Running head: STATISTICAL AND SUBSTANTIVE CHECKING

Statistical and Substantive Checking in Growth Mixture Modeling

Bengt Muthén

University of California, Los Angeles

Abstract

This commentary discusses the Bauer and Curran (2002) investigation of growth mixture modeling. Single-class modeling of non-normal outcomes is compared to modeling with multiple latent trajectory classes. New statistical tests of multiple-class models are discussed. Principles for substantive investigation of growth mixture model results are presented and illustrated by an example of high school dropout predicted by low mathematics achievement development in grades 7 - 10.

Statistical and Substantive Checking In Growth Mixture Modeling

Introduction

The development of growth mixture modeling (Muthén & Shedden, 1999; Muthén, 2001a, b; Muthén, Brown, Masyn, Jo, Khoo, Yang, Wang, Kellam, Carlin, & Liao, J., 2002) met a long-standing substantive need for more developmentally-meaningful analysis of longitudinal data. The Bauer and Curran (2002) paper's scrutiny of this technique is timely because it may help protect against poor applications now that the technique has left the initial phase of "toy applications" aimed at methods illustrations and has entered the phase of serious substantive applications. This commentary on the Bauer-Curran (BC from now on) paper is intended to clarify some issues in BC and further help promote good uses of growth mixture models (GMMs from now on).

BC points out that a researcher may presume that a GMM with multiple latent trajectory classes generated the data when the data have in fact been generated by a single-class growth model with non-normal outcomes. BC is concerned about being able to distinguish between the two alternatives. In choosing between them, BC correctly states "The dilemma for the applied researcher is that the fit statistics most commonly used to evaluate growth mixture models do not adequately discriminate between these two possibilities." Because of this, BC warns that "researchers should be cautious in the use and interpretation of growth mixture models, particularly when evaluating predictors of class membership." The issue of the two alternative explanations is classic in finite mixture statistics (for a historical overview, see, e.g.

McLachlan & Peel, 2002, pp. 14-17), but is perhaps little known in psychology. While the statistical literature has focused on a univariate outcome, the BC paper makes a contribution by investigating consequences for multivariate outcomes in a growth modeling context.

This commentary focuses on three matters. First, if it is truly the case that a researcher cannot distinguish statistically between the two alternatives, is that a serious problem? Second, is it true that a researcher cannot distinguish statistically between the two alternatives – what are the statistical options for attempting to distinguish between the two alternatives? Third, what are the substantive options for choosing between the two alternatives?

It will be shown that the strongly non-normal data considered by BC are not well fitted by a GMM so that the faulty conclusions BC is concerned about would not be made when using proper statistical testing. An example illustrates how the flexibility of the GMM allows an elaboration of conventional modeling that can give further insight into the data structure. It is argued that substantive considerations are key in deciding whether the added flexibility of GMM gives meaningful and useful results.

Equivalent Models

How serious is it if a researcher cannot distinguish statistically between two model alternatives? It is well-known in statistical modeling that some models represent a given data set equally well. This is particularly true for more exploratory models. A classic example is exploratory factor analysis (EFA) where an orthogonal rotation such

as Varimax and an oblique rotation such as Promax reproduce the same correlation matrix for the outcomes. A researcher may be bothered by these two alternative explanations of the data because one says that the factors are uncorrelated and the other says that they are correlated. Factor (un)correlatedness cannot be proven by EFA. Confirmatory factor analysis (CFA) may be able to address the issue, bringing in further knowledge about measurement. Does this mean that EFA should be cautioned against in favor of CFA? Although there may be good reasons to do CFA when more knowledge is available, this does not invalidate EFA but simply means that the Varimax solution defines the factors differently than Promax and both alternatives are valid. For example, to solve a set of math and reading achievement items, a person draws on math and reading skills. One can consider a factor that is connected with good performance on the math items, recognizing that some degree of reading skills is often involved in math items. Alternatively, one can define a math factor as what requires purely mathematical ability, purging this factor of any reading content. In the first case the math factor can be reasonably thought of as correlated with the reading factor, while in the second case the factors are uncorrelated. In sum, Varimax and Promax are just two alternative ways of viewing the same reality. The choice can be based on which view is most useful for a certain practical purpose.

A similar situation arises in latent profile analysis, i.e. latent class analysis with continuous outcomes. Bartholomew and Knott (1999, pp. 154-155), points out a well-known psychometric fact that a covariance matrix generated by a latent profile model can be perfectly fitted by a factor analysis model. A covariance matrix from a k-class model can be fitted by a factor analysis model with $k - 1$ factors. Molenaar

and von Eye (1994) show that a covariance matrix generated by a factor model can be fitted by a latent class model. This should not be seen as a problem, but merely as two ways of looking at the same reality. The factor analysis informs about underlying dimensions and how they are measured by the items, while the latent profile analysis sorts individuals into clusters of individuals who are homogeneous with respect to the item responses. The two analyses are not competing, but are complementary.

Concluding from these two examples of equivalent models, one could argue that BC does not demonstrate a problem with GMM as long as researchers are aware that a single-class non-normal-outcomes model is an alternative view that may fit the data equally well. But, do the alternatives really fit the data equally well? The next section turns to this question.

Statistical Model Selection Procedures

The message in BC is to some extent confounded with limitations of commonly used finite mixture model selection procedures such as the Bayesian Information Criterion (BIC). It is true that a researcher cannot rely on BIC-type information to distinguish between the two alternatives that BC is concerned with. However, is BIC the best that we can do? This section briefly describes two new approaches and shows that they to some extent alleviate BC's concerns. A key notion is the need for checking how well the mixture model fits the data, not merely basing a model choice on k classes fitting better than $k-1$ classes. It should be emphasized that there are many possibilities for checking model fit against data in mixture settings and methodology for this is likely to expand considerably in the future (see, e.g., the residual diagnostic

approaches proposed in Wang & Brown, 2002).

New Mixture Tests

Lo, Mendell and Rubin (2001) proposed a likelihood-ratio based method for testing $k-1$ classes against k classes. The Lo-Mendell-Rubin likelihood ratio test (LMR LRT from now on) avoids a classic problem of chi-square testing based on likelihood ratios. This concerns models that are nested, but where the more restricted model is obtained from the less restricted model by a parameter assuming a value on the border of the admissible parameter space, in the present case a latent class probability being zero. It is well-known that such likelihood ratios do not follow a chi-square distribution. LMR considers the same likelihood ratio but derives its correct distribution. A low p value indicates that the $k-1$-class model has to be rejected in favor of a model with at least k classes. The LMR LRT procedure was implemented in Mplus (Muthén & Muthén, 1998-2002) in Version 2.12 of August 2002. This implementation uses the usual Mplus mixture modeling assumption of within-class conditional normality of the outcomes given the covariates. When non-normal covariates are present, this allows a certain degree of within-class non-normality of the outcomes. The LMR LRT procedure has been studied for GMMs by Monte Carlo simulations (Masyn, 2002). More investigations of performance in practice are, however, of interest and readers can easily conduct studies using the Mplus Monte Carlo facility for mixtures.

The LMR LRT is a breakthrough for helping to select the best-fitting number of classes. However, the test is unlikely to be suitable when the alternative is a

single-class model with strongly non-normal outcomes because of the LMR LRT assumption of within-class normality conditional on covariates. When testing 1 versus 2 GMM classes, the 1-class model specifies conditional normality, which with sufficiently non-normal outcomes is likely to be rejected in favor of more classes.

Muthén and Asparouhov (2002) proposed a new approach for testing the fit of a k class mixture model for continuous outcomes. As opposed to the LMR LRT, this procedure concerns test of a specific model's fit against data. The procedure relies on testing if the multivariate skewness and kurtosis estimated by the model fits the corresponding sample quantities. The sampling distributions of the skewness and kurtosis (SK from now on) tests are assessed by computing these values over a number of replications in data generated from the estimated mixture model. Obtaining low p values for skewness and kurtosis indicates that the k-class model does not fit the data. Univariate and bivariate test results are also provided for each variable and pair of variables. These tests may provide a useful complement to the LMR LRT. The SK tests were implemented in Mplus (Muthén & Muthén, 1998-2002) in Version 2.12 of August 2002. Currently the SK tests are not available with missing data. Given the inherent sensitivity to outliers, the SK testing should be preceeded by outlier investigations. The SK procedure needs further investigation, but is offered here as an example of the many possibilities of testing a mixture model against data (see also Wang & Brown, 2002).

Mixtures And Non-Normal Outcomes

The theme of checking fit of a mixture model against data is now elaborated in the context of non-normal outcomes. BC considers outcomes with a rather high degree of nonnormality using two alternative univariate skew/kurtosis values of 1/1 and 1.5/6. As is shown below, however, growth mixture models with normal components often do not generate very high non-normality. This points to the promise of the SK tests.

True 2-Class Model With Close To Normal Outcomes.

The following univariate skew and kurtosis values were obtained in a sample of $n = 2000$ generated by a 2-class linear GMM for six time points with within-class multivariate normality for the outcomes and well-separated intercept growth factor means that are 2 standard deviations apart,[1]

$$Skewness = (\ 0.247 \quad 0.268 \quad 0.205 \quad 0.145 \quad 0.080 \quad 0.058\ ), \tag{1}$$

$$Kurtosis = (\ -0.188 \quad -0.055 \quad 0.015 \quad 0.043 \quad 0.018 \quad -0.028\ ). \tag{2}$$

Such data would typically be considered close to normal from a practical point of view. Using Monte Carlo simulations, it can be shown that the model parameters can be well recovered at the sample sizes of $n = 200$ and $n = 600$ considered by BC. This is important to emphasize because a reader of BC[2] may get the mistaken impression that GMM has difficulties with approximately normal data, with convergence mainly for smaller samples due to capitalizing on chance.

The example above is not atypical in real data with clear trajectory classes. For example, Muthén, Leuchter and Morgan (2002) studied depression development

before and after treatment in a randomized clinical trial of 51 individuals receiving either a placebo or medication. The analysis considered the development of a responder class for which depression decreases rapidly and stays low throughout the 8 weeks of the trial. The responder trajectories are quite different from those of the non responders. Nevertheless, the univariate skewness and kurtosis values are modest.

True Single-Class Model With Non-Normal Outcomes.

Although GMM often fits means, variances, and covariances well, GMM often underestimates skewness and kurtosis when analyzing highly non-normal data. Table 1 shows univariate skew and kurtosis values for the cases without covariates considered in BC. A 2-class linear GMM is fitted to data generated by a single-class model under the two BC skew/kurtosis alternatives 1/1 and 1.5/6. The model parameter values generating the data are exactly those used in BC. The columns labelled Mixed Skew and Mixed Kurtosis are skew and kurtosis computed from the estimated 2-class mixture and mixed over the two classes. They show that the 2-class GMM that BC estimated gives a considerable underestimation of the univariate skew and kurtosis values in the BC sample. The 2-class model does not fit the BC data so the two alternatives are in fact not equivalent as BC implies.[3] These observations motivated the development of the SK tests. Results from SK testing are briefly considered next.

Skewness-Kurtosis Testing With Mixtures and Non-Mixtures

Muthén and Asparouhov (2002) applied the SK tests to the true 2-class model of Section 3.2.1, demonstrating good type I error level (sample sizes of $n = 50$ to $n = 1000$). This was the case both with class-invariant and class-varying growth factor

variances. For the same model, the LMR LRT test of one versus two classes did not reach an acceptable power level until $n = 1000$. Adding covariates to the model, however, dramatically improved LMR LRT power.

Muthén and Asparouhov (2002) also studied the SK test performance on the data considered in BC, that is, data generated by the true single-class model with non-normal outcomes. A Monte Carlo study with 500 replications estimated the power to reject the 2-class alternative at the two skewness/kurtosis settings of BC (1/1 and 1.5/6). For $n = 600$ the power was estimated as 1.00 for both the multivariate skewness and kurtosis tests. For $n = 200$ the power was estimated as $0.67 - 0.77$ for skewness and $0.082 - 0.226$ for kurtosis. This means that with a sample size of a little more than $n = 200$, there is sufficient power for the skewness test to reject the 2-class alternative. In conclusion, using this new test, BC would not have made the faulty conclusions that they were concerned about.

Muthén and Asparouhov (2002) further studied the extent to which one can avoid concluding a 2-class model when the true model is single-class with mild skew/kurt of 0.1/0.5 (these values are similar to those seen in the LSAY data exampe below). This showed that there was insufficient power to reject the 2-class model at $n = 600$. With mild non-normality, the SK test does not help. On the other hand, in such cases, Muthén and Asparouhov (2002) found that, unlike the case in BC, the influence of the covariates in the 2-class solution was not distorted relative to the 1-class model that generated the data. Also, the LMR LRT might be helpful in such situations.

Substantive Theory and Auxiliary Information For Predicting and Understanding Model Results

The BC section Implications for Applied Research discusses the role of substantive theory in guiding the interpretation of the model. The BC paper missed the opportunity to contribute a thorough discussion of how psychological theory can guide GMM and move GMM from initial analyses of an exploratory nature towards more confirmatory uses. As discussed in Muthén (2002a), GMM has even more confirmatory potential than CFA. This is because people, not only parameters, can be given fixed values – i.e. fixed class membership for individuals showing typical class behavior or having known class membership from auxiliary information. To illustrate how substantive ideas can be brought to bear on the analyses, an example concerning mathematics achievement development in grades 7 - 10 is discussed in some detail, following a brief discussion of general issues related to substantive evidence. Further applications are discussed in Muthén (2001a, b) while Muthén (2002b) gives an overview of the general latent variable framework in which the modeling fits.

Substantive Evidence in Favor of Mixtures

In addition to a statistical assessment of the model as discussed earlier, the model can be investigated using substantively-based evidence. Auxiliary information can be used to more fully understand model results even at an exploratory stage where little theory exists. Once substantive theory has been formulated, it can be used to predict an intervowen set of events that can then be tested.

Substantive theory building typically does not rely on only a single outcome

measured repeatedly, accumulating evidence for a theory only by sorting into classes observed trajectories on a single outcome variable. Instead, many different sources of auxiliary information are used to check the theory's plausibility. For example, in the depression trial example considered earlier, the placebo responder effect has not only been observed in many different studies, but has also been associated with changes in brain function early in treatment (Leuchter, Cook, Witte, Morgan, and Abrams, 2002). Mental health research may find that a pattern of a high level of deviant behavior at ages where this is not typical is often accompanied with a variety of negative social consequences so that there is a distinct subtype. A good education study of failure in school also considers what else is happening in the student's life, involving predictions of accompanying problems of different, psychological nature. Gene-environment interaction theories may predict emergence of problems as a response to adverse life events at certain ages. These are the situations where GMM is particularly useful. GMM can include the auxiliary information in the model and test if the classes formed have the characteristics on the auxiliary variables that are predicted by theory. Auxiliary information may take the form of antecedents, concurrent events, or consequences. These are briefly discussed in turn below.

Antecedents.

Auxiliary information in the form of antecedents (covariates) of class membership and growth factors should be included in the set of covariates to correctly specify the model, find the proper number of classes, and correctly estimate class proportions and class membership (Muthén, 2002a). The fact that the "unconditional

model" without covariates is not suitable for finding the number of classes has not been fully appreciated.

An important part of GMM is the prediction of class membership probabilities from covariates. This gives the profiles of the individuals in the classes. The estimated prediction of class membership is a key feature in examining predictions of theory. If classes are not statistically different with respect to covariates that according to theory should distinguish classes, crucial support for the model is absent.

Class-variation in the influence of antecedents (covariates) on growth factors or outcomes also provides a better understanding of the data. As a caveat one should note that if a single-class model has generated the data with significant positive influence of covariates on growth factors, GMM that incorrectly divides up the trajectories in say low, medium, high classes may find that covariates have lower and insignificant influence in the low class due to selection on the dependent variable. If a GMM has generated the data, however, the selected subpopulation is the relevant one to which to draw the inference. In either case, GMM provides considerably more flexibility than what can be achieved with conventional growth modeling. As an example, consider the Muthén and Curran (1997) analysis of a preventive intervention with a strong treatment-baseline interaction. The intervention aimed at changing the trajectory slope of aggressive-disruptive behavior of children in classrooms grades 1 - 7. No main effect was found, but Muthén-Curran used multiple-group latent growth curve modeling to show that the initially more aggressive children benefited from the intervention in terms of lowering their trajectory slope. The Muthén-Curran technique

is not, however, able to capture a non-monotonic intervention effect that exists for children of medium-range aggression, but is absent for the most or least aggressive children. In contrast, such a non-monotonic intervention effect can be handled using GMM with the treatment/control dummy variable as a covariate having class-varying slopes (see Muthén, Brown, Masyn, Jo, Khoo, Yang, Wang, Kellam, Carlin, & Liao, 2002). There are probably many cases where the effect of a covariate is not strong, or even present, except in a limited range of the growth factor or outcome.

Concurrent Events and Consequences (Distal Outcomes).

Modeling with concurrent events and consequences speaks directly to standard considerations of concurrent and predictive validity. In generalized GMM available in Mplus, concurrent events can be handled as time-varying covariates that have class-varying effects, as time-varying outcomes predicted by the latent classes, or as parallel growth processes. Consequences can be handled as distal outcomes predicted by the latent classes or as sequential growth processes. Examples of distal outcomes in GMM include alcohol dependence predicted by heavy drinking trajectory classes (Muthén & Shedden, 1999) and prostate cancer predicted by prostate-specific antigen trajectory classes (Lin, Turnbull, McCullogh & Slate, 2002).

One may argue that being able to predict a distal outcome from trajectory class membership does not necessarily constitute evidence of a GMM. For example, if data have been generated by a conventional single-class growth model where increasing growth factor intercept and slope values gives an increasing probability of the distal outcome, a GMM might point to a 2-class solution with a high and a low class where

the high class has a higher distal outcome probability. When statistical evidence is lacking, substantive considerations are therefore key in the analysis.

A very useful feature of GMM even if a single-class non-normal growth model cannot be rejected is that cutpoints for classification are provided. For instance, individuals in the high class, giving the higher probability for the distal outcome, are identified, while this information is not provided by the conventional single-class growth analysis. It is true that this classification is done under a certain set of model assumptions (e.g. within-class conditional normality of outcomes given covariates), but even if the classification is not indisputable, it is nevertheless likely to be useful in practice. In single-class analysis one may estimate individuals' values on the growth factors and attempt a classification but it can be very difficult to identify cutpoints and the classification is inefficient. The added classification information in GMM versus conventional single-class growth modeling is analogous to the earlier discussion of latent class and latent profile analysis adding complementary information to factor analysis. In addition, GMM classification is an important tool for early detection of likely membership in a problematic class as will be discussed in the example below.

An Example

This section briefly reports on a growth mixture example studied in more detail in Muthén (2002a). This example considers mathematics achievement data from the Longitudinal Study of American Youth (LSAY), a national sample of students in public schools in the US. Here, data from grades 7 - 10 are used. The interest is in relating achievement development to dropping out of high school. Based on the

educational literature the following covariates are included: female, hispanic, black, mother's education, home resources, student's educational expectations measured in 7th grade (1 = HS only, 2 = Vocational training, 3 = some college, 4 = Bachelor's, 5 = Master's, 6 = Dr, PhD), student's thoughts of dropping out measured in 7th grade, whether or not the student have ever been arrested measured in 7th grade, and whether or not the student have ever been expelled. Corresponding to individuals with complete data on the covariates, the analyses consider a subsample of $n = 2757$ of the total of $n = 3116$ individuals. The overall dropout rate in the sample is 14.7%, or 458 individuals. Mplus Version 2.12 was used for the analyses.

Statistical Checking

The univariate skewness and kurtosis sample values in the LSAY data are as follows,

$$Skewness = (\ 0.168 \quad 0.030 \quad 0.063 \quad -0.077\ ), \tag{3}$$

$$Kurtosis = (\ -0.551 \quad -0.338 \quad -0.602 \quad -0.559\ ). \tag{4}$$

In line with the earlier discussion of the LMR LRT, due to the low non normality in the outcomes it is plausible that this tests is applicable in the LSAY analysis for testing a 1-class model versus more than 1 class. In the LSAY analysis, this test points to at least two classes with a strong rejection ($p = 0.0000$) of the 1-class model. The SK tests carried out on the listwise present subsample of $n = 1538$ reject the 1-class model (p values are 0.0000 for both multivariate skewness and multivariate kurtosis), but do not reject 2 classes (p values are 0.4300 and 0.5800). The LMR LRT for 2 versus 3 or more classes obtained a high p value (0.6143) in support of 2 classes. Taken

together, the statistical evidence points to at least 2 classes.[4] Adding the distal outcome of dropping out of high school to the model, however, the LMR LRT rejects the 2-class model in favor of at least 3 classes ($p = 0.0060$). Because the interest is in using the growth mixture model to predict high school dropout, the 3-class solution is chosen. The 3-class solution produces a distinct low class of 19%, a middle class of 28%, and a high class of 52%.

The skewness and kurtosis tests find that already a 2-class GMM fits the data. In such a situation the LMR LRT is useful for testing multi-class alternatives against each other as was done in this application. The earlier discussion of mildly non-normal data, however, suggests that the BC alternative of a single-class model with non-normal outcomes is still possible. Substantive considerations need to guide the analysis and interpretations and this will be considered next.

Substantive Checking

This section reports on analysis results using a conventional 1-class growth model and GMM. Substantive meaningfulness based on educational theory, auxiliary information, and practical usefulness is discussed.

Conventional 1-class Growth Modeling.

As a first step, the conventional 1-class growth model results are considered. Briefly stated, a linear growth model fits reasonably well and has a positive growth rate mean of about 1 standard deviation across the four grades. The covariates with significant influence (sign in parenthesis) on the initial status are: female (+), hispanic

(-), black (-), mother's education (+), home resources (+), expectations (+), dropout thoughts (-), arrest (-), and expelled (-). The covariates with significant influence (sign in parenthesis) on the growth rate are: female (-), hispanic (-), home resources (+), expectations (+), expelled (-).

3-class GMM Including a Distal Outcome.

For the 3-class model it is interesting to consider what characterizes the class of poorly developing students apart from their problematic mathematics achievement. The multinomial logistic regression for class membership indicates that relative to the high class the odds of membership in the low class is significantly increased by being male, black, having low home resources, having low 7th-grade educational expectations, having had 7th-grade thoughts of dropping out, having been arrested, and having been expelled. The low class appears to be a class of students with problems both in and out of school. The profile of the low class is reminiscent of individuals at risk for dropping out of high school (see, e.g. Rumberger & Larson, 1998 and references therein). Many of these students are "disengaged" to use language from high school dropout theories. Interestingly, comparing the middle class to the high class, the disengagement covariates of low educational expectations, 7th-grade dropout thoughts, having been arrested, and expelled are no longer significant. This suggests that the low class is a distinct class, more specifically characterized as disengaged and at risk for high school dropout. The two higher classes may or may not make a substantively meaningful distinction among students, but their presence helps to isolate the low class.

Further bolstering the notion that the low class is prone to high school

dropout, the probability of dropping out as estimated from the 3-class model is distinctly different in the low class. The probabilities are: 0.692 for the low class; 0.076 for the middle class; and 0.006 for the high class. In other words, more than 2/3 of the students in the low class are likely to drop out. Other concurrent and distal outcomes were also added to the 3-class model to further understand the context of the low class, including responses to the 10-th grade question "How many of your friends will drop out before graduating from high school?" (1 = none, 2 = a few, 3 = some, 4 = most.) Treating this as an ordered polytomous outcome influenced by class and the covariates resulted in estimated probabilities for response in either of the three highest categories (few, some, most): 0.259 for the low class; 0.117 for the middle class; and 0.030 for the high class. Considerably more students in the low class have friends who are also thinking of dropping out. In contrast, heavy alcohol involvement in grade 10 was not distinctly different in the low class.

Practical Usefulness.

An educational researcher is likely to find it interesting that the analyses suggest that dropout by grade 12 can be predicted already by end of grade 10 with the help of information on problematic math achievement development. From the point of view of intervention, it is valuable to explore the question of whether a dependable classification into the low class can be achieved earlier than grade 10. GMM can help answer this question. For example, by grade 7 the covariates and the first math achievement outcome are available and given the estimated 3-class model, new students can be classified based on the model and their grade 7 data. GMM allows the

investigation of whether this information is sufficient or if math achievement trend information provided by adding grade 8 information, or grade 8 and 9 information, is needed before a useful classification can be made.

Have these analyses proven that there is a "failing class" of low-performing students who are likely to drop out of high school? No, other alternatives, including that of a single-class model with mildly non-normal outcomes, are still possible. The conventional single-class analysis reported on earlier showed that low initial status and low growth was associated with low home resources, low expectations, dropout thoughts, being arrested, and being expelled. These are the same factors that influence low class membership in the 3-class GMM, so that in line with BC one can argue that the GMM may merely be making an artificial division of the growth factors into a low, medium, and high range. Contrary to the BC results, however, if a single-class model generated the data in this case, the GMM does not fail to find significance of the covariates in their class membership prediction. The two alternatives are not contradictory, but GMM provides an elaboration. Whether or not the division into classes is meaningful is largely a substantive question. An argument in favor of there being a distinct "failing class" is obtained from the distal outcome of high school dropout. The fact that the dropout percentage is dramatically higher for the low class than for the other two, 69% versus 8% and 1%, suggests that the three classes are not merely gradations on an achievement development scale, but that the low class represents a distinct group of students.

The mathematics achievement example illustrates that when put in a

substantive research context drawing on existing theories and auxiliary information, a GMM analysis can give substantively meaningful insights. Educational researchers can look at the evidence and decide if they feel that the low class finding supports the notion of a distinct subgroup of students. They can consider to which extent the low class construction is useful for practical purposes such as prediction of the distal outcome of high school dropout. The initial exploratory analyses can lead to designs including further measures that can shed more light on the hypothesized low class, leading towards more confirmatory analyses.

Conclusion

This commentary on BC has focused on new statistical tests combined with substantive considerations in order to settle on a model that fits the data well and that is useful. It was shown that the non-normal single-class data that BC generated was not well fitted by a GMM so that the alternative 2-class interpretation that BC was concerned about would not have been made on these data. In general, however, BC's point is well taken. There are presumably situations where it is very difficult to tell the two alternatives apart. For example, the BC data may be well fitted by GMM that allows within-class nonnormality of outcomes. In this connection, the large skew and kurtosis values used in the BC data are perhaps more commonly seen in real data when there are strong floor or ceiling effects, a situation not covered in BC. Non-normal GMM taking into account floor and/or ceiling effects was considered in Muthén (2001c).

The commentary argues that there are many examples of equivalent models in

statistics and the equivalence does not necessarily cause a problem, but merely provides different ways of looking at the same data. Substantive theory, auxiliary information, and practical usefulness will continue to have to guide the statistical analysis. GMM allows a very flexible way to look at data that is useful even in cases where the notion of trajectory classes is not well established. Such exploratory GMM analyses can be valuable for theory building, leading to more confirmatory GMM applications.

References

Bartholomew, D. J., & Knott (1999). *Latent variable models and factor analysis* (2nd ed.). London: Arnold.

Bauer, D. J., & Curran, P. J. (2002). Distributional assumptions of growth mixture models: Implications for over-extraction of latent trajectory classes. Forthcoming in *Psychological Methods.*

Leuchter, A. F., Cook, I. A., Witte, E. A., Morgan, M., & Abrams, M. (2002). Changes in brain function of depressed subjects during treatment with placebo. *American Journal of Psychiatry, 159,* 122-129.

Lin, H., Turnbull, B. W., McCulloch, C. E., & Slate, E. (2002). Latent class models for joint analysis of longitudinal biomarker and event process data: Application to longitudinal prostate-specific antigen readings and prostate cancer. *Journal of the American Statistical Association, 97,* 53-65.

Lo, Y., Mendell, N. R., & Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika, 88,* 767-778.

Masyn, K. (2002, June). *Latent class enumeration revisited: Application of Lo, Mendell, and Rubin to growth mixture models.* Paper presented at the meeting of the Society for Prevention Research, Seattle, WA.

Molenaar, P. C., & von Eye, A. (1994). On the arbitrary nature of latent variables. In A. von Eye & C. C. Clogg (Eds.), *Latent variable analysis: Applications for Developmental Research* (pp. 226-242). Thousand Oakes, CA: Sage Publications.

Muthén, B. (2001a). 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. (2001b). Second-generation structural equation modeling with a combination of categorical and continuous latent variables: New opportunities for latent class/latent growth modeling. In L. M. Collins & A. Sayer (Eds.), *New Methods for the Analysis of Change* (pp. 291-322). Washington, DC: APA.

Muthén, B. (2001c). *Two-part growth mixture modeling.* Unpublished manuscript.

Muthén, B. (2002a). *Growth mixture modeling.* Manuscript in preparation.

Muthén, B. (2002b). Beyond SEM: General latent variable modeling. *Behaviormetrika, 29,* 81-117.

Muthén, B., & Asparouhov, T. (2002). *Mixture testing using multivariate skewness and kurtosis.* Manuscript in preparation.

Muthén, B., Brown, C. H., Masyn, K., Jo, B., Khoo, S. T., Yang, C. C., Wang, C. P., Kellam, S., Carlin, J., & Liao, J. (2002). General growth mixture modeling for randomized preventive interventions. *Biostatistics, 3,* 459-475.

Muthén, B., & Curran, P. (1997). General longitudinal modeling of individual differences in experimental designs: A latent variable framework for analysis and power estimation. *Psychological Methods, 2,* 371-402.

Muthén, B., Leuchter, A., & Morgan, M. (2002). *Assessing medication effects in the presence of placebo response: A clinical trial application of growth mixture*

modeling. Unpublished manuscript.

Muthén, B., & Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics, 55,* 463-469.

Muthén, L., & Muthén, B. (1998-2002). *Mplus User's Guide.* Los Angeles: Muthén & Muthén.

Rumberger, R. W., & Larson, K. A. (1998). Student mobility and the increased risk of high school dropout. *American Journal of Education, 107,* 1-35.

Wang, C. P., & Brown, C. H. (2002). *Residual diagnostics for growth mixture models: Examining the impact of a preventive intervention on multiple trajectories of aggressive behavior.* Manuscript submitted for publication.

Author Note

Bengt Muthén, Graduate School of Education & Information Studies, Social Research Methodology Division, University of California, Los Angeles.

This research was supported under grant K02 AA 00230 from NIAAA. I thank Katherine Masyn for expert research assistance and Tihomir Asparouhov, the members of my Research Apprenticeship Course, and Hendricks Brown for helpful comments.

Correspondence concerning this article should be addressed to Bengt Muthén, University of California, Los Angeles, Graduate School of Education and Information Studies, Social Research Methodology Division, 2023 Moore Hall, Mailbox 951521, Los Angeles, California 90095-1521. E-mail: bmuthen@ucla.edu.

Footnotes

[1]The parameter values are as follows. A linear growth model is considered with time scores $0, 1, \ldots, 5$, class 1/class 2 intercept factor means of 0/2, slope factor means of 0.25/0.25, intercept factor variances of 1/1, slope factor variances of 0.25/0.25, intercept-slope covariances 0/0, R^2 for the outcomes of 0.8, and class 1 probability of 0.7.

[2]See section Nonnormality and the Estimation and Fit of Latent Trajectory Classes, which contains a discussion of doing GMM on multivariate normal data

[3]A 3-class model also strongly underestimated the sample skewness and kurtosis in the BC data.

[4]AIC points to at least 3 classes, while BIC points to 2 classes. The 1-class log likelihood, number of parameters, AIC, and BIC values are: $-30,021.955$, 27, $60,097.909$, and $60,257.791$. The 2-class log likelihood, number of parameters, AIC, BIC, and entropy values are: $-29,676.457$, 63, $59,478.914$, $59,851.971$, and 0.552. The 3-class log likelihood, number of parameters, AIC, BIC, and entropy values are: $-29,566.679$, 99, $59,331.359$, $59,917.591$, and 0.620.

Table 1

Sample and Model-Estimated Univariate Skewness and Kurtosis for a 2-class Growth Mixture Model (Averages Across 10 Replications at n = 600)

Skewness 1, Kurtosis 1				
Outcome	Sample Skew	Mixed Skew	Sample Kurt	Mixed Kurt
y1	0.951	0.429	0.878	0.161
y2	1.027	0.566	1.003	0.177
y3	1.006	0.552	0.873	0.122
y4	0.994	0.580	0.942	0.169
y5	1.018	0.534	0.997	0.148

Skewness 1.5, Kurtosis 6				
Outcome	Sample Skew	Mixed Skew	Sample Kurt	Mixed Kurt
y1	1.567	0.893	5.798	1.873
y2	1.317	0.884	4.232	1.749
y3	1.219	0.922	3.402	1.803
y4	1.545	0.912	5.924	1.926
y5	1.604	0.845	6.187	1.836