Article

Introduction to Latent Class Analysis With Applications

Journal of Early Adolescence
2017, Vol. 37(1) 129–158

DOI: 10.1177/0272431616648452
jea.sagepub.com
SAGE

Mariano Porcu[1] and Francesca Giambona[1]

Abstract

Latent class analysis (LCA) is a statistical method used to group individuals (cases, units) into classes (categories) of an unobserved (latent) variable on the basis of the responses made on a set of nominal, ordinal, or continuous observed variables. In this article, we introduce LCA in order to demonstrate its usefulness to early adolescence researchers. We provide an application of LCA to empirical data collected from a national survey carried out in 2010 in Italy to assess mathematics and reading skills of fifth-grade primary school pupils (10 years in age). The data were used to measure pupils' supplies of cultural capital by specifying a latent class model. This article aims to describe and interpret results of LCA, allowing users to replicate the analysis. All LCA examples included in the text are illustrated using the Latent GOLD package, and command files needed to reproduce all analyses with SAS and *R* are available as supplemental online appendix files along with the example data files.

Keywords

latent class analysis, pedagogical, cross-sectional, latent GOLD, SAS, R

In the field of early adolescence, researchers are frequently interested in studying phenomena that often cannot be directly observed (e.g., socialization, temperament, eating disorder, cultural capital). Indeed, to analyze

[1]Department of Social Sciences and Institutions, University of Cagliari, Italy

Corresponding Author:
Mariano Porcu, Department of Social Sciences and Institutions, University of Cagliari, Viale Sant'Ignazio 78, Cagliari, Italy.
Email: mrporcu@unica.it

certain phenomenon, not all variables of interest may exist in practice nor are able to be directly measured. For instance, it could be interesting to measure unobservable variables by grouping pupils based on observed variables. Furthermore, it could be particularly interesting to look for a latent phenomenon as some pupils may not yet be able to manifest their preferences, attitudes, or behaviors.

A latent phenomenon or variable (LV) is a random variable that in principle (or in practice) cannot be directly observed; rather, its value can be deduced (throughout a mathematical model) from observed (measured) variables. Individual preferences, behaviors, and attitudes are examples of LVs that can be conceptualized as either categorical or continuous. For instance, it could be difficult to know whether an individual is a drug abuser if they do not admit to it; nevertheless, if we know a person's preferences and what a person thinks about drug use, we could potentially classify a person as, for example, a drug abuser.

Latent class analysis (LCA) belongs to latent class (LC) models, often referred to as finite mixture models. It is a statistical procedure for identifying class membership probabilities among statistical units (e.g., children), using the responses provided to some chosen set of observed variables. For example, we may wish to classify pupils in a school based on their attitude toward a specific discipline, to identify categories of pupils such as artistic, humanistic, or mathematical.

LCA was introduced by Lazarsfeld (1950) in order to classify individuals based on the values with which they identified in a set of dichotomous observed variables. More than 20 years later, Goodman (1974) developed an algorithm for obtaining maximum likelihood estimates of the model parameters, and proposed extensions for polytomous variables and multiple LVs. In 1979, Haberman showed the connection between LC models and log-linear models for contingency tables with missing cell counts. Many important extensions of the classical LC model have been proposed since then (see Clogg, 1981, 1995), such as models containing covariates, local dependencies, ordinal and continuous variables, more than one LV, and repeated measures. A general framework for categorical data analysis with discrete LVs was proposed by Hagenaars (1990) and then extended by Vermunt (1997).

In LCA, categorical and continuous observed variables are considered to classify each individual into one of the categories of a LV. LCA is a specification of the most general class of latent variable models (LVMs). Bartholomew, Knott, and Moustaki (2011) provided a framework that identifies four main types of LVMs. In this framework, observed variables are often defined as *indicators* or *manifest* variables. According to Bartholomew (1987, 1999), the four main types of LVMs can be specified as (a) factor analysis (FA), (b)

latent trait analysis (LTA), namely item response theory (IRT), (c) latent profile analysis (LPA), and (d) LCA. A fundamental distinction in Knott and Bartholomew's (1999) classification is between continuous and discrete LVs. Prior to applying a LVM, a researcher has to justify whether the conceptualization of the underlying LV is continuous or categorical.

In FA and LTA, the LVs are considered as continuous, whereas in LPA and LCA, the LV is considered to be discrete. Sometimes, LCA is referred to as a person-centered analysis (i.e., the interest is in finding heterogeneous groups of individuals) in contrast to a variable-centered analysis such as FA (i.e., the focus is on relationships among variables; Bauer & Curran, 2004; Molenaar & von Eye, 1994).

LCA and cluster analysis (CA; Kaufman & Rousseeuw, 2005; Tryon, 1939) are similar as both of them are procedures that group individuals into homogeneous classes. However, CA is not based on an underlying statistical model and does not provide information about the probability that a given individual belongs to a certain class. CA simply relies on a matrix (or an array) of distances among the classified objects and the researcher is left with choosing the proper metrics to calculate such distances. Moreover, CA does not provide information about item response behavior in terms of probabilities (e.g., given that an individual said "yes" to a certain item, what is the probability that she or he belongs to a certain class?). Furthermore, in LCA, a statistical model is proposed for the population from which the sample under study came, and consequently the LC model can be tested.

LCA differs from IRT as in the latter, the LV is assumed to be continuous, whereas in LC models, the LV is assumed to be categorical and consisting of two or more nominal or ordered classes (Hagenaars & McCutcheon, 2002). IRT models use the observed responses (to a number of items) to measure a continuous LV and the strength of the relationship between item scores, that is, the probability of responding into a particular category and LVs (De Ayala, 2009; Hambleton & Swaminathan, 1985; Sijtsma & Molenaar, 2002).

LCA assumes a parametric statistical model and uses observed data to estimate parameter values for the selected model. Each individual has a certain probability of membership to each latent class. Observations within the same latent class are homogeneous on certain criteria, whereas those in different latent classes are dissimilar from each other. In this way, latent classes are represented by distinct categories of a *discrete* LV.

In LCA, we estimate the *class membership probabilities* (i.e., the probability for an individual's membership in a certain class) and the *item response probabilities* conditional upon class membership (i.e., the probability for an individual to provide a certain response to a specific item given that she or he has been classified in a specific latent class). According to the *item response*

probabilities, observations are grouped (clustered) into classes. LCA can be used to find separate classes of individuals according to the responses provided to the items of a questionnaire (Collins and Lanza ., 2010; Magidson & Vermunt 2000; Lazarsfeld, 1950). LCA results can be used to classify individuals into their most likely (latent) group, where groups are the categories of a categorical LV.

Some extensions of LCA have been advanced. Among these, we shall mention latent class regression analysis (Hagenaars & McCutcheon, 2002), which has been introduced to consider the effect of some covariates on the probability of membership to each class, and multilevel LCA to classify individuals when data are nested (e.g., students nested within schools; Henry & Muthén, 2010; Vermunt, 2003, 2008), considering jointly the effect of covariates.

A number of empirical research studies have used LCA to classify individuals into the categories of a (latent) categorical variable on the basis of the observed variable values. For example, in psychology, LCA has been used to assess temperament (Stern, Arcus, Kagan, Rubin, & Snidman, 1995) and depression (Lanza, Flaherty, & Collins, 2003). In educational studies, teaching style has been modeled using LCA (e.g., Aitkin, Anderson, & Hinde, 1981). In the social sciences, LCA has been used to express poverty as a multidimensional construct (Dewilde, 2004); in behavioral research, for alcohol use (Auerbach & Collins, 2006; Coffman, Patrick, Palen, Rhoades, & Ventura, 2007; Jackson, Sher, Gotham, & Wood, 2001), or other individual behaviors (Chung, Flaherty, & Schafer, 2006; Chung, Park, & Lanza, 2005;; Lanza & Collins, 2002, 2006; Velicer, Colleen, Anatchkova, Fava, & Prochaska, 2007).

In the field of early adolescence, researchers could use LCA to address a variety of research questions, but few researchers have done so thus far. Butera, Lanza, and Coffman (2014) used LCA to assess the effect of early sex on delinquency (measured as a categorical LV) using data from the National Longitudinal Study of Adolescent Health for 1,890 adolescents in 11th and 12th grade. Collins and Lanza (2010) provided some examples of LCA on adolescent delinquency, social, and behavioral adolescent issues. Lanza et al. (2003) used eight binary indicators of adolescent depression to group respondents into five classes of depressed early adolescents. Lacourse et al. (2010) used an LCA to classify Canadian adolescents into classes of behaviorally disordered children.

The aim of this article is to introduce LCA and demonstrate how to analyze, present, discuss, and interpret the results of an LCA in the field of early adolescence. The usefulness of LCA for early adolescence researchers is demonstrated through an empirical application using a selected sample of

data. Relying on some variables observed directly as responses provided to items within a questionnaire, we propose an LCA to group a sample of early adolescent students into classes that represent different values of the LV (i.e., *cultural capital*).

Overview of Cultural Capital

Cultural capital is a sociological LV widely recognized and introduced by Pierre Bourdieu. Bourdieu (1986) distinguished between three types of capital:

1. Economic capital: refers principally to economic resources (cash, assets);
2. Social capital: resources based on group membership, relationships, networks of influence, and support; and
3. Cultural capital: forms of knowledge, skills, education, and advantages that provide a higher status in society. Parents provide their children with cultural capital by transmitting their own attitudes and knowledge needed to succeed in an educational system.

Theorists and researchers in the field of education have paid great attention to the concept of cultural capital due to its positive effect on students' educational success (see, for example, Cheadle, 2008; Sullivan, 2001; van de Werfhorst & Hofstede, 2007). Cultural capital can be viewed as a multidimensional concept that cannot be measured directly with a single variable. Moreover, how to measure cultural capital in the best way is a questionable issue (Kingston, 2001; Lareau & Weininger, 2003). Early studies focused on indicators of high status culture and possession of cultural objects at home (e.g., paintings, works of art), or participation in cultural events (e.g., art shows, music concerts, museums; Aschaffenburg & Maas, 1997). However, this approach has been criticized for being too narrow (Lareau & Weininger, 2003); consequently, indicators of reading habits, cultural communication, educational resources at home, and extracurricular activities have been taken into account (Covay & Carbonaro, 2010). According to this theoretical perspective, we consider cultural capital for early adolescence as a LV measurable using an LCA.

The following research questions were principally addressed in the example application:

Research Question 1: Are there classes of early adolescents differently supplied with cultural capital?

Research Question 2: Is there a latent class structure that adequately represents the heterogeneity in cultural capital endowment among early adolescents?
Research Question 3: Are some of the observed covariates predictive of individuals' membership in each class?

Specifically, this article will focus on

1. describing the data used in the analysis (indicator variables and covariates),
2. demonstrating a step-by-step analysis of the data,
3. interpreting the results of a standard LCA application, and
4. introducing a simple example of LCA with covariates.

All procedures will be illustrated using Latent GOLD, but the equivalent SAS and *R* codes are provided as additional online appendix materials.

Data Description

The data considered in the following application came from a 2010 survey administered in Italy to assess the mathematics and reading skills of fifth-grade (10-11 years old) primary school pupils. The data have been provided by the Institute for the Evaluation of the School System (INVALSI). INVALSI annually evaluates the efficiency and effectiveness of the Italian education system as a whole and for individual schools, by assessing the skills and knowledge of students enrolled in both private and public institutions. Presently, INVALSI assesses student achievement at different stages of the educational system: the second and fifth year of the primary school (about ages 7 and 10, respectively); the first and third year of lower secondary school (ages 11 and 13); and the second year of upper secondary school (age 15). The original sample contained nearly 40,000 observations clustered in about 1,400 schools and collected nationwide.[1] For the pedagogical purposes of this article, we randomly selected a subsample of 2,095 pupils considering only records with no missing variables. In that way, no hierarchical structure of data had to be considered in the analysis.

In 2010, the INVALSI survey was administered using two different sources to collect information: the *pupil's biographic form* and the *pupil's questionnaire*. The pupil's biographic form was completed by the school staff and gathered information held by the school (e.g., if the student started school 1 year earlier) and some family data that are not asked directly of pupils (e.g., parents' highest education, parents' employment status). The pupil's questionnaire

includes binary (*correct* = 1, *incorrect* = 0) test items in mathematics (44) and reading (69), as well as information about pupils' home possessions, family habits, daily reading time, and weekly extracurricular activities. A description of the variables used on the survey and the percentage of observations in each category are summarized in Table 1.

Depending on the responses to the items on the questionnaire, LCA modeling could be useful for identifying possible categories of the latent attribute "cultural capital," that is, early adolescents will be grouped into classes (the categories of a categorical variable) according to their endowment of that intangible asset.

Application: An Illustrated Example Using Latent GOLD

The first routine to fit a latent class model, maximum likelihood latent structure analysis (MLLSA), was limited to a small number of nominal observed variables. Currently, standard available routines can handle a long list of variables made up of different scale types (nominal, ordinal, discrete, and continuous). For example, the log-linear and event history analysis with missing data using the EM algorithm (LEM) routine (Vermunt, 1997) provides a command language that can be used to specify a large variety of models for categorical data, including LC models. Contrary to these command language driven packages, Latent GOLD provides a Graphical User Interface—GUI (i.e., *SPSS*-like)—that is especially developed for LCA (for a detailed introduction to Latent GOLD, see Vermunt & Magidson, 2005a, 2005b). It implements a variety of LC models, deals with variables of different scale types, allows multiple covariates to be included and to specify a multilevel model when data are nested with a hierarchy (e.g., pupils within schools).

In this illustrated example, data on fifth-grade pupils will be used to estimate an LC model of cultural capital with Latent GOLD.

Specifically, we demonstrate how to load data, fit standard LC models, explore which models fit the data in the best way, interpret results, and display results graphically. In Online Appendix A, we include instructions to replicate the application using SAS and *R*.

Loading Data

Within Latent GOLD, users can easily load data from a variety of input file formats: SPSS data file (filename.sav), ASCII text file (fileneame.dat, filename.csv, or filename.txt), ASCII array (array format) file (filename.arr), and

Table 1. Summary of Survey Variables.

Variable	Description	Value	%
Demographic and family background			
SEX	Gender	1 = male	49.88
		2 = female	50.12
AREA	Residence	1 = North	41.96
		2 = Center	20.24
		3 = South	37.8
MEDUC	Mother's educational level	1 = primary	3.68
		2 = low-secondary	39.57
		3 = high-secondary	41.15
		4 = tertiary	15.61
FEDUC	Father's educational level	1 = primary	3.53
		2 = low-secondary	46.4
		3 = high-secondary	35.04
		4 = tertiary	15.04
CITZEN	Nationality	1 = Native Italian	92.79
		2 = Other	7.21
Mathematic and reading achievement			
MATHSCOR	Raw score in mathematics	% of correct answers	69.36
READSCOR	Raw score in reading	% of correct answers	63.88
Pupil's home possessions			
QUIETPLACE	A quiet place to study	1 = yes	86.78
		2 = no	13.22
COMPUTER	A computer available for the homework	1 = yes	77.95
		2 = no	22.05
DESK	A pupil's own desk	1 = yes	87.59
		2 = no	12.41
WEB	Internet connection	1 = yes	79.47
		2 = no	20.53
ENCY	An encyclopedia	1 = yes	75.37
		2 = no	24.63
BOOKS	Number of books available	1 = none or <10	10.55
		2 = 11-200	74.61
		3 = >200	14.84
Pupil/family habits			
JOYREAD	Daily hours spent in reading for amusement	1 = none or <1 hour	66.63
		2 = 1-2 hours	23.72
		3 = >2 hours	9.64
SPORT	Sport activities not in school time	1 = none	18.33
		2 = one or more	81.67
CULTACT	Highbrow extracurricular activities (music, theater, etc.)	1 = none	62.63
		2 = one or more	37.37
LANG	Language spoken at home	1 = Italian	79.52
		2 = Other	20.48

Latent GOLD save file (filename.lgf). For this application, we have prepared an ASCII text file (i.e., put name of file with extension here) named *lca_data*. To load the data file, go to the menu option and choose File > Open and choose the file type (i.e., text files).

Fitting the Model

For our analysis, we use the following 10 items described in Table 1: QUIETPLACE, COMPUTER, DESK, WEB, ENCY, BOOKS, JOYREAD, SPORT, CULTACT, and LANG. We use these items because they are related to the LV cultural capital. Table 1 shows the frequencies of these 10 items in the questionnaire, which shows a substantial amount of variability: About two thirds of pupils have a quiet place to study, 78% have a computer available for doing homework, 88% have their own desk, and so on. An important question to ask when inspecting your data is whether there are patterns of responses that occur more frequently than others. To address this question, it is necessary to construct a contingency table considering all 10 items used as observed variables. Each cell of the contingency table is the combination of responses to the 10 items, namely, the *response pattern*, and contains the frequency (proportion) of pupils who provided a specific response pattern.

Hence, in the present application, the contingency table contains 2,304 cells (as we consider eight dichotomous variables and two variables with three categories, that is, $2^8 \times 3^2 = 2{,}304$); thus, some cells have zero counts. Bootstrapping methods (Nylund, Asparouhov, & Muthén, 2007) have been proposed in the context of LCA as a way to deal with such "sparsely populated" tables when assessing global fit of the model. To solve this problem, a parametric bootstrapping method may be used by selecting the "bootstrapping chi2" option in the main menu of Latent GOLD (note that in the application shown here, the estimates do not change very much).

The analysis of the response patterns is always a useful screening tool to gain more knowledge of the empirical data (e.g., to assess which response patterns occur most frequently). In this sense, the aim of an LCA is to try to fit a set of latent classes that represents the response patterns in the data, and to provide a sense of the latent class membership. In other words, LCA gives the probability of obtaining a certain response pattern (and a certain value of the categorical variable) for a given individual.

Formally, let $j = 1, \ldots, J$ be the observed (manifest) variables with $r_j = 1, \ldots, R_j$ the corresponding response categories. $\mathbf{Y}$ is the *vector* of response patterns; y represents a *particular* response pattern (e.g., for our data set, we have 10 manifest variables, y = no, no, yes, no, yes, none or <10, none or <1 hour, none, none); and C represents the number of latent classes, $c = 1, \ldots C$.

In traditional LC models, two parameters have to be estimated in order to estimate the probability that an individual choose a certain category of a manifest variable:

1. γ_c the probability of membership in latent class c (*class membership probabilities*);
2. $\rho_{i|c}$ the probability of response *i* to item *j* conditional on membership in latent class c (*item response probabilities*);

$$P(\mathbf{Y} = y) = \sum_{c=1}^{C} \gamma_c \prod_{j=1}^{J} \prod_{r_j=1}^{R_j} \rho_{j,r_j|c}^{I(y_j = r_j)}, \quad (1)$$

where $I(y_j = r_j)$ is an indicator function that equals 1 when the response to item $j = r_j$ and 0 otherwise. Equation 1 expresses the probability of observing a particular vector of responses as a function of the probabilities of membership in each latent class (γ) and the probabilities of observing each response conditional on latent class membership (ρ). Each unit (individual) is in one and only one class (i.e., the classes are mutually exclusive and exhaustive) such that the sum of the probabilities of membership in each class sum to 1, $\sum_{c=1}^{C} \gamma_c = 1$.

The parameter ρ expresses the relationship between each item (manifest variable) and each latent class (i.e., item response probabilities indicate how well each individual can be classified into specific latent classes, given their item response values). Sometimes, item response probabilities are called *measurement parameters* for their usefulness in the measurement of the LV and for the interpretation of the latent classes. Each individual (pupil) provides only one response to each manifest variable *j* and, consequently, the vector of item response probabilities for a particular manifest variable conditional on a particular latent class sums to 1 (for more technical details on LCA mathematical definitions, see Collins & Lanza, 2010). Ultimately, LC parameters are estimated to characterize each class (i.e., the class profile), the response patterns related to the latent classes, and the size of each class.

When there are missing data, parameter estimates are likely to be biased. Generally, missingness should be handled by modern missing data methods such as maximum likelihood estimation and multiple imputations (Schafer & Graham, 2002). Because LC models are multiple group models where group membership must be inferred from the data, Enders and Gottschall (2011) argued that it is not possible to use any multiple imputation method to preserve group differences in the mean and covariance structure across hidden latent classes. Thus, they recommend treating group membership as a latent

class variable in a mixture modeling framework, where a latent class intercept quantifies the relative probability of membership in each class. So, the imputation phase should be carried out using only the manifest variables because the imputation does not have a priori knowledge of class membership. In Latent GOLD, the missing values option allows for the inclusion of records containing missing values on covariates and predictors as well as records containing missing values on the indicators. Including cases with missing values on covariates and predictors causes the mean to be imputed for the scale type numeric and the effect of the missing value category to be equated to zero for the scale type nominal. Missing values on indicators and dependent variables are handled directly in the likelihood function (for details on treatment of missing data in Latent GOLD, see Vermunt & Magidson, 2005 a,b).

Assessing Local Independence

Local independence is the underlying assumption of LVMs (Lazarsfeld & Henry, 1968). Observed items are conditionally independent of each other given that an individual belongs to a certain latent class. Local dependence for a *K*-class model exists if the model does *not* fit the data. A diagnostic statistic that measures the extent to which the association between two manifest variables is reproduced by an LC model is given by the bivariate residual (BVR) statistic (Magidson & Vermunt, 2001).

A formal measure of the extent to which the observed association between two variables is reproduced by a model is given by the BVR statistic (Magidson & Vermunt, 2001). Each BVR corresponds to a Pearson χ^2 statistic (divided by the degrees of freedom) where the observed frequencies in a two-way cross-tabulation of the variables are contrasted with those expected counts estimated under the corresponding LC model. A BVR value substantially larger than 1 suggests that the model falls short of explaining the association in the corresponding two-way table (Magidson & Vermunt, 2004). If it happens, the model may fail in reproducing the relation between variables, and a model with a different number of classes may provide a significant improvement over the fitted model.

To fit an LC model within Latent GOLD, select *Cluster* from the *Menu*. Then, the *LC Cluster Analysis* dialog box opens. For our analysis, we use the 10 variables QUIETPLACE, COMPUTER, DESK, WEB, ENCY, BOOKS, JOYREAD, SPORT, CULTACT, and LANG as observed (manifest) variables of the LV cultural capital. Select the 10 variables in the *Variables list* box and click the *Indicators* button to put them into the *Indicator list* box. With the *SCAN* option, identify each category or value for the item indicator variables. If you double click on any

Table 2. Summary of Model Results Used for *Model Selection*.

Model	LL	BIC	AIC	# pars	L^2	*df*
1-class	−11,689.168	23,470.104	23,402.336	12	2,094.2826	2,083
2-classes	−11,427.505	23,046.192	22,905.010	25	1,570.9559	2,070
3-classes	−11,308.271	22,907.141	22,692.543	38	1,332.4892	2,057
4-classes	−11,271.902	22,933.818	22,645.805	51	1,259.7511	2,044

Note. LL = log-likelihood; BIC = Bayesian information criterion; AIC = Akaike information criterion; # pars = number of estimated parameters; L^2 = square of the likelihood; *df* = degrees of freedom.

variable, you can view the categories of each indicator and verify the values. After highlighting all 10 variables and right-clicking, select the appropriate scale type (see Screenshot 1 in the supplementary file Online Appendix B).

Specifying the Number of Classes

When conducting an LCA, a major decision in the process of model specification is choosing the number of latent classes to retain. This is done by considering parsimony and the interpretability of the solution (i.e., the labeling and substantive meaning of the latent classes). To determine the number of latent classes, we will estimate different models, each specifying a different number of latent classes. We start with the model with only one class and continue increasing it to determine if the set of available model diagnostics points to a certain number of classes to retain. Based on the model selection diagnostics available in Latent GOLD, we estimated up to four latent classes.

To specify the number of classes estimated in Latent GOLD, go into the *Variables* tab and in the box titled *Clusters* (below the *Indicators* push button), type "1-4" to request the estimation of 1-, 2-, 3-, and 4-class models. Now, click on *Estimate* (at the right bottom of the analysis dialog box). Table 2 summarizes for each model the following information: log-likelihood, Bayesian information criterion (BIC), Akaike information criterion (AIC), number of estimated parameters, the likelihood ratio χ statistic (L2; one of several statistics that can be used to assess how well the model fits the data), and the model degrees of freedom (*df*). Model selection in LCA (i.e., the choice of the number of latent classes to retain) is done by analyzing several indices reported in Table 2 and the interpretability of classes. To date, there is no common agreement about what are the best criteria for determining the number of latent classes in mixture models (i.e., the wide class of models to which LCA belongs); this lack of agreement is the most critical drawback in the application of these statistical procedures (Nylund et al., 2007).

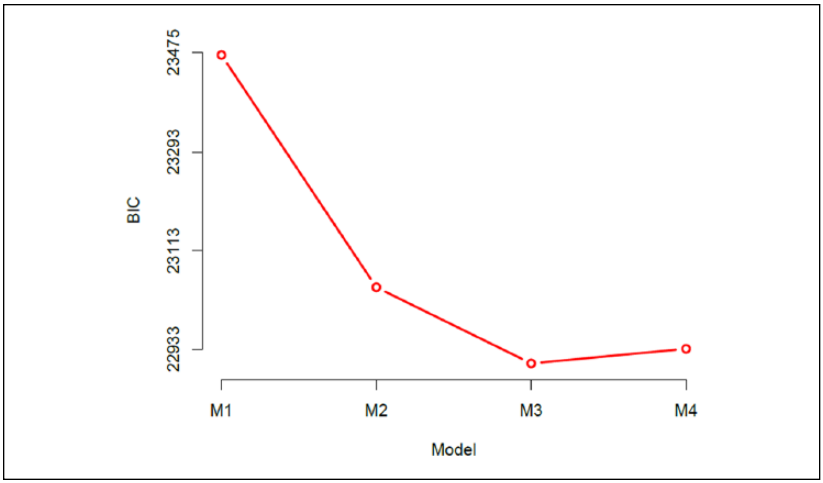

Figure 1. Screen plot of BIC value by each model specification.
Note. BIC = Bayesian information criterion; M1 = model with only one class; M2 = model with two classes; M3 = model with three classes; M4 = model with four classes.

By highlighting the data file name in Latent GOLD, a summary of all the models estimated is displayed. Additional statistics can be displayed by clicking on the associated check-box in the Model Summary Display. As shown in Table 2, the column labeled L^2 indicates the amount of shared association among the variables, which remains unexplained after estimating the model. L^2 measures the lack of model fit with the lower the value, the better the relative fit of a given model to another model. However, L^2 may not be reliable when there are large numbers of empty cells or sparse data (Magidson & Vermunt, 2004).

As we want to obtain information on the most appropriate number of classes, to retain and have model fit measures that are reliable in the case of sparse data, we specifically look at the information criteria measured by AIC and BIC. The number of appropriate latent classes was chosen by comparing models in terms of relative fit measures and primarily by interpretability of the latent classes. You can also select the data set file name and right-click for further indices (see Screenshot 2 in supplementary file Online Appendix B).

Based on the results shown in Table 2 and displayed in Figure 1, we chose to retain the model with three latent classes. Specifically, both BIC and AIC worsen as more than three latent classes are extracted from the data, which signifies a possible over-extraction in the number of latent classes in the observed data. Modeling with three latent classes provides the most

parsimonious solution, and the solution could be interpreted meaningfully. That is, the three latent classes that underlie cultural capital endowment could be labeled in terms of pupils who are well supplied, moderately supplied, and poorly supplied.

In Latent GOLD, the BVRs are provided under the Residuals Tab. In our example, BVRs are larger than 1 for some two-way tables involving "DESK" or "BOOKS," which suggest a flaw or weakness in the 3-class solution. Traditionally, when lack of fit is flagged for a given LC solution, it is improved by adding another latent class (Magidson & Vermunt, 2004). When a 4-, or even better, 5-class model is specified, the BVRs are smaller than the 3-class solution. This suggests that the 5-class solution may provide an improvement in fit over the 3-class solution. However, the 5-class solution is more difficult to interpret and not as meaningful as the 3-class solution.

An alternative is to specify a non-traditional LC model; that is, an LC factor model to include more than one LV. LC factor models were proposed as a general alternative to the traditional exploratory LC modeling (see Magidson & Vermunt, 2001, 2004, for more details of LC factor models).

Viewing and Interpreting Results

Next, for the 3-class model, we will assess the statistical significance associated with each variable. Within Latent GOLD, click on (+) from the "3-class" model to display the output, then select *Parameters, Profile*, and *ProbMeans*.

Interpreting Parameters

When you click on *Parameters*, a summary of the Parameter estimates and related statistics is displayed (see Table 3 and Screenshot 3 in the supplementary file Online Appendix B).

For each indicator, the *p* value is shown to be less than .05, indicating that the null hypothesis (i.e., all effects associated with that indicator are zero) should be rejected. Thus, for each indicator, the knowledge of the response of a certain individual to that indicator contributes significantly in discriminating between clusters. The R^2 values, in the last column of the table, indicate how much of the variance of each indicator is accounted for by this 3-class model. For example, 67.07% of the variance of the variable labeled COMPUTER was explained, whereas only the 0.028% of the variance of variable labeled DESK was explained. Note that when using the model display menu and right-clicking on the output, other statistics can be displayed in the output tab (e.g., standard errors, *z* statistics).

Table 3. Estimated Parameters for 3-Classes Model Solution.

Variable	Values	LC1	LC2	LC3	Wald	*p*	R^2
QUIETPLACE	1 = yes	0.3515	−0.0336	−0.3180	44.1899	<.001	.0358
	2 = no	−0.3515	0.0336	0.3180			
COMPUTER	1 = yes	1.1316	1.2572	−2.3887	13.545	<.001	.6707
	2 = no	−1.1316	−1.2572	2.3887			
DESK	1 = yes	0.3442	−0.1250	−0.2191	29.0435	<.001	.0277
	2 = no	−0.3442	0.1250	0.2191			
WEB	1 = yes	0.6366	0.3945	−1.0310	124.1653	<.001	.3311
	2 = no	−0.6366	−0.3945	1.0310			
ENCY	1 = yes	0.5592	−0.3204	−0.2388	79.0814	<.001	.1207
	2 = no	−0.5592	0.3204	0.2388			
BOOKS	1 = none or <10	−1.9327	1.9507	−0.0181	34.9081	<.001	.0602
	2 = 11-200	0.0588	0.5096	−0.5684			
	3 ≥200	1.8738	−2.4603	0.5865			
JOYREAD	1 = none or <1 hour	−0.4473	0.4610	−0.0136	44.0174	<.001	.0405
	2 = 1-2 hours	0.1942	−0.0821	−0.1121			
	3 ≥2 hours	0.2531	−0.3789	0.1258			
SPORT	1 = none	−0.3310	0.3254	0.0056	38.3935	<.001	.0533
	2 = one or more	0.3310	−0.3254	−0.0056			
CULTACT	1 = none	−0.2282	0.2500	−0.0218	26.4739	<.001	.0351
	2 = one or more	0.2282	−0.2500	0.0218			
LANG	1 = Italian	0.2919	−0.2344	−0.0575	32.0384	<.001	.0375
	2 = Other	−0.2919	0.2344	0.0575			

Note. LC1 = Latent Class 1; LC2 = Latent Class 2; LC3 = Latent Class 3.

Profile: Output and Plots

At this point, we need to sketch out the profile of these latent classes. This step will help us to define the response pattern for each class and to properly label the classes. To complete this step, we need to look at the *item response probability* and at the *probability of class membership* for each pupil.

Item Response Probabilities

With *Profile*, we view the parameters re-expressed as item response probabilities. Table 4 reports the Profile results. Overall, Class 1 contains about 57% of pupils, Class 2 contains 27%, and the remaining 16% is in Class 3.

Table 4. Item Response Probabilities.

Variable	Values	LC1	LC2	LC3
QUIETPLACE	1 = yes	0.9177	0.8377	0.7451
	2 = no	0.0823	0.1623	0.2549
COMPUTER	1 = yes	0.9242	0.9400	0.0106
	2 = no	0.0758	0.0600	0.9894
DESK	1 = yes	0.9234	0.8251	0.7962
	2 = no	0.0766	0.1749	0.2038
WEB	1 = yes	0.9123	0.8651	0.2703
	2 = no	0.0877	0.1349	0.7297
ENCY	1 = yes	0.8846	0.5690	0.6085
	2 = no	0.1154	0.4310	0.3915
BOOKS	1 = none or <10	0.0101	0.2905	0.1273
	2 = 11-200	0.7617	0.7077	0.7557
	3 = >200	0.2282	0.0018	0.1170
JOYREAD	1 = none or <1 hour	0.5718	0.8271	0.7257
	2 = 1-2 hours	0.3026	0.1339	0.1832
	3 = >2 hours	0.1257	0.0390	0.0910
SPORT	1 = none	0.1126	0.3205	0.1992
	2 = one or more	0.8874	0.6795	0.8008
CULTACT	1 = none	0.5531	0.7631	0.6516
	2 = one or more	0.4469	0.2369	0.3484
LANG	1 = Italian	0.8608	0.6834	0.7546
	2 = Other	0.1392	0.3166	0.2454
Overall		0.5648	0.2720	0.1631

Note. LC1 = Latent Class 1; LC2 = Latent Class 2; LC3 = Latent Class 3.

The item response probabilities (the γ in Equation 1) show the differences in response patterns that help us distinguish the classes. For example, consider the variable labeled QUIETPLACE (i.e., availability of a quiet place to study at home) where the probability to answer "yes" is 0.92 for the first class, 0.84 for the second class, and 0.74 for the third class. Looking at the pattern of responses for all the classes, we get an overall picture of the meaning of the three classes, which helps us to label them appropriately and meaningfully. Generally, pupils in Class 1 are much more likely to respond "yes" to the variables labeled QUIETPLACE, COMPUTER, DESK, WEB, ENCY, and BOOKS than the other two classes; they speak Italian at home (LANG); and have more than 10 books at home.

Starting with these item response probabilities, we have thus classified pupils into three classes corresponding to groups of pupils that are *well,*

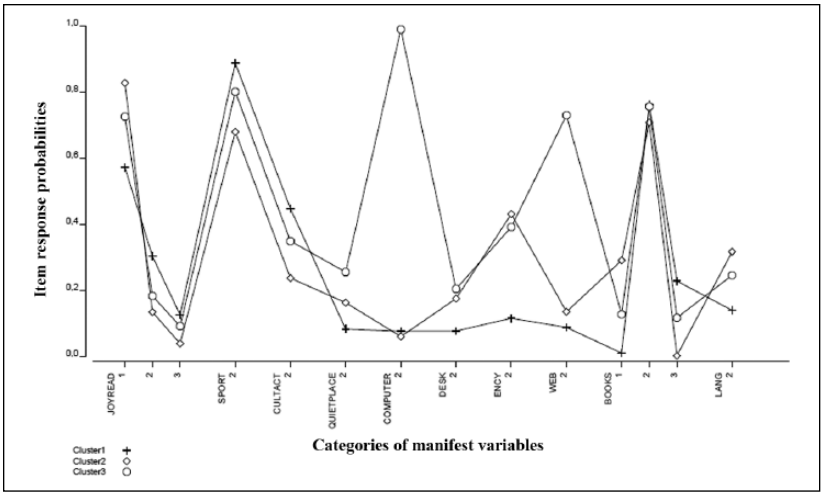

Figure 2. Profile plot for all clusters.
Note. The *y*-axis represents the item response probabilities for Class 1 (labeled Cluster in Latent GOLD), Class 2, and Class 3. The *x*-axis represents the categories for each variable. The conditional probabilities for variables are displayed and connected to form a line graph.

moderately, and *poorly* supplied with cultural capital. Looking at the two opposite classes, pupils in Class 1 show a response pattern that indicates more intense cultural activities and a larger possession of goods at home. Pupils in Class 2 show a response pattern with low levels in both cultural assets. Finally, pupils in Class 3 have more intense cultural activities than those in Class 2 (sport, cultural activities, and so on), but they have smaller possession of goods at home (a quiet place to study, a computer, etc.). We can also take the results provided in Table 4 and plot item response probabilities. This would be done by having the *x*-axis representing the items and the *y*-axis representing the probability to score a certain category to a specific item, given that a pupil belongs to a particular cultural capital class.

To create plots that display the item response probabilities in Latent GOLD, you need to click on *Prf-Plot*. Figure 2 shows the profile plot for the 3-class model and gives an immediate view of the three classes that are represented as three different lines. Latent GOLD labels the latent classes as Clusters, but we prefer to call them latent classes. The profile for any particular cluster may be highlighted by clicking on the symbol next to any of the three clusters (i.e., Cluster 1, Cluster 2, or Cluster 3) at the bottom of the plot. For example, to highlight the profile for Cluster 3, click on the symbol of

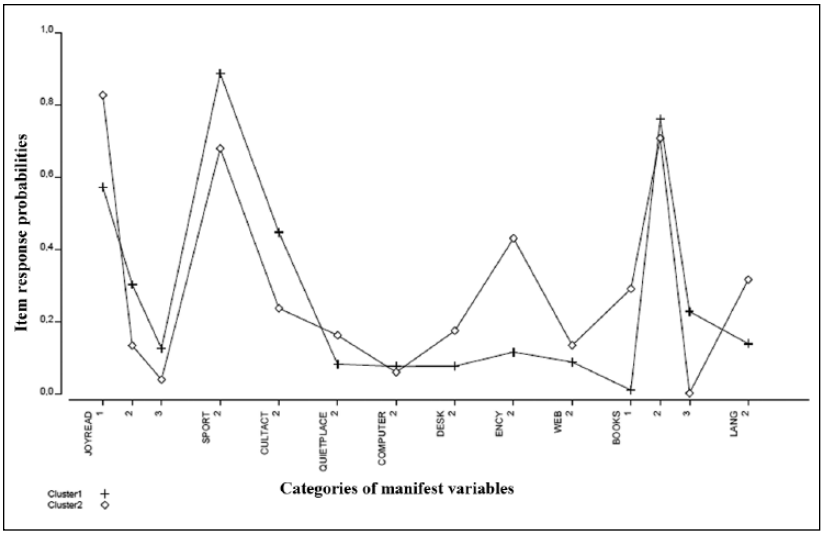

Figure 3. Profile plot for Cluster 1 versus Cluster 2.
Note. The *y*-axis represents the item response probabilities for Class 1 (labeled Cluster in Latent GOLD), Class 2, and Class 3. The *x*-axis represents the categories for each variable. The conditional probabilities for variables are displayed and connected to form a line graph.

Cluster 3. Furthermore, right-clicking on the plot allows you to choose the category of the item. Figures from 2 to 5 report all Profile Plots: (a) All clusters, (b) Cluster 1 versus Cluster 2, (c) Cluster 2 versus Cluster 3, and (c) Cluster 1 versus Cluster 3.

ProbMeans: Output and Tri-Plot

The *ProbMeans* output re-sets the parameters in terms of row percentages rather than column percentages (see Screenshot 4). The first row of the table contains the overall probability for any individual included in a specific cluster (the size of each cluster), which is also reported in the first row of numbers in the Profile table (see Screenshot 4 in supplementary file Online Appendix B). Graphically, the probabilities are displayed using the tri-plot (Figure 6).

The same information provided by ProbMeans can be displayed with the *Tri-Plot* graphic option in Latent GOLD. This option has the advantage of yielding a barycentric coordinate display of the categories of all indicators, where the vertices of the triangle represent the three clusters. For example, looking at the variable QUIETPLACE, 60% of pupils who indicated that they have a quiet place to study are in Cluster 1, while 26% are in Cluster 2, and 14% are in Cluster 3.

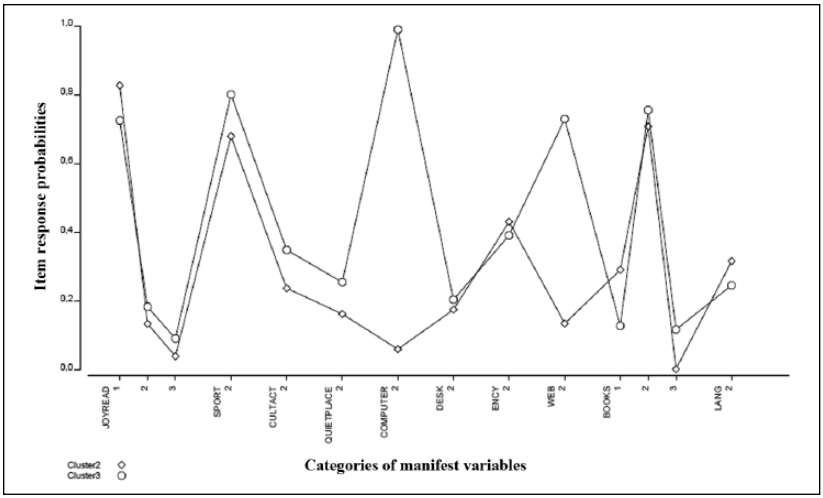

Figure 4. Profile plot for Cluster 2 versus Cluster 3.
Note. The *y*-axis represents the item response probabilities for Class 1 (labeled Cluster in Latent GOLD), Class 2, and Class 3. The *x*-axis represents the categories for each variable. The conditional probabilities for variables are displayed and connected to form a line graph.

Classifying Pupils Into Classes: The Probability of Latent Class Membership

Latent GOLD creates an output file that contains the original data used in the analysis followed by the probability of membership to each latent class (Class 1, Class 2, . . .). Next, each individual is assigned to the corresponding modal class. In the Output Tab, check the box for Standard Classification. Click on *Classification—Posterior* to view the Classification Output. Click on Classification on the left of the panel. A snapshot of the output is shown in Figure 4. The first row of the Classification Output shows that four pupils have the following response pattern:

QUIETPLACE = COMPUTER = DESK = WEB = ENCY = BOOKS = JOYREAD = SPORT = CULTACT = LANG = 1. These four pupils are classified into Class 2 as the probability of being in this class is the highest (.9689). Under the column labeled *modal*, we see the label "2" to indicate this classification. Each pupil has to be classified into one of these three mutually exclusive classes so that the probabilities sum up to 1 for each pupil. Furthermore, one important point to note here is that for some pupils, the class membership is well determined. For some pupils, it is a bit more ambiguous, as there is no single class to which they certainly belong.

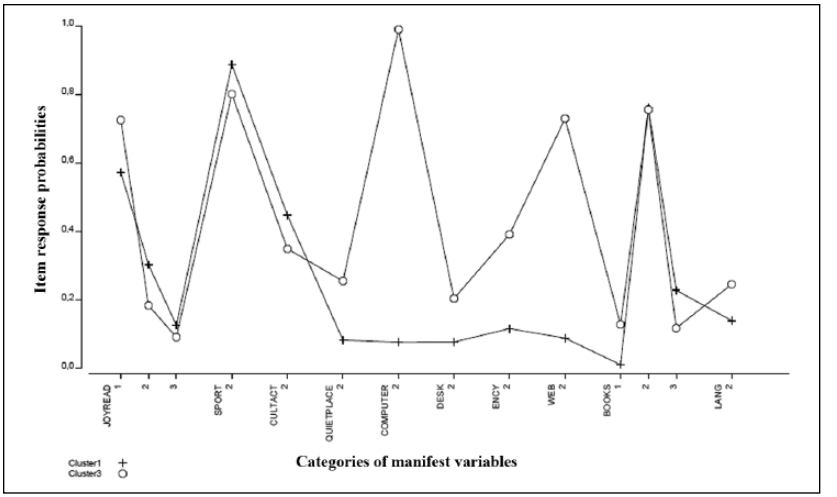

Figure 5. Profile plot for Cluster 1 versus Cluster 3.
Note. The *y*-axis represents the item response probabilities for Class 1 (labeled Cluster in Latent GOLD), Class 2, and Class 3. The *x*-axis represents the categories for each variable. The conditional probabilities for variables are displayed and connected to form a line graph.

At this point, we can select Save Results from the main menu and the new data set, containing the old variables and the new five columns (Observed Frequencies, Modal, Cluster 1, Cluster 2, and Cluster 3) will be saved. We can look now at the number of response patterns and number of pupils within each latent class. In Latent Class 1 (Cluster 1), there are 174 response patterns and 1,268 pupils; in Latent Class 2 (Cluster 2), there are 140 and 460, respectively; and in Latent Class 3 (Cluster 3), there are 170 and 367, respectively (see Screenshot 5 in supplementary file Online Appendix B).

How to Interpret the Latent Classes?

We determined that the LCA could group pupils into three latent classes with respect to their cultural capital. We labeled these classes as

1. Well supplied (Latent Class 1, LC1);
2. Moderately supplied (Latent Class 3, LC3); and
3. Poorly supplied (Latent Class 2, LC2).

In particular, pupils in LC1 have more cultural goods at home (the item response probability to answer *yes* for QUIETPLACE, COMPUTER, DESK, WEB,

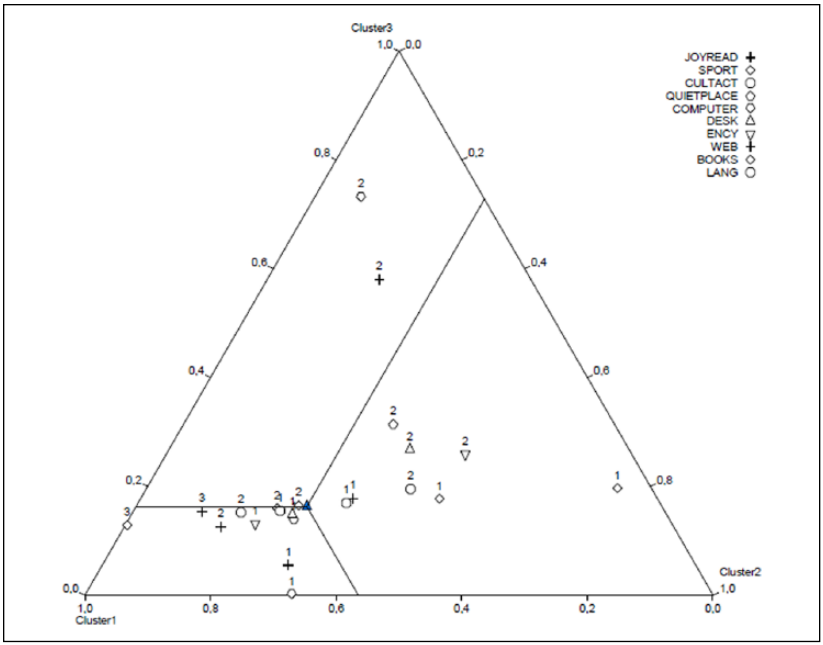

Figure 6. Tri-plot.
Note. Class membership probabilities (ProbMeans in Latent GOLD output) plotted graphically.

and ENCY is greater than 85%). Pupils classified in LC2 are also well equipped with cultural goods at home. Nevertheless, pupils in LC1 are more involved in *cultural activities*, as they have greater probabilities to spend more than 1 hour in reading for amusement (JOYREAD), to play at least one sport (SPORT), or to attend cultural activities (CULTACT; compared with LC2 and LC3, the item response probability is greater), and they speak Italian at home (lang).

Regarding *cultural activities*, LC2 and LC3 are quite similar except for SPORT as pupils in LC3 have a greater probability to play at least one sport (than pupils in LC1). Conversely, LC1 and LC2 are quite similar for *cultural goods possession* except for ENCY. The main difference between LC2 and LC3 is with regard to the probability to have resources at home, such as a computer (COMPUTER) and Internet connection (WEB); pupils in LC3 do not have these resources at home.

The 3-latent classes' solution has been selected not only on the basis of statistics but also in order to better interpret the latent classes. Indeed, if we selected fewer latent classes (e.g., two), we might lose information about the

Table 5. Latent Class Classification and Parents' Highest Educational Level.

Parents' highest educational level	Latent class			Total
	LC1	LC2	LC3	
Primary	12 (40.00)	12 (40.00)	6 (20.00)	30
Low-secondary	314 (47.50)	209 (31.62)	138 (20.88)	661
High-secondary	608 (65.31)	184 (19.76)	139 (14.93)	931
Tertiary	334 (70.61)	55 (11.63)	84 (17.76)	473
Total	1,268	460	367	2,095

Note. Values in parentheses represent %. LC1 = Latent Class 1; LC2 = Latent Class 2; LC3 = Latent Class 3.

Table 6. Frequency of Cases Occurring Within Latent Class Classification by Residence Area.

Residence area	Latent class			Total
	LC1	LC2	LC3	
North	530 (60.30)	179 (20.36)	170 (19.34)	879
Center	274 (64.62)	88 (20.75)	62 (14.62)	424
South	464 (65.31)	193 (19.76)	135 (14.93)	792
Total	1,268	460	367	2,095

Note. Values in parentheses represent %. LC1 = Latent Class 1; LC2 = Latent Class 2; LC3 = Latent Class 3.

middle category; whereas, choosing more than three classes (e.g., four or five) does not add relevant information about students' cultural capital supply (rather, we include more parameters in the model).

Finally, we can cross the observed frequencies of pupils in each class with some relevant characteristics about the pupils or their parents. For example, Tables 5 and 6 show, for each class, the observed frequencies of pupils classified by parents' highest educational level and the residence of pupils. We observe that pupils in LC1 have parents with higher educational levels (high including secondary 65%, tertiary 71%), and mainly live in Central Italy (65%; followed by Northern, 60%; and Southern, 59%).

This step is a post-classification of pupils, to explore if some variable of interest could affect the probability of membership in each latent class and, consequently, if it is reasonable to specify a latent class regression model.

The parents' level of education (see Table 5) seems to be a good predictor of the latent class membership probability.

Some Extensions of the LCA: Adding Covariates

Several extensions have been proposed for the basic LC model. One of the most important is the inclusion of covariates, which predict the LV. In this case, a latent class regression model (LCRM) can be specified. An LCRM (Bandeen-Roche, Miglioretti, Zeger, & Rathouz, 1997; Bouwmeester, Sijtsma, & Vermunt, 2004; Dayton & Macready, 1988; Hagenaars & McCutcheon, 2002; Heijden, Dressens, & Bockenholt, 1996) is a finite mixture model designed to identify a number of categorical classes of a LV on the basis of individual responses to a set of indicator variables. It considers, jointly, the effect of covariates on the probability of belonging to a certain latent class. Thus, covariates can be added into the latent class model to predict the latent class membership probability.

In our application, pupils' covariates could be introduced in the analysis (e.g., the educational level of the mother/father or the geographical area of pupils' residence). In this way, LCA classifies pupils into classes according to (a) the probability that a pupil provides a specific response pattern to items and (b) a given vector of covariates.

The probability of class membership is related to the values or levels of the covariates through multinomial logistic regression, where the dependent variable is latent rather than directly observed (Agresti, 2002). When a grouping variable is included in LCA with covariates, the multinomial logistic regression parameters are estimated for each group. As with any regression framework, covariates can be categorical, continuous, or a combination of both. Categorical covariates must be coded as dummy variables in order to preserve their categorical information and estimate the model (in Latent GOLD, this step is not required, as the user specifies at the beginning the scale of measurement of each variable included).

The *item response probabilities* (ρ parameters) in Equation 1 do not depend on the values or levels of covariates. To express the item response probabilities as function of covariates, Equation 1 changes to the following:

$$P\left(\mathbf{Y} = y \mid X = x\right) = \sum_{c=1}^{C} \gamma_c\left(\mathrm{x}\right) \prod_{j=1}^{J} \prod_{r_j=1}^{R_j} \rho_{j,r_j|c}^{I\left(y_j = r_j\right)}, \tag{2}$$

where $\gamma_c(\mathrm{x}) = P(L = c \mid X = x)$ is the *probability of membership* to latent class c given the x value of the X covariate (that is a standard baseline category

as in a multinomial logistic model). With a single covariate X, $\gamma_c(x)$ can be expressed as

$$\gamma_c(x) = P(L = c \mid X = x) = \frac{\exp(\beta_{0c} + \beta_{1c}x)}{1 + \sum_{c'=1}^{C-1} \exp(\beta_{0c} + \beta_{1c}x)}. \quad (3)$$

where $c' = 1, \ldots, C-1$.

Model estimation is performed by means of logistic regression with the only difference being that the outcome is latent rather than observed. When covariates are included, not only are the sets of parameters ρ and γ estimated (i.e., the probabilities of observing each response conditional on latent class membership, ρ; and the probability of observing a particular vector of responses as a function of the probabilities of membership in each latent class, γ), but also a set of β parameters are provided, that is, the logistic regression of coefficients for covariates, predicting class membership.

To include a covariate in Latent GOLD, click on the main menu, then cluster, and select the covariate to add in the box labeled *covariates* in the bottom right. We consider as a covariate the highest parents' educational level labeled HIGHEDU2 (in years) and we re-estimate the 3-class model. Table 7 displays the estimated 3-latent class model with the 10 observed variables and the highest parents' level of education. To test the effect or importance of the covariate to the 3-class model, we can use the log-likelihood ratio test (LR) by comparing the two log-likelihoods (LL) calculated for the model without covariate to the LL for the model with the covariate. Specifically, LR = −2(−11,308 − [−11,219]) = 178, with 2 *df* (*df* = difference between the number of estimated parameters). As the LR test statistic has a χ distribution, we can easily verify (comparing this value with the corresponding value of the chi-square distribution) that it is significant at $p < .0001$, indicating that parents' educational level is a significant predictor of the conditional latent class membership.

The probability of membership in LC1 increases as the parents' education level increases. The opposite occurs for LC2 and LC3, especially for LC2. Pupils whose parents are more educated have a higher chance to be more supplied with cultural capital (i.e., the probability of membership to LC1 increases and, conversely, the probability of membership to LC2 decreases). Figure 7 displays the pattern of class membership based on increasing levels of parents' education. The inclusion of this covariate provided new response patterns as pupils will now be classified on the account of item responses to the 10 cultural capital items and on parental education (758 response profiles have now been identified; see Screenshot 6 in supplementary file Online Appendix B).

Table 7. Estimated Parameters 3-Class Model With Highest Parents' Educational Level as Covariate.

Variable	Values	LC1	LC2	LC3	Wald	*p*	R^2
QUIETPLACE	1 = yes	0.291	0.063	−0.354	42.480	<.001	.030
	2 = no	−0.291	−0.063	0.354			
COMPUTER	1 = yes	1.024	1.272	−2.296	13.826	<.001	.596
	2 = no	−1.024	−1.272	2.296			
DESK	1 = yes	0.290	−0.018	−0.271	26.641	<.001	.020
	2 = no	−0.290	0.018	0.271			
WEB	1 = yes	0.684	0.453	−1.137	104.732	<.001	.355
	2 = no	−0.684	−0.453	1.137			
ENCY	1 = yes	0.487	−0.174	−0.312	69.167	<.001	.078
	2 = no	−0.487	0.174	0.312			
BOOKS	1 = none or <10	−3.691	2.517	1.175	32.795	<.001	.067
	2 = 11-200	1.006	−0.097	−0.908			
	3 = >200	2.686	−2.419	−0.266			
JOYREAD	1 = none or <1 hour	−0.387	0.289	0.098	45.288	<.001	.034
	2 = 1-2 hours	0.213	−0.098	−0.116			
	3 = >2 hours	0.174	−0.191	0.017			
SPORT	1 = none	−0.429	0.334	0.095	49.856	<.001	.064
	2 = one or more	0.429	−0.334	−0.095			
CULTACT	1 = none	−0.265	0.216	0.049	40.686	<.001	.045
	2 = one or more	0.265	−0.216	−0.049			
LANG	1 = Italian	0.350	−0.216	−0.135	39.623	<.001	.043
	2 = Other	−0.350	0.216	0.135			
HIGHEDU		0.766	−0.677	−0.089	119.551	<.001	

Note. Values in parentheses represent %. LC1 = Latent Class 1; LC2 = Latent Class 2; LC3 = Latent Class 3.

Multilevel LCA

Another important extension is the Multilevel LCA (MLCA) for nested data when observations are not independent (LCA assumes independent observations). The idea underlying MLCA is that the parameters of the regression model may differ across unobserved subgroups (Hagenaars & McCutcheon, 2002; Vermunt, 2003, 2008).

In Latent GOLD (advanced version), the Multilevel Model option can be used to define LC models for nested data (as in the framework of the present application, pupils could have been clustered within classes, and classes within schools; or in other fields, patients nested within hospitals, or citizens nested within regions).

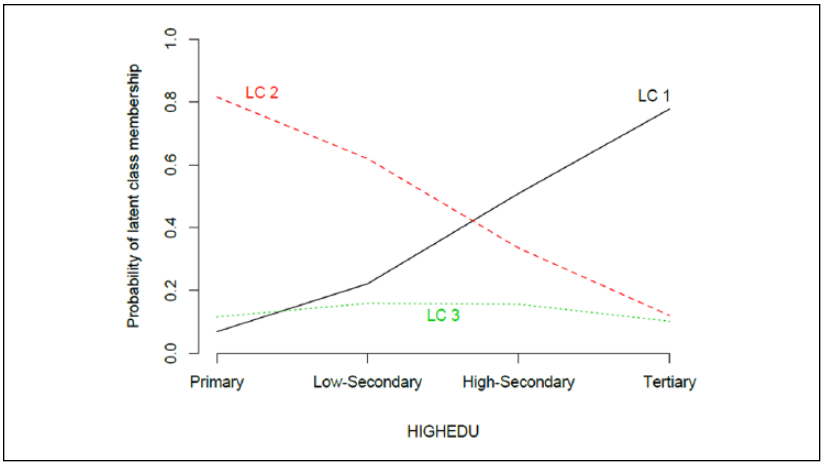

Figure 7. Probability of latent class membership of pupils in each of the three classes on the basis of the parents' highest level of education.
Note. LC = latent class.

Using our data from Model Menu, select Advanced; next, select Multilevel Model Group ID and then select the second level units. For our data, the second level is the variable labeled "region" (i.e., pupils are clustered into 21 regions of residence). Furthermore, we can select the number of classes for the region. We fit the 3-class model with different numbers of classes of region. Latent GOLD reports the log-likelihood value, the number of estimated parameters, and the BIC and AIC information criteria for the estimated models. Results show that a 3-class Multilevel Model was the most parsimonious model (see Screenshot 7 in supplementary file Online Appendix B). Consequently, when data are nested, it is necessary, preliminarily, to check if group parameters of the regression model may differ across subgroups.

Some Concluding Remarks

In this article, we have introduced the way in which an LCA can be used in the field of early adolescence. Using a sample of data from a 2010 survey administered in Italy to assess mathematics and reading skills of fifth-grade (10 years old) pupils nationwide, we carried out an LCA with Latent GOLD (SAS and *R* codes are provided as well in a supplemental file in the online appendix). In this way, pupils were classified into mutually exclusive categories representing different levels of the LV, cultural capital. Ten manifest variables were considered, and on the basis of their item response probabilities and on

the value of the latent class membership probability, pupils were classified into three classes. Furthermore, a covariate, parents' highest education level, was added into the 3-class model and shown to enhance latent class membership probability.

It is hoped that this example application has provided researchers with details on how to use and conduct an LCA when studying a latent phenomenon in the field of early adolescence. In general, LCA is a useful procedure that can be used for classifying pupils based on a single LV and can be extended to multilevel contexts and include covariates.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Note

1. Detailed information on the sampling design are available in a series of thematic and technical reports at the Institute for the Evaluation of the School System (INVALSI) website (www.invalsi.it).

References

Agresti, A. (2002). *Categorical data analysis*. New York, NY: John Wiley.

Aitkin, M., Anderson, D., & Hinde, J. (1981). Statistical modelling of data on teaching styles. *Journal of the Royal Statistical Society: Series A*, *144*, 419-461.

Aschaffenburg, K., & Maas, I. (1997). Cultural and educational careers: The dynamics of social reproduction. *American Sociological Review*, *62*, 573-587.

Auerbach, K. J., & Collins, L. M. (2006). A multidimensional developmental model of alcohol use during emerging adulthood. *Journal of Studies on Alcohol*, *67*, 917-925.

Bandeen-Roche, K., Miglioretti, D. L., Zeger, S. L., & Rathouz, P. J. (1997). Latent variable regression for multiple discrete outcomes. *Journal of the American Statistical Association*, *92*, 1375-1386.

Bartholomew, D. J. (1987). *Latent variable models and factor analysis*. New York, NY: Oxford University Press.

Bartholomew, D. J., Knott, M., & Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach* (3rd ed.). New York, NY: John Wiley.

Bauer, D. J., & Curran, P. J. (2004). The integration of continuous and discrete latent variable models: Potential problems and promising opportunities. *Psychological Methods*, *9*, 3-29.

Bourdieu, P. (1986). The forms of capital. In J. Richardson (Ed.), *Handbook of theory and research for the sociology of education* (pp. 241-258). New York, NY: Greenwood.

Bouwmeester, S., Sijtsma, K., & Vermunt, J. K. (2004). Latent class regression analysis for describing cognitive developmental phenomena: An application to transitive reasoning. *European Journal of Developmental Psychology*, *1*, 67-86.

Butera, N. M., Lanza, S. T., & Coffman, D. L. (2014). A framework for estimating causal effects in latent class analysis: Is there a causal link between early sex and subsequent profiles of delinquency? *Prevention Science*, *15*, 397-407.

Cheadle, J. E. (2008). Educational investment, family context, and children's math and reading growth from kindergarten through the third grade. *Sociology of Education*, *81*, 1-31.

Chung, H., Flaherty, B. P., & Schafer, J. L. (2006). Latent-class logistic regression: Application to marijuana use and attitudes among high-school seniors. *Journal of the Royal Statistical Society: Series A*, *169*, 723-743.

Chung, H., Park, Y., & Lanza, S. T. (2005). Latent transition analysis with covariates: Pubertal timing and substance use behaviours in adolescent females. *Statistics in Medicine*, *24*, 2895-2910.

Clogg, C. C. (1981). New developments in latent structure analysis. In D. J. Jackson & E. F. Borgotta (Eds.), *Factor analysis and measurement in sociological research* (pp. 215-246). Beverly Hills, CA: SAGE.

Clogg, C. C. (1995). Latent class models. In G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), *Handbook of statistical modeling for the social and behavioral sciences* (pp. 311-359). New York, NY: Plenum Press.

Coffman, D., Patrick, M. E., Palen, L., Rhoades, B. L., & Ventura, A. (2007). Why do high school seniors drink? Implications for a targeted approach. *Prevention Science*, *8*, 241-248.

Collins, L. M., & Lanza, S. T. (2010). *Latent class and latent transition analysis: With applications in the social, behavioural, and health sciences*. New York, NY: John Wiley.

Covay, E., & Carbonaro, W. (2010). After the bell: Participation in extracurricular activities, classroom behavior, and academic achievement. *Sociology of Education*, *83*, 20-45.

Dayton, C. M., & Macready, G. B. (1988). Concomitant variable latent class models. *Journal of the American Statistical Association*, *83*, 173-178.

De Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford Press.

Dewilde, C. (2004). The multidimensional measurement of poverty in Belgium and Britain: A categorical approach. *Social Indicators Research*, *68*, 331-369.

Enders, C. K., & Gottschall, A. C. (2011). Multiple imputation strategies for multiple group structural equation models. *Structural Equation Modeling*, *18*, 35-54.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, *61*, 215-231.

Hagenaars, J. A. (1990). *Categorical Longitudinal Data-Loglinear Analysis of Panel, Trend and Cohort Data* . Newbury Park: SAGE.

Hagenaars, J. A., & McCutcheon, A. (2002). *Applied latent class analysis*. Cambridge, UK: Cambridge University Press.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer Academic.

Heijden, P. G. M., Dressens, J., & Bockenholt, U. (1996). Estimating the concomitant-variable latent-class model with the EM algorithm. *Journal of Educational and Behavioral Statistics*, *21*, 215-229.

Henry, K., & Muthén, B. (2010). Multilevel latent class analysis: An application of adolescent smoking typologies with individual and contextual predictors. Structural, Equation Modeling, *17*, 193–215.

Jackson, K. M., Sher, J. J., Gotham, H. J., & Wood, P. K. (2001). Transitioning into and out of large-effect drinking in young adulthood. *Journal of Abnormal Psychology*, *100*, 378-391.

Kaufman, L., & Rousseeuw, P. J. (2005). *Finding groups in data. An introduction to cluster analysis*. Hoboken, NY: John Wiley.

Kingston, P. W. (2001). The unfulfilled promise of cultural capital theory. *Sociology of Education*, *74*, 88-99.

Knott, M., & Bartholomew, D. J. (1999). *Latent variable models and factor analysis. Kendall's library of statistics* (2nd ed.). London, England: Edward Arnold.

Lacourse, É., Baillargeon, R., Dupéré, V., Vitaro, F., Romano, E., & Tremblay, R. E. (2010). Two-year predictive validity of conduct disorder subtypes in early adolescence: A latent class analysis of a Canadian longitudinal sample. *Journal of Child Psychology and Psychiatry*, *51*, 1386-1394.

Lanza, S. T., & Collins, L. M. (2002). Pubertal timing and the onset of substance use in females during early adolescence. *Prevention Science*, *3*, 69-82.

Lanza, S. T., & Collins, L. M. (2006). A mixture model of discontinuous development in heavy drinking from ages 18 to 30: The role of college enrollment. *Journal of Studies on Alcohol*, *67*, 552-561.

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: Vol. 2, research methods in psychology* (p. 663-685). Hoboken, NJ: Wiley

Lareau, A., & Weininger, E. B. (2003). Cultural capital in educational research: A critical assessment. *Theory and Society*, *32*, 567-606.

Lazarsfeld, Paul F. 1950a. "The Logical and Mathematical Foundation of Latent StructureAnalysis." Pp. 361–412 in Measurement and Prediction, edited by S. A. Stouffer, et al. Princeton, NJ: Princeton University Press.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. Boston, MA: Houghton Mill.

Magidson, J., & Vermunt, J. K. (2000). *Latent class analysis*. Cambridge, MA: Cambridge University Press.

Magidson, J., & Vermunt, J. K. (2001). Latent class factor and cluster models, bi-plots and related graphical displays. *Sociological Methodology*, *31*, 223-264.

Magidson, J., & Vermunt, J.K. (2004). Latent class models. In: D, Kaplan. (Ed.), *The Sage Handbook of Quantitative Methodology for the Social Sciences, Chapter* 10, 175-198. Thousand Oaks: SAGE.

Molenaar, P. C. M., & 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 Oaks, CA: SAGE.

Nylund, K., 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*, 535-569.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*, 147-177.

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. Thousand Oaks, CA: SAGE.

Stern, H. S., Arcus, D., Kagan, J., Rubin, D. B., & Snidman, N. (1995). Using mixture models in temperament research. *International Journal of Behavioral Development*, *18*, 407-423.

Sullivan, A. (2001). Cultural capital and educational attainment. *Sociology*, *35*, 893-912.

Tryon, R. C. (1939). *Cluster analysis*. Ann Arbor, MI: Edwards Brothers.

van de Werfhorst, H. G., & Hofstede, S. (2007). Cultural capital or relative risk aversion? Two mechanisms for educational inequality compared. *British Journal of Sociology*, *58*, 391-415.

Velicer, W. F., Colleen, A. R., Anatchkova, M. D., Fava, J. L., & Prochaska, J. O. (2007). Identifying cluster subtypes for the prevention of adolescent smoking acquisition. *Addictive Behaviors*, *32*, 228-247.

Vermunt, J. K. (1997). *LEM: A general program for the analysis of categorical data—User's manual*. Tilburg, The Netherlands: Tilburg University.

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

Vermunt, J. K. (2008). Latent class and finite mixture models for multilevel data sets. *Statistical Methods in Medical Research*, *17*, 33-51.

Vermunt, J. K., & Magidson, J. (2005a). *Latent GOLD 4.0 user's guide*. Belmont, MA: Statistical Innovations.

Vermunt, J. K., & Magidson, J. (2005b). *Technical guide for Latent GOLD 4.0: Basic and advanced*. Belmont, MA: Statistical Innovations.

Author Biographies

Mariano Porcu is PhD in applied statistics and professor in social statistics at the Department of Social Sciences and Institutions, University of Cagliari. His most recent research interests are the framework of university and school system evaluations. Specifically, he studies research activity related to the measurement of student achievement and the composite indicators used to evaluate the performance of educational institutions.

Francesca Giambona is a PhD in applied statistics and a post-doctoral researcher at the Department of Social Sciences and Institutions, University of Cagliari. The focus of her academic research is on school system evaluation, information and communication technology (ICT) and students' achievement, well-being, and students' mobility.