The Methodology Center

SAS Graphics Macros for Latent Class Analysis Users’ Guide

Version 2.0.1

John Dziak
The Methodology Center

Stephanie Lanza
The Methodology Center

Please send questions and comments to mchelpdesk@psu.edu

The suggested citation for this users’ guide is

Dziak, J. J., & Lanza, S. T. (2015). *SAS graphics macros for latent class analysis users’ guide* (Version 2) was supported by P50 DA010075 and P50 DA039838 from the National Institute on Drug Abuse. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute on Drug Abuse or the National Institutes of Health. We thank Dr. Xianming Tan, Dr. Mariya Shiyko, Dr. Runze Li, Amanda Applegate, and Aaron Wagner for their help, comments and suggestions.

Contents

1. About the SAS Graphics Macros for Latent Class Analysis

1.1 Overview

The suite of SAS macros, LcaGraphicsV2.sas, has been designed to help investigators explore and summarize the results from latent class analyses (LCA) conducted using the freely available procedure PROC LCA (Lanza et al., 2015). The following three macros are currently offered in the suite:

- **IdentificationPlot** to assess model identification;
- **ItemResponsePlot** to graphically summarize characteristics of each latent class; and
- **OddsRatioPlot** to show the confidence interval for odds ratios corresponding to covariates.

The general procedure for using the macros in the SAS system is described below, followed by a detailed description of each macro.

Changes from Version 1.0

- Bug fixes

1.2 General Procedure for Using the Macros

A SAS macro function is a special block of SAS commands that are first defined, and then called when needed. The procedure for using the LCA graphics macro functions is very straightforward.

1. Before running a macro, the SAS %INCLUDE statement must be used to read the macro code. For this product, all three macros are included in the same file, `LcaGraphicsV1.sas`. The following syntax should be included either before or after PROC LCA is called, and can be modified to show the local path where the file has been saved.

   ```
   %INCLUDE “S:\LcaGraphicsV2.sas”;
   ```

2. Each of the three macros in the LCA graphics suite must be provided information from PROC LCA in the form of a SAS dataset. The required datasets can be created using the `OUTPARAM=, OUTSTDERR=,` and `OUTSEEDS=` options in PROC LCA Version 1.2.4 or above. These are described further in the PROC LCA & PROC LTA users’ guide. The

example PROC LCA call shown below specifies that the LCA parameter estimates, corresponding standard errors (if available), and model fit information resulting from multiple seeds are to be saved in the working datasets named `param1, stderr1, and seeds1,` respectively.

```
     PROC LCA DATA=mydata OUTPARAM=param1 OUTSTDERR=stderr1
     OUTSEEDS=seeds1;
```

Note that the `OUTSEEDS` option is only available in conjunction with the `NSTARTS` option (see the PROC LCA & PROC LTA users' guide).

3. A macro function is called using a percent sign, the macro name, and any further required information in parentheses. For example, for the identification plot macro (described later), the calling syntax is

```
      %IdentificationPlot(SeedsDataset=name);
```

where *`name`* is the name of the output dataset produced using the `OUTSEEDS` option. The above syntax will invoke the IdentificationPlot macro.

2. Identification Plot

This plot uses a bar chart to show the frequency distribution of the distinct log-likelihood values resulting from multiple sets of randomly generated starting values. An examination of how many sets converge to the maximum likelihood solution (i.e., the highest log-likelihood value) can be a useful way to judge how well-identified an LCA solution is. The bar plot is shown twice, once as a text-based display in the output window (produced by the macro invoking PROC CHART) and again as a graph in the graphics output window (produced by the macro invoking PROC GCHART).

In PROC LCA versions 1.2.4 and above, the `NSTARTS` option is available to fit a model using multiple sets of starting values. This is intended to help users avoid reporting parameter estimates that correspond to a “local” (as opposed to the “global”) maximum of the likelihood function, an issue that can easily arise when fitting mixture models. It is important to note that even in very well-identified models, such as a simple model with high latent class separation and a large sample size, not all randomly generated starting values are assured to converge on the maximum likelihood solution. Therefore, the user often needs to use his or her judgment to determine how well-identified a model is. The proportion of times a model converges to the highest log-likelihood value, which can be discerned from this plot and is also provided in PROC LCA standard output, can be used as a measure of confidence that one has identified the global maximum. One possible rule of thumb might be to use 50 random starting values. You could consider a model to be sufficiently identified when, say, at least 25% of them cause the algorithm to converge to the highest log-likelihood value. This cutoff is only a heuristic; an investigator is free to use personal judgement and to consider the meaningfulness and theoretical interpretability of models. (Note that there are unusual situations in which a model is unidentified for clear statistical reasons; in particular, if it has negative degrees of freedom. This macro is not helpful in those situations; they require that the model be changed.)

The ideal result is a plot with a single bar, meaning that all of the starting values resulted in the same final estimate. This suggests that the best solution for this model with these data has probably been found. (This would not be a justified conclusion if `NSTARTS` was very small, though; for example, if only 2 random starts were used they might agree by chance.) If the identification plot shows many different solutions and the best solution is no more common than the others, then it is still very feasible that there might be a higher log-likelihood value that has not been found yet. In this case, the model cannot be identified well using the specified dataset.

In addition to rerunning the model specifying a larger number of starting values, it may be wise to try a simpler model (e.g., a model with fewer latent classes).

In the Identification Plot, the log-likelihood values are rounded to two decimal places, so that, for example, a solution with a log-likelihood of -5432.001 is considered to be the same as one with a log-likelihood of -5432.002. This is necessary because, as in all computational routines, the EM algorithm for estimating the parameters of LCA models does not give unlimited decimal precision in a finite number of iterations.

Requirements:

The macro requires an `OUTSEEDS` dataset from PROC LCA, run using the `NSTARTS` option.

Output:

The macro produces a frequency bar chart of log-likelihood values resulting from multiple random sets of starting values.

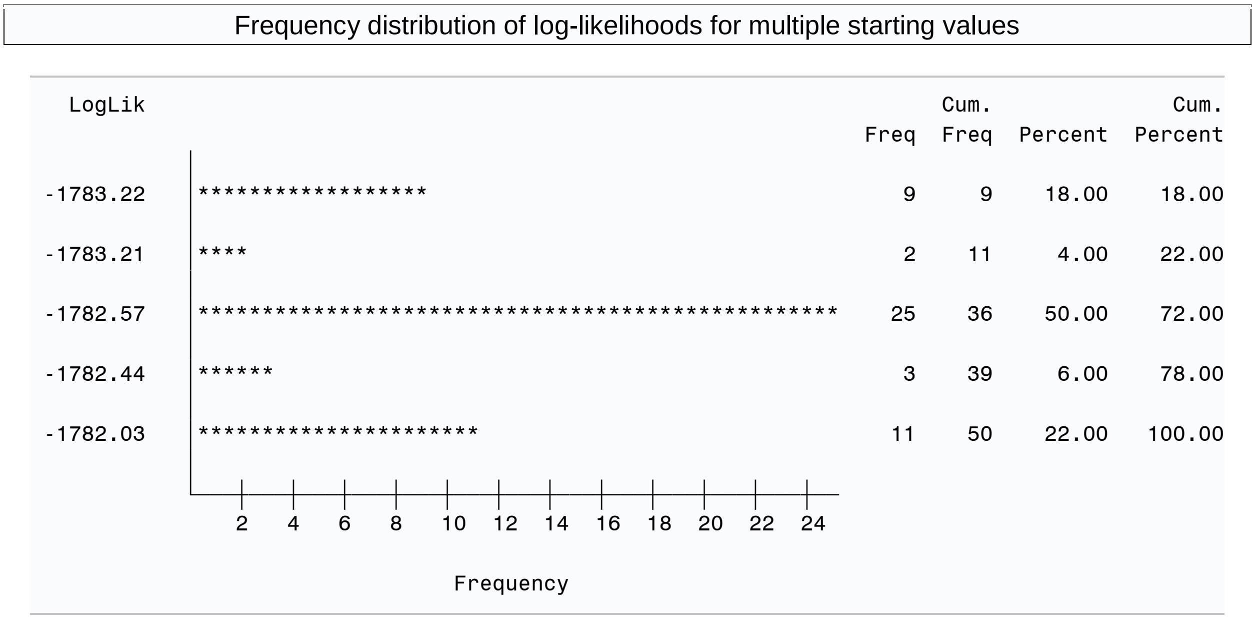
Syntax:

```
%IdentificationPlot(SeedsDataset=datasetname);
```

Example code:

```
PROC LCA DATA=ExampleData OUTPARAM=out1 OUTSTDERR=out2 OUTSEEDS=out3;
NCLASS 4;
ITEMS Alcohol Tobacco Marijuana Inhalant Cocaine OtherHard ;
CATEGORIES 2 2 2 2 2 2;
ID ID;
SEED 100000;
RHO PRIOR = 1;
NSTARTS 50;
RUN;
%IdentificationPlot(SeedsDataset=out3);
```

Example results:

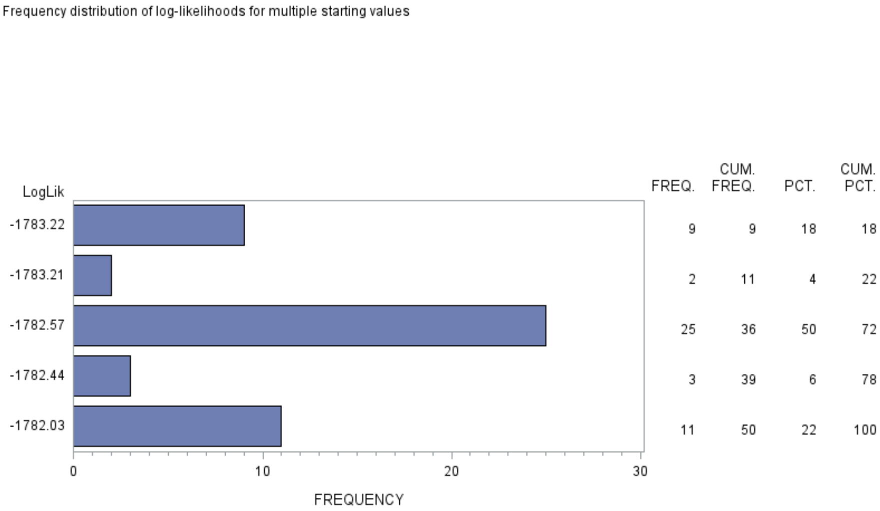


Frequency distribution of log-likelihoods for multiple starting values

```
  LogLik                                                                        Cum.                Cum.
                                                                         Freq   Freq   Percent   Percent

-1783.22  |******************                                               9      9     18.00     18.00
          |
-1783.21  |****                                                             2     11      4.00     22.00
          |
-1782.57  |**************************************************              25     36     50.00     72.00
          |
-1782.44  |******                                                           3     39      6.00     78.00
          |
-1782.03  |**********************                                          11     50     22.00    100.00
          |
          ----+---+---+---+---+---+---+---+---+---+---+---+-
              2   4   6   8  10  12  14  16  18  20  22  24

                           Frequency
```



Frequency distribution of log-likelihoods for multiple starting values

The example plots above illustrate a model with possible identifiability problems. Most of the starting values converge to a solution that is poorer than the best solution available. The dataset that was used to run this example was a simulated (artificial) dataset, but if plots like these were found in an actual empirical study, the researcher may wish to do some further investigation and decision-making. The researcher may judge that the model and solution are reasonable and that at least some of the starting values agreed with the best, and conclude that identifiability is good enough. If the researcher does not feel that identifiability is good enough, a simple solution would be to try a smaller number of classes; this usually improves the percentage agreement. In fact, running the above code with NCLASS 3 gave 100% agreement.

3. Item Response Plot

This plot shows the characteristics of each latent class by graphically summarizing the pattern of conditional item-response probabilities. It is similar to the profile plots constructed by Latent GOLD (Vermunt & Magidson, 2005) and Mplus (Muthén & Muthén, 2007). These plots can help an investigator to describe each class intuitively in terms of how likely members are to provide a particular response and to determine which items differentiate the latent classes.

One plot is provided for each response category. Thus, if all of the items are dichotomous (for example, each item is coded 1 for ”yes” and 2 for ”no”), there will be one plot showing the probability of responding “yes” to each item given latent class membership and one showing the probability of responding “no” to each item. These two plots provide essentially the same information since these probabilities sum to one for each class-item combination. If items have different numbers of response categories, the item with the maximum number of response categories will determine the number of plots that will be produced. In multiple-groups LCA, one set of plots will be shown for each group unless measurement invariance is imposed (measurement invariance can be imposed using the `MEASUREMENT` option; see PROC LCA & PROC LTA user’s guide for details).

Requirements:

The macro requires an `OUTPARAM` dataset from PROC LCA.

Output:

The macro produces a graphical representation of the item-response probabilities (rho parameter estimates) for each response category for each item, conditional upon latent class.

Syntax:

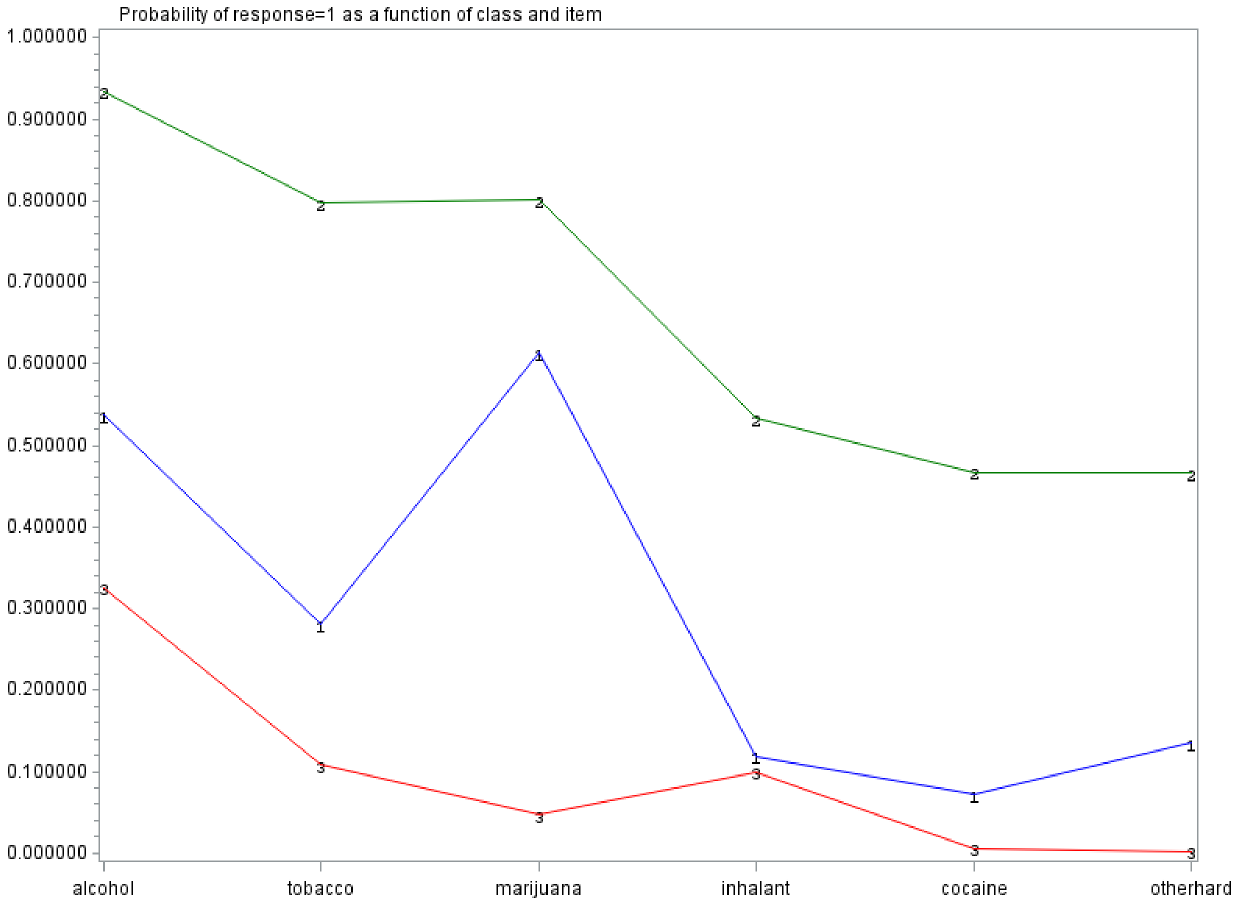
```
%ItemResponsePlot(ParamDataset=datasetname);
```

Example code:

```
PROC LCA DATA=ExampleData OUTPARAM=out1 OUTSTDERR=out2 OUTSEEDS=out3;
NCLASS 3;
ITEMS Alcohol Tobacco Marijuana Inhalant Cocaine OtherHard ;
CATEGORIES 2 2 2 2 2 2;
ID ID;
SEED 100000;
RHO PRIOR = 1;
NSTARTS 50;
RUN;
%ItemReponsePlot(SeedsDataset=out3);
```

Example results:

A plot of the probability of response 1 to each item (interpreted as “yes” in this simulated dataset) is shown below. A similar plot of the probability of response 2 is also generated; however, because the items in this dataset are all dichotomous, the second plot is not needed, because the probability of 2 is simply one minus the probability of 1. (Informally speaking, the second plot will be like the first plot but upside down.)



Each represents one latent class and appears in a different color. The integers shown on each line correspond to the latent class (e.g., classes 1 through 3 in the plot above). Values represent the point estimates of item-response probabilities (rho parameters) for each class. Values for each latent class are connected by line segments to make it easier to perceive the overall set of rho parameters as a single ‘profile.’.

In this dataset, classes, 1, 2, and 3 generally correspond to medium, high, and low profiles of use of the various substances specified, although marijuana in particular is used almost equally by the high-use and medium-use classes, and inhalants are uncommon in the medium and low profiles.

4. Odds Ratio Plot

For LCA models that include covariates, this plot can be used to show the 95% confidence interval for the odds ratio for each latent class (relative to the reference class), corresponding to a one-unit increase for each covariate. If the `BINARY` option is invoked to estimate the effect of a covariate on membership in a particular latent class relative to membership in any of the other latent classes, the Odds Ratio Plot shows the 95% confidence interval for the odds ratio. When the LCA model involves two or more groups, the confidence intervals are shown for each group.

Requirements:

This macro requires an `OUTPARAM` dataset and an `OUTSTDERR` dataset from PROC LCA. The model and the estimates must be such that PROC LCA was successfully able to compute both logistic regression coefficients and standard errors. Sometimes this may require the use of a `BETA PRIOR` and/or a `RHO PRIOR` (please see the PROC LCA & PROC LTA user's guide for more information).

Output:

The macro produces a graphical representation of confidence intervals for the odds ratios (i.e., for the exponentiated beta coefficients from the logistic regression of class membership on the covariates). The y-axis is arranged on a log scale so that the confidence intervals will be easier to compare.

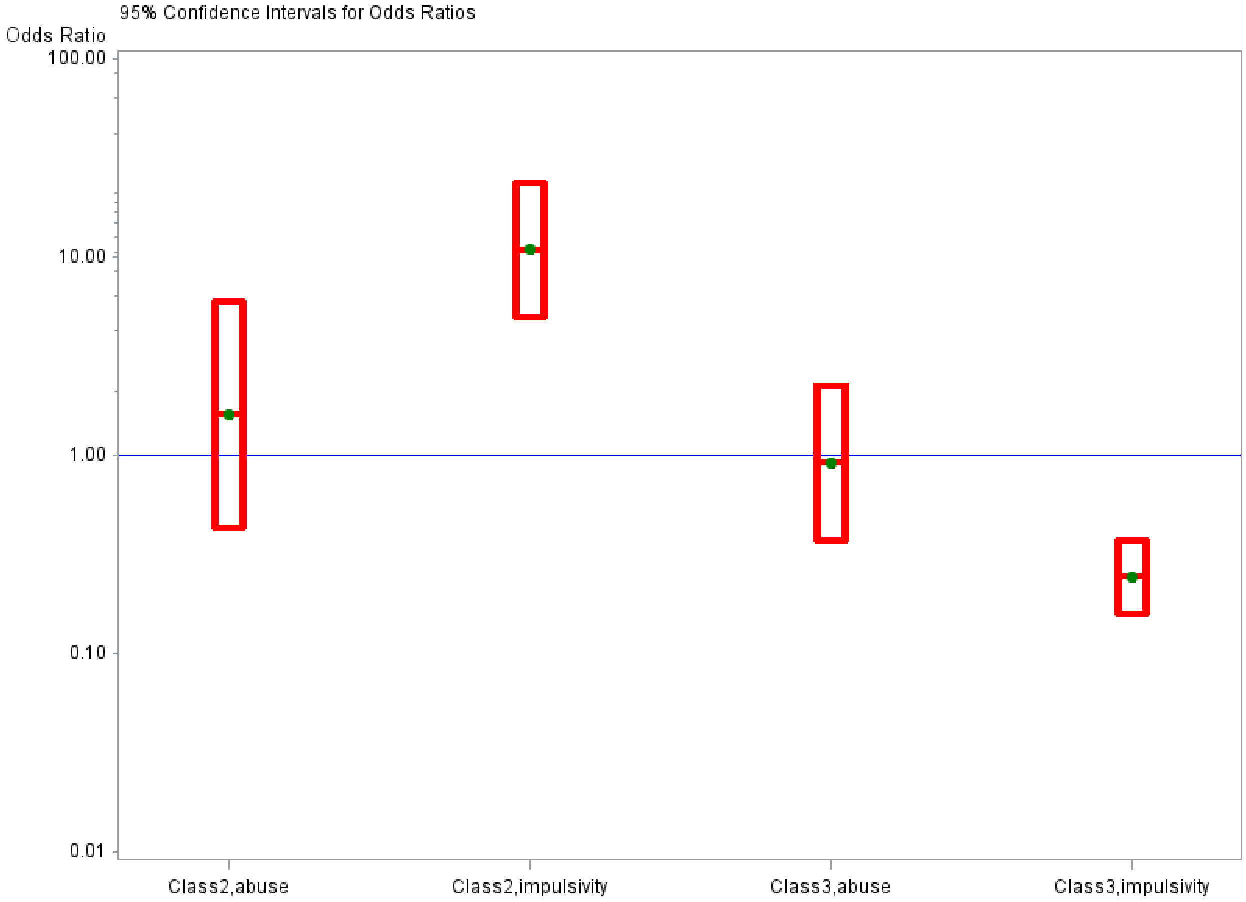
Syntax:

```
%OddsRatioPlot(ParamDataset=datasetname, StdErrDataset=datasetname);
```

Example code:

```
PROC LCA DATA=ExampleData OUTPARAM=out1 OUTSTDERR=out2 OUTSEEDS=out3;
NCLASS 3;
ITEMS Alcohol Tobacco Marijuana Inhalant Cocaine OtherHard ;
CATEGORIES 2 2 2 2 2 2;
ID ID;
SEED 100000;
COVARIATES Abuse Impulsivity;
RHO PRIOR = 1;
BETA PRIOR = 1;
NSTARTS 50;
RUN;
%OddsRatioPlot(ParamDataset=out1, StdErrDataset=out2);
```

Example results:



Recall that Class 1, the reference class, represents medium use; classes 2 and 3 represent high and low use, respectively. Child abuse history does not appear to be related to drug use class. However, higher impulsivity appears to be strongly related to higher odds of being in class 2, and lower odds of being in class 3, relative to class 1.

5. About the Examples

The examples presented here are all based on a simulated (artificial) dataset. The simulated dataset is included in the download. Of course, it should not be used for substantive research, since it is fabricated data for demonstrating software, and its findings may or may not be valid in the real world.

The variables are as follows:

- Six dichotomous (1=yes, 2=no) items indicating possible drug use behaviors: Alcohol, Tobacco, Marijuana, Inhalant, Cocaine, OtherHard.
- Abuse: A dichotomous (0=no, 1=yes) item indicating history of having been abused as a child
- Impulsivity: A numerical item indicating degree of tendency towards impulsive behavior in everyday life

6. References

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

Lanza, S. T., Dziak, J. J., Huang, L., Wagner, A., & Collins, L. M. (2015). *PROC LCA & PROC LTA users' guide* (Version 1.3.2). University Park: The Methodology Center, Penn State. Retrieved from http://methodology.psu.edu

Muthén, L.K. and Muthén, B.O. (1998-2007). *Mplus user's guide* (5th Ed.). Los Angeles, CA: Muthén & Muthén.

PROC LCA & PROC LTA (Version 1.3.2) [Software]. (2015). University Park: The Methodology Center, Penn State. Retrieved from http://methodology.psu.edu

Vermunt, J. K., & Magidson, J. (2005). *Latent GOLD 4.0 user's guide*. Belmont.