A Short Guide to using Latent Class Analysis

*Debbie Cooper and Comfort Ajoku, ONS*

# 1.0 Introduction

The aim of this guide is to briefly describe Latent Class Analysis (LCA) and how it can be used as well as to provide an applied example. The applied example will illustrate how to carry out an LCA using R (R Foundation for Statistical Computing, 2011) and how to interpret the LCA. Some guidance on carrying out LCA in SAS will also be provided. It is hoped that this will enable colleagues across the GSS to implement LCA in their work.

# 2.0 What is Latent Class Analysis?

LCA provides a flexible and powerful approach to categorical data analysis (McCutcheon and Hagenaars, 1997). It is a type of model-based cluster analysis that generally uses the expectation-maximisation (EM) algorithm for model estimation (for further methodological detail see McCutcheon, 1997 and Linzer & Lewis, 2011).

In numerous studies, particularly in social research, researchers are interested in latent variables (variables that cannot be measured directly) e.g. personal well-being or quality of life. These variables tend to be measured by means of a number of indicator (observed) variables. For example, the Office for National Statistics (ONS) uses four indicator variables to measure personal well-being (a latent variable) in the UK1:

Figure 1. Indicator variables used to measure personal well-being in the UK

**Personal Well-being**

**Latent Variable**

**Anxious**

**Happy**

**Satisfied**

**Worthwhile**

**Indicator**

**Variables**

In LCA, the indicator variables are categorical variables. LCA is used to identify patterns of responses to the indicator variables to create a set of mutually exclusive latent classes, that is, groups of individuals or other units of analysis. Individuals in the same latent class will have similar response patterns to the indicator variables whilst individuals across latent classes tend to have different response patterns to each other. In other words, LCA splits respondents into homogenous groups (latent classes).

LCA has numerous advantages over traditional cluster analysis techniques such as hierarchical cluster analysis and K-means clustering. Some of these advantages include:

* It is model-based unlike other types of cluster analysis which tend to be distance-based. An advantage of this is that there are more formal criteria for choosing the final model when using LCA (for further information see Vermunt & Magidson, 2002).
* It is relatively easy to deal with variables having different scale types (Vermunt & Magidson, 2002).
* In traditional cluster analysis techniques persons are assigned to clusters on an all-or-none basis. On the other hand, LCA allows membership of a person to each cluster to a certain degree allowing for fractional cluster membership (captured by posterior possibilities).

# 3.0 Applied Example

In order to better explain LCA, I will be providing an applied example of LCA using ONS personal well-being data. The aim of this example is not to provide users of personal well-being data with the UK personal well-being profiles2 but only to illustrate how to carry out and interpret an LCA. **The results shown in this paper are not official statistics** (see [Office for National Statistics (2016)](https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/bulletins/measuringnationalwellbeing/oct2015tosept2016) for the latest personal well-being official statistics).

LCA can be carried out in many software programs such as SAS®3, R (R Foundation for Statistical Computing, 2011), STATA3 (StataCorp LP, 2015), Mplus (Muthén & Muthén, 2011) and Latent Gold (Vermunt & Magidson, 2013), amongst others. For the purposes of this example LCA has been carried out using R and the code used to specify the LCA in R will be described. Although the applied example will be coded in R, an explanation regarding how to evaluate the resulting models will be provided and this should be relevant regardless of the software program used. For information regarding how to carry out LCA in SAS please refer to Appendix 1.

This section is split into 4 subsections. The first describes the formula and command required to run an LCA in R. Following this, there is a subsection which briefly describes the personal well-being dataset used in the applied example. Next, the code specified to run the LCA on the personal well-being dataset in R is provided and described. The results of this analysis are then provided along with an explanation of how to interpret the LCA results.

## 3.1 Specifying a Basic Latent Class Analysis in R

LCA can be carried out using the R package poLCA (Linzer & Lewis, 2013; Linzer & Lewis, 2011). The formula definition for a basic LCA model is as follows:

f <- cbind(Y1, Y2, Y3) ~ 1

Y1, Y2 and Y3 are the categorical variables to be included in the LCA. The “~ 1” instructs poLCA to estimate a basic latent class model.

In order to run the poLCA, the basic command used in this paper is as follows:

poLCA (formula, data, nclass = 2, maxiter = 50000, graphs = FALSE, na.rm = TRUE,

nrep = 10, verbose = TRUE)

Description of the options specified in the command above:

**formula:** the formula definition ‘f’ specified above

**data:** the name of the data frame to be used in the LCA

**nclass:** the number of latent classes to be calculated in the model. The default is 2 latent classes. poLCA assumes one set of latent classes every time it is run. Therefore, in order to obtain multiple models, each assuming a different number of latent classes, the command must be run a number of times each time specifying a different number of latent classes to be assumed. This will become clearer in the applied example provided below.

**maxiter:** this is the maximum number of iterations for convergence. If convergence is not achieved before reaching this number of iterations an error message will appear and the analysis will terminate.

**graphs:** this specifies whether a graph showing the parameter estimates should be produced. The default is FALSE. It takes quite long for R to run analysis with 4 or more latent classes on large datasets. Therefore, it might be quicker to run the analysis without producing graphs and, if required, only producing a graph for the best fitting model once the resulting models have been compared.

**na.rm:** this specifies how poLCA handles cases with missing values. If specified as TRUE, those cases are removed by means of listwise deletion before model estimation. If specified as FALSE, cases with missing values are retained. The default is TRUE. Linzer and Lewis (2011) suggest that it is not necessary to delete cases with missing values before estimating the model because poLCA excludes cases with missing values from the calculation.

**nrep:** this option is used to specify the number of times the model should be estimated using different starting values. It is preferable to set nrep to greater than 1 to ensure that the algorithm finds a global rather than local maximum of the log-likelihood function.

**verbose:** this indicates whether the results of the model should be output to the screen or not. The default is TRUE.

There are a number of additional options which can be included in this command. For a full list of these options please refer to Linzer and Lewis (2011).

## 3.2 Dataset Information

As illustrated in Section 2, four personal well-being questions4 are asked in order to measure personal well-being in the UK, these are:

Overall, how satisfied are you with your life nowadays?

Overall, to what extent do you feel that the things you do in your life are worthwhile?

Overall, how happy did you feel yesterday?

Overall, how anxious did you feel yesterday?

The four personal well-being variables are measured on a scale of 0 to 10, where 0 is ‘not at all’ and 10 is ‘completely’. For the purposes of this analysis, the responses to these questions are categorised as follows:

≥ 0 and ≤ 4 = 1 (Low)

≥ 5 and ≤ 7 = 2 (Medium)

≥ 8 and ≤ 10 = 3 (High)

*Please note that these categories are different to those used in the national statistics published by the ONS.* Fewer categories than those used for official statistics were required for this analysis in order for meaningful results to be obtained. In this case, the use of too many categories results in poor model fit as there is not enough differentiation between the categories.

## 3.3 Latent Class Analysis of Personal Well-being

An LCA of personal well-being data was carried out using the R package poLCA (Linzer & Lewis, 2013; Linzer & Lewis, 2011). The code to run the LCA was specified as follows:

library (foreign)

library (MASS)

library (scatterplot3d)

library (poLCA)

data <- read.spss("D:PWB\_LCA/ wellbeing.sav")

data1 <- as.data.frame(data)

data2 <- data1[,16:19]

f <- cbind(satisth2, happyth2, anxiouth2, worthth2)~1

wellbeing2 <- poLCA (f, data2, nclass=2, maxiter=50000, graphs=FALSE, nrep=10,

verbose =TRUE)

wellbeing3 <- poLCA (f, data2, nclass =3, maxiter = 50000, graphs = FALSE, nrep = 10, verbose = TRUE)

wellbeing4 <- poLCA (f, data2, nclass =4, maxiter = 50000, graphs = FALSE, nrep = 2, verbose = TRUE)

wellbeing5 <- poLCA (f, data2, nclass =5, maxiter = 50000, graphs = FALSE, nrep = 2, verbose = TRUE)

wellbeing6 <- poLCA (f, data2, nclass =6, maxiter = 50000, graphs = FALSE, nrep = 2, verbose = TRUE)

wellbeing7 <- poLCA (f, data2, nclass =7, maxiter = 50000, graphs = FALSE, nrep = 2, verbose = TRUE)

wellbeing8 <- poLCA (f, data2, nclass =8, maxiter = 50000, graphs = FALSE, nrep = 2, verbose = TRUE)

The first step involves specifying the libraries. The ‘foreign’ package is used to import the SPSS data file into R. poLCA depends on two packages: ‘MASS’ and ‘scatterplot3d’ therefore these were specified in addition to the poLCA package. Following specification of the libraries:

* The “data <- read.spss” line reads in an SPSS file containing the personal well-being variables to be used in the analysis.
* “Data1” sets “Data” as a frame so that it can be used for analysis.
* “Data2” is a subset of “Data1” containing only the personal well-being variables to be used in the analysis.
* The function “*f”* takes the four personal well-being variables and models a basic latent class model with no covariates (as described in Section 3.1 the “~ 1” instructs poLCA to estimate a basic latent class model). Note that if the dataset has more variables than the f function, it will result in an error because the dimensions are not the same.
* “wellbeing2” carries out the latent class analysis for 2 classes (nclass = 2), “wellbeing3”, “wellbeing4” and “wellbeing5” run the analysis assuming 3, 4 and 5 latent classes, respectively.
* In each of the “wellbeing” commands:
  + the function “f” is specified
  + the data frame “data2” is specified for use in the analysis
  + the number of classes to be assumed in the model is specified
  + maxiter is set to 50000 to ensure convergence is obtained
  + graphs is set to FALSE so as not to produce any graphs (the analysis runs faster this way)
  + nrep is set to 10 in order to ensure that global rather than local maxima are found. As the number of latent classes calculated by the model increases, the analysis takes longer to run. Consequently, for models with 4 or more latent classes, one may wish to start by specifying a small number for nrep (in this case nrep=2 was used for “wellbeing4” and “wellbeing5”). It is possible to determine from the output whether global maxima have been found. If global maxima have been found there is no need to run the analysis with more repetitions. However, if it becomes evident that global maxima have not been found, it would be desirable to re-run the analysis specifying a higher number for nrep in order to obtain global maxima.
  + verbose is set to TRUE in order to output the results to the screen for interpretation

## 3.4 Interpretation of LCA Results

The “wellbeing2” analysis run in section 3.3 outputs the following results:

Model 1: llik = -934514.9 ... best llik = -934514.9

Model 2: llik = -934514.9 ... best llik = -934514.9

Model 3: llik = -934514.9 ... best llik = -934514.9

Model 4: llik = -934514.9 ... best llik = -934514.9

Model 5: llik = -934514.9 ... best llik = -934514.9

Model 6: llik = -934514.9 ... best llik = -934514.9

Model 7: llik = -934514.9 ... best llik = -934514.9

Model 8: llik = -934514.9 ... best llik = -934514.9

Model 9: llik = -934514.9 ... best llik = -934514.9

Model 10: llik = -934514.9 ... best llik = -934514.9

Conditional item response (column) probabilities,

by outcome variable, for each class (row)

$satisth2

Pr(1) Pr(2) Pr(3)

class 1: 0.0008 0.1414 0.8578

class 2: 0.1695 0.7222 0.1083

$happyth2

Pr(1) Pr(2) Pr(3)

class 1: 0.0174 0.1901 0.7925

class 2: 0.2519 0.5662 0.1819

$anxiouth2

Pr(1) Pr(2) Pr(3)

class 1: 0.7958 0.1541 0.0501

class 2: 0.4752 0.3419 0.1829

$worthth2

Pr(1) Pr(2) Pr(3)

class 1: 0.0014 0.1258 0.8728

class 2: 0.1232 0.6365 0.2403

Estimated class population shares

0.6368 0.3632

Predicted class memberships (by modal posterior prob.)

0.6278 0.3722

=========================================================

Fit for 2 latent classes:

=========================================================

number of observations: 303778

number of estimated parameters: 17

residual degrees of freedom: 63

maximum log-likelihood: -934514.9

AIC(2): 1869064

BIC(2): 1869244

G^2(2): 61111.46 (Likelihood ratio/deviance statistic)

X^2(2): 138365.7 (Chi-square goodness of fit)

**As specified in Section 3.0 these results are not official statistics.** They are only provided to illustrate how to carry out and interpret an LCA.

The results are interpreted as follows:

The first part of the output (Model 1 to Model 10 llik and best llik) shows that the latent class model was estimated ten times (as specified by nrep=10) using different starting values. The results assigned to “wellbeing2” will be those estimated for the model with the greatest value of the log-likelihood function (Linzer & Lewis, 2011). In this case, it seems as though the global maximum log-likelihood of -934514.9 was found on the first attempt at fitting the model. Therefore, the Model 1 results will be assigned to “wellbeing2”.

The next section of output provides conditional item response probabilities, by outcome variable, for each class. This output shows the probabilities of respondents in each latent class providing a low, medium or high response to the indicator variable in question. The rows represent the latent classes. The model assumed 2 latent classes in this case, therefore there are 2 rows. The columns indicate the categories (low, medium and high) of the indicator variable. For example, the conditional item response probabilities for the satisfaction variable produced in the 2-class model were as follows:

$satisth2

Pr(1) Pr(2) Pr(3)

class 1: 0.0008 0.1414 0.8578

class 2: 0.1695 0.7222 0.1083

“satisth2” is the name of the satisfaction variable used in this analysis. In the case of this analysis, as specified in Section 3.2 “low”, “medium” and “high” responses to each variable were coded as 1, 2 and 3 respectively. Therefore, in the output shown above, Pr(1) is the probability of a respondent providing a “low” response to the satisfaction variable.

The output shown above can be interpreted as follows: there is a 0.08% chance of a respondent in latent class 1 providing a “low” response to the satisfaction variable; a 14.14% chance of them providing a “medium” response and an 85.78% chance of them providing a “high” response. Therefore, overall respondents in latent class 1 are more likely to have high levels of satisfaction. On the other hand, respondents in latent class 2 are more likely to have medium levels of satisfaction (72.22%).

Taken together, the conditional probability results for all 4 personal well-being variables provided above indicate that respondents in latent class 1 are likely to have high levels of satisfaction (conditional probability = 85.78%), high levels of happiness (conditional probability = 79.25%), low levels of anxiety (conditional probability = 79.58%) and high levels of worth (conditional probability = 0.8728). On the other hand, respondents in latent class 2 are likely to have medium levels of satisfaction (conditional probability = 72.22%), medium levels of happiness (conditional probability = 56.62%), low levels of anxiety (conditional probability = 47.52%) and medium levels of worth (conditional probability = 63.65%).

The “estimated class population shares” section of the output provides the estimated proportions corresponding to the share of observations belonging to each latent class (Linzer & Lewis, 2011). Therefore, in the case of the 2-class model, the share of observations is estimated to be 63.68% in latent class 1 and 36.32% in latent class 2.

The “Predicted class memberships” is another way of estimating the size of the latent classes. This assigns observations to the latent classes using posterior probabilities. Generally, when the values for the “estimated class population shares” and “Predicted class memberships” are similar, this is an indication of good model fit. However, this congruence between values alone should not be used to assess model fit as there are other criteria which should be used to choose the best-fitting model. These are provided in the next section of output called “Fit for 2 latent classes”.

The “Fit for 2 latent classes” section of the output indicated the following results:

number of observations: 303778

number of estimated parameters: 17

residual degrees of freedom: 63

maximum log-likelihood: -934514.9

AIC(2): 1869064

BIC(2): 1869244

G^2(2): 61111.46 (Likelihood ratio/deviance statistic)

X^2(2): 138365.7 (Chi-square goodness of fit)

The “number of observations”, “number of estimated parameters”, “residual degrees of freedom” and “maximum log-likelihood” are reported in the output. The “number of observations” specifies the number of fully observed cases (i.e. cases without missing values) that were used in the analysis. The “number of estimated parameters” indicates the number of degrees of freedom used by the model. It is worth checking that the “residual degrees of freedom” is not negative. However, a negative residual will normally result in an error message.

The next set of values is useful for comparing models with different numbers of latent classes in order to assess model fit. The AIC is the Akaike Information Criterion whilst the BIC is the Bayesian Information Criterion. When models with different numbers of latent classes are compared, the model with the lowest AIC and BIC is normally chosen as the best-fitting model. This is because a lower value of the information criterion suggests a better balance between model fit and parsimony (Lanza, S.T & Rhoades, B.L., 2013). Sometimes, the AIC and BIC do not indicate the same model as the best-fitting one. In the case of basic exploratory LCA, the BIC is usually more appropriate because of the relative simplicity of the model (Lin and Dayton 1997; Forster 2000).

The Likelihood ratio and Chi-Square statistics may also be used to assess model fit. Like the AIC and BIC the aim is to select models that minimise the Likelihood ratio and Chi-Square statistics whilst also maintaining a low number of parameters. It is important to note that these tests are not useful with sparse data (which is a common problem in LCA) because they violate the chi-square distribution assumption (Linzer & Lewis, 2011; Nylund et. al, 2007). From their study investigating the various methods for assessing model fit, Nylund et. al (2007) concluded that the bootstrap likelihood ratio test is the best method for evaluating model fit, followed by the BIC. Currently, it doesn’t seem as though one can calculate the bootstrap likelihood ratio test using the poLCA package. Consequently, it is recommended that, when using poLCA, the BIC is used to choose the best-fitting model. A comparison of the BIC for all models estimated with the personal well-being data is shown in Table 1.

Table 1. Comparison of BIC obtained from the various latent class models estimated using personal well-being data

|  |  |
| --- | --- |
| **Number of Latent Classes estimated** | **BIC** |
| 2 | 1869244 |
| 3 | 1827588 |
| 4 | 1816278 |
| 5 | 1811968 |
| 6 | 1809868 |
| 7 | 1809400 |
| 8 | 1809478 |

The results in Table 1 indicate that the model with the lowest BIC is the 7-class model. Consequently, out of all the competing models, the model with 7 latent classes fits the data best and would be chosen as the final model for interpretation.

# 4.0 Extensions and Uses of Latent Class Analysis

The LCA described above is an exploratory LCA. This is probably the simplest form of LCA. There are many extensions to this which create a multitude of uses for LCA. For example, by including restrictions you can carry out a confirmatory LCA. Covariates could also be added to an LCA. Another extension of LCA is Latent Transition Analysis (SAS software for this is also available from [The Methodology Centre](https://methodology.psu.edu/downloads/proclcalta)) which can be used to assess changes in latent classes over time. Analysis with continuous variables can also be carried out; this is usually referred to as Latent Profile Analysis.

LCA is an extremely useful way of conducting more holistic analysis because it is a type of person-centred analysis as opposed to variable-centred analysis. Apart from being used in final analysis of data, LCA has also been used to assess the quality of questionnaires and the resulting estimates e.g. it has been used to assess classification error (see Biemer & Wiesen, 2002; Biemer 2011); to assess mode effects (see Biemer, 2001); and to evaluate questionnaire items (see Kreuter, Yan & Tourangeau, 2008).

# 5.0 Conclusion

LCA is a powerful technique which has a multitude of applications to official statistics. It is becoming very popular because of its many uses and the ease with which it can be carried out. Producers of official statistics are encouraged to apply LCA to their work in order to be able to provide higher quality and more meaningful statistics to users. For those colleagues interested in learning more about LCA, the following sources may be of interest:

McCutcheon, A.L., 1987. Latent Class Analysis. California: Sage Publications.

Magidson, J. and Vermunt, J.K. Latent Class Models. [In: Kaplan, D. eds. The SAGE Handbook of Quantitative Methodology for the Social Sciences.](http://srmo.sagepub.com/view/the-sage-handbook-of-quantitative-methodology-for-the-social-sciences/SAGE.xml) California: Sage Publications. pp.175-198.

Please contact Debbie Cooper ([Debbie.cooper@ons.gsi.gov.uk](mailto:Debbie.cooper@ons.gsi.gov.uk)) with any queries.

# References

Biemer, P. P., 2001. Nonresponse Bias and Measurement Bias in a Comparison of Face to Face and Telephone Interviewing. *Journal of Official Statistics*, 17(2), p.295-320.

Biemer, P. P. and Wiesen, C., 2002. Measurement Error evaluation of self-reported drug use: a latent class analysis of the US Household National Survey on Drug Abuse. *Journal of the Royal Statistical Society A,* 165(1), p.97-119.

Biemer, P. P., 2011. Latent Class Analysis of Survey Error. New Jersey: Wiley.

Forster M.R., 2000. Key Concepts in Model Selection: Performance and Generalizability.

*Journal of Mathematical Psychology*, 44, 205-231.

Chanfreau, J., Cullinane,C., Calcutt, E. and McManus, S. , 2014. Wellbeing in Wales Secondary analysis of the National Survey for Wales 2012-13. [pdf]. Welsh Government Social Research. Available at: <http://gov.wales/docs/caecd/research/2014/140430-national-survey-wellbeing-wales-2012-13-en.pdf> [Accessed 25 January 2017].

Kreuter, F., Yan, T. and Tourangeau, R., 2008. Good item or bad—can latent class analysis tell?: the utility of latent class analysis for the evaluation of survey questions. *Journal of the Royal Statistical Society A*. 171(3).

Lanza, S.T., Collins, L.M., Lemmon, D.R. and Schafer, J.L. 2007. PROC LCA: A SAS Procedure for Latent Class Analysis. *Structural Equation Modeling*, 14(4), p.671–694.

# Lanza, S.T. and Rhoades, B.L., 2013. Latent Class Analysis: An Alternative Perspective on Subgroup Analysis in Prevention and Treatment. *Prevention Science*, 14(2), p.157-168.

Lin T.H. and Dayton C.M., 1997. Model Selection Information Criteria for Non-Nested Latent

Class Models. *Journal of Educational and Behavioral Statistics*, 22(3), 249-264.

Linzer, D. A. and Lewis J. B., 2013. "poLCA: Polytomous Variable Latent Class Analysis." R package version 1.4. <http://dlinzer.github.com/poLCA.>

Linzer, D. A. and Lewis J. B., 2011. "poLCA: an R Package for Polytomous Variable Latent Class Analysis." *Journal of Statistical Software*. 42(10): 1-29. <http://www.jstatsoft.org/v42/i10>

McCutcheon A.L. and Hagenaars, J.A.,1997. Simultaneous Latent Class Models for Comparative Social Research. In: Langeheine, R. and Rost, J. (eds) Applications of Latent Trait and Latent Class Models. New York: Waxmann. Pgs. 266-277.

Muthén, L. K., & Muthén, B. O. (1998-2011). Mplus User's Guide. Sixth Edition. Los Angeles, CA: Muthén & Muthén.

Office for National Statistics., 2016. Personal well-being in the UK: Oct 2015 to Sept 2016 [online] Available at: <https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/bulletins/measuringnationalwellbeing/oct2015tosept2016> [Accessed 25 January 2017].

Office for National Statistics., 2016. Annual Population Survey [online]. Available through: UK Data Service Discover <https://discover.ukdataservice.ac.uk/series/?sn=200002> [Accessed 25 January 2017].

Office for National Statistics., 2017. Surveys using the 4 Office for National Statistics personal well-being questions [online]. Available at: <https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/methodologies/surveysusingthe4officefornationalstatisticspersonalwellbeingquestions> [Accessed 25 January 2017].

R Development Core Team, 2011. R: A Language and Environment for Statistical Computing. 3.0.2. [online]. Available at: <http://www.R-project.org> [Accessed 19 January 2016].

SAS Institute Inc. 2010. SAS/STAT™ 9.22 User’s Guide. Cary, NC: SAS Institute Inc. [online] Available at: <https://support.sas.com/documentation/cdl/en/statug/63347/PDF/default/statug.pdf> [Accessed 19 January 2016].

StataCorp. 2015. Stata Statistical Software: Release 14. College Station, TX: StataCorp LP.

The Methodology Centre. 2015. SAS Procedures for Latent Class Analysis & Latent Transition Analysis. [online]. Available at: <https://methodology.psu.edu/downloads/proclcalta> [Accessed 18 January 2016].

Vermunt, J. K., & Magidson, J., 2002. Latent class cluster analysis. In: J. A. Hagenaars, & A. L. McCutcheon eds. *Applied latent class analysis.* New York: Cambridge University Press, pp.89-106.

Vermunt, J. K. and Magidson, J., 2013. Technical Guide for Latent GOLD 5.0: Basic, Advanced, and Syntax. Belmont Massachussetts: Statistical Innovations Inc. [online]. Available at: <http://www.statisticalinnovations.com/latent-gold-5-1/> [Accessed 19 January 2016].

# Appendix 1 – SAS® code for proc LCA

As described in Section 3, LCA can also be carried out in SAS. The Methodology Centre at Pennsylvania State University developed a SAS plugin which enables users to run LCA in SAS. It is important to note that the LCA procedure in SAS has not been written and is not supported by SAS Institute Inc.

The procedure used for running LCA in SAS is called proc LCA. In order to carry out an LCA in SAS, the code should be specified as follows:

**proc** **LCA** data = *dataset\_for\_analysis*;

NCLASS ***number\_of\_latent\_classes\_to\_be\_estimated***;

ITEMS *categorical\_variables\_to\_include\_in\_analysis*;

Categories **number\_of\_categories\_in\_each\_variable**;

SEED **specifies\_seed\_for\_random\_number\_generator**;

**run**;

The following is an example of how this code would be specified for the personal well-being data described in Section 3.2:

**proc** **LCA** data = PWB\_LCA.wellbeing;

NCLASS **5**;

ITEMS happyth2 anxiouth2 worthth2 satisth2;

Categories **3** **3** **3** **3**;

SEED **100000**;

**run**;

The first line calls the ‘wellbeing’ dataset which is the dataset containing the personal well-being variables for analysis. Next, the number of classes to be estimated is specified. In this case we are running a latent class analysis which estimates 5 latent classes. The “ITEMS” option specifies the variables within the ‘wellbeing’ dataset that should be included in the LCA. Following this, the “Categories” option specified the number of categories that each variable contains. In the case of the personal well-being dataset, each variable (happyth2, anxiouth2, worthth2 and satisth2) contains 3 categories (low, medium and high) as described in Section 3.2. Finally, a seed is specified, this is a starting value for the random number generator (the user can pick any starting value). There are a number of other options which can be included in the proc LCA e.g. MAXITER to specify the maximum number of iterations as was done with poLCA. For a full list of options and further information on carrying out LCA in SAS please refer to Lanza et. al (2007).

The proc LCA output includes the Likelihood Ratio G2, degrees of freedom, AIC and BIC. The same methods for assessing model fit as those discussed in Section 3.4 apply here.