, Cerezo, & LeLeiko, 2011)? Do differential longitudinal trajectories of glycemic control exist among youth with type 1 diabetes (Helgeson et al., 2010) or do differential trajectories of adherence among youth newly diagnosed with epilepsy exist (Modi, Rausch, & Glauser, 2011), and if so, do psychosocial and demographic variables predict these patterns? Do patterns of perceived stressors among youth with type 1 diabetes differentially affect glycemic control (Berlin, Rabideau, & Hains, 2012)? Each of these

questions is relevant to pediatric psychology and has been explored using LVMM. The purpose of this two-part article is to offer a nontechnical overview and introduction to cross-sectional (Part 1) and longitudinal mixture modeling (Part 2; Berlin, Parra, & Williams, 2014) to facilitate applications of this promising approach within the field of pediatric psychology. We begin with a general overview of LVMM to highlight notable strengths of this analytic approach, and then provide step-by-step examples illustrating three prominent types of mixture modeling: Latent class, latent profile, and growth mixture modeling.

Conceptually, LVMM is a person-centered analytic tool that focuses on similarities and differences among people instead of relations among variables (Muthén & Muthén, 1998). The primary goal of LVMM is to identify

homogenous subgroups of individuals, with each subgroup possessing a unique set of characteristics that differentiates it from other subgroups. In LVMM, subgroup membership is not observed and must be inferred from the data.

Most broadly, LVMM refers to a collection of statistical approaches in which individuals are classified into unobserved subpopulations or latent classes. The latent classes are represented by a categorical latent variable. Individuals are classified into latent classes based on similar patterns of observed cross-sectional and/or longitudinal data. For any given variable(s), the observed distribution of values may be a ''mixture'' of two or more subpopulations whose membership is unknown. As such, the goal of mixture modeling is to probabilistically assign individuals into subpopulations by inferring each individual's membership to latent classes from the data. As a by-product of mixture modeling, every individual in the data set has his/her own probabilities calculated for his/her membership in all of the latent classes estimated (when summed they equal 1). Latent classes are based on these probabilities, and each individual is allowed fractional membership in all classes to reflect the varying degrees of certainty and precision of classification. Said differently, by adjusting for uncertainty and measurement error, these classes become *latent* (Asparouhouv & Muthén, 2007; Muthén, 2001).

LVMM is part of a latent variable modeling framework (Muthén, 2001; Muthén & Muthén, 1998) and is flexible with regard to the type of data that can be analyzed. Observed variables used to determine latent classes can be continuous, censored, binary, ordered/unordered categorical counts, or combinations of these variable types, and the data can be collected in a cross-sectional and/or longitudinal manner (Muthén & Muthén, 1998). Consequently, a diverse array of research questions involving latent classes can be investigated. For example, hypotheses can focus on predicting class membership, identifying mean differences in outcomes across latent classes, or describing the extent to which latent class membership moderates the relationship between two or more variables. The literature has used many names to describe mixture modeling, or finite mixture modeling as it is known in the statistics literature (McLachlan & Peel, 2000). Names vary according to the type of data used for indicators (continuous vs. categorical, akin to cross-sectional latent profile analysis vs. latent class analysis, etc.), whether continuous latent variables are included with categorical latent class variables (cross-sectional factor mixture models, longitudinal growth mixture models), whether the data were collected cross-sectionally or longitudinally (latent class vs. latent transition), and whether variability is allowed within the latent classes (latent class growth modeling vs. growth mixture modeling; Muthén, 2008). Although there are many types of models that can be examined, we begin in Part 1 by focusing on cross-sectional examples using latent class analysis and latent profile analysis. In Part 2, we focus on longitudinal LVMM and present examples of latent class growth modeling and growth mixture modeling. For both articles, we organize our discussion and examples using the four steps recommended by Ram and Grimm (2009): (a) problem definition, (b) model specification, (c) model estimation, and (d) model selection and interpretation.

An important issue when considering whether to use LVMM is sample size. As with other analytic techniques, the proper sample size is important for obtaining adequate statistical power as well as reducing bias related to parameter and standard error estimates. An insufficient sample size can be particularly problematic when conducting mixture analyses because it is often associated with (a) convergence issues, (b) improper solutions, and (c) the inability to identify small but meaningful subgroups. Unfortunately, determining the sample size needed to conduct a mixture analysis is not straightforward. ''Rules of thumb'' (e.g., 5 or 10 observations per estimated parameter) are commonly used to justify a particular sample size. However, research indicates that these rules are not particularly useful and likely lead to over- or underestimating sample size requirements (for discussion, see Wolf et al., 2013). This is because ''the sample size needed for a study depends on many factors, including the size of the model, distribution of the variables, amount of missing data, reliability of the variables, and strength of the relations among the variables'' (Muthén & Muthén, 2002, pp. 599–600).

In recent years, the Monte Carlo simulation method has emerged as a promising approach for estimating sample size in the context of structural equation modeling in general and LVMM in particular (Muthén & Muthén, 1998–2010; Muthén & Muthén, 2002; Wolf et al., 2013). This approach can estimate the sample size needed for a specified model by simulating the analysis a large number of times. Monte Carlo simulation research is likely to be encouraging for pediatric psychologists who do not have large samples because it demonstrates that small samples can be sufficient depending on several factors such as model complexity and missing data (Wolf et al., 2013). Fortunately, several examples of Monte Carlo simulations designed to estimate sample size are currently available (Muthén & Muthén, 1998–2010 [example 12.3 in particular]; see also Muthén & Muthén, 2002; Wolf et al., 2013).

Problem Definition

The problem definition stage includes three steps. The first step is to formulate hypotheses about the nature of unobserved subgroups. This is done by considering theory and previous research. Second, raw individual-level data and descriptive statistics across primary study variables are examined to help researchers determine the best estimator for their data (e.g., maximum likelihood estimation, weighted least-squares estimator for censored or categorical data) and whether there is a need to take into account nesting of data (via multilevel modeling and/or adjusting the standard errors) and non-normal distributions (e.g., through robust strategies, like robust maximum likelihood estimation). The third step is to determine whether to include covariates and allow continuous measures to correlate. For longitudinal LVMM, this step establishes a single-group model that best represents the nature of change over time. If SEM is used, goodness-of-fit statistics and other indices are then reviewed to establish the best way of modeling relationships among study variables. We encourage those interested in an overview of Structural Equation Modeling (SEM) specific to pediatric psychology to review the article by Nelson, Aylward, and Steele (2008).

Model Specification

In the model specification stage, researchers determine how many classes will be investigated. Ram and Grimm (2009) recommend estimating one more class than is expected. Alternatively, researchers may take an exploratory approach to model specification and estimate as many classes as the data will allow (i.e., additional classes are estimated until a statistically proper and/or practical solution is no longer obtained). The exploratory approach may be more or less justifiable depending on theory and previous research. At this stage, researchers also should decide which parameters are expected to be stable across groups and which parameters are expected to vary across groups. For example, in estimating a latent profile model in which the continuous indicators of the latent class are allowed to correlate, researchers must make decisions about whether the strength/direction of these correlations (and/or covariance and variances) will be freely estimated or fixed to be equal across classes. These decisions can be based on theory, previous research, and/or practical considerations (model convergence, etc.). Generally, more restrictive models (e.g., having various parameters equal across classes) tend to have fewer statistical problems, and as such may be wise starting points for investigators. These initial analyses may then be followed-up by assessing the extent to which freeing parameters (preferably one at a time) affects model fit and the substantive meaning of the solutions obtained.

Model Estimation

In this stage, data are fit to models specifying different numbers of classes. Before fitting data to the models, a decision is made about which estimation method will be used. Guidance about the most appropriate estimation method can be found in most introductory SEM texts.

One important aspect of model estimation in the context of LVMM is the concept of local maxima or a local solution. In nontechnical terms, this means that care must be taken to ensure that the researcher's statistical software has provided the ''best'' solution to estimate how the data fit each particular model. This ''best'' solution is generally determined by a number called the log-likelihood, with the ''best'' solution providing the log-likelihood closest to 0, or said differently, being at the maximum (the plural of which is maxima). In the context of LVMM models, multiple maxima of the likelihood often exist, this is in part due to where the software begins the estimation and the start values used. The potential consequence is that the final solution may be a ''local solution'' and the best given those start values—but not the ''best'' global solution given a range of possible start values. For all LVMM, it is therefore important to use multiple sets of starting values to find the global maximum (i.e., replicate the highest log-likelihood). Most commercially available software do this automatically, with many providing messages if the log-likelihood is not replicated. If the best log-likelihood value is not replicated in at least two final-stage solutions, this may be a sign of a local solution and/or problems with the model. In cases in which the log-likelihood is not replicated, the investigators should increase the number of random starts until they are confident that they are not at local maxima.

Model Selection and Interpretation

The final stage of conducting LVMM involves a series of steps to identify the best fitting model. This is one of the most challenging aspects of the analyses and has been described as ''an art – informed by theory, past findings, past experience, and a variety of statistical fit indices'' (Ram & Grimm, 2009, p. 571). Ram and Grimm (2009) provide a helpful flowchart for making decisions about model selection. Their first step is examining the output of each model estimated for potential problems (e.g., software-generated

error messages and warnings, estimation problems, local maxima, negative variances, out-of-range values, correlations $> \pm 1$). Second, models with different numbers of classes are compared using information criteria (IC)-based fit statistics. These include the Bayesian Information Criteria (BIC; Schwartz, 1978), Akaike Information Criteria (AIC; Akaike, 1987), and Adjusted BIC (Sclove, 1987). Lower values on these fit statistics indicate better model fit. Third, the accuracy with which models classify individuals into their most likely class is examined. Entropy is a type of statistic that assesses this accuracy, and can range from 0 to 1, with higher scores representing greater classification accuracy. Fourth, statistical model comparison likelihood ratio tests and bootstrapping procedures should be used, such as the Lo–Mendell–Rubin test (LMR; Lo, Mendell, & Rubin, 2001) and the Bootstrap Likelihood Ratio Test (BLRT; McLachlan & Peel, 2000). The LMR and BLRT tests compare the improvement between neighboring class models (i.e., comparing models with two vs. three classes, and three vs. four, etc.) and provide p-values that can be used to determine if there is a statistically significant improvement in fit for the inclusion of one more class. Among the information criterion measures, the BIC is generally preferred, as is the BLRT for statistical model comparisons (Nylund, Asparouhov, & Muthen, 2007). An additional consideration is the size of the smallest class. Although a four-class model might provide the best fit to the data, if this additional class is composed of a relatively small number (e.g., proportionally, $<1.0\%$ and/or numerically $n < 25$) of members, the researcher must be able to defend what is gained by the addition of this class given the possibility of low power and precision relative to the other, larger classes (Lubke & Neale, 2006). In summary, deciding on the number of classes can be difficult, and should involve consideration of the research question, fit indices, the substantive meaning of each solution, parsimony, and/or theory (Bauer & Curran, 2003).

Use of Latent Variables Representing Class Membership

While describing and determining the optimal number of classes may be of substantive interest, researchers are often interested in investigating hypotheses related to predictors of latent classes and whether there are significant mean differences across the latent classes on outcome variables. These hypotheses will often use "auxiliary" variables that are not included in the model to retain some "independence" between the classes and the variables of interest.

If these predictor and/or criterion variables were included in the model, they would influence the formation of the latent classes and would, in essence, become indicators of those latent classes (Asparouhov & Muthén, 2013; Clark & Muthén, 2009). This may or may not be a problem given the research question(s). In those instances in which researchers want to form classes independent of hypothesized predictors or outcomes, a common strategy (after having chosen the preferred model) is to export each individual's posterior probabilities for each class using the most likely class membership (i.e., the class with the highest/maximum posterior probability; Nagin, 2005) and then use traditional analyses, such as logistic regression or analysis of variance. This strategy is equivalent to fixing individuals' probabilities of their highest class to 1 and all others equal to 0. However, this strategy can be problematic because it may introduce error and decrease precision, and by doing so, turns the latent class (which corrects for "error" by modeling this uncertainty) into an observed variable.

One alternative to analysis of variance and logistic regression using the most likely class membership is posterior probability-based multiple imputation (pseudo-class draws; Asparouhov & Muthén, 2007; Wang, Brown, & Bandeen-Roche, 2005). Pseudo-class draws take into account differing individual probabilities of latent class membership by taking random samples in which individuals are permitted to flip into neighboring classes at a rate specified by the posterior probabilities. Pseudo-class draws are similar to multiple imputation in missing data analysis (Little, Jorgensen, Lang, & Moore, 2014), except in this case, the latent classes are what is missing. Using this strategy, tests of categorical latent variable multinomial logistic regression (to predict classes) and equality tests of means across latent classes (to assess mean differences) can be computed based on pseudo-class draws, thereby providing less biased estimates by retaining the "latent" nature of the classes (Asparouhov & Muthén, 2013; Wang et al., 2005). Simulation studies show that this approach works well when entropy is high and class separation is large (Clark & Muthén, 2009). Although conceptually superior to the maximum posterior probability approach, pseudo-class draws have recently been criticized as potentially attenuating the relations between class and outcomes (Lanza, Tan, & Bray, 2013). Predictors and distal outcomes of latent classes are active areas of inquiry, with some emerging techniques such as the three-step (Asparouhov & Muthén, 2013; Vermunt, 2010) and stepwise (Lanza, Flaherty, & Collins, 2003) approaches showing promise in certain situations in terms of potentially being more robust and less biased than either pseudo-class draws or

maximum posterior probability strategies. Interested readers are encouraged to review Asparouhov and Muthén's 2013 simulations comparing these approaches. All of these approaches are available in Mplus and can be easily implemented by declaring auxiliary variables in the syntax (see Supplementary file online for examples) and placing, for example, for pseudo-class draws, either an ''(e)'' for mean difference or a ''(r)'' for predictors of class. Having provided an overview of mixture modeling and strategies to determine the optimal number of classes, we now turn our attention to illustrative examples. The Mplus syntax for these analyses is available as a Supplementary file online.

Example Data
Participants

Data were obtained from the Early Childhood Longitudinal Study, Kindergarten Class of 1998–1999 (ECLS-K) data file. The ECLS-K is a longitudinal study that followed a nationally representative sample of children, their parents, teachers, and schools from across the United States. Data were collected in the fall and the spring of children's kindergarten year (1998–1999), the fall and spring of first grade (1999–2000), the spring of third grade (2002), the spring of fifth grade (2004), and the spring of eighth grade (2007). Children in the ECLS-K came from public and private schools and attended both full-day and part-day kindergarten programs. Children also came from diverse socioeconomic and racial/ethnic backgrounds; however, the examples presented later will concentrate on non-Hispanic Black girls and boys, given their heightened risk for obesity/overweight (Davison & Birch, 2001; Ogden, Carroll, Kit, & Flegal, 2012). Child race and gender were assessed during baseline interviews with parents. At the first time point, there were $n = 3,169$ non-Hispanic Black children (50.2% male), and during the eighth-grade assessment, there were $n = 951$ non-Hispanic Black children (50.4% male).

Measures
Body Mass Index

Heights and weights were assessed at six time points: Fall and spring of the kindergarten year (1998–1999), spring of first grade (1999–2000), spring of third grade (2002), spring of fifth grade (2004), and spring of eighth grade (2007). These data were used to calculate age- and sex-specific body mass index (BMI) scores using tables provided by the Centers for Disease Control and Prevention/ National Center for Health Statistics (CDC, 2010). The CDC suggests that BMI is a reliable proxy indicator of adiposity for most children and teens, given research showing that BMI scores correlate to direct measures of body fat (Mei et al., 2002). For descriptive purposes, the present project will employ the $\geq$85th to <95th percentile for age- and gender-specific BMI cutoff points for overweight classifications and the $\geq$95th percentile of age- and gender-specific cutoff points for obese weight status classifications (CDC, 2010). In addition, a standardized BMI score (BMI z-score) was calculated for each child participant following guidelines established by the CDC.

Socioeconomic Status

The ECLS-K computed a composite standardized socioeconomic status score using information on parent education, occupation, and income that was gathered during parent interviews. Higher scores reflect higher levels of educational attainment, occupational prestige, and income.

School Food Environment

The availability of foods at school that are associated with an increased risk for overweight and obesity was assessed by asking youth whether they can purchase sweets (e.g., candy, ice cream, cookies), salty snacks (e.g., potato chips, corn chips, popcorn, crackers), and soda, sports drinks, or fruit drinks that are not 100% juice (e.g., Coke, Gatorade, Hi-C) at their school.

Dietary Intake

Youth were asked a series of questions assessing how frequently they consumed the following specific foods during meals or as snacks in the 7 days before the survey: (a) milk; (b) 100% fruit juices; (c) soda, sports drinks, or fruit drinks that are not 100% juice; (d) green salad; (e) potatoes, not including French fries, fried potatoes, or potato chips; (f) carrots; (g) vegetables, not including green salad, potatoes, or carrots; and (h) fruit, not including fruit juices. Answer choices for all food items ranged from none (e.g., ''I did not eat carrots during the past 7 days'') to four or more times per day.

Physical Activity

Four questions included in the eighth-grade child questionnaire assessed engagement in physical activity. Youth were asked if they participated in school sports during the current school year (answer choices were ''Did not participate'' and two options we combined, ''Participated'' and ''Participated as an officer, leader, or captain''). Youth also indicated the frequency of their engagement in nonschool sports (''Rarely or never,'' ''Less than once a week,'' ''Once or twice a week,'' or ''Every day or almost every day''). Days of exercise was determined by asking youth to indicate how many of the past 7 days they exercised or participated in vigorous physical activity for a minimum of

20 min (range: 0–7 days). Days of physical education classes was determined by responses to a question asking youth to select the number of days (ranging from 0 to 5) they attend physical education classes during an average week when they are in school.

Sedentary Behavior

Sedentary behavior habits were determined by three, two-part questions on the ECLS-K child survey. These questions asked youth to indicate how many hours per day they usually spend watching television (including videotapes and DVDs), playing computer or video games, and using the Internet, on weekdays and weekends. Youth were also asked to indicate (Yes or No) whether they have a television in their bedroom.

Cross-Sectional Latent Variable Mixture Model Examples

For our examples, all models were estimated in Mplus version 7 (Muthén & Muthén, 1998–2012), under missing data theory using all available data and robust (Full Information) maximum likelihood estimation. This strategy for handling missing data is a modern method of modeling with missing data that makes use of all available data points (see Little et al., 2014). This approach also adjusts the standard errors and scales chi-square statistics to account for non-normally distributed data. Alternative modern approaches to handling missing data were considered but not chosen because they are not available within a mixture modeling framework (i.e., using auxiliary variables to predict missingness in conjunction with Full Information Maximum Likelihood) or would limit the availability of indices to determine the optimal number of classes (e.g., model comparison tests are not available with multiple imputation techniques).

Latent Class Analysis Example

The goal of latent class analysis is to classify individuals from a heterogeneous population into smaller, more homogenous, subgroups called latent classes. Because individuals' memberships in latent classes are not observed directly, they must be inferred from their individually varying patterns of responses present in the data. Latent class models can be depicted graphically (see Figure 1, where c is a categorical latent variable, which gives rise (points) to the *binary indicators*. The arrows pointing from c to the variables imply that the item probabilities/thresholds vary across the latent classes of ''c''). We now provide a step-by-step description of a latent class analysis.
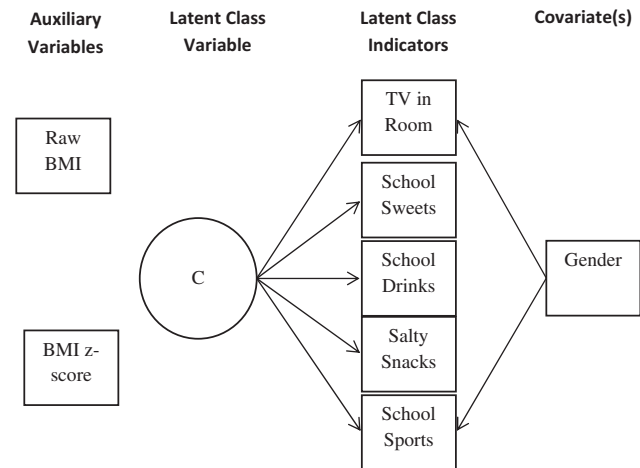

Figure 1. A graphical representation of a latent class model.

Problem Definition

The availability of sweet and salty snacks and sugar-sweetened beverages in the schools (Fox, Dodd, Wilson, & Gleason, 2009; Hollar et al., 2010), participation in physical activities (Janssen & Leblanc, 2010), and television in a child's room (Dennison, Erb, & Jenkins, 2002; Feng, Reed, Esperat, & Uchida, 2011) have been associated with weight status in children and adolescents. The extent to which individually varying patterns across the presence versus absence factors exist and contribute to higher BMI among non-Hispanic Black youth is unclear, hence the purpose of this example.

Model Specification

An exploratory approach was taken to identify the number of school/home obesity risk environment latent classes that may exist for Black youth in the eighth grade. Latent classes with higher probabilities of risk were hypothesized to have higher BMI values. Because gender differences in BMI values were expected, gender was included in the model to control for this potential confound. The influence of gender on binary latent class indicators was assumed to be equal across classes. Because variables included in the estimation of latent classes influence the formation of these classes, BMI was considered as an auxiliary variable.

Model Estimation

Several models were fit to the data, specifying one through four latent classes. The IC, entropy, and likelihood ratio tests are presented in Table I. For each model (k number of classes), replication of the best log-likelihood was verified to avoid local maxima. For models with greater than two classes, it was verified that the null model (k-1) log-likelihood for the BLRT tests was equal to the best log-likelihood

Table I. *Latent Class Example: Information Criteria, Entropy, Likelihood Ratio Tests, and Tests of Mean Differences Across Classes, Average Class Probabilities for Most Likely Class Membership by Latent Class*

Fit statistics	1 Class	2 Class	3 Class	4 Class
Log-likelihood (number of replications)	−2,575.487 (50/50)	−2,446.885 (50/50)	−2,439.481 (43/50)	−2,436.700 (4/100)
AIC	5,164.97	4,919.77	4,916.96	4,915.40
BIC	5,198.98	4,982.93	5,009.28	5,017.43
SSA-BIC	5,176.75	4,941.64	4,948.93	4,950.74
Entropy	N/A	0.66	0.79	0.83
LMR test	N/A	251.10	14.46	5.46
LMR, *p*-value	N/A	<0.0001	0.01	0.55
BLRT test	N/A	257.20	14.81	5.56
BLRT *p*-value for	NA	<0.0001	0.05	0.72
Error messages?	No	No	Yes	Yes
z-BMI differences across class	N/A	χ^2 (1) = 0.09, p = 0.77	χ^2 (2) = 3.940, p = 0.14	χ^2 (3) = 3.984, p = 0.26
Two-class model	1	2		
1, n = 258.9, 27.1%	0.885	0.115		
2, n = 693.1, 72.8%	0.088	0.912		
Three-class model	1	2	3	
1, n = 82.23, 8.6%	0.861	0.000	0.139	
2, n = 173.2, 18.1%	0.009	0.873	0.118	
3, n = 696.6, 73.2%	0.017	0.068	0.915	
Four-class model	1	2	3	4
1, n = 82.12, 8.63%	0.860	0.000	0.129	0.011
2, n = 143.7, 15.1%	0.011	0.867	0.115	0.007
3, n = 693.6, 72.9%	0.018	0.057	0.917	0.008
4, n = 32.6, 3.4%	0.000	0.000	0.138	0.862

Note. AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; SSA-BIC = Sample-Size-Adjusted BIC; LMR = Lo–Mendell–Rubin test; BLRT = Bootstrap Likelihood Ratio Test.
N = 952.

value of the model with one less class. In Mplus, the three- and four-class solution generated the following error messages: "ONE OR MORE PARAMETERS WERE FIXED TO AVOID SINGULARITY OF THE INFORMATION MATRIX. THE SINGULARITY IS MOST LIKELY BECAUSE THE MODE IS NOT IDENTIFIED, OR BECAUSE OF EMPTY CELLS IN THE JOIN DISTRIBUTION OF THE CATEGORICAL VARIABLES IN THE MODEL." Because none of the youth in Class 1 and all of the youth in Class 2 participated in school sports, there was no variability in these two estimates, so the standard error of their respective thresholds was fixed to 0. Too many of these "boundary estimates" may be a sign of local maximum, and/or the extraction of too many classes (Geiser, 2012). Furthermore, the log-likelihood was not replicated initially. To address this issue, the number of initial and final stage random sets of starting values was increased, resulting in a model in which the log-likelihood value was replicated four times.

Model Selection and Interpretation

Determining the optimal class solution is not typically a clear-cut process, as researchers must often reconcile conflicts between the various indices and/or their guiding theories. To determine the most optimal number of classes for our example, we began by reviewing the IC indices [AIC, BIC, and sample-size-adjusted (SSA)-BIC] presented in Table I. The various indices each suggested a different optimal number of classes, with the BIC, AIC, and SSA-BIC suggesting a three-, four-, and two-class model, respectively. Statistical model comparisons (the LMR and BLRT) in this case both suggested that the three-class model provides a significantly better fit than a two-class model, and that a four-class model does not provide a statistically significant improvement over the three-class model. On inspection of Figure 2, it appeared that the four-class model contributes one additional small class with about 3.4% of the sample.

Other tools to aid in model selection include the entropy values and mean class assignment probabilities (Table I). For our example, the best entropy value was for the four-class model. On review of the average class (diagonal) probabilities, all were more than 0.85, suggesting some degree of adequacy. In the three-class model, there appears to be a relatively higher proportion of youth in Class 3 whose most likely class is not Class 3.
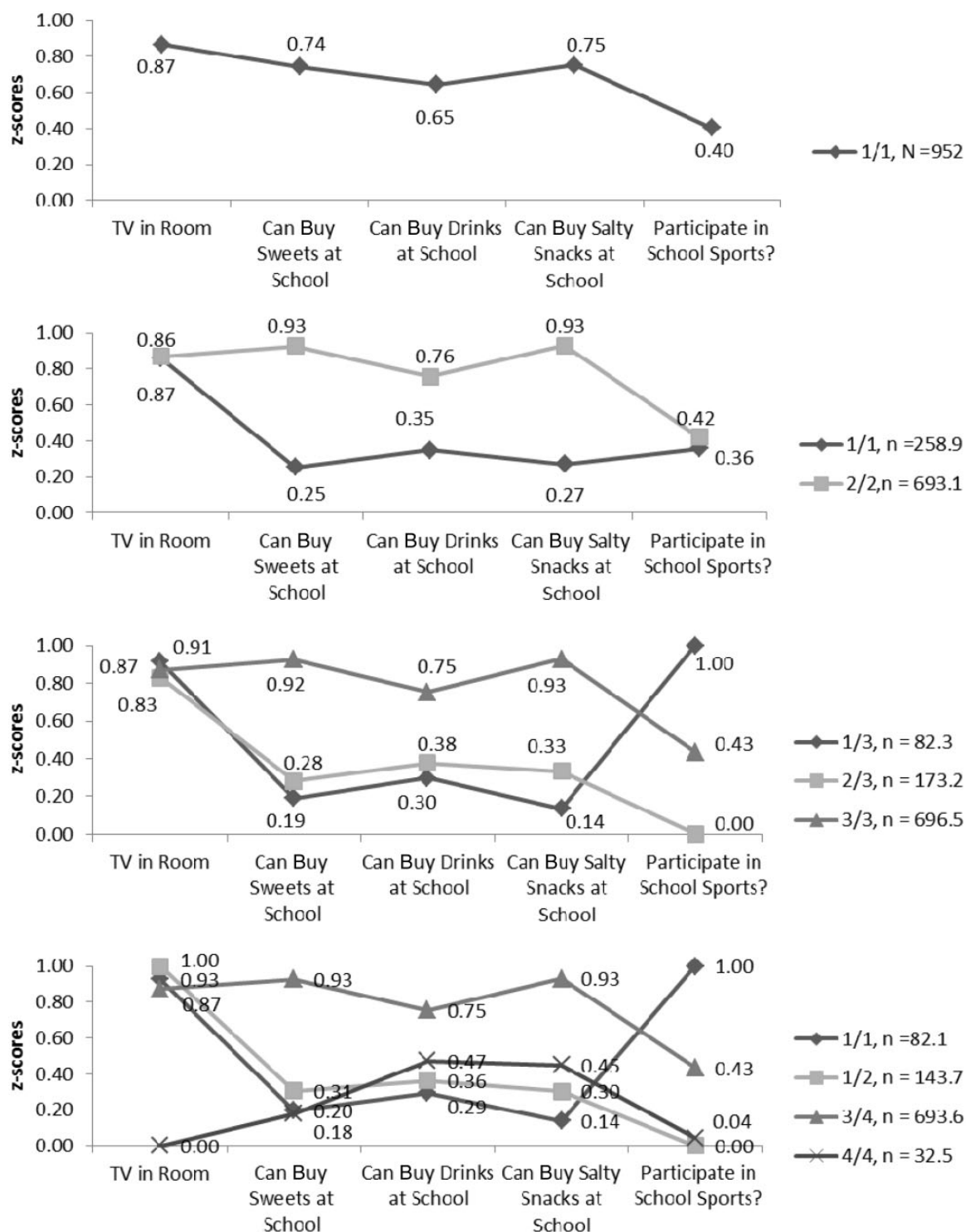
Figure 2. One-, two-, three-, and four-class models of home and school obesity risk.

The differences across these values for a three- versus four-class model appeared negligible. This, coupled with a preference for the BIC, BLRT values, and parsimony, led us (with these data and no compelling theoretical constraints) to choose a three-class model. If additional validity or cross-validation data were available for this example, they could be used to support our decision. Validity data could come in the form of theoretically important predictors of classes and/or mean differences in variables thought to be important in distinguishing the classes.

Having decided on the three-class model, the next step is to interpret these classes. As the indicators are binary (yes/no), the values in Figure 2 can be interpreted as the percent of that class' members who positively endorsed that particular item. For example, the majority of participants reported that they had televisions in their rooms (see one-class model: 87%). As such and given the patterns of risk, the largest class was named High Environmental/Moderate Behavioral Risk (Class 3, 73% of sample), which was characterized by high rates of television in bedrooms (87%), moderate levels of school sport participation (43%), and the highest availability of sweet/salty snacks and drinks at school (between 75% and 93%). The second largest class (Class 2, 18%), named Low Environmental/High Behavioral Risk, was characterized by having a television in the bedroom (83%), no school sport participation (0%), and the low to moderate availability of sweet/salty snacks (28%/33%) and drinks at school (38%). The smallest class (Class 1, 9%) was named Low Environmental/Low Behavioral Risk, characterized by having a television in the bedroom (91%), high participation in school sports (100%), and the lowest availability of sweet/salty snacks (28%/33%) and drinks (38%) at school. For our example data, it was hypothesized that these differential risk patterns are associated with different BMI z-scores; however, no overall differences were found with regard to BMI z-scores for the three-class model ($\chi^2(2) = 3.940$, $p = 0.14$): Class 1: $M = 1.05$, $SE = 0.13$; Class 2: $M = 0.79$, $SE = 0.09$; and Class 3: $M = 0.92$, $SE = 0.04$. This suggests that the impact of these factors might be small in magnitude given their distal influence and/or that a more comprehensive assessment of these constructs is needed (rather than yes/no answers to single items).

Latent Profile Analysis Example

The goal of latent profile analysis is to classify individuals from a heterogeneous population into smaller, more homogenous subgroups based on individuals' values on continuous variables. It is important to note that extensions of latent profiles are not limited to continuous variables, but can include combinations of continuous, count, and categorical variables as indicators of latent class, as well as allowing these indicators to relate to one another. While mixture models are flexible with regard to the inclusion of noncontinuous variables, this tends to increase model complexity and may introduce issues with convergence if the scales are too dissimilar. Similar to a latent class analysis (LCA), a latent profile model can be depicted graphically (Figure 3), where the arrows pointing from the categorical latent variables ''c'' to the variables implies that the item means

of continuous indicators can vary across the latent classes of ''c.'' Below, we illustrate an example of a latent profile analysis, using the last wave of data collection.

Problem Definition

Low levels of physical activity, high rates of sedentary behavior (playing video games, watching television, etc.), and suboptimal nutrition (fast food with little fruit/vegetable intake, consumption of sugar-sweetened soda, etc.) are associated with increased adiposity in youth (Davison & Birch, 2001). The extent to which individually varying patterns of risk across these variables exist and relate to youth's BMI, gender, and socioeconomic status (SES) is unclear.

Model Specification

An exploratory approach was taken to identify patterns of physical activity, sedentary behaviors, and nutritional risk that may exist for non-Hispanic Black youth in the eighth grade. Latent profiles characterized by low physical activity, high sedentary behavior, and suboptimal nutrition are hypothesized to have higher BMI. To determine the extent to which SES and gender predict latent class, pseudo-class draws were used for a posterior-probability-based multinomial logistic regression of the latent class variable on SES and gender. Pseudo-class draws and equality tests of means via the chi-square statistic were used to determine differences in BMI z-scores across the latent classes.

Model Estimation

Several models were fit to the data, specifying one through four latent profiles. Data were not transformed or standardized for these analyses. The IC, entropy, and likelihood ratio tests are presented in Table II. For each model, replication of the best log-likelihood was verified to avoid local maxima. For models with more than two classes, it was verified that the null model log-likelihood for the BLRT was equal to the best log-likelihood value of the model with one less class.

Model Selection and Interpretation

The IC indices (AIC, BIC, and SSA-BIC) are presented in Table II, all suggested that four or more classes were preferred. Consistent with the IC, the BLRT suggested that each successive model above a one-class model provided statistical improvement (e.g., four-class was better than a three-class, which was better than a two-class, model). The LMR test, however, suggested that a one-class model was the preferred solution. At this point, review of the entropy, interpretability of the various solutions, sample sizes, and theoretical considerations is useful.

For our example, all entropy values were acceptable with the two- and three-class solutions having the highest

values. The diagonal class probabilities presented in Table II were all acceptable and ranged from 0.903 to 0.994, averaging 0.973, 0.970, and 0.946 (diagonals), and off-diagonal averages of 0.03, 0.02, and 0.02, for the two-, three-, and four-class solutions, respectively. These values suggest either a two- or three-class model. While the data were modeled in their original scales, these profiles are presented in Figure 2 using z-scores given the large differences in scales of measurement. Unlike the previous latent class example, the values plotted in the figure represent the means of each indicator (rather than the percentage of class members endorsing a particular item). On inspection of Figure 2 and class sample sizes in Table II, it appears that the four-class model contributes one additional small class ($n = 18.1$) consisting of about 1.9% of the sample that is marginally different from the other classes. Although not presented here, researchers interested in the statistical differences across the means of each profile/class's indicators could treat additional copies of these variables as auxiliary, and conduct the appropriate statistical tests for equality of means. All things considered, the three-class model was chosen based on the fit statistics, data, sample size, and parsimony. Although more classes may provide a statistical benefit, a three-class model is adequate.

The largest class was named Average Activity, Moderate Screen, Above Average Diet (Class 1, 89.4%). The second largest (Class 2, 8.3%) was named Average Risk/Resource, as the majority of the scores were around a z-score of 0. The smallest class (Class 3, 2.3%) named Mixed Risk/Resource was an unusual mix of high levels of school sports, screen-time, salad, and soda. For our example data, it was hypothesized that differential risk patterns may be associated with different BMI z-scores; however, no differences were found overall with regard to BMI z-scores for the three-class model, $\chi^2 (2) = 0.29$, $p = 0.87$. With regard to predictors of latent class, classes were equal in terms of gender (p ranged from 0.15 to 0.98), and the Mixed Risk/Resource Class had lower reported SES ($p = 0.04$) relative to Average Risk/Resource.

Conclusions

LVMM is a powerful yet underutilized research tool. Practical guidance for conducting LVMM analyses may facilitate the use of this analytic technique by pediatric psychologists. This article was designed to provide pediatric psychologists with step-by-step instructions for carrying out two types of LVMMs. Concrete examples of latent class analysis and latent profile analysis were described.
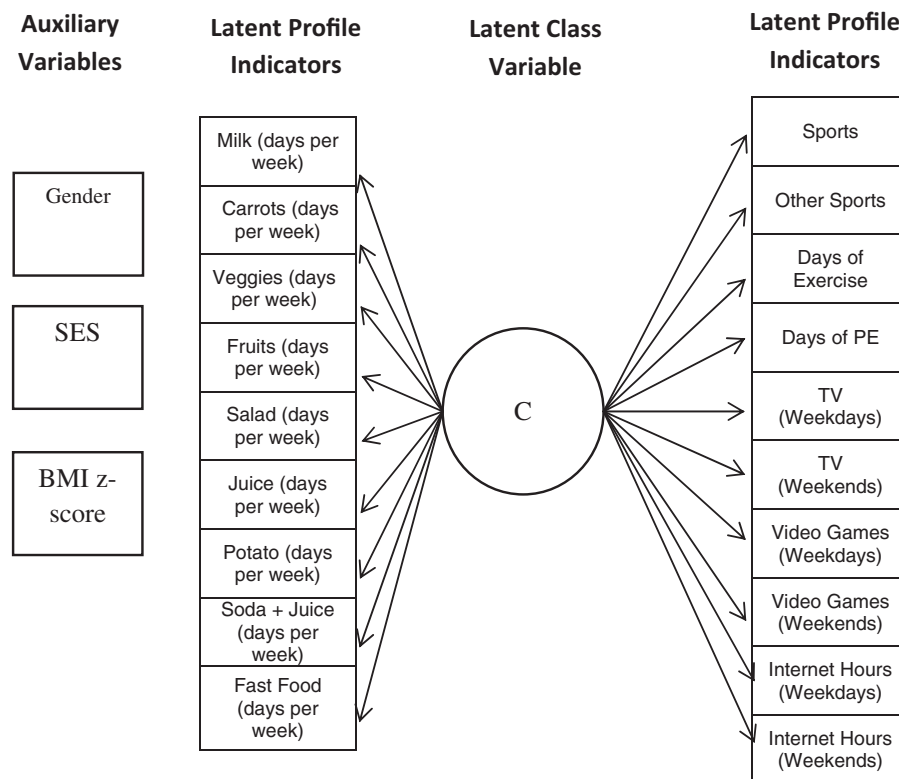

Figure 3. Two-, three-, and four-class models of physical activity, sedentary behaviors, and nutritional obesity risk behaviors.
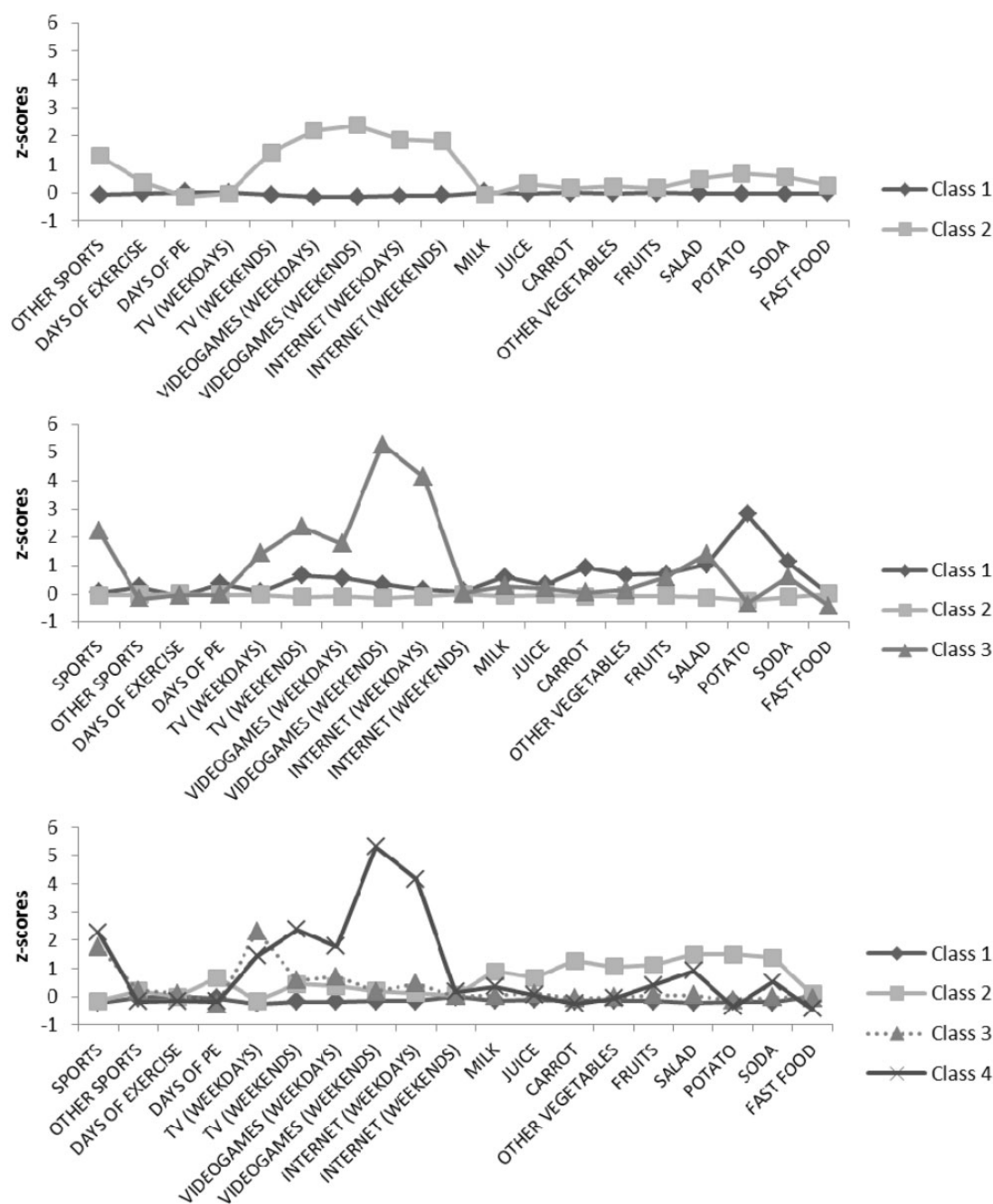
(continued)

Figure 3. Continued.

For each of these examples, we delineated the specific procedures for conducting the analyses and discussed key decisions researchers often must make to estimate an LVMM. We also highlighted several issues/challenges that often arise at different stages of the model fitting processes and provided possible solutions. We hope the latter information will facilitate researchers' ability to work through common mixture modeling problems. It is important to note that LVMM, like all analytic techniques, has limitations. These include difficulties deciding on the most

optimal number of classes, model convergence issues, and the need for relatively large sample sizes. Important other considerations that are relevant to pediatric psychology research and LVMM include, but are not limited to, small pediatric sample sizes, heterogeneity of important sample factors (i.e., different types of cancer, injury, age-range), time-varying assessments, time-varying covariates (such as surgery, medication), and clustering of nonindependent individuals in certain contexts (i.e., group interventions, families), and nonrandom attrition (such as due

Table II. *Latent Profile Example: Information Criteria, Entropy, Likelihood Ratio Tests, and Tests of Mean Differences Across Classes; Average Class Probabilities*

Fit statistics	1 Class	2 Class	3 Class	4 Class
Log-likelihood (number of replications)	−34,735.728	−34,175.709 (8/10)	−33,724.719 (2/100)	−33,424.066 (8/20)
AIC	69,547.456	68,471.417	67,613.438	67,056.132
BIC	69,732.042	68,762.868	68,011.754	67,561.314
SSA-BIC	69,611.355	68,572.311	67,751.326	67,231.015
Entropy	N/A	0.967	0.969	0.931
LMR test	N/A	1,112.663	896.040	597.347
LMR, *p*-value	N/A	0.2914	0.1277	0.7109
BLRT	N/A	1,120.039	901.979	601.306
BLRT *p*-value for	NA	<0.0001	<0.0001	<0.0001
Two-class model	1	2		
1, *n* = 886.7, 93.3%	0.994	0.006		
2, *n* = 64.3, 6.7%	0.049	0.951		
Three-class model	1	2	3	
1, *n* = 850.5, 89.4%	0.937	0.063	0.000	
2, *n* = 79.3, 8.3%	0.007	0.992	0.001	
3, *n* = 21.2, 2.3%	0.000	0.020	0.980	
Four-class model	1	2	3	4
1, *n* = 753.4, 79.2%	0.975	0.017	0.008	0.000
2, *n* = 100.3, 10.5%	0.090	0.903	0.006	0.001
3, *n* = 79.1, 8.3%	0.062	0.012	0.924	0.002
4, *n* = 18.1, 1.9%	0.012	0.001	0.006	0.980

Note. AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; SSA-BIC = Sample-Size-Adjusted BIC; LMR = Lo–Mendell–Rubin test; BLRT = Bootstrap Likelihood Ratio Test.
N = 952.

to mortality). Although beyond the scope of this introduction, many of these considerations (time-varying assessments/covariates, nesting, censored data, sample heterogeneity, etc.) can either be modeled directly or as a research question that lends itself to an LVMM (illness and demographic heterogeneity). Despite these limitations and considerations, we believe that LVMM is an analytic tool that can be useful to pediatric psychologists who wish to identify subgroups of individuals who share similar data patterns and determine the extent to which subgroup membership relates to variables of interest.

Supplementary Data

Supplementary data can be found at: http://www.jpepsy.oxfordjournals.org/

Conflicts of interest: None declared.

References

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika, 52*, 317–332. doi:10.1007/BF02294359

Asparouhouv, T., & Muthén, B.O. (2007, November 16). Wald test of mean equality for potential latent class predictors in mixture modeling [Online technical appendices]. Retrieved from http://www.statmodel.com/download/MeanTest1.pdf.

Asparouhov, T., & Muthén, B. (2013, February 7). Auxiliary variables in mixture modeling: A 3-step approach using Mplus [Online web notes]. Retrieved from http://statmodel.com/examples/webnotes/AuxMixture_submitted_corrected_webnote.pdf.

Bauer, D. J., & Curran, P. J. (2003). Over extraction of latent trajectory classes: Much ado about nothing? Reply to Rindskopf (2003), Muthén (2003), and Cudeck and Henly (2003). *Psychological Methods, 8*, 384–393. doi:10.1037/1082-989X.8.3.384

Berlin, K. S., Lobato, D. J., Pinkos, B., Cerezo, C., & LeLeiko, N. S. (2011). Patterns of medical and developmental comorbidities among children presenting with feeding problems: A latent class analysis. *Journal of Developmental and Behavioral Pediatrics, 32*, 41–47. doi:10.1097/DBP.0b013e318203e06d

Berlin, K. S., Parra, G. R., & Williams, N. A. (2014). An introduction to latent variable mixture modeling (part 2): Longitudinal latent class growth and growth

mixture models. *Journal of Pediatric Psychology, 39*, 188–203. doi:10.1093/jpepsy/jst085

Berlin, K. S., Rabideau, E. M., & Hains, A. A. (2012). Empirically derived patterns of perceived stress among youth with type 1 diabetes and relationships to metabolic control. *Journal of Pediatric Psychology, 37*, 990–998. doi:10.1093/jpepsy/jss080

Centers for Disease Control and Prevention, U.S. Department of Health and Human Services, National Center for Health Statistics. (2010). *CDC growth charts*. Retrieved from http://www.cdc.gov/growthcharts/.

Clark, S., & Muthén, B. (2009). Relating latent class analysis results to variables not included in the analysis. Retrieved from https://www.statmodel.com/download/relatinglca.pdf.

Davison, K. K., & Birch, L. L. (2001). Childhood overweight: A contextual model and recommendations for future research. *Obesity Reviews, 2*, 159–171. doi:10.1046/j.1467789x.2001.00036.x

Dennison, B. A., Erb, T. A., & Jenkins, P. L. (2002). Television viewing and television in bedroom associated with overweight risk among low-income preschool children. *Pediatrics, 109*, 1028–1035. doi:10.1542/peds.109.6.1028

Feng, D., Reed, D. B., Esperat, M. C., & Uchida, M. (2011). Effects of TV in the bedroom on young Hispanic children. *American Journal of Health Promotion, 25*, 310–318. doi:10.4278/ajhp.080930-QUAN-228

Fox, M. K., Dodd, A. H., Wilson, A., & Gleason, P. M. (2009). Association between school food environment and practices and body mass index of US public school children. *Journal of the American Dietetic Association, 109*, 108–117. doi:10.1016/j.jada.2008.10.065

Geiser, C. (2012). *Data analysis with Mplus*. New York, NY: Guilford.

Helgeson, V. S., Snyder, P. R., Seltman, H., Escobar, O., Becker, D., & Siminerio, L. (2010). Brief report: Trajectories of glycemic control over early to middle adolescence. *Journal of Pediatric Psychology, 35*, 1161–1167. doi:10.1093/jpepsy/jsq011

Hollar, D., Messiah, S. E., Lopez-Mitnik, G., Hollar, T. L., Almon, M., & Agatston, A. S. (2010). Healthier options for public schoolchildren program improves weight and blood pressure in 6-to 13-year-olds. *Journal of the American Dietetic Association, 110*, 261–267. doi:10.1016/j.jada.2009.10.029

Janssen, I., & LeBlanc, A. G. (2010). Review systematic review of the health benefits of physical activity and fitness in school-aged children and youth.

International Journal of Behavioral Nutrition and Physical Activity, 7, 40. doi:10.1186/1479-5868-7-40

Lanza, S. T., Flaherty, B. P., & Collins, L. M. (2003). Latent class and latent transition analysis. In J. A. Schinka, & W. F. Velicer (Eds.), *Handbook of psychology: Research methods in psychology* (pp. 663–685). Hoboken, NJ: Wiley.

Lanza, S. T., Tan, X., & Bray, B. C. (2013). Latent class analysis with distal outcomes: A flexible model-based approach. *Structural Equation Modeling: A Multidisciplinary Journal, 20*, 1–26. doi:10.1080/10705511.2013.742377

Little, T. D., Jorgensen, T. D., Lang, K. M., & Moore, W. G. (2014). On the joys of missing data. *Journal of Pediatric Psychology, 39*, 151–162. doi:10.1093/jpepsy/jst048

Lo, Y., Mendell, N., & Rubin, D. (2001). Testing the number of components in a normal mixture. *Biometrika, 88*, 767–778. doi:10.1093/biomet/88.3.767

Lubke, G. H., & Neale, M. C. (2006). Distinguishing between latent classes and continuous factors: Resolution by maximum likelihood. *Multivariate Behavioral Research, 4*, 499–532. doi:10.1207/s15327906mbr4104_4

McLachlan, G., & Peel, D. (2000). *Finite mixture models*. New York, NY: Wiley.

Mei, Z., Grummer-Strawn, L. M., Pietrobelli, A., Goulding, A., Goran, M. I., & Dietz, W. H. (2002). Validity of body mass index compared with other body-composition screening indexes for the assessment of body fatness in children and adolescents. *American Journal of Clinical Nutrition, 75*, 978–985.

Modi, A. C., Rausch, J. R., & Glauser, T. G. (2011). Patterns of nonadherence to antiepileptic drug therapy in children with newly diagnosed epilepsy. *JAMA, 305*, 1669–1676. doi:10.1001/jama.2011.506

Muthén, B. (2001). Latent variable mixture modeling. In G. A. Marcoulides, & R. E. Schumacker (Eds.), *New developments and techniques in structural equation modeling* (pp. 1–33). Mahwah, NJ: Lawrence Erlbaum Associates.

Muthén, B. (2008). Latent variable hybrids: Overview of old and new models. In G. R. Hancock, & K. M. Samuelsen (Eds.), *Advances in latent variable mixture models* (pp. 1–24). Charlotte, NC: Information Age Publishing, Inc.

Muthén, L. K., & Muthén, B. O. (1998–2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

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

Nagin, D. S. (2005). *Group-based modeling of development*. Cambridge, MA: Harvard University Press.

Nelson, T. D., Aylward, B. S., & Steele, R. G. (2008). Structural equation modeling in pediatric psychology: Overview and review of applications. *Journal of Pediatric Psychology, 33*, 679–687. doi:10.1093/jpepsy/jsm107

Nylund, K. L., Asparouhov, T., & Muthen, B. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling. A Monte Carlo simulation study. *Structural Equation Modeling, 14*, 535–569. doi:10.1080/10705510701575396

Ogden, C. L., Carroll, M. D., Kit, B. K., & Flegal, K. M. (2012). Prevalence of obesity and trends in body mass index among US children and asolescents, 1999–2010. *Journal of the American Medical Association, 307*, 483–490. doi:10.1001/jama.2012.40

Ram, N., & Grimm, K. J. (2009). Growth mixture modeling: A method for identifying difference in longitudinal change among unobserved groups. *International Journal of Behavioral Development, 33*, 565–576. doi:10.1177/0165025409343765

Schwartz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*, 461–464. doi:10.1214/aos/1176344136

Sclove, L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika, 52*, 333–343. doi:10.1007/BF02294360

Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis, 18*, 450–469. doi:10.1093/pan/mpq025

Wang, C. P., Brown, C. H., & Bandeen-Roche, K. (2005). Residual diagnostics for growth mixture models: Examining the impact of preventive intervention on multiple trajectories of aggressive behavior. *Journal of the American Statistical Association, 100*, 1054–1076. doi:10.1198/016214505000000501

Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample size requirements for structural equation models an evaluation of power, bias, and solution propriety. *Educational and Psychological Measurement, 73*, 913–934. doi:10.1177/0013164413495237