Exploring the Measurement Profiles of Socioeconomic Background and their differences in Reading Achievement: A Two-level Latent Class Analysis

Kajsa Yang Hansen & Ingrid Munck; Department of Education, University of Gothenburg

Background and Study Purposes

A large body of SES-Attainment studies has established the relationship between student's socioeconomic background and academic achievement, making SES a single powerful predictor of student's performance. Typically SES is measured as a composition of parental education, occupation and income. However, a variety of alternative measurement of SES has also been developed (White, 1982; Sirin, 2005). The general trend of SES measurement is moving from SES being seen as a single index to regarding it as a multi-dimensional concept that imposes its effects at different levels of observation. The measures of SES also have extended from the aforementioned aspects to a set of rather broad indicators of home possessions and activities. Due to the variability in the measures and measurement methods of SES, the estimated effects of SES on academic achievement vary and become difficult to compare and evaluate. It is thus necessary to reexamine the measurement of SES by exploring the psychometric properties of SES.

Throughout the history of IEA studies, three major types of the measurement of SES have been found in the SES-effect studies: (1) SES is taken as an observed index or a component of its indicators; (2) SES is taken as a uni-dimensional latent variable; (3) SES is taken as a multi-dimensional multilevel concept. These views of SES followed tightly with the development of computer technology and statistical analytical software.

In the early 70s´Jöreskog developed the first computer program for analysis of Linear Structural Relationships (LISREL, Jöreskog & Söbom, 1979). Munck (1979) criticized the use of a component model for measuring SES or home status, and fitted a confirmatory factor analysis model to the SES indicators available in the Six Subject Study, namely, parental education, father´s occupation and number of siblings. This way of measuring SES achieved a major increase of explanatory power of SES in the variation of reading achievement, compared to the regression approach.

With the help of two-level structural equation modeling technique, Yang (2003) identified an economic capital factor and a cultural capital factor at individual-level and a general capital factor at school-level. However, this general latent structure of SES was found to differ largely across countries and across observation levels. Such differences are reflections of the country-specific social, cultural, and economic situation. It was also due to the differences in the availability of SES data.

Muthén & Muthén (2009) implemented the multilevel mixture modeling technique in the Mplus program, which opened new possibilities for exploring the psychometric profiles of SES. Unlike the previous view of SES simply being an observed index or a latent continuous construct, this measurement approach conceptualizes SES by forming distinct categories or typologies of it. The latent class approach takes into consideration the measurement error in the response patterns of the SES indicators, and bases the categorizations on the prior and posterior probability distributions under the conditioned maximum likelihood estimation (Muthén, 2007).

The aim of the proposed study is thus to apply the two-level latent class analysis to identify the unobserved categories of individuals, based on a set of SES measures in IEA PIRLS 2006 data, and to examine the reading achievement differences regarding the different latent classes.

Methodology and Data Sources

Sample and Variables

The Swedish data of the IEA PIRLS 2006 were used in the current analysis. All together there are 4 393 4th graders and 147 schools in the sample. Variables that indicate the family socioeconomic status (SES) and school composition of its student intake characteristics were selected from Student Questionnaire (StQ), Home Questionnaire (HQ) and School Questionnaire (SQ), in order to identify latent SES classes of individuals and schools accordingly. International standardized students' reading achievement (ASRREA01) was included to examine the mean achievement differences among different latent classes. A set of auxiliary variables at student-levels were also included to describe the characteristics of the estimated latent classes.

Table 1 and Table 2 showed the descriptive statistics of all variables included in the analysis. The number of observations of certain categories in some variables is fairly low, which may cause the low cell sizes in the cross-tabulation of the categories among different variables and impossibility to estimate the logliklihood statistics. Therefore, the categories with few observations were reclassified into one larger category of the variable.

Insert Table 1 about here

In Table 1, SES indicators and the covariates ASBGLNG1 and INTAKECH at the individual respective the school levels were presented. The SES indicators like parental highest educational level, number of books at home, number of children's book at home, and family well-off status, parental occupation, were taken from HQ. A sum score of the 5 internationally common possession items from StQ (i.e., PC, Desk, Own books Newspapers and Own room) was calculated and recoded into a 3-scale variable representing the level of educational aids at home (EDUAIDS).

Among these SES indicators, two categories of variables can be distinguished. One category, according to Bourdieu (1984, 1997), is the objectified state of cultural capital, which signifies the cultural preference in people's everyday life. These variables are about the amount of books at home and the highest educational level of parents. Another category is related to family wealth (i.e. economic capital). Information on family well-off status, parental occupations and educational aids for their children indicated the economic and status aspects of SES. Frequencies of individuals in each category of the variables as well as number of missing in each variable were also presented in Table 1.

The variable ASBGLAN1 (i.e., use the test language at home) was from StQ and used as covariate at individual-level to further describe the estimated latent SES classes. INTAKECH is a factor score variable, being derived from School Questionnaire variables that characterize school intakes economic and ethnic background. These variables are percentage students that are from economically disadvantaged family (ASBGPST1); percentage students who are from affluence family (ASBGPST2); percentage students who do not speak the test language at home (ASBGPST3); and percentage students who receive some instructions in their home language (ASBGPST4). This intakes' character variable was used as covariate to affect the school-level latent classes.

Other background variables that were in the further description of individual latent classes were presented in Table 2, all of which were in their original scale and individual-level variables taken from StQ or HQ. When the number of latent classes was determined and the class membership of each individual in the sample can therefore be saved. The frequencies of the variables in Table 2 can thus be compared among the latent classes. The results of the comparison can validate the interpretation of the latent SES classes, and can further describe the class characteristics accordingly.

Analytical Method

Latent class analysis (LCA, see e.g., Hagenaars & McCutcheon, 2002) was used in the current study to identify small number of homogenous groups of individual students and schools inferred from the properties of the set of SES measures. A latent class is characterized by a pattern of the conditional

probability that indicates the degree of association of the observed dependent variables to the each of the latent classes, analogous to the factor loadings. The conditional probability signifies the probability of an individual in a latent class responded in a certain pattern in the set of observed indicators. By looking at the pattern of responses for all SES measures, one gets an overview of the nature of each of the latent classes and helps us to interpret the relationship between these latent classes and other outcome variables such as reading achievement.

Insert Table 2 about here

An important aspect of the LCA is to investigate the characteristics of individuals within each latent class by relating the latent classes to the so called auxiliary variables, e. g., the covariates concurrent outcomes and distal outcomes. The mean comparison test of the auxiliary variables in the latent classes can be done to get a further understanding of the characters of individuals in each of the latent groups.

Like the conventional covariance structure modeling, the observations are assumed to be independent of each other in LCA. However, due to the stratified multi-stage clustered sampling design in IEA PIRLS study, this assumption was violated. For example, students in the same school will achieve similarly, compared to students from different schools. Therefore, Multi-level technique was needed to account such independency (e.g., Asparouhov & Muthén, 2008; Vermunt, 2003). Multi-level latent class model (MLCM) was applied in the current study and it allowed not only examining the multilevel structure of the data but also allow simultaneously detecting the subgroups of individuals according to some observed data properties. In other words, the MLCM takes into account the nested structure of the data and allows the latent class intercepts differ across second level units. In this way it can detect if and how the second level unites influence the lower level latent classes (for a detailed example, Henry & Muthén, 2009). In order to find the most fit random effect model, the modeling process followed an exploratory manner. The parameters in the MLCM were estimated by the maximum likelihood estimator with robust standard error, which is available in Mplus with the combination of missing data analysis. The current analyses were carried out by computer program Mplus (Muthén & Muthén, 1998-2010).

Process of analysis

First, individual-level LCA was performed with 6 SES measures, namely the recoded 3-scale categorical variables ASBHWELL, ASBHHEDUP, ASBHHOCP, ASBHBOOK, ASBHCHBK, and EDUAIDS. To determine the optimal number of latent classes, the LCA model was estimated in a stepwise manner, i.e., a 2-class LCA model was fitted in the first step, and an additional class was added sequentially, until the model fitted the data well. Each model was estimated by different sets of starting values to prevent local maximum in iteration processes. Based on the Bayesian Information Criterion (BIC; Schwartz, 1978; see also, Kass & Wasserman, 1995), the models were compared. BIC is the best indicator for enumeration of latent classes over the rest of information criterion measures. The model with lowest BIC was preferred. Further assessment for whether the number of classes chosen was correct, the Vuong–Lo–Mendell–Rubin likelihood ratio test (LRT) of each model was applied. LRT compares the logliklihood differences between the model with k classes with the one with k-1 classes. A significant improvement in logliklihood difference implies that the k-class model fit the data better. TECH 11 in Mplus gives such LRT test (Muthén & Muthén, 1998-2010; Nylund, Asparouhov, & Muthén, 2007).

When the number of latent classes has been decided at individual-level, auxiliary variables i.e., reading achievement, were related to the latent class variable to test the mean difference of reading achievement across the latent classes using posterior probability-based multiple imputations.

In the next step, the individual level LCA model was extended to a two-level mixture model to capture the randomness in the probability of membership of the individual-level latent class across schools, i.e., the school variations in the latent class intercepts. A school-level latent factor was thus

identified by the continuous latent class intercepts, and a school-level covariate, INTAKECH (i.e., the between-school differences in intake's SES composition) was included to predict the random effect factor. Finally, the school-level latent factor and the covariate was related to reading achievement to account for the between school differences in the outcome variable. The current analysis was carried out by Mplus version 6 (Muthén & Muthén, 1998-2010).

Findings

Results from Individual-Level LCA Analysis

Exploratory LCA model was fitted at individual-level to classify individuals into possible latent groups. Figure 1 showed the changes in the Bayesian Information Criterion (BIC) in LCA models. The BIC value was decreasing dramatically from 1-class LCA model to 4-class LCA model, and stabilized between 4-class and 5-class model.

Insert Figure 1 about here

It should be noted that entropy level in each of the LCA models was rather low, round .67. Simulation study suggested that an entropy level of .80 or over is desirable to ascertain the precision and the usefulness of the estimated latent classes in classifying individuals. However, the lower level of entropy can be expected in the situation that the latent class analysis is done in an exploratory manner, which is the case here (Lee Van Horn, *et al.,* 2009). Moreover, entropy level is influenced by measurement properties of indicators, for example, misclassification in categorical variables is one of the common causes of such a problem. However, this problem can be remedied by indicator evaluation in comparative latent class analysis (Kreuter, Yan & Tourangeau, 2008).

Insert Table 3 about here

Table 3 showed the *p*-values for LRT test and the adjusted LRT test that were achieved from a series of LCA models. The BIC difference between 4- and 5-class LCA models was marginal. The p-values for LRT and adjusted LRT were non-significant, which suggested, however, that the 4 classes was the sufficient number of latent groups of individuals according to the combined characteristics of the SES measures in the LCA model.

Conditional probabilities were estimated for each category of SES measure and were included in to each latent class. Table 4 showed the posterior probability of answering "3", for each measure of SES in different latent classes. The response alternative 3 for the individual-level SES indictors are "finished university education or higher", "more than 100 books", "more than 100 children's books at home", "well-off", "professional work", "have all 5 home possession items" (see Table 1 for detailed descriptions of variables and labels).

It was easy to notice in Table 4 that the conditional probabilities for answering 3 in all SES indicators were much higher in latent class 1, compared to the rest of the classes. And the opposite pattern was found in latent class 2. The probabilities to answer 3 in the indicators of economic aspect of SES (i.e., well-off status, home possession index, and occupation) were fairly similar in Class 1 and 3. However, those of the cultural aspects of SES (i.e., books at home and parental education) were much lower in latent classes 3. The probabilities for the individuals in the latent class 4 who answered 3 in the SES indicators were at inter-medium to low level except for those of book variables.

Insert Table 4 and Figure 2 about here

It is thus the combination of the levels of posterior probabilities of SES indicators determined the latent classes. Figure 2 showed the graphical profiles of each latent class of individuals according to the estimated probabilities. For latent class 1, the posterior probabilities were very high on all 6 SES indicators, which implied that the individuals who belong to this latent class are very likely coming

from economically and culturally affluent family. The latent class 1 can be called economic and cultural affluence group. For the latent class 2, the posterior probability was low on all the 6 SES indicators, meaning that the individuals in this group are very likely from economically and culturally disadvantaged family. And this latent class can be named as disadvantaged group. The pattern of posterior probabilities for latent class 3 were rather similar as those of economic capital indicator in latent class 1, their cultural capital were however much lower. This latent class was thus named economic affluence group. Finally, the latent class 4 can be seen as a book-loving group since the amount of books and children's books at home made the distinction between latent class 2 and 4.

Based on the most probable membership to the latent class, individuals were distributed into the four latent classes. Class 1 contained the most 4th graders in the sample, about 1939 students (44%). The second largest group was class 4 who has 30% of the samples, 1337 individuals. And the number of individuals in class 2 and 3 were much less, being 583 (13%) and 835 (12%) respectively.

Auxiliary variable reading achievement (ASRREA01) was brought into the LCA model. It should be noted that an auxiliary variable is not involved in the determination of the latent classes, rather to use as a further description of the latent classes. In Mplus, the mean achievement differences in reading score among the latent groups were tested. The average reading achievement is 569.9 for Class 1, 513.9 for Class 2, 550.2 for Class 3, and 536.6 for Class 4. The mean achievement differed significantly among these classes. Compared to observed statistic properties in the achievement variable in Table 1, class 3 was the average achieving group in reading.

To get more detailed information on these 4 SES latent groups, the estimated class membership were saved and merged with the original data file in SPSS format. Figure 3 showed the percentage distribution of the highest category in some background variables in the 4 latent SES classes (for information on the highest category, see Table 2).

As shown in Figure 3, the patterns of distribution of these observed background variables in each of the latent classes held the same as those was found for the latent classes in Figure 2. For the economic and cultural affluence group (i.e., the latent class 1), the percentage of individuals in the highest category of the set of background variables were at the top of all other groups. And the distributions for the disadvantaged group held the lowest profile. For this group, the percentage individuals whose parents were born in the country and speak the test language at home were the lowest among all latent SES classes. This suggested that the disadvantaged group was an immigrant-concentrated group, who has the lowest reading achievement level. The book-loving group was very similar to the immigrant group except for the large amount of books at home and higher amount of native born. It can be assumed that the difference in the mean level of reading achievement was due to the differences in the above-mentioned aspects between the two latent groups.

In the next step, a covariate ASBGLNG1 and the reading achievement variable were included in the model (see Appendix I for the model diagram). In this model, the covariate variable was involved in the determinacy of the latent classes. However, the nature of the 4 latent classes kept intact so as the average reading achievement level in each of the latent classes. The probability for individuals whose home language being other than the test language being included in class 2 was over 2 times higher than that in the reference class 1. There was no difference in the probability of the membership among other classes concerning the language used at home. This result was confirmed by the patterns found in figure 3.

In sum, the individual-level LCA found a stable 4-class solution for the classification of students according to their SES indicators, and the reading achievement differed significantly among the four latent groups. The distribution pattern of the highest category in some background variables in each of the estimated latent class confirmed the interpretation of the nature of the latent classes, namely, economic and cultural affluence group, immigrant-dominant group, economically well-off group and the last, book-loving group. And for these latent classes, significant cross-class differences were observed in reading achievement.

Results from Two-Level Mixture Modeling

The data collection in PIRLS 2006 was based on the stratified cluster sampling design, which implied that the data achieved has a hierarchical structure. Multilevel latent class analysis approach thus has to be applied to account for the biased standard error estimation. Moreover, due to the increasing degree of residential segregation and more frequent practices of choice of schools, increasing amount of between-school differences in school SES composition has been observed (see e.g., Björklund et al., 2005). This made, however, that the probability of belonging to one of the 4 latent SES classes to differ for student intakes in different schools. Thus, the probability for students belongs to, for example, the immigrant-dominant SES class is likely to vary significantly across schools.

To capture the randomness of the individual-level probability of class membership, a between-level latent variable was thus identified. The identification process was rather exploratory, meaning that the nature of the second level latent variable (i.e., whether it is a categorical or a continuous latent variable) was determined by model fit statistics. It was showed that a continuous latent variable of the random effect model at school-level fitted the data best (see Appendix II for the model fit statistics in all the tested models, including individual-level LCA models). It is often the case that only one factor is identified representing the randomness of the probabilities of the individual latent classes, due to the fact that the variation of these estimated thresholds tend to be very small.

The random effect model was depicted by Figure 4 below. However, the model was not the best fit model in terms of BIC measure. According to BIC measures for the 3 random effect models (see Appendix II); the highly constrained model should be chosen as the final model. However, the meaningfulness of the model representation of reality, as being the case here, was used as criteria to select the final model.

And the continuous latent variable f was assumed to be influenced by the SES composition of student intake body at school INTAKECH – the school-level covariate. In turn, both affected the average reading achievement level at school. In the within part of the model, the filled dots attached below the individual-level categorical latent variable cw represented the three random means of the four classes of cw, the latent SES class variable. They were referred to as cw#1, cw#2, and cw#3 shown in circles at the school-level model, because they are continuous latent variables that vary across schools. The random mean of the latent class 4 was zero because it was set to the reference group. The within-level part of the model was very relaxed, allowing the cross-class differences in ASBGLNG1 as well as the impact of ASBGLNG1 on reading achievement.

Insert Figure 4 about here

The model results showed that the nature of the 4 latent SES classes at the individual-level were rather stable, with only slight changes of the percentage student in each groups, after taking into account hierarchical nature of the data as well as the covariates ASBGLNG1 and INTAKECH at the student- and school-level respectively. Figure 5 showed the thresholds for category 1 in each SES indicator to belong to each of the four latent SES classes.

Insert Figure 5 about here

Comparing this distribution with the one showed in Figure 2, it can be noted that the order of the classes was different. However, four rather similar latent classes were identified. Based on the estimated mean probabilities for the lowest category in each SES indicators in each of the four latent classes, we found that the Class 4 being the low SES immigrant-dominant class, similar to the class 2 in Figure 2. The Class 1 in Figure 5 could be characterized as the economic and cultural affluence group, similar as the class 1 in Figure 2. And the Class 2 in Figure 5 was very similar with the Class 1 in probabilities for the economic indicators of SES, making this latent class an economically well-off group. This group was correspondent to the class 3 in Figure 2. Finally, the Class 3 found in the two-level random effect model was similar to the book-loving group in the individual –level analysis.

Table 5 showed the mean reading achievement for each latent class and the distribution of students in each latent class. The economic and cultural affluent group (Class 1) achieved the highest in the reading test and the immigrant-dominated group (Class 4) the lowest. For the two other latent SES classes (Class 2 and Class 3) the average achievement in reading test differed only about 2 points after taking into account the school SES composition of its students and the use test language at home, which may imply that the higher amount of books at home compensate the effect from otherwise disadvantage home background and make the students achieve a higher level of reading ability, almost 30 points higher than the group of students who has the similar level of SES but much less books, as those in Class 4.

Insert Table 5 about here

It was showed in Table 6 that the two-level LCA that the individual-level covariate i.e., the uses the test language at home variable (ASBGLNG1, a variable indicating student's immigrant background) has different impact on both the probability to belong to each latent SES group and the reading achievement. With the latent class 4 as reference group, it is over 10 times grater chance for a native student belong to Class 3 than to Class 4, over 6 times greater chance for a native student belong to Class 2 than to Class 4, and almost 8 times grater chance for a native student belong to Class 1 than to Class 4. And after controlling for such group differences, the only impact of the use test language at home on reading achievement that was found is in the latent class 4.

Insert Table 6about here

The random effect factor f at the school-level captured the differences in the average probability of belonging to each of the 4 latent classes has no impact on school achievement; neither did the school-level covariate, i.e., the school SES composition of student intakes. However, the school SES composition of student intakes was found to affect the randomness of the average probability of belonging to each of the latent SES classes.

Conclusion and Discussion

Two-level latent class analysis (TLCA) has showed in the current study an alternative approach to classify individuals into different unobserved groups according to a set SES indicators, taking into account simultaneously the hierarchical data structure and the randomness of the individual classification across collective level units (i. e., students within schools).

Substantively, the present study explored the psychometric profiles of SES and examined the reading achievement differences according to the latent profile belongingness by applying such TLCA techniques. Unlike the previous view of SES simply being an observed index of all its indicators or as being a latent continuous construct, this measurement approach conceptualized SES by forming distinct and homogenous categories or typologies of it. 4 latent classes of individuals were found according to the profile of SES indicators. Reading achievement differed significantly among the 4 latent SES classes, namely, the economic and cultural affluence class, the immigrant-dominant class, the economic well-off class and the book-loving class.

One of the interesting findings of the study is that the belongingness of different latent SES classes was affected by the non-native indicator. It is showed that speaking the test language at home does not affect student's reading achievement as long as the student is not in an immigrant-dominant environment. This raised the concern of the increasing segregation in both schools and society in Sweden with respect of social and immigration background (Skolverket, 2009).

Due to the increasing segregation in Swedish compulsory schools, the school intakes may have different chance to be categorized into different latent SES classes. The school-level continuous factor captured the probability variation across schools. This assumption was confirmed by the significant impact found of the school SES composition on the random effect factor. However, no effect was

found from these aspects on school reading achievement.

The explanation for not finding any significant impacts of the two above-mentioned aspects reading achievement could be that the random effect factor at school-level and the school SES composition are highly confounded aspects and when the correlation between the two was controlled, no direct effect could be detected. It can also be that the school SES composition was measured with errors and when used in the analysis, it is very often the case that the measurement error may attenuate the correlation and makes the relationship non-significant. Therefore, more covariates at both individual and school levels are warranted in the future improvement of the study.

It should also be noted that the entropy level in the current single level and two-level latent class models were rather low, given the preferred level of at least .80. This may imply that the accuracy and usefulness of the classification of the latent classes is rather low. Moreover, the final model did not have an optimal fit, rather the substantive meaningfulness was used as the model selection criterion. Further attempt of adding more covariates in the model, as mentioned before, and to test different parameter constrains, may be remedial for the problems.

In order to get better understanding of the non-significant findings in the current study and to get confirmation about the feasibility and accuracy of the method in the current study in classifying students into unobserved SES groups, multiple countries that with similar circumstances in society and schools have to bring into the analysis. Other steps also could be done to improve the quality and accuracy of the model and estimates. One is to include auxiliary variables into the MLCA model. These auxiliary variables will not involve in the latent class estimation, rather they help to describe the nature of the latent classes. One advantage to have auxiliary variables in the model is that it can simultaneously test the latent class differences in these auxiliary variables, which is a much convenient way, compare to the approach we used in this study.

References

Asparouhov, T. M., B. (2008). Multilevel mixture models. In G. R. Hancock, & Samuelsen, K. M (Ed.), *Advances in latent variable mixture models* (pp. 27-51). Charlotte, NC: Information Age Publishing, Inc.

Björklund, A., Clark, M., Edin, P.-E., Fredriksson, P., & Krueger, A. B. (2005). *The Market Comes to Education in Sweden. An Evaluation of Sweden's Surprising School Reforms*: Russel Sage Foundation.

Bourdieu, P. (1984). *Distinction: a social critique of the judgment of taste*. Cambridge, Mass.: Harvard University Press.

Bourdieu, P. (1997). The Forms of Capital. In A. H. Halsey, Lauder, H., Brown, P. & Wells, Stuart, A. (Ed.), *Education: culture, economy, and society* (pp. 46-58). Oxford: Oxford University Press.

Hagenaars, J., & McCutcheon, A. (Eds.). (2002). Applied latent class analysis models. New York: Cambridge University Press

Henry, K. M., B. . (2009). Multilevel latent class analysis: An application of adolescens smoking typologies with individual and contextual predictors. *Structural Equation Modeling*.

Jöreskog, K. G. & Sörbom, D.(1979). *Advances in factor analysis and structural equation models.* Cambridge, MA: Abt Books.

Kreuter, F., Yan, T., & Tourangeau, R. (2008). Good Item or Bad – Can Latent Class Analysis Tell?: the Utility of Latent Class Analysis for Evaluation of Survey Questions. *Journal of Royal Statistical Society, 171*(3), 723-738.

Lee Van Horn, M., Jaki, T., Masyn, K., Landesman Ramey, S., Smith, J. A., & Antaramian, S. (2009). Assessing Differential Effects: Applying Regression Mixture Models to Identify Variations in the Influence of Family Resources on Academic Achievement. *Developmental Psychology, 45*(5), 1298-1313.

Munck, I.M.E. (1979): Model building in comparative education: Applications of the LISREL method to cross-national survey data. IEA studies Nr10, Stockholm: Almqvist & Wiksell International.

Muthén, L. K. & Muthén, B. O. (1998-2009). *Mplus User's Guide. Fifth Edition.* Los Angeles, CA: Muthén & Muthén.

Muthén, B. (2007) Latent Variable Hybrids - Overviews of Old and New Models. In Advances in Latent Variable Mixture Models. Information Age Publishing

Nylund, K. E., Asparouhov, T. & Muthén, B. O. (2007). Deciding on the Number of Classes in Latent Class Analysis and Growth Mixture Modeling: A Monte Carlo Simulation Study. *Structural Equation Modeling, 14*(4), 535–569

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A Meta-analytic review of research 1990-2000. *Review of Educational Research,* 75*(3),* 417-453.

Schwartz, G. (1978). Estimating the dimension of a model. The Annals of Statistics, 6, 461–464.

Skolverket (2009). Vad påverkar resultaten i svensk grundskola? Kunskapsöversikt om betydelsen av olika faktorer. Stockholm: Skolverket.

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

White, K. R. (1982). The relation between socioeconomic status and academic achievement. *Psychological Bulletin*, 91, 461.

Yang, Y. (2003). *Measuring Socioeconomic Status and its Effects at Individual and Collective Levels: A Cross-Country Comparison.* Acta Universitatis Gothoburgensis, Gothenburg Studies in Educational Sciences 193, Gothenburg.

Table 2. Auxiliary variables in the current Latent Class Analysis.

Variables		Source	Highest category (no. of categories)	Missing
ASBHMJF	Main job of father	HQ	Professional (11c)	727
ASBHMJM	Main job of mother	HQ	Professional (11c)	682
ASBGTA5	Own room	StQ	Yes (dummy)	84
ASBHLEDF	Highest education of father	HQ	Beyond the ISCED level 5A (7c)	832
ASBHLEDM	Highest education of mother	HQ	Beyond the ISCED level 5A (7c)	861
ASBGBOOK	How many books at home	StQ	Over 100 books (5c)	146
ASBGTA4	Daily newspaper	StQ	Yes (dummy)	97
ASDHHER	Index home educational resources	StQ	High (3c)	344
ASBGBRNM	Mother born in the country	HQ	Yes (dummy)	200
ASBGBRNF	Father born in the country	HQ	Yes (dummy)	159
ASBGLNG1	Use the test language at home	StQ	Yes (dummy)	162

Table 3. The p-values for Vuong-Lo-Mendell-Rubin test (LRT) and Entropy level for Latent class analysis models of individual-level SES.

	1- vs. 2-class	2- vs. 3-class	3- vs. 4-class	4- vs. 5-class
p-value for LRT	.0000	.0002	.0006	.7407
p-value for adjusted LRT	.0000	.0002	.0007	.7414

Table 5. The mean reading achievement and the distribution of students in each latent class according to the most likely class membership estimate.

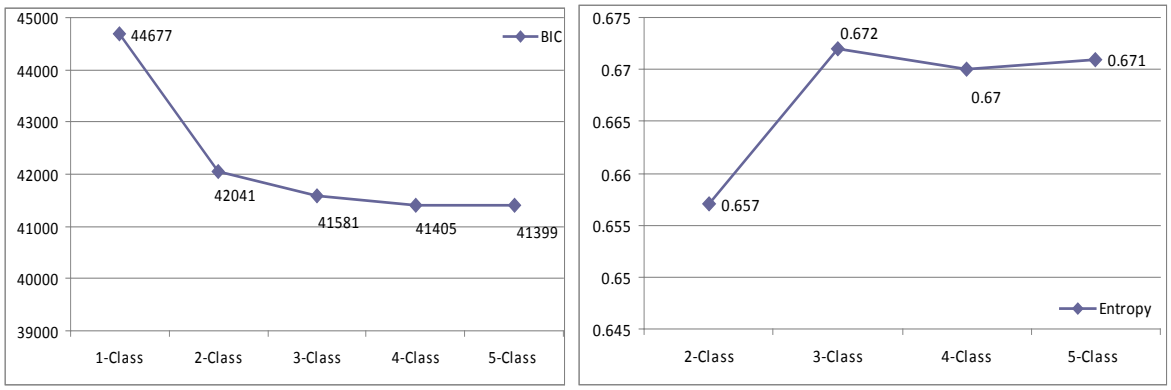
	Mean outcome	No. of cases	Percentage
Class 1	556.646	1687	39.9
Class 2	522.461	497	11.7
Class 3	520.349	1435	33.9
Class 4	492.659	611	14.5

Table 6. Effects of individual-level covariate ASBGLNG1 on latent class variable and on the reading achievement.

		Odds Ratio		Regression Coefficient
Class1	ASBGLNG1 → cw#1	7.94	ASBGLNG1 → EYPV1	0.06(ns)
Class2	ASBGLNG1 → cw#2	6.51	ASBGLNG1 → EYPV1	0.14(ns)
Class3	ASBGLNG1 → cw#3	10.14	ASBGLNG1 → EYPV1	0.04(ns)
Class4	Reference group	-	ASBGLNG1 → EYPV1	0.10

Note: “ns” = not statistically significant. cw#1, cw#2, and cw#3 are thresholds for latent class 1, 2 and 3 respectively.

Figure 1. Changes in the Bayesian Information Criterion (BIC) and Entropy in LCA models of individual-level SES.



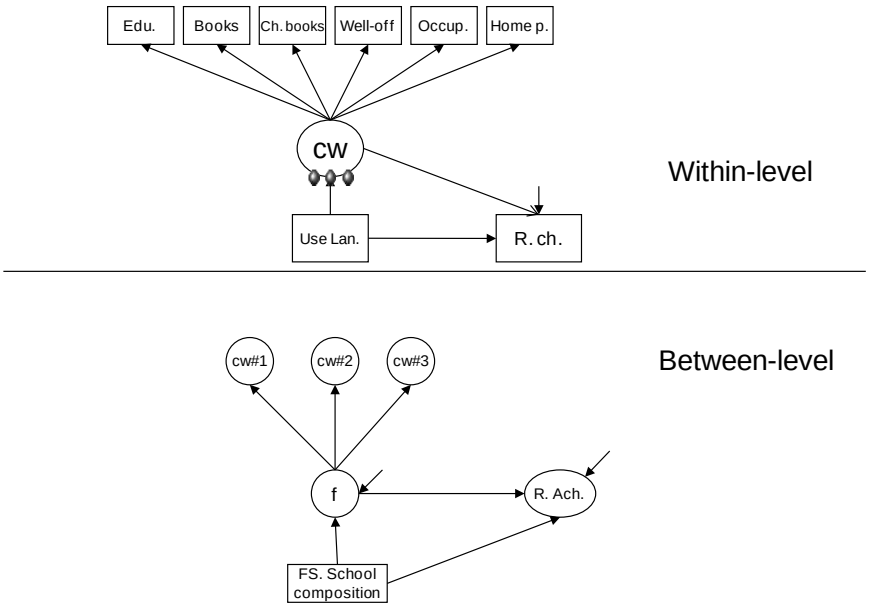
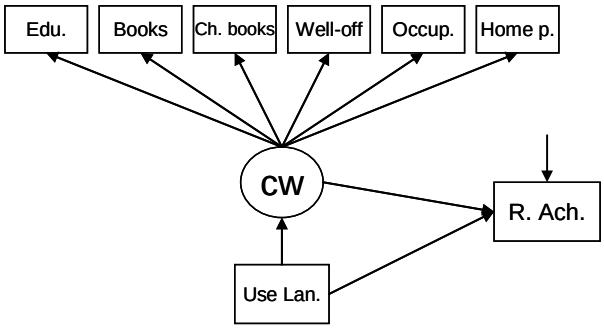


Figure 4. Two-level Mixture model of socioeconomic status and Reading achievement.

Appendix I. The path diagram for individual-level Latent class Model with a covariate variable ASBGLNG1 and an outcome variable EYPV1.



Appendix II. Model evaluation of LCA and MLCA models that were estimated in the current analysis.

Fit Criteria	Individual-level latent classes				
Level-1 model only (Complex with weight)	1-Class	2-Class	3-Class	*4-Class*	5-Class
No. of free parameters	12	25	38	51	64
Loglikelihood	-22288	-20915	-20631	-20488	-20430
BIC	44677	42041	41581	41405	41399
Entropy		.66	.67	.67	.67
p-value for LRT					
2-Level model with level-2 latent class variable (cb1), weighted					
No. of free parameters			38	*51*	64
Loglikelihood			-20463	*-20325*	-20275
BIC			41244,834	*41078*	41087
Entropy			0,68	*0,68*	0,64
2-Level model with level-2 latent class variable (cb2), weighted					
No. of free parameters			41	55	
Loglikelihood			-20463	-20327	
BIC			41270	41115	
Entropy			0,72	0,47	
2-Level model with level-2 latent class variable (cb3), weighted					
No. of free parameters			44	59	
Loglikelihood			-20463	-20327	
BIC			41295	41149	
Entropy			0,54	0,68	
2-level model, at 2-level, a continuous factor f with covariates x and w, and outcome variable. Constrained model at within level (weighted)					
No. of free parameters				*65*	
Loglikelihood				*-42461*	
BIC				*85465*	
Entropy				*0.67*	
2-level model, at 2-level, a continuous factor f with covariates x and w, and outcome variable. Relaxed x cw correlation at within level (weighted)					
No. of free parameters				67	
Loglikelihood				-42460	
BIC				85479	
Entropy				0,67	
2-level model, at 2-level, a continuous factor f with covariates x and w, and outcome variable. Relaxed all constrains at within level (weighted)					
No. of free parameters				73	
Loglikelihood				-42450	
BIC				85511	
Entropy				0,67	

Note: 1. Covariate at student-level x=ASBGLNG1. Covariate at school-level w=INTAKECH. Outcome variable: y=EYPV1. 2. The models in Italic were the ones were accepted as the final models. For the individual-level model, a 4-class model was the best fit model. For the two-level LCA model, a constrained random effect model with 4 classes at level-1 and one continuous factor at level-2, controlled by covariates from both level and

Table 1. The descriptive information of all variables included in the LCA models of SES.

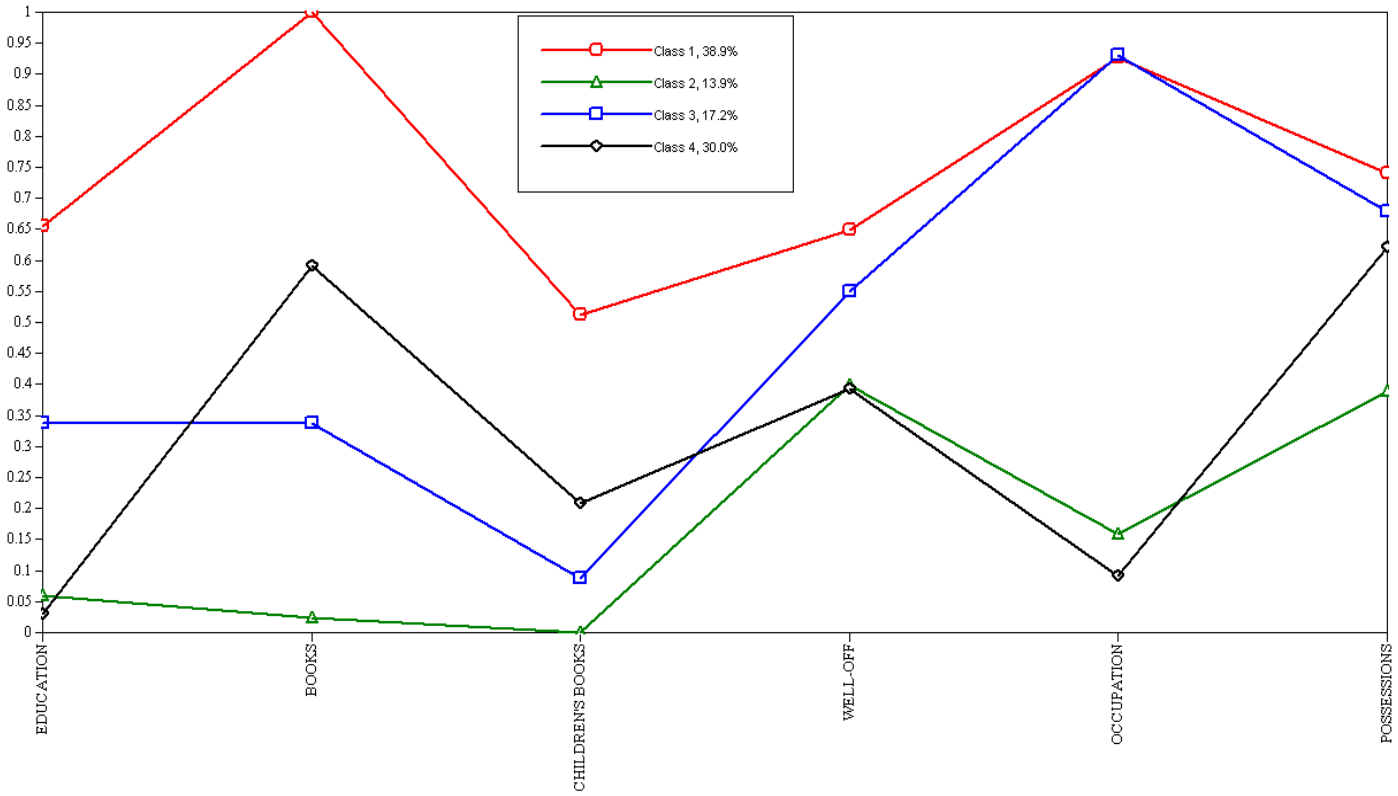
VARIABLES	Label	Source	Scale	Value labels (percent individuals in each category)	Missing
ASBHWELL	Being well-off family	HQ	3-scale	1 = Not well-off (7%); 2 = Average (41%); 3 = Well-off (52%).	401
ASBHHEDUP	parental educational level	HQ	3-scale	1 = finished upper secondary school (29.2%); 2 = finished post-secondary but not university (36.6%); 3 = finished university or higher (34.2%).	696
ASBHBOOK	Books at home	HQ	3-scale	1 = less than 25 (10.4); 2 = 26-100 (26.7%); 3 = more than 100 (62.9%).	303
ASBHCHBK	Children's book at home	HQ	3-scale	1 = less than 25 (15.1%); 2 = 26-100 (57.2%); 3 = more than 100 (27.7%).	296
ASBHHOCP	Parental highest occupation	HQ	3-scale	1 = skilled worker, general labour or never worked outside home for pay (8.6%); 2 = small business owner or clerical (33.7%); 3 = professional (57.7%).	451
EDUAIDS	Index of sum score of the 5 common home possessions items (PC, desk, own room, newspaper, books)	StQ	3-scale	1 = low 0-3 (8%); 2 = average 4 (27.5%); 3 = high 5 (64.5%).	65
ASBGLNG1	Use the test language at home	StQ	dummy	1 = yes (94.8%); 0 = no (5.2%)	162
INTAKEBCH	School intake background characters with respect to 4 variables: ASBGPST1-Percentage disadvantaged student at school; ASBGPST2-Percentage affluence students at school; ASBGPST3-Percentage students who do not speak the test language at home; ASBGPST4-Percentage of students who receive some instructions in their home language	Derived from SQ	Continuous Factor score	Mean = .06; St.D = .93. The 4 variables that the factor score was based on are on 4-scale 1=0-10%; 2 = 11-25%; 3 = 26%-50%; 4 = more than 50%.	0
ASRREA01	Plausible value: overall reading achievement	StQ	Continuous	Mean = 548.7; St.D = 63.5	0

Note: The percentages presented in the parentheses are valid percent and the dataset was weighted by House Weight (HOUWGT).

Table 4. The conditional probability of answering “category 3” of the SES indicators in each latent class.

VARIABLES	Label (category 3)	Latent class 1	Latent class 2	Latent class 3	Latent class 4
ASBHHEDUP	Parental educational level (finished university or higher)	.65	.06	.34	.03
SBHBOOK	Books at home (more than 100 books)	1.00	.02	.34	.59
ASBHCHBK	Children's book at home (more than 100 books)	.51	.00	.09	.21
ASBHWELL	Being well-off family (well-off)	.65	.40	.55	.39
ASBHHOCP	Parental highest occupation (professional)	.93	.16	.93	.09
EDUAIDS	Index of sum score of the 5 common home possessions items (PC, desk, own room, newspaper, books) (5, high)	.74	.39	.68	.62

Figure 2. Graphical description of the latent classes according to the profiles of the individual SES background measures.



Total individuals in each latent class based on their most likely class membership	Class 1:1939 (44%)	Class 2: 583 (13%)	Class 3: 535 (12%)	Class 4: 1337 (30%)

Figure 3.Graphical description of the students' characteristics in each of the latent SES classes by a set of background variables of their parents.

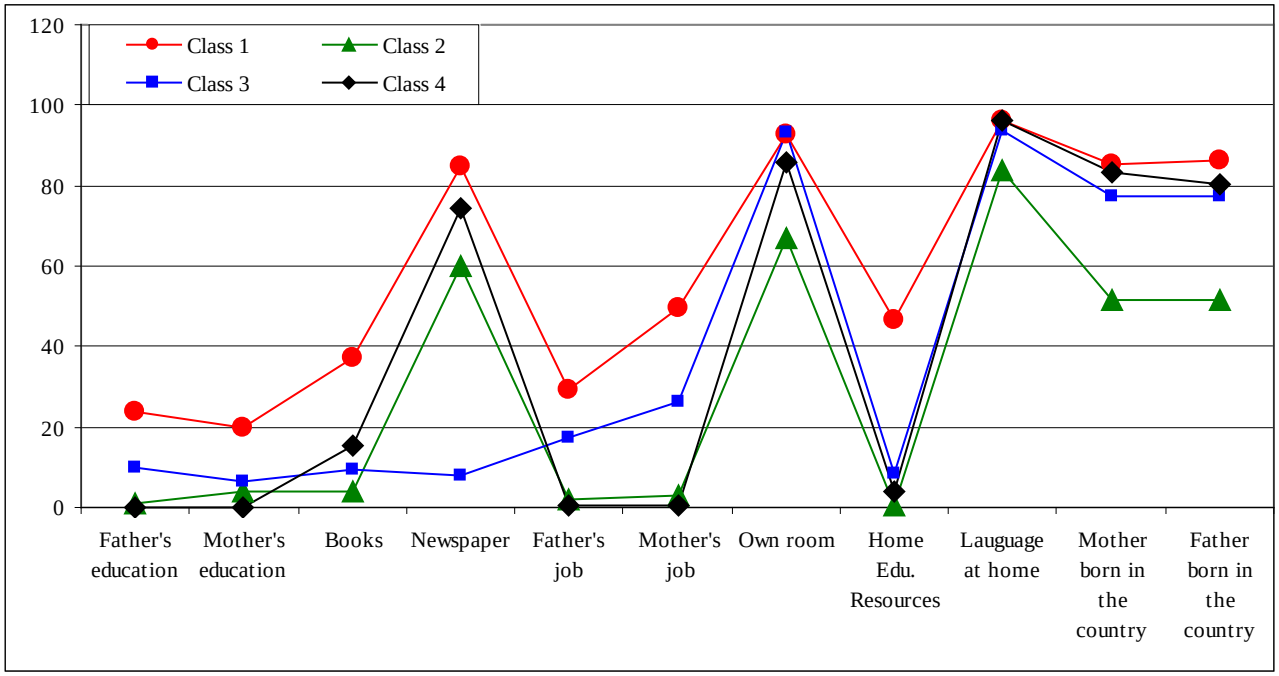
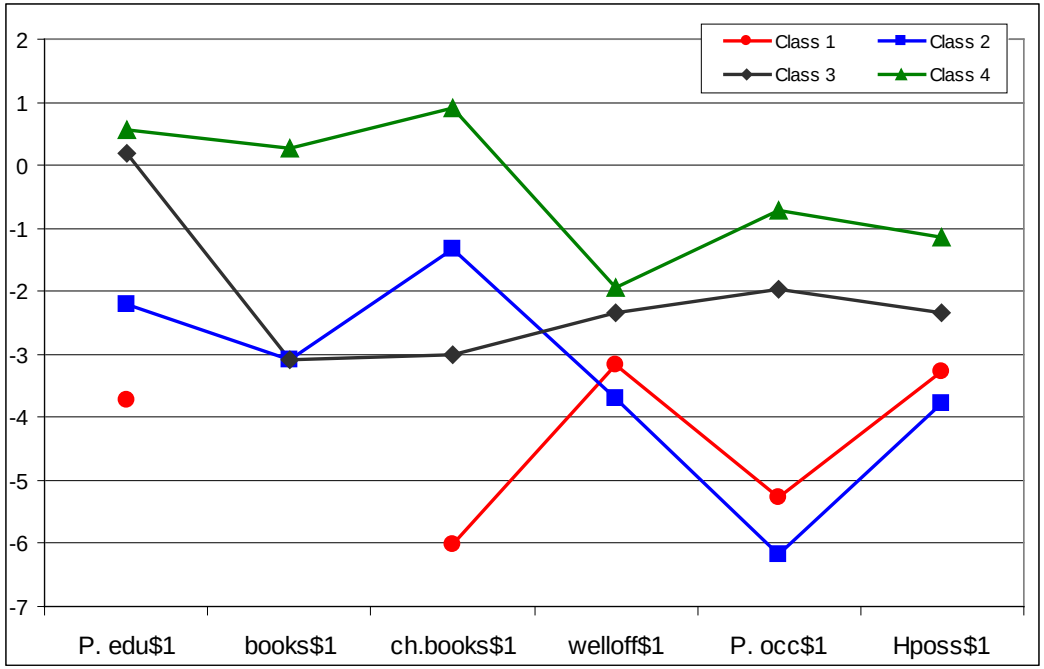


Figure 5.. The thresholds for category 1 in each SES indicator to belong to each latent class.



Note: the threshold for books$1 in Class 1 is -30.718 and was not included in the figure. This is mainly for the purpose of comparison between Figure 5 and Figure 2, because otherwise the patterns of the distribution of the thresholds in other classes will be flat.